Linear Dynamics of Evidence Integration in Contextual Decision Making

Joana Soldado Magraner

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
of
University College London.

The Gatsby Computational Neuroscience Unit
University College London

Declaration

I, Joana Soldado Magraner, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

December 6, 2018

Abstract

Individual neurons in Prefrontal Cortex (PFC) exhibit a vast complexity in their responses. Central in Neuroscience is to understand how their collective activity underlies powerful computations responsible for higher order cognitive processes. In a recent study (Mante et al., 2013) two monkeys were trained to perform a contextual decision-making task, which required to selectively integrate the relevant evidence –either the color or the motion coherence of a random dots stimulus– and disregard the irrelevant one. A non-linear RNN trained to solve the same task found a solution that accounted for the selective integration computation, which could be understood by linearizing the dynamics of the network in each context. In this study, we took a different approach by explicitly fitting a Linear Dynamical System (LDS) model to the data from each context. We also fitted a novel jointly-factored linear model (JF), equivalent to the LDS but with no dynamical constraints and able to capture arbitrary patterns in time. Both models performed analogously, indicating that PFC data display systematic dynamics consistent with the LDS prior. Motion and color input signals were inferred and spanned independent subspaces. The input subspaces largely overlapped across contexts along dimensions that captured coherence and coherence magnitude related variance. The dynamics changed in each context so that relevant stimuli were strongly amplified. In one of the monkeys, however, the integrated color signal emerged via direct input modulation. The integration took place within subspaces spanned by multiple slow modes. These strongly overlapped along a single dimension across contexts, which was consistent with a globally identified decision axis. Interestingly, irrelevant inputs were not dynamically discarded, but were also integrated, although in a much lower extent. Finally, the model reproduced the main dynamical features of the population trajectories and accurately captured individual PSTHs. Our study suggests that a whole space of sensory-related input signals invariantly modulates PFC responses and that decision signals emerge as the inputs are shaped by a changing circuit dynamics. Our findings imply a novel mechanism by which sensory-related information is selected and integrated for contextual computations.

Impact statement

The results presented in this thesis are expected to make the most significant contribution to basic research in neuroscience. The goal of this work was to provide a theoretical framework that explains fundamental principles of brain computation. In particular, we focused on understanding highly adaptive neural systems, endowed with the flexibility to quickly shift their mode of operation based on the demands of the environment. For that, we studied the dynamics of populations of neurons in prefrontal cortex (PFC) under contextual decision making computations. We proposed a model that provides a mechanistic interpretation for how these type of computations are implemented in the neural substrate.

The principles inferred from this model can be tested in other neural systems operating under different regimes. Upon improvement of its generalization capability, the model can be rendered powerful enough to make predictions about different brain states and their link to behavior. A better understanding of basic neural processing will with no doubt accelerate the discovery of medical treatments for neurological diseases. In particular, those tightly linked to malfunctions at the circuit level, such as epilepsy.

The data that we used in this work was collected in the context of a previous study and had already been analyzed in detail. In fact, the publication that derived from it was of high impact in the field. Nevertheless, we found that many aspects of the data were worth reanalyzing –and doing it so under a new perspective. We believe that our approach has lead to valuable new insights into the problem in question. We therefore expect that our initiative will motivate other researchers to reuse data, which can help replace unnecessary animal experimentation.

Outside the biological sciences, disciplines such as machine learning and artificial intelligence can also benefit greatly from this work. The domain of applicability includes both academic and non-academic or industrial settings involving any form of complex systems with emergent collective behavior. The same principles that govern brain computation can in theory be applied to all sorts of intelligent systems. Therefore, we expect that our work will contribute to the design of new algorithms that can support advanced forms of generalized AI.

Finally, this work can contribute to the understanding of foundational questions in the fields of physics, mathematics and philosophy. The way our brain operates is unavoidably linked to the way we construct physical models, mathematical theories and other types of logic frameworks. Establishing the boundaries of our own rationale is akin to understanding the limitations of any human-based formal system.

To my sisters

"All models are wrong but some are useful"

Acknowledgments

I thank my supervisor, Maneesh Sahani, for his continuous support throughout my whole PhD. The scientific journey we have embarked ourselves on has been a continuous intellectual *tour de force*. His guidance has been invaluable and I have learned enormously from him. In particular, I take with me two specific lessons. The first, learning how to extract deep mathematical intuitions from neural data and relate them back to biology. The second, which was completely unexpected for me, learning to ignore all the external pressures that prevent us from achieving the best of our science. I will always be thankful to him for having given me the opportunity to work with him and for having taken me under his wing, in spite of my initial struggles with the very same subject he was teaching. Finally, his comments and remarks throughout the whole writing process helped improve this thesis enormously. I also thank him for his direct contributions to it. Chapter 2 is based on initial notes from him and the central pieces of the code used to fit one of the models are based on his own implementation.

I thank all the members of the Gatsby Unit for contributing to create such a unique research environment. I feel honored I was given the opportunity to enter the Gatsby program. The Gatsby Unit is the most stimulating and academically challenging place I have ever been. I have learned a great deal from both faculty members and fellow students. I have also made many good friends in there, which I hope I will keep in touch with. Sofia Jativa, Eszter Vertes, Lea Duncker, Alex Antrobus, Maria Lomeli, Vincent Adam, Heiko Strathmann and many others.

I thank my collaborators at Stanford University, Krishna Shenoy and Chandramouli Chandrasekaran and also Bill Newsome. It has been an honor to work closely with them. I thank both Maneesh and Krishna for having given me the opportunity to spend some time at Krishna's lab, which has allowed me to benefit from yet another highly stimulating environment.

I thank all the members of the Institute of Neuroinformatics in Zurich, for having given me the opportunity to start a career in this field. In particular, my previous supervisors, Valerio Mante, Michael Pfeiffer and Jerry Chen, for their guidance and their trust in my capabilities to conduct my own independent research. The two years I spent at the INI as a master student were one of the most fruitful periods in my life, both academically and personally. The INI was an inspirational and highly stimulating place, with a unique familiar and nurturing environment. My passion for neuroscience bloomed from the interactions with the people I met in there. Both at the INI and at the Gatsby I had countless fascinating conversations about science, philosophy, ethics and life in general.

I have also highly enjoyed interacting with the experimental groups at the Sainsbury Wellcome Center, where the Gatsby Unit is based now. I have benefited greatly from being embedded in this mixed theoretical and experimental setting. I had the same experience during my master's years at the INI, which is next door to the HIFO, the

Brain Research Institute. This rich environment allowed me to establish contacts that resulted in very fruitful collaborations. I feel grateful to have spent these two periods of my academic life in these two amazing places. I also thank the admin personnel from the two institutions. They were impressively diligent and always helpful.

I thank Arbie and Felix, the two monkey subjects in the study. I genuinely hope that this work takes us a step closer towards the understanding of all brains and minds, which in the long run, will help improve the well-being of all sentient beings.

I thank my funding institution, the Gatsby Foundation, for supporting my PhD education. I also thank the international students division at the University of Zurich, which supported me during my master's program in Zurich. I will always be extremely thankful to them for having granted me with financial stability, which gave me so much peace of mind.

I thank my parents, for having provided me with the best childhood one could dream of, for supporting me throughout my whole life without question and for allowing me to pursue my own path, wherever that has taken me. I thank the rest of my family, for being always such good company and an infinite source of happiness. I thank all the good friends I have left in Spain and in multiple other countries, which I know I will keep for life. Finally, I thank my sisters, Inma and Saray, for being always there for me. They have been a continuous source of inspiration and admiration, both personally and scientifically. We started our journey in this life together and a part of them will always remain with me.

Contents

List of Figures

List of Tables

Including Chapters

Chapter 1

In this chapter we first provide an historical overview of the origins of perceptual decision-making studies in neuroscience. We then discuss previous work on the role of Prefrontal Cortex (PFC) in perceptual decision-making. Finally we turn our focus to contextual computations for decision-making, elaborating on the way PFC is believed to mediate such important computations.

Chapter 2

In this chapter we review existing methods that have been applied in neuroscience with the goal of extracting low-dimensional structure from neural recordings. After this, we introduce the two models that we used to analyze the PFC data from the study by Mante et al. (2013), recorded during contextual decision-making. The first, a Linear Dynamical Systems model (LDS) and the second, a novel jointly-factored linear model (JF).

Chapter 3

In this chapter we start by introducing the work by Mante et al. (2013), which provided important insights into the mechanisms mediating contextual decision-making computations in PFC's neural populations. We summarize the main findings of the study and comment on the challenges and open questions. We then introduce a new modeling perspective, based on the methods described in Chapter 2, that allow us to gain a deeper understanding of the PFC data. This is rooted on a dynamical systems framework, but takes a model fitting approach. Finally, we perform a comprehensive analysis of the results obtained by the models, with the goal of drawing new insights into the nature of the computations implemented by this circuit.

Chapter 4

In this chapter we elaborate on the results obtained in this thesis, discussing in great detail the implications of our study in the light of current and previous findings. Finally, we comment on the limitations of our own methodology and propose further extensions and analysis to consider in future work.

Chapter 1

Introduction

In this chapter, we start by providing a brief historical overview of the origins of perceptual decision-making studies in neuroscience. The roots of the field date back to the thinkings of eminent philosophers such a Descartes. The stem of its foundations followed the ideas of the fathers of modern neuroscience, Sherrington and Cajal. The full branching occurred during the last part of the 20th century and bloomed in the early 2000s. This brings us to the present time, as perhaps, we might have just begun to collect some of its fruits.

1.1 Perceptual decision making

During the Enlightenment, many of the core beliefs that had been established since the times of ancient Greece started to get shattered. The thinkers of the new modernist era began to accept that all natural phenomena could be explained on purely materialistic grounds. This included human behavior. It was Descartes that first challenged the widely accepted ideas of philosophers such as Aristotle, who argued that the nonmaterial soul was the source of human motivations and intentions (Glimcher, 2003).

Descartes, inspired by the scientific revolution during the Enlightenment, proposed that mechanistic explanations, grounded in the material world, could be used to explain basic behaviors such as reflexes. Therefore, it should be possible to predictably link stereotyped sensory stimuli with simple motor responses. However, Descartes still adopted a dualistic view as he thought that more complex behaviors could not be explained on purely physiological grounds. These non-deterministic set of behaviors had to have their origin in the nonmaterial soul.

A few centuries later, Charles Sherrington, following Descartes ideas, continued the motor reflex tradition. He was convinced that fully mechanistic models could explain basic actions. In particular, he believed that a physical connection had to exist in our bodies, which linked sensory events and motor outputs via an intermediate integrative process. He looked for such "integration" centers in spinal circuits by studying the most basic form of behaviors: motor reflexes. These basic deterministic "decision" processes could be understood from the direct connection between sensory receptors and motor neurons. Sherrington believed that understanding reflexes was the first step towards understanding much more complicated forms of decision making (DM) (Glimcher, 2003).

Around the same time, right before the beginning of the twentieth century, Santiago Ramon y Cajal sketched the first basic plan for the nervous system organization (Cajal, 1893) and postulated his neuron doctrine –the principle that individual neurons are the elementary signaling elements of the nervous system (Waldeyer, 1891; Cajal, 1954). Cajal's neuron doctrine motivated most of the research in neuroscience in the second half of the twentieth century. This included seminal studies such as the ones by Hubel and Wiesel (1959), in which individual cortical neurons in some brain areas could be characterized as highly specific sensory receptors, lawfully responding to distinct external stimuli. Similarly, neurons in several cortical areas were found to respond to specific motor actions, such as saccadic eye movements. The growing understanding of these specialized sensory and motor systems made it possible to apply Sherrington's approach to study complex sensory-motor transformations in the cortex (Glimcher, 2003). This included motor actions triggered by perceptual judgements.

1.2 Neural correlates of perceptual decisions.

In the last two decades of the twentieth century several groups started studying decision-making processes by monitoring neuronal activity of monkeys trained to perform perceptuomotor tasks (Glimcher, 2003). The experiments involved situations in which a controlled sensory stimulus would trigger specific motor actions. In particular, Bill Newsome's group at Stanford University developed a visuomotor behavioral paradigm, the random dots motion coherence task (RDM), designed to test whether neuronal activity in certain regions of the brain reflected the perception of the visual stimulus and could also account for the eventual action, or decision, of the monkeys (Newsome et al., 1989; Britten and Newsome, 1998). The task consisted of discriminating the overall direction of motion of a pattern of randomly moving dots when a percentage of them were set to move coherently in a particular direction. The monkeys had to report their judgment performing a saccade to a target placed in the direction of movement –given an alternative "incorrect" target placed opposite to the direction of movement. The area Newsome and colleagues were recording from, the middle temporal area (medial temporal area MT or V5), was known to be involved in motion perception. Importantly, it was high enough in the visual hierarchy to act as a region where sensory-to-motor transformations could occur –following Sherrington's steps looking for reflex "integration" centers in the spinal cord. In a series of physiological studies, they found that motion direction selective neurons in MT were modulated by the "quality" of the visual stimulus, which was controlled by the coherence level of the random dots. More interestingly, they identified single neurons whose sensitivity for direction of motion was better than the monkey's own sensitivity, as their physiological threshold for motion direction discrimination was lower than the monkey's own psychophysical threshold. This meant that, had the monkey considered the motion direction information from only those neurons, its performance would have been greater. The finding suggested that large pools of neurons in MT are read out by downstream areas, regardless of the quality of the information they transmit, to transform the motion information into decisions.

In later studies, the same group identified another region, the lateral intraparietal area (LIP), as a candidate center where the decision-making computations could be implemented. The proposed decision-making mechanism involved the integration of signals from motion direction cells in MT by circuits in LIP (Shadlen and Newsome, 1996; Shadlen et al., 1996). Evidence for this computation came from the ramping-like time courses of recorded neural responses in LIP –after they had been averaged across trials. The activity of single neurons in this region reflected both the direction of a saccade and the quality of the sensory information that instructed the command. The findings suggested that this area accumulates sensory signals relevant to the selection of an action that triggers a specific eye movement –a saccade towards a given direction– (Shadlen and Newsome, 2001). In the original study, however, the researchers emphasized that decisions might not be formed in LIP, as it could be that this brain region was inheriting the signals from another area or group of areas (Shadlen and Newsome, 1996). Direct manipulation studies would have to be perform in order to establish a causal link between LIP activity and decision-making behaviors.

Further work helped characterize the potential mechanisms used to trigger the decisions. Roitman and Shadlen (2002) designed a reaction time task which allowed them to establish a stronger link between LIP activity and decisions. They found that the rate of buildup of activity in LIP neurons was not only correlated to the strength or quality of the motion stimulus, but also to the reaction time of the monkeys. The steeper the firing rate change was, the faster the monkeys responded. Furthermore, the activity seemed to reach a fixed level right before the initiation of a saccade, suggesting that the decision process was terminated upon the crossing of this fixed threshold.

In subsequent studies it was hypothesized that the activity of action planning areas such as LIP is directly related to the odds of selecting a choice versus an alternative (Gold and Shadlen, 2001). In particular, it was proposed that the firing rates (FR) of neurons reflected an accumulated decision variable (DV) computed based on the noisy sensory evidence. The DV approximated the log of the likelihood ratio (logLR or log odds) favoring one hypothesis over another. These findings motivated the use of two-choice reaction-time models –which had been successfully applied in psychophysical perceptual DM tasks– to describe the decision formation process reflected in the FR responses. These included drift-diffusion models (DDMs), which describe the accumulation of noisy sensory evidence as a random walk process with mean drift rate given by the strength of the evidence – and under the assumption of Gaussian noise. The incorporation of a threshold would terminate the process, generating a categorical choice (Ratcliff and Rouder, 1998; Gold and Shadlen, 2007). Using similar mechanisms, race-to-threshold models considered pools of neurons accumulating evidence towards different alternatives and mutually inhibiting each other. These types of models have been used to describe how different neural populations in LIP integrate directional signals from pools of MT neurons and race against each other towards a threshold to trigger a saccade in a given direction (Mazurek et al., 2003).

Finally, an important experiment was made to prove a direct, causal relationship between LIP neuron's activity and decision. Hanks et al. (2006), using an electrical microstimulation protocol, were able to bias the monkey's choices by altering the activity of groups of neurons in both areas MT and LIP. Their study established a causal role of these areas in the decision formation process and confirmed that LIP acts as an integration center where motion directional signals are accumulated and transformed into decisions. However, as we will see in the next section, this whole framework has been recently challenged.

1.3 The role of PFC in perceptual decision making

Following the findings in area LIP, researchers continued studying the visual-saccadic system with the goal of characterizing the elements of the decision-making circuitry engaged in visuomotor tasks. LIP, which is found in the posterior parietal cortex (PPC), was known to have direct projections to PFC, in particular to the dorsolateral prefrontal cortex (dlPFC). The prefrontal cortex is recognized as a higher order area, however, reciprocal connections exists between the two regions. In fact, dlPFC and PPC share multiple functional properties involving attention or working-memory and present very similar activity profiles and latencies of activation during such tasks. However, PFC seems to play a stronger role in this type of cognitive functions, as its inactivation causes more severe impairments (Katsuki and Constantinidis, 2012). In working-memory tasks for spatial location involving delayed motor executions, neurons in dorsolateral PFC had been shown to display persistent activity during the delay period. This activity, interestingly, was often selective of particular locations. Kim and Shadlen (1999) hypothesized that such persistent neurons may be involved in the transformation of sensory evidence from visual cortex towards a behavioral plan to shift the gaze. Indeed, similar response profiles as in LIP were identified in dlPFC and in the frontal eye fields (FEF) (Hanes and Schall, 1996; Kim and Shadlen, 1999). FEF was known to be part of a big network for controlling eye movements, exerting a direct influence in the generation of both saccades and visual smooth pursuit (Krauzlis, 2004). The new findings suggested that in such areas, the developing oculomotor commands may indeed reflect the formation of the monkey's judgement for motion direction (Gold and Shadlen, 2000; Ding and Gold, 2012).

Modern studies have identified many other cortical and subcortical structures as part of the decision-making circuitry in the brain (Siegel et al., 2015; Hanks et al., 2015; Yartsev et al., 2018). Decision signals seem to arise from the gradual transformation of sensory information as it travels through this distributed but highly interconnected network. Therefore, the view of LIP as a core integrative center for decision-making has gradually been abandoned. Even the causal relationship between LIP activity and choice-generation has been recently disputed. The concern had been raised that microstimulation studies –as applied in Hanks et al. (2006)– could excite passing fibers projecting to other regions in the brain. Therefore, behavioral effects following the microstimulation could arise from the activation of such regions and not from the activity of the particular area being stimulated. Katz et al. (2016), using inactivation techniques –which do not present this problem–, failed to impact the monkey's choices when inhibiting LIP. Finally, recent studies have also questioned the nature of the computations being implemented in this area (Latimer et al., 2015). In particular, it was pointed out that the smooth and ramping-like responses observed in LIP could arise from averaging single trial spike trains that experience sharp FR transitions at different times [1]. This study challenged the mechanism implied by a drift-diffusion model for explaining the emergence of decision-related activity in individual neurons, on a trial-by-trial basis.

Neurons in PFC often present "mixed" selectivity, meaning that they encode a mixture of different task-related variables (Rigotti et al., 2013). These include external stimuli, choice-related signals and other behaviorally relevant variables such as economic values. This latter type of non-sensory signals are the focus of a large body of new research studying economic values of choice or value-guided DM. In parietal areas such as LIP and superior colliculus (SC), it was found that neurons reflect the relative value of different alternatives and also the prior probability that a particular action would be reinforced (Glimcher, 2003). Reward value or utility encodings are also found in PFC and the basal ganglia. These regions have been shown to play a crucial role in value-guided decision-making –in particular PFC's orbitofrontal cortex (OFC). The activity of neurons in these areas seem to reflect the computation of a utility function, which combines prior beliefs for certain outcomes, likelihood estimates derived from sensory evidence and value assessments. Neuroeconomic models of choice have been extensively used to explain how neurons modulate their responses as a function of these variables. However, a limitation of these models is that they do not capture the dynamics of decision formation (Hunt et al., 2015). On the contrary, dynamical models such as the DDM are able to reproduce the temporal evolution of neural responses throughout the whole decision period. The dynamical framework implies that economic variables represented in the FRs of neurons arise from the unfolding, in time, of the very same decision process (Hunt et al., 2015). Other dynamical models related to the DDM are recurrent neural networks (RNN) (Bogacz et al., 2006). The appeal of these type of models is that they capture the temporal evolution of the population of neurons as a whole. Different computations, from sensory processing to choice generation, arise from

[1] The authors in Kim and Shadlen (1999) were already aware of this possibility: that the average response from many trials might give the appearance of a gradual evolution of a signal reflecting the monkey's plan. They indeed looked for discrete changes in the firing patterns of FEF neurons using an existing algorithm. However, no evidence for abrupt transitions in the FRs was found.

the dynamics of the population and can be understood from the portrait of attractor dynamics implemented by the network. Furthermore, given the degree of complexity in PFC's individual responses –which typically present "mixed" selectivity–, the population perspective can provide a clearer picture of the type of variables that are encoded in the area. This modeling framework, as we will see in the next section, has proven to be very successful for understanding complex decision processes, such as contextual decision making.

1.4 The role of PFC in contextual decision making

A problem that we constantly face is that of having to adapt our behavior when certain conditions change in our environment. How are we able to recognize in which situations we must perform a given action? and how can we do this so flexibly, quickly accommodating to an ever-changing world? The prefrontal cortex is believed to play a crucial role in these types of behaviors (Fuster, 2015; Miller and Cohen, 2001). PFC is involved in planning, selective attention and executive control. It has also been shown to hold the representation of goals, contexts and task rules (Wallis et al., 2001; Tanji and Hoshi, 2008). PFC is indeed required in order to switch behaviors according to contexts or rules (Buckley et al., 2009), and hence, is directly responsible for implementing contextual decision making computations. Furthermore, it is crucial for ignoring the presence of distractors during working-memory tasks (Katsuki and Constantinidis, 2012), which highlights its importance to be able to disregard irrelevant information. In the domain of visuomotor saccadic control, where it has been extensively studied, PFC plays a key role mediating the selection of motor actions. In particular, it is believed to act as a bridge between sensory and motor areas, functioning as an integrative center where sensory information is transformed into motor plans that lead to actions or choices (Kim and Shadlen, 1999). Furthermore, it also seems to affect the type of signals it receives by directly modulating early sensory areas. For instance, via top-down control of visual attention (Noudoost et al., 2010; Zhou and Desimone, 2011; Gregoriou et al., 2014). These observations have led researchers to propose that in context-dependent settings, PFC gates the information that it receives based on its relevance. Recently, however, this view has been challenged. Mante and colleagues demonstrated in Mante et al. (2013) that during contextual decisions, relevant inputs are not pre-selected. Instead, the same PFC circuit that performs the sensory-motor transformations seems to disregard the irrelevant information. Their study extended the classic RDM paradigm, incorporating a color dimension into the random dots display. The monkeys had to discriminate, based on the context, either the overall motion or the color of the dots. The activity of the PFC population they recorded from displayed a marked tuning for both color and motion, regardless or the context the monkey was in. However, decision signals seemed to be influenced only by the sensory modality that was relevant in a given context.

This finding alone was of great importance, however, in order to fully understand the system one might want to ask, how are these contextual behaviors actually realized by neural circuits in PFC? The researchers took a step forward and proposed a mechanism that explained how different operations could be flexibly generated by the circuit in order to implement the contextual switch. The modeling framework that was considered takes the view that the dynamics of the circuit supports such computations. In particular, as explained in Machens et al. (2005), it is believed that networks adapt to environmental demands by switching between distinct dynamical behaviors. The large, complex and highly non-linear circuits in the brain exhibit a wide range of different dynamics and could therefore support a whole galaxy of computations. Within a given context, the network implements operations using the dynamics specific of the context. For instance, evidence integration can be realized following the mechanisms implied by a drift-diffusion model (DDM), as in Shadlen et al. (1996); Mazurek et al. (2003), which was used to describe the decision formation process in individual cells [2]. This very same operation can be implemented at the population level, using recurrent neural network models (RNNs) with specific attractor dynamics (Seung, 1996; Bogacz et al., 2006; Mante et al., 2013). Such type of models can incorporate different levels of biophysical realism, which range from abstract rate-based models to networks with integrate and fire neurons (Machens et al., 2005). These can be both excitatory and inhibitory, obeying Dale's law, and incorporate different synaptic currents (Wang, 2002). Regardless of the level of abstraction chosen, the goal of this modeling approach is to provide a mechanistic account for flexible choice generation and importantly, to capture the dynamics of the decision formation process at the level of the whole population. These models have proven to be extremely useful in order to gain intuitions about the operations that may be implemented by neural circuits. However, a limitation that they present is that they are typically hand-crafted and that they cannot reproduce the heterogenous responses found in areas such as PFC. In this thesis, we aim to bridge this gap, by proposing a dynamical model that is able to lawfully describe the whole period of decision formation –providing a mechanistic account of the process– and at the same time, can be fit to capture all the complexity of the individual neuron responses.

[2] although, as we explained, the validity of this model has been recently challenged (Latimer et al., 2015)

Chapter 2

Methods review

In this chapter we review existing methods that have been applied in neuroscience with the goal of extracting low-dimensional structure from recordings of populations of neurons. To the extent of the analysis performed in this thesis, we found that a linear account was a very good description of the neural data under our study. We therefore restrict this review to linear models and methods that capture linear relationships to experimental covariates or inputs. After this, we introduce the two models we used to analyze PFC data from the study by Mante et al. (2013). The first, a novel jointly-factored linear model (JF) which incorporates input covariates, and the second, a Linear Dynamical Systems model (LDS), where input influences are also learned. The mathematical formulation we use to describe each method may differ from that used in the original sources, but it will allow us to work on a common formalism to compare across models[1].

2.1 Linear dimensionality reduction methods

The goal of dimensionality reduction methods is to identify a low-dimensional representation of high-dimensional data. In the case of neural recordings, the underlying assumption is that the activity of individual neurons in a population share some common structure. This could be due to anatomical constraints in the circuit or because of common factors influencing the activity of individual neurons, leading to the emergence of patterns of co-variation between them (Pang et al., 2016). In this section we perform a non-comprehensive review of dimensionality reduction techniques previously applied in neuroscientific studies. For an exhaustive review we refer to Cunningham and Ghahramani (2015) and for a general overview of different methods to Cunningham and Yu (2014); Pang et al. (2016).

[1] This chapter is based on initial notes by Maneesh Sahani.

2.1.1 Matrix factorization and PCA

Considering a matrix $Y \in \mathbb{R}^{N \times T}$, here representing T observations of N-dimensional vectors, where N is the number of neurons in our recordings and T could be trials, conditions or time. A least squared error rank-D approximate factorization

$$\underset{N\times T}{Y} \approx \underset{N\times D}{C}\ \underset{D\times T}{X}$$

may be found by SVD or equivalently by selecting the leading D eigenvectors of $YY^\intercal$ in C and choosing $X = C^\intercal Y$, where C is an orthogonal matrix. This solution is the well known matrix approximation lemma or Eckart-Young-Mirsky theorem. If Y is centered, with the row means subtracted, this is equivalent to performing PCA. The eigenvalues (or singular values) give the variance (or square-root variance) in Y captured by each component. We refer to C as the 'basis' for Y, or the 'loadings' and to X as the 'scores', 'coordinates' or 'latents'. We can think about this type of decomposition as explaining the activity of a neuron n during a bunch of observations $1...T$ via a linear combination of D latent causes, which are shared across all neurons. Note that this solution is not unique for a given rank D, but it does uniquely nest components in terms of the reduction in approximation error (so that the first D' columns of C are the solution to the rank-D' problem). A low-rank approximation of the data –in fact, an optimal one, as given by the Eckart-Young-Mirsky theorem– can therefore be computed by projecting the original data onto the subspace spanned by the columns of C using the projection matrix $CC^\intercal$

$$\underset{N\times T}{\hat{Y}} \approx \underset{N\times D}{C}\ \underset{D\times N}{C^\intercal}\ \underset{N\times T}{Y}$$

Thinking about the latents as a compressed version of Y, this expression returns an approximate reconstruction of the data. The low-rank approximation found is such that the squared error between the original data and the reconstructed data is minimized. This is often used to define the PCA objective

$$L_{PCA} = ||Y - CC^\intercal Y||_F^2 \qquad s.t.\ C^\intercal C = I$$

where F indicates the Frobenius norm. In this setting, $C^\intercal$ is viewed as a decoding matrix that compresses the data and C as the matrix required for reconstruction or encoding (Kobak et al., 2016).

2.1.2 LDA and demixing PCA

Demixing PCA (dPCA) is a method related to both PCA and linear discriminant analysis (LDA) (Kobak et al., 2016). It is based on a similar objective function to PCA (see previous section), but the decoding and encoding mappings differ and the reconstruction is not on the neural activity, but on the data averaged over a subset of the task parameters

$$L_{dPCA} = \sum_{l}^{L} ||Y_l - C_l B_l Y||_F^2$$

where marginalization for the subset l involves averaging out the rest of the parameters $\neg l$. This provides a representation of the data that separates the effects associated to parameter l –that is, separating the possible parameter values or classes $1...M$– without mixing variance from the other parameters $\neg l$. Having different encoders and decoders gives flexibility in the mappings found. In this way, a demixed representation of parameter l classes is obtained along the decoder axis. At the same time, performing the reconstruction on the averages allows to find data representations along the encoding axis that capture parameter class means. The decoder and encoder axes together minimize the reconstruction error between the original data and the parameter class means. The dPCA objective arises from assuming the following decomposition on the (centered) data matrix

$$Y = \sum_{l} Y_l + Y_{noise}$$

All the terms of the decomposition are uncorrelated, therefore, the covariance of the data can be linearly decomposed into the sum of covariance matrices from each average, or marginalization

$$Cov = \sum_{l} Cov_l + Cov_{noise}$$

This is equivalent to the variance/covariance decomposition done in ANOVA/MANOVA.

The rows of B_l form a basis, the demixed PCs, that capture variance specific to parameter l. Therefore, a whole set of directions are obtained which reflect variance related to parameter l. As we will see, these dimensions will also be ordered by the amount of variance explained. To identify the encoding and decoding matrices C_l and B_l –whose product is of rank $D < N$– each individual term of the loss can be optimized separately by casting it as a regression problem with a rank constraint. This is called reduced rank regression (RRR). We also use RRR for the type of decomposition introduced in our new method, the Joint Factorization model. Therefore, it is worth describing it in detail. Following the explanation in Kobak et al. (2016):

1. We note first that each loss $||Y_l - WY||^2$, for a general matrix $W = C_l B_l$, denotes a classic linear regression problem with ordinary least squares (OLS) solution $W_{OLS} = Y_l Y^{\intercal}(YY^{\intercal})^{-1}$. Requiring that W is low-rank, W_D, the solution can be found via SVD.

2. This can be seen by writing $Y_l - W_D Y = (Y_l - W_{OLS}Y) + (W_{OLS}Y - W_D Y)$. The first term are the residuals given an arbitrary linear transformation of Y. As it is

well known, the residuals are orthogonal to the regressors Y. The second term, which is a linear combination of the regressors $(W_{OLS} - W_D)Y$, will therefore be orthogonal to the residuals. Given the orthogonality of the two terms, we can rewrite $||Y_l - W_D Y||_F^2 = ||Y_l - W_{OLS}Y||_F^2 + ||W_{OLS}Y - W_D Y||_F^2$, which separates the loss into the error due to the least squares fit plus an additional penalty due to the enforcement of the rank constraint. The first term does not depend on W_D, therefore, the problem reduces to minimizing the second term.

3. We seek then the best rank-D approximation of $W_{OLS}Y$. As explained in the previous section, the Eckart-Young-Mirsky theorem guarantees that this is given by its first D principal components U_D. Therefore, $W_{OLS}Y \approx U_D U_D^\intercal (W_{OLS}Y)$, so $W_D = U_D U_D^\intercal W_{OLS}$.

4. Finally, we take $C_l = U_D$ and $B_l = U_D^\intercal W_{OLS}$.

Note that even though this procedure leads to components that are orthogonal for each task-parameter l, given by the rows in B_l, the components across different task parameters typically are not, as they are identified independently for each marginalization of the data.

In the case of working with sequentially recorded data, that is, data that has not been simultaneously recorded, the data matrix is replaced with a matrix containing the neuron's PSTHs. The algorithm to perform the optimization is as described before, but with the OLS full-rank solution rewritten so that a term that accounts for the noise covariance is shown explicitly. As the full covariance matrix in this case is not available, this is replaced by a diagonal matrix containing the individual neuron noise variances.

The relationship between dPCA and Linear Discriminant Analysis (LDA) can be understood by considering a separate LDA problem for each marginalization of the data Y_l (Kobak et al., 2016). The goal of LDA is to find a linear projection that has high variance in Cov_l and low variance in $Cov_{\neg l} = Cov - Cov_l$. This is achieve by maximizing

$$tr\left[B_l Cov_l B_l^\intercal \left(B_l Cov_{\neg l} B_l^\intercal\right)^{-1}\right]$$

where B_l contains the discriminant axes in rows. The solution is given by the leading eigenvectors of $Cov_{\neg l}^{-1} Cov_l$, or $Cov^{-1} Cov_l$. Casting LDA as a reduced-rank regression problem, the objective can be written as

$$L_{LDA} = \sum_{l}^{L} ||G_l - C_l B_l Y||_F^2$$

where G_l is an indicator matrix that assigns a parameter class for each data point. Comparing this expression with the dPCA objective, one can understand the main difference between the two methods. LDA looks for decoders that allow to reconstruct class identity (as encoded by G_l) whereas dPCA looks for decoders that allow to reconstruct class means (as encoded by Y_l). Within this formalism, it is possible to derive a decoder for dPCA, which is in this case given by the eigenvectors of $Cov^{-1}Cov_l^2$.

For a more detailed description of the method and the concepts just discussed in here, we refer the reader to Kobak et al. (2016).

Finally, for later comparison with other methods, we would like to make explicit the way dPCA decomposes a given data matrix (ignoring the noise term)

$$\begin{aligned}\underset{N\times T}{\hat{Y}^k} &\approx \sum_l \underset{N\times D}{C_l}\ \underset{D\times N}{B_l}\ \underset{N\times T}{Y^k} \\ &= \sum_l \underset{N\times D}{C_l}\ \underset{D\times T}{X_l^k}\end{aligned}$$

in this case Y contains the PSTHs of N neurons for a given condition k. The rank of the encoding and decoding matrices is given by D and can be different for each parameter l.

2.1.3 Targeted Dimensionality Reduction

Mante and colleagues (Mante et al., 2013) introduced a regression-based method to perform dimensionality reduction. Unlike PCA, which finds a basis with components sorted by the amount of total variance captured in the data, TDR identifies dimensions that specifically capture task-related variance, regardless of how much of the total variability they explain. The way it works is by performing linear regression on the neuronal responses using a set L of experimentally defined external parameters as regressors, plus a bias $[1, u_1, ..., u_L]$, which are believed to influence the activity of the neurons in the task. In particular, a different regression problem is solved for each of the N neurons i during R recorded trials and at a given moment in time t,

$$\underset{R\times 1}{y^i(t)} \approx \underset{R\times(L+1)}{U}\ \underset{(L+1)\times 1}{\boldsymbol{\beta}^i(t)}$$

where the matrix U contains in its rows the value of the external inputs for a given trial r, $[1, \boldsymbol{u}_r^\intercal]$. This was performed independently for each neuron, as the neurons were not simultaneously recorded, so the exact amount of trials varied across units.

Taking the components of $\boldsymbol{\beta}^i(t)$ across all neurons, an N dimensional regression weight vector $\boldsymbol{\beta}^l(t)$ is formed for each input and time. These are then "de-noised" by projecting them into a low D-dimensional space defined by PCA (see 2.1.1) on the averaged population responses

$$\underset{N\times 1}{\hat{\boldsymbol{\beta}}^l(t)} \approx \underset{N\times D}{C}\ \underset{D\times N}{C^\intercal}\ \underset{N\times 1}{\boldsymbol{\beta}^l(t)}$$

Out of all the vectors computed at each point in time, the one with the highest norm is selected, obtaining a unique vector per input l

$$\hat{\boldsymbol{\beta}}^l(t_{max}) = \max||\hat{\boldsymbol{\beta}}^l(t)||^2 \qquad t = 1..T$$

Finally, the set of $L+1$ selected vectors are orthogonalized via QR-decomposition. This defines an $(L+1)$-dimensional orthonormal basis $\hat{B} = [\hat{\boldsymbol{\beta}}^1(t_{max}), ..., \hat{\boldsymbol{\beta}}^{L+1}(t_{max})]$ where each axis represents an independent component of task-related variance (i.e. they are 'demixed' components). Population PSTHs, averaged for each experimental condition k, can then be projected into this subspace

$$\underset{(L+1)\times T}{X^k} = \underset{(L+1)\times N}{\hat{B}^\intercal} \underset{N\times T}{Y^k}$$

which give a small set of $L+1$ 'latent' variables X which summarize the evolution of the neural responses in time. The latent time courses define population trajectories in a low dimensional task-relevant subspace.

This method, in principle, is limited to have a unique dimension per parameter l. However, there is no reason to believe that input influences are allocated to single directions in neural space. Demixing PCA, introduced in the previous section, does not suffer from this limitation and seems to outperform TDR in wide range of data sets, both at capturing overall variance and at demixing the components (Kobak et al., 2016). A multidimensional extension of TDR was introduced by Jonathan Pillow, from Princeton University, during the 2017 Cosyne conference. This method might be close to the JF model we are about to introduce.

2.1.4 Probabilistic PCA and Factor Analysis

In PCA, as we saw in 2.1.1, a matrix $Y \in \mathbb{R}^{N\times T}$ containing T observations of N-dimensional vectors is factorized as

$$\underset{N\times T}{Y} \approx \underset{N\times D}{C} \underset{D\times T}{X}$$

Probabilistic PCA (PPCA) formalizes this problem in a probabilistic setting, by defining a generative model of the data. It is assumed that the observations arise from a set of underlying latent causes as

$$\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(C\boldsymbol{x}, \sigma^2 I) \qquad \boldsymbol{x} \sim \mathcal{N}(0, I)$$

with $\boldsymbol{x}$ being a vector containing the latents, as $D < N$ independent gaussian variables. These are linearly combined to give rise to the observations $\boldsymbol{y}$, under uncorrelated and isotropic gaussian noise $\sigma^2 I$. In the limit $\sigma \to 0$, standard PCA is recovered.

The resulting model for the observations –the marginal or observed likelihood– is a correlated gaussian. This can be seen after integrating out the latents from the joint distribution of latents and observations

$$p(\boldsymbol{y}) = \int p(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{x} = \int p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x} = \mathcal{N}(0, CC^\intercal + \sigma^2 I)$$

The correlation structure in $\boldsymbol{y}$ arises from the mixing of the latents through C. The covariance, as this last expression shows, is given by $CC^\intercal$. The ML values for the loadings C and noise σ^2 can be computed by optimizing the likelihood numerically, for instance, via gradient ascent. Alternatively the optimization can be done using EM, an iterative algorithm that we will introduce in a couple of sections.

Factor analysis (FA) is a very similar algorithm to PPCA, with the only distinction being that the observations noise covariance is still diagonal, but not required to be isotropic. Therefore, each dimension in the observations can be modeled with a different noise variance. This feature is particularly appealing in neuroscience settings, as neurons typically present a wide range of variabilities in their responses. PCA and PPCA can lead to biased estimates of the true correlational structure in the data, as these methods will tend to identify dimensions that align to the directions with high noise variance (Yu et al., 2009). To account for this, the activity of each neuron can be normalized, for instance, by z-scoring the responses (as we will do on the PFC data we will analyze in the next chapter). The FA model reads

$$\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(C\boldsymbol{x}, D) \qquad \boldsymbol{x} \sim \mathcal{N}(0, I)$$

where D is a diagonal matrix. The likelihood of the observations is

$$p(\boldsymbol{y}) = \mathcal{N}(0, CC^\intercal + D)$$

The anisotropy in the noise means that this model is not rotationally invariant, unlike PPCA. However, it makes it scale invariant, while PPCA is not. The parameters can be found by direct optimization of the likelihood or using EM. The likelihood can have multiple local optima and convergence to the global optimum is not guaranteed by these methods.

Factor analysis has been widely used in neuroscience, either as a tool to understand neural data (Williamson et al., 2016) or in the domain of brain machine interfaces (BMI) to decode neural activity in order to control neural prostheses, such as a computer cursor or a robotic arm (Santhanam et al., 2009).

An important extension of FA that has also been extensively applied to neural data is GPFA (Gaussian Process Factor Analysis) (Yu et al., 2009). Unlike PCA, PPCA and FA, which are static methods, GPFA captures the correlational structure of the data in time. The model can be seen as a collection of identical FA models, which are temporally linked via a gaussian process (GP). The GP prior connects the states of the latent variables across time by assuming an independent gaussian process for each dimension. The observations are generated via a linear combination of the latents after corruption with some gaussian noise, as in FA. The different smoothing properties of the neural trajectories can be captured by learning different characteristic time scales for each of the latent processes. The time scales parametrize a kernel function that specifies the covariance structure of all point across time.

Extensions of GPFA include PGPFA, a version which incorporates Poisson instead of Gaussian observations (Zhao and Park, 2017; Duncker and Sahani, 2018)

Other important latent variable methods that have been extensively used to model neural data are linear dynamical systems (LDS). These models are also designed to capture the correlational structure of the data in time. The main difference with respect to GPFA is that the latent variables are not evolving independently and that their state is assumed to follow a lawful time evolution, as specified by the parameters of a dynamical system. We will discuss these types of models at length in a couple of sections. There is a extensive list of related methods, such as exponential family PCA, variational GPFA, P-GPLVM, state space models with switching dynamics (HSLDS) (Lawrence, 2004; Wu et al., 2017; Petreska et al., 2011), etc which we will not cover in this brief review.

2.2 The Joint Factorization model

We start once again as in section 2.1.1 by considering a matrix $Y \in \mathbb{R}^{N \times T}$, which can be factorized as

$$\underset{N \times T}{Y} \approx \underset{N \times D}{C} \; \underset{D \times T}{X}$$

where N is the number of neurons in our recordings and T indicates time.

Let's now consider a set of K matrices $\{Y^k\}$, with K indicating the number of trials, or conditions, if working with averaged data. A joint low-rank factorization uses a common basis

$$Y^k \approx CX^k$$

which can easily be found by concatenating the matrices Y^k horizontally $\mathcal{Y} = [Y^1, ..., Y^K]$ and computing a factorization as above

$$\underset{N \times KT}{\mathcal{Y}} \approx \underset{N \times D}{C} \; \underset{D \times KT}{\mathcal{X}}$$

with X^k contained in the kth $D \times T$ block of $\mathcal{X}$.

This joint factorization allows us to treat whole matrices as data points.

2.2.1 Linear effects

Let's now suppose that in each condition k we can specify a set of L experimentally defined external parameters $\boldsymbol{u}_k^\intercal = [u_{1k}, ..., u_{Lk}]$. Our goal is to seek a joint factorization in which the latents for matrix Y^k depend linearly on $\boldsymbol{u}_k$

$$X^k = B_0 + \sum_l B_l u_{lk}$$

where B_0 and B_l are $D \times T$ matrices. Then

$$Y^k \approx C(B_0 + \sum_l B_l u_{lk})$$

A solution can be found by recasting the problem as a form of reduced-rank regression (RRR). For that, we define the experimental design matrix

$$\underset{(L+1)\times K}{U} = \begin{bmatrix} 1 & 1 & ... & 1 \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & ... & \boldsymbol{u}_K \end{bmatrix}$$

and

$$\underset{(L+1)T\times KT}{\mathcal{U}} = U \otimes I_{T\times T} = \begin{bmatrix} I_{T\times T} & I_{T\times T} & ... & I_{T\times T} \\ u_{11}I_{T\times T} & u_{12}I_{T\times T} & ... & u_{1K}I_{T\times T} \\ & & \vdots & \\ u_{L1}I_{T\times T} & u_{L2}I_{T\times T} & ... & u_{LK}I_{T\times T} \end{bmatrix}$$

where $\otimes$ is the Kronecker product and $I_{T\times T}$ a $T \times T$ identity matrix. Writing $\mathcal{B} = [B_0, B_1, ..., B_L]$ we obtain

$$\underset{N\times KT}{\mathcal{Y}} \approx \underset{N\times D}{C} \underset{D\times(L+1)T}{\mathcal{B}} \underset{(L+1)T\times KT}{\mathcal{U}}$$

We assume for now that the input influences $\mathcal{U}$ are known, so the goal is to identify $C\mathcal{B}$, which has rank D. This is a similar setting to RRR, but with a larger input space than the original U.

As we explained before 2.1.2, the solution to RRR can be seen by considering properties of the solution to (full-rank) least-squares linear regression. Let $\mathcal{W} = \mathcal{Y}\mathcal{U}^\intercal(\mathcal{U}\mathcal{U}^\intercal)^{-1}$ so that

$$\underset{N\times KT}{\mathcal{Y}} \approx \underset{N\times(L+1)T}{\mathcal{W}} \underset{(L+1)T\times KT}{\mathcal{U}}$$

in a least-squares sense. The goal is to find a rank-D factorization $C\mathcal{B}$ of $\mathcal{W}$ that minimises the squared approximation error in $\mathcal{Y}$ (but not necessarily the approximation error in $\mathcal{W}$ itself). This is obtained by a least-squares rank-D factorization of $\mathcal{W}\mathcal{U}$ (easily proven by noting that $\mathcal{Y} - \mathcal{W}\mathcal{U}$ is orthogonal to $\mathcal{W}\mathcal{U}$). In particular, we choose C to contain the leading D eigenvectors of $(\mathcal{W}\mathcal{U})(\mathcal{W}\mathcal{U})^\intercal = \mathcal{W}\mathcal{U}\mathcal{Y}^\intercal = \mathcal{Y}\mathcal{U}^\intercal(\mathcal{U}\mathcal{U}^\intercal)^{-1}\mathcal{U}\mathcal{Y}^\intercal$ and set $\mathcal{B} = C^{\intercal}\mathcal{W}$.

Having identified $C\mathcal{B}$, we can obtain the factorization of the data for a given condition k as

$$\underset{N\times T}{Y^k} \approx \underset{N\times D}{C} \underset{D\times(L+1)T}{\mathcal{B}} \underset{(L+1)T\times T}{\mathcal{U}^k}$$

Let's compare this result with the type of factorization obtained with TDR. If we had considered the average of the regression coefficients $\beta(t)$ across time –instead of taking the one with the largest norm– and used PCA to orthogonalise the vectors (possibly followed by a rotation to demix influences), instead of the QR decompostion, then this is almost equivalent to choosing $D = L + 1$ in the method described here (assuming centered data). However, in the current approach the basis would be based (roughly) on PCA of all the regression vectors at all times taken as a set – and so a single parameter that contributed high variance in different dimensions over time could dominate. To account for this, the activity of each neuron can be normalized, for instance, by z-scoring the responses (as we will do on the PFC data we will analyze in the next chapter).

Finally, as we will see below, these equations can be seen as a form of tensor decomposition for Y_{NTK}, which is a third-order tensor. A similar method has been applied in Seely et al. (2016). The goal of the study, however, was simply to identify the tensor unfoldings that offered a better reconstruction of the data. This was done by applying SVD to either $\underset{N\times KT}{\mathcal{Y}}$ or $\underset{K\times NT}{\mathcal{Y}}$ [2]. This study highlighted the advantage of exploiting the tensor structure of the neural data and, as we noted with our JF method, treat whole matrices as data points.

2.2.2 Collinear effects

In conventional experimental design, a fixed set of values of the inputs or control parameters u_{kl} will often be repeated across conditions k. Let the set of M_l possible values for the lth control parameter be p_{lm} for $m = 1...M_l$. Then we can write the experimental design matrix as

$$\underset{(L+1)\times K}{U} = \begin{bmatrix} 1 & 1 & ... & 1 \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & ... & \boldsymbol{u}_K \end{bmatrix} = \underset{(L+1)\times(1+\sum M_l)}{P} \underset{(1+\sum M_l)\times K}{Q}$$

where P contains all the possible parameter values:

[2] The unfolding $\underset{N\times KT}{\mathcal{Y}}$ is viewed as a reconstruction on the neurons based on a basis of K by T patterns and the one on $\underset{K\times NT}{\mathcal{Y}}$ is viewed as a reconstruction on the conditions based on a basis of N by T patterns. If the first set of basis achieves a lower reconstruction error, this implies a higher degree of correlations across neurons than across conditions and conversely for the second set. Externally driven systems are expected to be of the first type, and dynamical systems of the second. The type of structure expected to arise in brain areas that are not pure sensory systems, perhaps not that surprisingly, is that of a dynamical system which is input-driven, which implies a activity constraints across both neurons and conditions.

$$P = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & p_{11} & \dots & p_{1M_1} & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & p_{21} & \dots & p_{2M_2} & \dots & 0 & \dots & 0 \\ & & & & & \vdots & & & & & \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & p_{L1} & \dots & p_{LM_L} \end{bmatrix}$$

and the columns of the sparse matrix Q each contain exactly $(L+1)$ 1s that select the appropriate values for the kth condition.
Now, consider the case that the values p_{lm} are unknown (or, equivalently, allow for an arbitrary non-linear mapping from the experimentally-selected values to the linear weights u_{lk}). We refer to this case as expressing *collinear* effects, as the exact linear dependence on experimental parameters is broken, but different settings m of the same parameters l still influence Y_k along a common direction B_l (but possibly different at each point in time). Let us return to the original linear effect equation and rewrite it as follows:

$$Y^k \approx C(B_0 + \sum_l B_l u_{lk})$$

$$\begin{aligned} \Rightarrow \underset{NT\times 1}{\text{vec}(Y^k)} &\approx \underset{NT\times(L+1)}{[\text{vec}(CB_0), \text{vec}(CB_1), ..., \text{vec}(CB_L)]} \underset{(L+1)\times 1}{\begin{bmatrix} 1 \\ \boldsymbol{u}_k \end{bmatrix}} \\ \Rightarrow \underset{NT\times K}{\text{vec}_{[NT]}(\mathcal{Y})} &\approx \underset{NT\times(L+1)}{\text{vec}_{[NT]}(C\mathcal{B})} \underset{(L+1)\times K}{U} = \text{vec}_{[NT]}(C\mathcal{B})PQ \\ \Rightarrow \underset{NTK\times 1}{\text{vec}(\mathcal{Y})} &\approx \underset{NTK\times(L+1)(1+\Sigma M_l)}{\left(Q^\intercal \otimes \text{vec}_{[NT]}(C\mathcal{B})\right)} \underset{(L+1)(1+\Sigma M_l)\times 1}{\text{vec}(P)} \end{aligned}$$

where the operator $\text{vec}_{[r]}(A)$ rearranges the matrix A in a column-first order to have r rows – equivalent to `reshape(A,r,[])` in MATLAB.

This equation can be solved for P in a least squares sense if CB is known. However, the least-squares solution may yield a choice of P which violates the experiment design by introducing non-zero values in impossible places. Thus, we introduce a sparse matrix and a dense vector

$$S = \begin{bmatrix} 1 & 0 & 0 & \\ 0 & 0 & 0 & \\ \vdots & \vdots & \vdots & \dots \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \\ 0 & 1 & 0 & \\ \vdots & \vdots & \vdots & \dots \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \\ \vdots & \vdots & 1 & \dots \\ 0 & 0 & 0 & \end{bmatrix} \begin{matrix} \\ \\ \left.\begin{matrix} \\ \end{matrix}\right\}{\scriptstyle (1+\Sigma M_l)} \\ \\ \left.\begin{matrix} \\ \end{matrix}\right\}{\scriptstyle (1+\Sigma M_l)} \\ \\ \left.\begin{matrix} \\ \end{matrix}\right\}{\scriptstyle (1+\Sigma M_l)} \\ \\ \end{matrix} \qquad \boldsymbol{p} = \begin{bmatrix} 1 \\ p_{11} \\ \vdots \\ p_{1M_1} \\ p_{21} \\ \vdots \\ p_{2M_2} \\ \vdots \end{bmatrix}$$

such that $\operatorname{vec}(P) = S\boldsymbol{p}$. Now $\boldsymbol{p}$ may be found by solving

$$\Rightarrow \underset{NTK\times 1}{\operatorname{vec}(\mathcal{Y})} \approx \underset{NTK\times(L+1)(1+\Sigma M_l)}{\left(Q^\intercal \otimes \operatorname{vec}_{[NT]}(C\mathcal{B})\right)} \underset{(L+1)(1+\Sigma M_l)\times(1+\Sigma M_l)}{S} \underset{(1+\Sigma M_l)\times 1}{\boldsymbol{p}}$$

and so all of C, $\mathcal{B}$ and $\boldsymbol{p}$ can be found by alternating least squares (ALS).

The notation used in here can be rather overwhelming, but it simplifies greatly when using a tensor formalism, as we will in the next section.

2.2.3 Tensor formalism

The equations above may be expressed more compactly using Cartesian tensor notation. We write

$$\begin{aligned} \boldsymbol{Y}_{ntk} &= [Y_k]_{nt} \\ \boldsymbol{C}_{nd} &= [C]_{nd} \\ \boldsymbol{B}_{dtl} &= [B_{l-1}]_{dt} \\ \boldsymbol{U}_{lk} &= [U]_{lk} \\ \boldsymbol{P}_{m} &= [\boldsymbol{p}]_{m} \end{aligned}$$

and $\boldsymbol{Q}_{lmk}$ a sparse binary design tensor such that $\boldsymbol{U}_{lk} = \boldsymbol{Q}_{lmk}\boldsymbol{P}_m$. Here we adopt the Einstein convention that repeated indices imply summation. In this formalism the low-rank approximation is seen to be a constrained tensor decomposition

$$\boldsymbol{Y}_{ntk} = \boldsymbol{C}_{nd}\boldsymbol{B}_{dtl}\boldsymbol{U}_{lk} = \boldsymbol{C}_{nd}\boldsymbol{B}_{dtl}\boldsymbol{Q}_{lmk}\boldsymbol{P}_m$$

with $\boldsymbol{Q}$ and $\boldsymbol{P}$ possibly fixed.

Now, we can re-write the ALS algorithm suggested above in matrix form but without the need for Kronecker products, using notation such as $\boldsymbol{Y}_{(n)(tk)}$ to represent a "flattened" tensor (equivalent to MATLAB's `reshape(Y,[n,t*k])`. The operator LSV_D selects the first D left singular vectors of its argument.

$$\begin{aligned}
\boldsymbol{W}_{(nt)(l)} &= \boldsymbol{Y}_{(nt)(k)}\boldsymbol{U}_{lk}^{\intercal}(\boldsymbol{U}_{lk}\boldsymbol{U}_{lk}^{\intercal})^{-1} \\
\boldsymbol{C}_{nd} &= \mathrm{LSV}_D([\boldsymbol{W}_{ntl}\boldsymbol{U}_{lk}]_{(n)(tk)}\boldsymbol{Y}_{(n)(tk)}^{\intercal}) \\
\boldsymbol{B}_{d(tl)} &= \boldsymbol{C}_{nd}^{\intercal}\boldsymbol{W}_{(n)(tl)} \\
\boldsymbol{Z} &= [\boldsymbol{C}_{nd}\boldsymbol{B}_{dtl}\boldsymbol{Q}_{lmk}]_{(ntk)(m)} \\
\boldsymbol{P}_m &= (\boldsymbol{Z}^{\intercal}\boldsymbol{Z})^{-1}\boldsymbol{Z}\boldsymbol{Y}_{(ntk)}
\end{aligned}$$

and iterate to convergence.

Alternatively, it may be possible to use an approach resembling tensor low-rank regression (tensor RRR) (Rabusseau and Kadri, 2016) (but note that there might be an error in their method [3]) or use other forms of tensor decompositions, such as CP tensor decomposition or HO-SVD. This would allow to incorporate multilinear rank constraints. The rank constraint we are considering here apply to the first (neural) dimension, but it would be interesting to incorporate them, for instance, along the time dimension and also to impose different ranks for the different parameters.

The JF decomposition is related to other factorization methods based on CANDECOMP/PARAFAC tensor decomposition (Williams et al., 2018) or on non-negative matrix factorization (Mackevicius et al., 2018), which have been developed to analyze neural data sets. However, there are important differences. Our method focuses on a particular tensor unfolding $\boldsymbol{Y}_{(n)(tk)}$, instead of performing a decomposition on the whole three-way tensor as in Williams et al. (2018). This approach, in principle, does not favor extracting common structure across conditions. However, we do achieve this by imposing shared structure in the input specifications, via $\boldsymbol{U}_{lk} = \boldsymbol{Q}_{lmk}\boldsymbol{P}_m$, as we consider that a fixed set of input values are repeated across conditions. Furthermore, we perform a tensor decomposition within a regression setting. This semi-supervised approach, unlike their fully unsupervised method, allows to incorporate information about the task, which guides the factorization.

2.2.4 Extensions

Different specifications and extensions are considered in this section.

[3] There seems to be a problem in their proof, as they assume that the data came from exactly the model –in particular that the noise is iid (and so constant variance). In this case any covariance in the fibres of the output tensors must come from the shared regression input, and so PCA (which is effectively what they do) will indeed pick the right directions. However, as soon as you add shaped noise –covariance from a different source, or even independent but unequal variance as in FA– the method does not work (personal communication with Maneesh Sahani).

Following the setting where the input values are drawn from a fixed set and are repeated across conditions k, the assumption of collinearity may be avoided by replacing the single parameter u_{lk} by an M_l-length indicator vector. Therefore, the influences are categorical. Each value of the parameter l will then be associated with a different latent matrix B_{lm} (for $m = 1...M_l$). This model reduces (after centering the data matrices Y_k and dispensing with the base component B_0) to joint factorization of the *mean* data matrices $\bar{Y}_{lm} = \frac{1}{N_{u_{lk}=p_{lm}}} \sum_{k:u_{lk}=p_{lm}} Y^k$. This approach is related to dPCA, but instead of performing an independent factorization for each matrix l, all the mean matrices are factorized jointly.

The basis identified with this method can be rotated post-hoc or priors specifications can be introduce to promote sparsity in C and/or in the rows of B.

The effects of a "context" parameter can be modeled as inducing a switch between different patterns of linear dependence. In this case, observation matrices Y^k for different settings of the context parameter would first be (horizontally) concatenated and then the linear-effects analysis performed. The resulting matrices B_l would then contain the linear effects in the different contexts concatenated. Alternatively, one may seek a model in which the linear effects are the same in each context, with only B_0 differing. This case would combine a categorical effect for the context, with linear or collinear effects for the remaining parameters.

Finally, multiple extensions may be considered, like incorporating non-linear effects, defining prior specifications and noise models and assuming generalized linear outputs of the form $\mathcal{Y} \sim \text{ExpFam}[g(\mathcal{CBU})]$.

2.3 Linear Dynamical Systems

Linear dynamical systems (LDS) are probabilistic latent variable models also known as linear Gaussian state-space models (GSSMs). Linear dynamical systems are the continuous-state analogues of hidden Markov models (HMMs). The dynamics are typically considered stationary, with the parameters fixed over time. The data or observations $\boldsymbol{y}_t$ are assumed to be generated by a hidden low-dimensional linear dynamical process which is corrupted by Gaussian noise. This model can be characterized as a factor analysis over time (see 2.1.4), with hidden factors $\mathbf{x}_t$ evolving according to a dynamical rule. The observational model is also assumed to be linear and corrupted by Gaussian noise.

We follow the notation used in Macke et al. (2015) to describe the LDS equations, which specify the prior dynamics over the hidden state and the likelihood of the observations conditioned on the hidden state[4]

[4] we have also used Max Welling's notes on Kalman Filters and teaching material on LDSs by Zoubin Ghahramani, Geoffrey E. Hinton and others.

$$
\begin{aligned}
\mathbf{x}_t &= A\mathbf{x}_{t-1} + B\mathbf{u}_t + \boldsymbol{\eta}_t \\
\mathbf{y}_t &= C\mathbf{x}_t + \mathbf{d} + \mathbf{r}_t
\end{aligned}
$$

where $\boldsymbol{\eta}_t \sim \mathcal{N}(0, Q)$ are called the innovations and $\mathbf{r}_t \sim \mathcal{N}(0, R)$ are the noise in the observations. The transition matrix A dictates the dynamics and B specifies the subspace where the input biases $\mathbf{u}_t$ live. The inputs are also referred to in the literature as control parameters. The loading matrix C maps the hidden state into the observational space and $\boldsymbol{d}$ is a constant bias.

The observational space is N dimensional (in our case, given by the number of neurons recorded). $H < N$ is the size of the hidden state. We denote with K the number of trials (or conditions) and with T the length of each trial –we assume $K = 1$ in here for simplicity. Considering R as a diagonal matrix, the correlational structure of the data is fully determined by the low dimensional hidden state. The observational noise model allows to capture independent noise variances for each neuron –on the diagonal entries of covariance matrix R–, as in FA. We consider the first step as drawn from

$$\boldsymbol{x}_1 \sim \mathcal{N}(A\boldsymbol{x}_0 + B\boldsymbol{u}_1, Q_0)$$

where, for mathematical convenience, we assume that the dynamics acts already on this first step[5]. The noise covariance of the initial state Q_0 can also be set equal to Q to simplify calculations, as we will do in chapter 3.

The LDS can be related to the models previously described in this chapter, in particular, the JF model. As we will see in Chapter 3, the noise-free version of the LDS can be conceptualized as a form of low-rank factorization with additional constraints specified by the dynamics. Previous studies have also formalized the LDS equations within a matrix factorization based framework (Liu and Hauskrecht, 2016; She et al., 2018).

Assuming that the parameters are known, estimation of the hidden state is achieved via the widely used Kalman filtering and smoothing algorithms (Hamilton, 1986). Given the hidden state estimates, parameter estimates can be obtained by direct maximization of the likelihood via non-linear programming methods such as Newton-Raphson (Gupta and Mehra, 1974). Alternatively, this can be achieved via expectation maximization (EM), another widely used algorithm which maximizes a lower bound on the likelihood, the free energy. EM is an iterative procedure that involves performing inference on the posterior distribution over the latents during the E-step and computing parameter estimates in the M-step by maximazing the expected log joint distribution of latents and data over the posterior (Shumway and Stoffer, 1982). The EM steps always increase the likelihood and convergence is guaranteed to a stationary point for an exponential family.

[5] A “true” initial state $\hat{\boldsymbol{x}}_0$ can always be learned by setting $\boldsymbol{x}_0 = A^{-1}\hat{\boldsymbol{x}}_0$ to cancel out the dynamics.

However, the found optima can either be local or global maxima. The convergence can be slow in the latter stages and one may move indefinitely along a ridge, depending on the shape of the likelihood function Wu (1983). Alternatively, the parameters can be learned via other methods, such as spectral learning.

Considering the Markov structure implied by the LDS equations, the log joint probability of latents and observations can be written as

$$\log p(\{\boldsymbol{x}\},\{\boldsymbol{y}\}) = p(\boldsymbol{x}_1)\sum_{t=2}^{T}\log p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\sum_{t=1}^{T}\log p(\boldsymbol{y}_t|\boldsymbol{x}_t)$$

$$\begin{aligned}
\log p(\{\boldsymbol{y}\}|\{\boldsymbol{x}\}) &= -\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{y}_t-[C\boldsymbol{x}_t+\boldsymbol{d}])^\top R^{-1}(\boldsymbol{y}_t-[C\boldsymbol{x}_t+\boldsymbol{d}]) - \frac{1}{2}TN\log(2\pi)\\
&\quad -\frac{T}{2}\log|R|\\
\log p(\{\boldsymbol{x}\}) &= -\frac{1}{2}(\boldsymbol{x}_1-[A\boldsymbol{x}_0+B\boldsymbol{u}_1])^\top Q_0^{-1}(\boldsymbol{x}_1-[A\boldsymbol{x}_0+B\boldsymbol{u}_1])\\
&\quad -\frac{1}{2}\sum_{t=2}^{T}(\boldsymbol{x}_t-[A\boldsymbol{x}_{t-1}+B\boldsymbol{u}_t])^\top Q^{-1}(\boldsymbol{x}_t-[A\boldsymbol{x}_{t-1}+B\boldsymbol{u}_t])\\
&\quad -\frac{1}{2}TH\log(2\pi)-\frac{1}{2}(T-1)\log|Q|-\frac{1}{2}\log|Q_0|
\end{aligned}$$

The posterior over the latents can be specified given the likelihood and the prior using Bayes rule as

$$\log p(\{\boldsymbol{x}\}|\{\boldsymbol{y}\}) = \log p(\{\boldsymbol{y}\}|\{\boldsymbol{x}\}) + \log p(\{\boldsymbol{x}\}) + \mathrm{const}(\{\boldsymbol{x}\})$$

where $\mathrm{const}(\{\boldsymbol{x}\})$ contains terms that do not depend on $\boldsymbol{x}$.

EM updates

EM maximizes a lower bound on the log-likelihood, the free energy

$$\mathcal{F}(\theta) := E_q\left[\log p(\{\boldsymbol{x}\},\{\boldsymbol{y}\}|\theta) - \log q(\{\boldsymbol{x}\})\right] \leq \log p(\{\boldsymbol{y}\}|\theta)$$

with parameters θ. This bound is valid for any distribution over the latents $q(\{\boldsymbol{x}\})$ and is tight if and only if $q(\{\boldsymbol{x}\})$ equals the posterior $p(\{\boldsymbol{x}\}|\{\boldsymbol{y}\},\theta)$, which is our case. In the E-step the posterior distribution over the latents can be estimated via the Kalman smoother, which provides forwards and backwards recursive equations for inference. In the M-step, the free energy is maximized with respect to the parameters θ, given the expression

$$\mathcal{F}(\theta) = \sum_{k=1}^{K}\int p(\{\boldsymbol{x}\}_k|\{\boldsymbol{y}\}_k,\theta)\ \log p(\{\boldsymbol{x}\}_k,\{\boldsymbol{y}\}_k|\theta)$$

$$
\begin{aligned}
&= \sum_{k=1}^{K} \Bigg(\int p(\{\boldsymbol{x}\}_k | \{\boldsymbol{y}\}_k, \theta) \; \log p(\{\boldsymbol{y}\}_k | \{\boldsymbol{x}\}_k, \theta) \\
&\quad + \int p(\{\boldsymbol{x}\}_k | \{\boldsymbol{y}\}_k, \theta) \; \log p(\{\boldsymbol{x} | \theta\}_k) \Bigg) \\
&= \mathcal{Q}_{obs}(C, \boldsymbol{d}, R) + \mathcal{Q}_{dyn}(A, B, Q, Q_0, \boldsymbol{x}_0)
\end{aligned}
$$

where we have neglected the second term of the bound, the entropy, which is constant in the parameters. We have also made explicit the case of having multiple trials or conditions K. Given that our model is fully Gaussian, all parameters can be updated in closed form. In the appendix we provide the expressions for the updates, which can be found in standard machine learning books and teaching notes. Finally, the inputs $\mathbf{u}_t$ can also be learned. The corresponding inputs update equations are provided in the next section.

2.3.1 Learning input influences

The complete input time courses can be learned by considering a simple extension to the standard LDS equations. The free energy expression can be optimized with respect to the inputs by treating them as additional parameters. The corresponding updates can be computed during the M-step. The input biases during condition k, for a given input parameter l at time step t are

$$
u_{l,t}^{k} = \left(\boldsymbol{b}_l^{\intercal} Q^{-1} \boldsymbol{b}_l\right)^{-1} \boldsymbol{b}_l^{\intercal} Q^{-1} \left([\boldsymbol{x}_t^k - A\boldsymbol{x}_{t-1}^k] - B_{\neg l} \boldsymbol{u}_{\neg l,t}^k \right)
$$

where $\neg l$ indicates that the dimension l has been excluded. As we did with the JF model, we can consider the case that a fixed set of values of the inputs u_l^k are repeated across conditions. The input biases can be learned by grouping trials together as

$$
u_{l,t}^{a} = \left(\boldsymbol{b}_l^{\intercal} Q^{-1} \boldsymbol{b}_l\right)^{-1} \frac{1}{K_a} \boldsymbol{b}_l^{\intercal} Q^{-1} \sum_{k \exists K_a}^{K_a} \left([\boldsymbol{x}_t^k - A\boldsymbol{x}_{t-1}^k] - B_{\neg l} \boldsymbol{u}_{\neg l,t}^k \right)
$$

where we consider that the input of type a for parameter l has been presented in a subset K_a of the trials. In our analysis, a would indicate coherence level.

2.3.2 Linear Dynamical Systems with Poisson observations

PLDS is an extensions to the LDS that incorporates Poisson observations Macke et al. (2011, 2015). The model is therefore better suited to capture spike count data. However, given that the noise in the observations is no longer Gaussian and that the mapping from hidden state to observations is non-linear, performing inference and learning is much more complicated in this model. In particular, inference must be computed relying on variational approximation methods, so the optimization process is much slower than for the LDS.

$$\begin{aligned} \mathbf{x}_t &= A\mathbf{x}_{t-1} + B\mathbf{u}_t + \boldsymbol{\eta}_t \\ \mathbf{y}_t &\sim \text{Poisson}\left(\exp\left[C\mathbf{x}_t + \mathbf{d}\right]\right) \end{aligned}$$

where $\boldsymbol{\eta}_t \sim \mathcal{N}(0, Q)$

2.3.2.1 Likelihood for averaged vs simultaneously recorded count data

In Mante's data set most neurons were not simultaneously recorded. Therefore, we derived an expression to fit the PLDS considering the case of having unequal number of trials for the different observations. To account for this discrepancy, we weighted the contribution to the likelihood from every neuron i by the number of trials recorded. We found that this procedure was almost equivalent to fitting the standard model to the condition-averaged count data.

Writing $z_{it}^k = (C\mathbf{x}_t^k)_i + d_i$, we obtained the expression

$$\begin{aligned} \mathcal{L} &= \sum_{t,k,i,m_i^k} \log P(y_{it}^{m_i^k} | \mathbf{x}_t^k) \\ &= \sum_{t,k,i,m_i^k} \log\left(\frac{\exp(z_{it}^k)^{y_{it}^{m_i^k}}}{y_{it}^{m_i^k}!} \exp\Big(-\exp(z_{it}^k)\Big)\right) \\ &= \sum_{t,k,i}\Big\{(\sum_{m_i^k} y_{it}^{m_i^k}) z_{it}^k - \sum_{m_i^k} \exp(z_{it}^k) - \sum_{m_i^k} \log(y_{it}^{m_i^k}!)\Big\} \\ &\qquad\qquad \text{total spike counts} \\ &= \sum_{t,k,i}\left\{\left(\frac{\sum_{m_i^k} y_{it}^{m_i^k}}{M_{it}^k}\right) z_{it}^k - \frac{M_{it}^k}{M_{it}^k} \exp(z_{it}^k) - \left(\frac{\sum_{m_i^k} \log(y_{it}^{m_i^k}!)}{M_{it}^k}\right)\right\} \\ &\qquad\qquad \text{mean spike counts} \end{aligned}$$

where M_i^k indicates the number of trials recorded for neuron i under condition k.

Chapter 3

Linear Dynamics of PFC circuits during Contextual Decision Making

In our everyday lives, we constantly face situations where we must adapt our actions based on the immediate demands and pressures of our environment. How can humans and other animals flexibly adapt their behavior in the short time scales involved? and in particular, how does the brain generate the appropriate commands to produce those actions? Given the overwhelming stream of sensory information that we are constantly facing, it is a mystery how the relevant information can make its way through the mesh of interconnected networks in our brain, finding the trigger that will select the appropriate computations to drive behavior. Understanding how neural circuits are endowed with such flexibility is the question we would like to resolve. An important step towards it was made a few years ago thanks to a study by Valerio Mante and colleagues, from Bill Newsome's lab at Stanford University (Mante et al., 2013). The goal of the study was to understand how populations of neurons in prefrontal cortex (PFC) select sensory inputs and accumulate relevant evidence towards a choice. In this thesis, we build on their work by first, providing new insights into the nature of the computations implemented in this PFC circuit, and second, by offering a different methodological approach that helps unravel the complexity of the neural data observed under such computations.

3.1 Context-dependent computations by recurrent dynamics in PFC

In the study by Mante et al. (2013) two monkeys were trained to discriminate noisy sensory evidence, either the direction of motion or the color of a random-dot visual stimulus. An external cue that changed depending on the context indicated which of the two stimulus features was relevant (figure 3.1). The monkeys were instructed to report their choices with a saccade to one of two visual targets. Extracellular recordings of neurons in PFC, mainly from the Frontal Eye Fields (FEF) and surrounding areas, were performed during behavior. The period of the task the researchers focused on was the stimulus presentation, when the process of sensory evidence accumulation must take place. This corresponded to the time from 100ms to 850ms, starting 100ms after dots onset and ending 100ms after dots offset. The activity of the individual neurons recorded presented a rich diversity of responses, reflecting a highly heterogeneous and "mixed" representation of task variables – which is characteristic of PFC (Rigotti et al., 2013). These included the decision of the monkey and the motion and color coherence values –indicating both the strength and the direction of the sensory evidence. Importantly, the two sensory input signals were reflected in the population activity regardless of the context, that is, regardless of whether they were relevant for the decision or not. This was surprising, given that FEF is predominantly a motor related area (Kim and Shadlen, 1999). However, it has been implicated in a multitude of other functions, such as attention (Brooks and List, 2006; Zhou and Desimone, 2011; Gregoriou et al., 2014), sensory integration and motor planing (Kim and Shadlen, 1999; Glaser et al., 2018). It is also known to respond to a diversity of visual features (Cassanello et al., 2008; Barborica and Ferrera, 2003). FEF neurons, importantly, are selective to motion coherence –reflecting both motion strength and direction– and also to motion coherence magnitude –reflecting motion strength, but not direction– (Ding and Gold, 2012). In this study, however, the focus was on the directional coherence signals.

The fact that the irrelevant information made its way to FEF ruled out gating of inputs as a possible mechanism for the selective accumulation of evidence. A PCA-like analysis using targeted dimensionality reduction (TDR, see section 2.1.3) revealed a separable representation of task variables at the level of the population. The trajectories in high dimensional neural space were projected onto a low-dimensional subspace that reflected the variance of a set of task-related variables: choice, strength and direction of motion evidence, strength and direction of color evidence and context (figure 3.2). The main observations extracted from the population analysis were that:

First, trajectories moved in opposite directions along the decision dimension, reflecting the choice of the monkey. This gradual movement along the decision axis was suggestive of a slow integrative process that accumulates the sensory evidence during the presentation of the visual stimulus. The direction of movement (left or right in figure 3.2) indicates where the relevant evidence pointed to, either towards the receptive field (RF) of the recorded neurons (Tin or choice1) or away from it (Tout or choice2). Positive color/motion inputs, by convention, indicated evidence towards choice1 and correspondingly negative inputs indicated evidence towards choice2 (figure 3.1b).

Second, it was found that the input also exerts a strong effect on the trajectories, deflecting them along the input axes, with the amplitude of such deflections given by the coherence level or strength of the evidence. The direction (up or down in figure 3.2) reflected the sign of the evidence, that is, indicated where the color/motion evidence pointed to (choice1 or choice2). The presence of arcs was taken to imply that the signals along the input axes are transient, as the trajectories deviate first and eventually go back to the starting point. However, the effect on the choice axis persisted even after

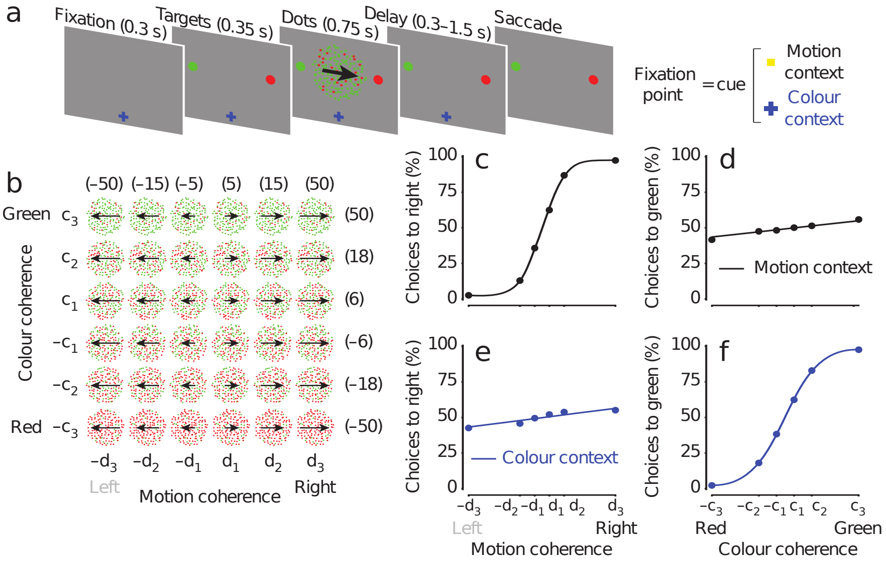


Figure 3.1: a) Structure of the contextual decision making task b) Stimulus space, consisting of six color and motion coherence values, giving 36 different conditions for each possible color and motion combination. Coherence values were grouped in 3 different levels according to coherence strength (weak, medium, strong). The signs, either positive or negative, indicate the direction of the evidence: Tin, pointing towards the RF of the recorded neurons or Tout, pointing away from it (c-f) Psychometric curves. The monkeys were able to perform the contextual DM task, given that the irrelevant inputs had almost no impact on the monkey's choices (d,e). Reprinted by permission from Mante et al. (2013)

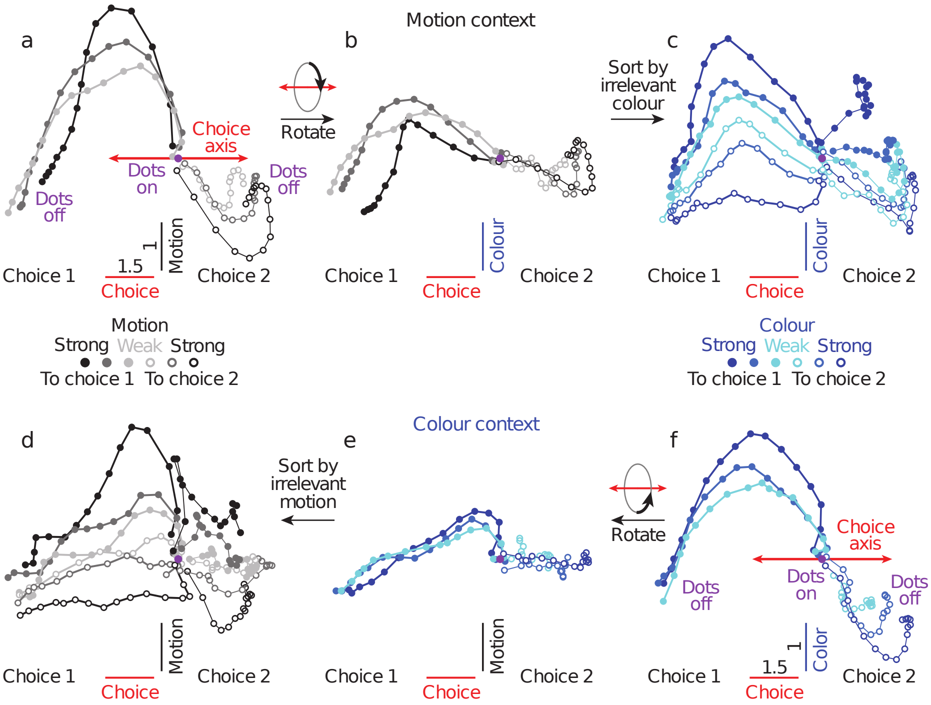


Figure 3.2: Population trajectories, during the period of dots presentation, projected onto the task-relevant subspace spanned by the decision, the color and the motion directions. The first/second row corresponds to the data for the motion/color contexts. Each trajectory reflects the average population activity in a given condition, with the effects of either color (first column) or motion (third column) averaged out. The strength of the evidence in each condition is indicated by the gray and blue scales. Filled/hollow circles designate Tin/Tout types of evidence. Reprinted by permission from Mante et al. (2013)

the visual stimulus has been withdrawn (dots off in fig 3.2).

Third, a single set of axes was found for the two contexts (see section 2.1.3), which provided similar neural representations across contexts. When computed separately for each context, the decision axes were found to point in the same direction as the axis computed using the data from the two contexts. In the case of the color and motion directions, their estimation was a bit more problematic, as inputs, when relevant, are hard to disambiguate from the decision –given that they are integrated towards a decision. This was resolved by orthogonalising the inputs and decision directions found via regression. The input vectors estimated independently in each context were shown to live within the subspace spanned by the single set of axes estimated across contexts. This suggested that the same stable activity pattern is responsible for integrating relevant evidence in both contexts (along the choice axis) and that similar activity patterns represent the momentary motion and color evidence in both contexts (motion and color axes).

Fourth, motion and color inputs result in comparable deflections along the motion and color axes, respectively, whether they were relevant or not. As mentioned before, the striking observation is that, even though these deflections are observed in both contexts, each input appears to only influence the choice in its own context. For instance, motion input influences choices in the motion context but not choices in the color context. It is worth noting though, as the authors mention in the paper, that the irrelevant input strength is found to be reduced compared to when it is relevant. However, the magnitude of the reduction seems to be too small to account for the behavioral effects. This can be seen by comparing the influence of medium coherences when they are relevant and high coherences when irrelevant. In those situations, the deflections along the input axes are of comparable magnitude across contexts. However, the effect on behavior is very different, as medium coherences, when irrelevant, have almost no impact on the monkey's choices (figure 3.1).

Fifth, the responses in both contexts were found to live in different regions of state space, separated along a context dimension. Nevertheless, the rest of the directions –as explained before– were largely invariant across contexts.

The observations summarized in here applied to the two monkeys (monkeys A and F), except for some key features in the representation of the color signal. In monkey F, there was some color modulation along the color input axis, but the effect was very weak in the two contexts. Furthermore, color and decision axes were highly correlated. Therefore, the possibility of a gating effect for color in this monkey could not be ruled out (as explained by an early selection model, see figure 3 in Mante et al. (2013)). The movement of trajectories along the choice axes was also different in monkey F. In the two monkeys, the trajectories seem to be subject to an overall drift towards choice1 that creates an asymmetry between the responses in the two choices. In monkey F, this effect seems to be much stronger (see fig. B.47, extended data figure 7 in Mante et al. (2013)). This condition-independent bias was interpreted as a form of "urgency signal",

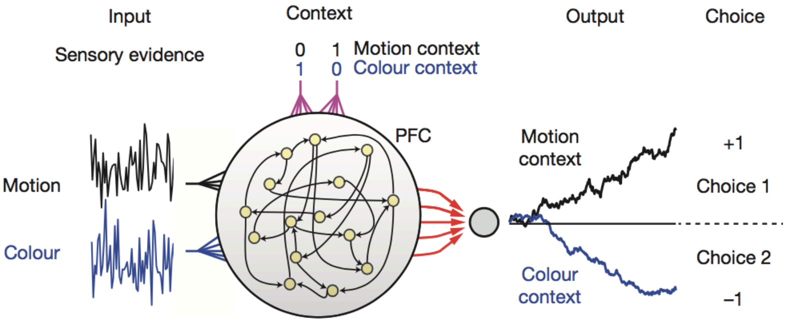


Figure 3.3: A nonlinear RNN represents the circuit in PFC. Color and motion noisy sensory inputs are fed into the network through two input weight vectors. Inputs are assumed constant along the trial and are drawn at each time step from a gaussian distribution with mean set to the coherence values. Two binary contextual lines indicate the current context and are on throughout the whole trial. The read-out unit is trained – via back-propagation-through time– to output the sign of the relevant input at the end of the trial. The network implements a mechanism that washes away the irrelevant input information and that gradually integrates the relevant input signal towards a choice. Reprinted by permission from Mante et al. (2013)

which guarantees that a decision boundary is reached before the termination of the trial (Churchland et al., 2008; Cisek et al., 2009; Hanks et al., 2011). Finally, it is worth mentioning that the performance of this monkey was worse than that of monkey A.

The fact that the very same input into PFC could lead to movement along the axis of choice in one context but not in the other was difficult to explain using existing models of context-dependent selection and integration. In fact, the observed PFC responses ruled out several of these theories. In order to explain the observations, a recurrent neural network with nonlinear rate-based units (nlRNN) was trained to solve an analogous task to the one solved by the monkeys. Importantly, the network was not engineered to solve the task, but its architecture incorporated some elements that matched the observations from the PFC population analysis. For instance, the design was based on the assumption that input directions are fixed across contexts (figure 3.3).

After training, the model was able to solve the contextual task and qualitatively reproduced the monkey's performance (see fig. B.46 (e-h), extended data figure 2 in Mante et al. (2013)). Note, however, that the irrelevant inputs had no influence at all on the choices, unlike what was observed in the monkeys, meaning that the irrelevant information was perfectly disregarded. The model population trajectories captured the main dynamical features of the PFC population responses. In particular, they reflected the same striking phenomenon as the PFC responses: inputs appear to influence choice only when they are relevant. The task-related axes in this case were defined directly from the synaptic weights of the network, with input and context axes corresponding to the input and context learned weights. Each individual input weight controlled the amount of variance in a given neuron attributed to the external input influence. Therefore, the weights were the analogue of the regression coefficients found in the experimental case. The direction of choice, however, was defined using the dynamical solution found as a result of the training. In order to solve the task, the network implemented two contextually dependent sets of fixed points that formed two approximate line attractors. The choice axes was defined as the average of the two attractor directions. Applying the regression-based TDR analysis similar input and choice directions could be recovered. As mentioned before, the line attractors were contextually dependent, i.e. never exist together in the same context, which was provided with the activation the network contextual lines. Furthermore, the line attractors were bounded by two stable fixed points formed at their extremes. When inputs were turned off, trajectories quickly relaxed back to the line attractor and very slowly drifted along it towards one of the attractors at the extremes.

Finally, to understand the selection mechanism of inputs for integration along the line attractors, the local dynamics of model responses around the identified fixed points was analyzed. The selection mechanism did not depend on the projection of the inputs onto the line attractors, as this projection was found to be similar in both contexts, but still, inputs were integrated differently. The selection of the inputs relied on a context-dependent relaxation of the network dynamics which reversed the movement along the line attractor caused by the irrelevant inputs. The relaxation dynamics occurred on a path that was orthogonal to a specific direction in state space, the so called 'selection vector'. It was the projection of the inputs onto this 'selection vector', and not onto the line attractor, that gave the amount of integration. However, this amount was still reflected as a displacement along the line attractor. The direction of the selection vector and the direction of the line attractor are a property of the recurrent synaptic weights learned by the model during training and corresponded, respectively, to the left and right zero-eigenvectors of the underlying linear system in each context.

We end this summary by pointing at some of the features that the RNN model could not reproduce in detail. The most prominent one was the absence of deflections along the input directions. In the network, inputs were kept on during the whole trial, therefore, this reinforced the idea that inputs in PFC are likely transient, possibly because they are being attenuated after a decision is reached. Another substantial difference was that the experimental trajectories seem to converge to the same end point along the decision axis. Training a model under instability conditions (in the absence of noise) allowed it to capture this feature. In this situation, the model implemented a saddle node surrounded by two stable fixed points, not a line attractor, which made all trajectories quickly drift towards one of the two stable attractors, converging to the same point. Other differences, such as the asymmetry in the trajectories between choice1 and choice2, were explained by training the model under "urgency signal" conditions.

3.1.1 Conclusions from Mante et al. study

The key points that were made in the study by Mante et al. are:

1. A single stable dimension in neural space exists that reflects decision in both contexts. This finding is consistent in the two monkeys. The gradual movement of

the trajectories following this direction is suggestive of a slow integrative process that accumulates relevant evidence along a choice axis.

2. Input directions are largely invariant across contexts. Specifically, the input axes estimated independently in each context lie almost entirely within the task-relevant subspace of inputs and decision estimated using the two contexts.
3. The same stable activity pattern is responsible for integrating relevant evidence in both contexts (along the choice axis) and similar activity patterns represent the momentary motion and color evidence in both contexts (motion and color axes).
4. Input magnitudes are comparable across contexts, although the strength of each input is reduced when irrelevant. However, the magnitude of the reduction seems too small to account for the behavioural effects. The absence of contextual gating of sensory inputs rules out early selection models for selective integration. In monkey F, however, this was true only for the motion modality, leaving open the possibility of a gating effect for color.
5. Inputs arriving into PFC are transient, possibly because they are being attenuated after a decision is reached.
6. PFC population trajectories in the two contexts live in different regions of state space and are separated along a contextual axis.
7. A form of urgency signal seems to be biasing the population responses towards the positive choice direction, reflecting an overall tendency of the FRs to increase during the trial.
8. Within a dynamical systems framework, a new model is proposed that explains the selective integration of sensory inputs: a non-linear RNN, trained to solve the task, implements two contextually-dependent approximate line attractors, along which relevant inputs are integrated. The exact mechanism underlying the contextual integration could be understood by linearising the dynamics of the network in each context.

3.1.2 Open questions and new approach

The work by Mante et al. offered invaluable insights for understanding how PFC circuits could flexibly generate complex computations. However, only indirect links could be established between the contextual integration mechanism implemented by the RNN model and the pattern of activations found in PFC's population data. This left the interpretation of some aspects of the data still open for discussion. The specific questions that we find were not entirely resolved are the following:

1. First of all, is a dynamical systems framework suitable for understanding the activity of population of neurons in PFC? and in particular, is a dynamical systems model adequate to describe the temporal structure present in PFC's responses?

2. The theoretical analysis of the nlRNN is based on the assumption that the dynamics in each context are well approximated by a system linearized locally around the identified fixed points. However, is a linear model a good approximation of the global dynamics in each context? and can this simple linear model still capture the complexity of single neuron PSTHs?

3. Is there a better way of estimating input directions? which first, explicitly disentangles the effects of direct inputs from the signals resulting from their integration, and second, which does not require performing any ad hoc decision of the type made in the study (see 2.1.3) –where inputs were chosen as the inferred input regression vectors with the highest norm in time.

4. Is it sufficient to use a two-dimensional subspace to explain the effects of color and motion? that is, are there other input-related activity patterns present in the data? This was not considered in the study.

5. Are input dimensions invariant across contexts? that is, are inputs into the system generating similar modes of activation in each context? or put it in other words, do inputs target the same subpopulation of neurons across contexts and is the pattern of modulation across neurons preserved in the two contexts?

6. The data already suggests that there is no complete selective gating of inputs across contexts, but how exactly do inputs arriving to the system differ across contexts? Is there a substantial change in strength? How do inputs modulate the network throughout the whole period of stimulus presentation?

7. Do inputs have a transient nature? that is, are they attenuated towards the end of the trial? or are the patterns of modulation observed along the input dimensions simply reflecting a change in the projection of the population state vector onto these dimensions?

8. How is the selective integration achieved? Does the real network implement a contextual line attractor? Is the dynamics of the circuit changing across contexts at all, or can the data be explained with a system that is purely input driven?

9. Is there a unique integration direction across contexts? or at least, is the decision at the end of the trial reflected via the same patterns of activation in the two contexts? More than that, is the integration pattern reflected along a single dimension in each context, or should we be thinking about integration subspaces?

10. How can the differences in the data from monkey F be explained? Can we find additional evidence to accept or rule out the early selection model for color? What is the nature of the strong "urgency signal" present in the data from this monkey?

We believe that in order to address such questions, one has to directly interrogate the data. Therefore, we took the approach of inferring the dynamics directly from it. The focus of this study will be to try to determine what changes in the circuit, so that a different computation is implemented in each context.

In order to achieve this, we fitted a linear dynamical system (LDS) model independently to the data from each context. The idea was to compare the solutions learned under each context in order to asses which elements of the computation are different and which ones are preserved across contexts. It is important to note in here that we are not explicitly modeling the capability of the circuit to switch computations across contexts. We are simply analyzing the dynamical properties of the network in each case. Understanding how the contextual switch is implemented in the circuit is beyond the scope of this study, but we believe that our findings will provide valuable insights to address this question.

To start our quest, we first wanted to test the assumption that firing rates in PFC evolve following a simple dynamical rule –an assumption that underlies our choice of model. For that, we fitted to the data a different linear model, which is not constrained by any dynamical prior. The new method is based on a joint-factorization of the data which considers a linear dependence on external variables or input influences. We refer to it as the Joint Factorization model. Finally, we compare performance and inferred solutions under both models, with the goal to resolve the questions raised in here.

3.2 Methods

In this section we describe the two models fitted to the PFC data, a Linear Dynamical System model and a Joint Factorization model. We will list in here the exact specifications and constraints used to incorporate information about the task. For general settings and fitting procedures see chapter 2 (2.2 and 2.3). Before starting, we wanted to point out an important distinction between the two models. The LDS is probabilistic in nature, whereas the JF model is based on a factorization of the data structure, without any explicit noise model defined. Given this, we will perform our analysis on the "noiseless" LDS prior, ignoring the covariance terms inferred during the optimization, and therefore, focusing on the estimated mean dynamics. Furthermore, the deterministic component of the LDS prior, as we will show, is included in the JF model class. That is, the two models are nested, with the LDS in this case having less flexibility, as it is endowed with the dynamical prior. When comparing the solutions and performance of the two models, we will always consider the noiseless LDS prior, so that both models are put in equal grounds (for further discussion see the supplementary material B.1.1.5). The rest of the analysis performed on the the LDS will also be based on the prior deterministic component. Once the models are fit, we will be able to resolve the questions

1. How does each model perform? Can these simple linear models accurately capture the data?
2. How well do they generalize? Does the dynamic prior help?
3. Is the temporal structure present in the PSTHs consistent with the assumptions of the dynamical prior?

3.2.1 LDS model

We fit an LDS model independently to the PFC data from each context

$$\begin{aligned}
\mathbf{x}_k^{(cx)}(t) &= A^{(cx)}\mathbf{x}_k^{(cx)}(t-1) + B^{(cx)}\mathbf{u}_k^{(cx)}(t) + \boldsymbol{\eta} & \qquad \boldsymbol{\eta} &\sim \mathcal{N}(0, Q^{(cx)}) \\
\mathbf{y}_k^{(cx)}(t) &= C\mathbf{x}_k^{(cx)}(t) + \mathbf{d} + \mathbf{r} & \qquad \mathbf{r} &\sim \mathcal{N}(0, R)
\end{aligned}$$

where $\mathbf{y}_k^{(cx)}(t)$ is the vector of observations, containing the trial-averaged z-scored FRs at time t for the $N = 727$ neurons recorded under condition k in context cx (a few neurons from the original data set were excluded as they had very sparse firing rates). Conditions $k = 1...K$ with $K = 36$ are specified by the possible color and motion coherence combinations presented in a given trial: six different values for each color and motion indicating coherence level (weak-medium-strong) and direction (in or out of the RF, see 3.1). Context can be $cx = 1$ for color and $cx = 2$ for motion. The period of the task that we are modeling is the same the study by Mante et al. focused on, during the stimulus presentation, when the evidence accumulation process must take place. More

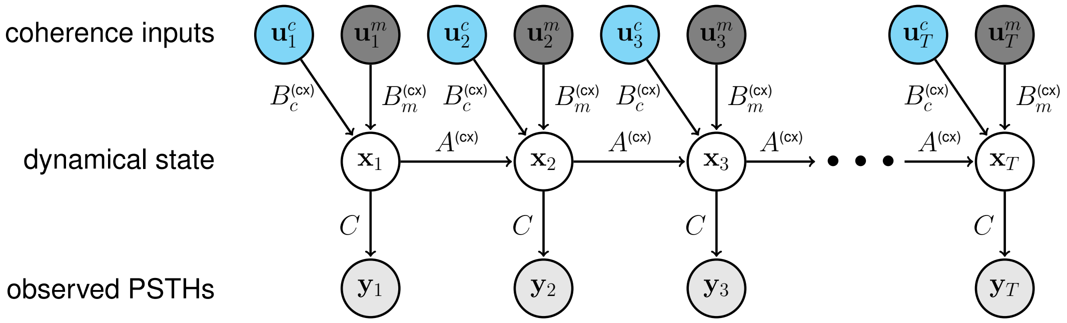


Figure 3.4: LDS model. Observations (PSTHs) are generated by a hidden and low-dimensional linear dynamical system evolving in time. The dynamics are given by the transition matrix A. Color (blue) and motion (black) coherence-related inputs $\boldsymbol{u}^c, \boldsymbol{u}^m$ bias the state of the system at each time step through the input subspaces B_c, B_m. The deterministic evolution of the state is perturbed by gaussian noise, with covariance Q. The loading matrix C maps the hidden state into the high dimensional observational space. Additional gaussian noise is added to the observations. The parameters are context dependent cx except for the loadings C, a constant output bias $\boldsymbol{d}$ and the observations noise covariance matrix R.

concretely, the data from each trial starts at 100ms after dots onset and ends 100ms after dots offset (at 850ms), lasting 750ms. The FRs are binned using a 50ms squared sliding window, so each "average trial" or condition is $T = 15$ time steps long.

The observations are assumed to be generated by a low-dimensional (dim=H) hidden dynamical system whose state is given by $\mathbf{x}_k^{(cx)}(t)$ (fig. 3.4). The posterior over the state given the observations is inferred during the expectation step of the EM algorithm. Parameters are learned in the M step (2.3). The inputs can be assumed known, given by the coherence values presented in condition k and set to be constant along the whole trial $\boldsymbol{u}^k = \left[\mathbf{u}_c^k; \mathbf{u}_m^k\right]$, or the whole time series $[\mathbf{u}^k(1)...\mathbf{u}^k(T)]$ can also be learned. The dimensionality of the input subspace is given by $D = D_c + D_m$. The first step is drawn from

$$\boldsymbol{x}_k^{(cx)}(1) \sim \mathcal{N}(A^{(cx)}\boldsymbol{x}^{(cx)}(0) + B^{(cx)}\boldsymbol{u}_k^{(cx)}(1), Q^{(cx)})$$

We will omit the context index from now on for simplicity.

The transition matrix $A_{H\times H}^{(cx)}$, input subspaces $B_{H\times D}^{(cx)}$ and innovations covariance noise $Q_{H\times H}^{(cx)}$ are learned independently for each context. The loading matrix $C_{N\times H}$, output bias $\mathbf{d}_{N\times 1}$ and output noise variances ($R_{N\times N}$ is diagonal) are shared across contexts. The columns of $C_{N\times H}$ are constraint to be orthonormal, such that $C^{\intercal}C = I$.

Note that we are fitting the LDS to trial *averaged* data, so the "noise" terms are capturing *residual* trial-to-trial noise that was not fully averaged away and also reflect model *mismatch*. In any case, as explained at the beginning of this section, we will focus on the LDS prior state vector equation in the deterministic setting, ignoring the noise terms

$$\boldsymbol{x}^k(t) = A\boldsymbol{x}^k(t-1) + B\boldsymbol{u}^k(t)$$

we can express the state of the system at any time t as a function of the initial state and the inputs history

$$\boldsymbol{x}^k(t) = A^t\boldsymbol{x}(0) + \sum_{t'=1}^{t} A^{t-t'} B\boldsymbol{u}^k(t')$$

$$\boldsymbol{x}^k(t) = \underbrace{A^t\boldsymbol{x}(0)}_{base} + \underbrace{\sum_{t'=1}^{t} A^{t-t'} B_m\boldsymbol{u}_m^k(t')}_{motion} + \underbrace{\sum_{t'=1}^{t} A^{t-t'} B_c\boldsymbol{u}_c^k(t')}_{color} \tag{3.1}$$

where we have rewritten the equation so that the motion and color influences, via the subspaces B_m and B_c, appear explicitly. Bringing the hidden state vector into the observational space we obtain

$$\boldsymbol{y}^k(t) = C\boldsymbol{x}^k(t) + \boldsymbol{d}$$

$$\boldsymbol{y}^k(t) = \underbrace{CA^t\boldsymbol{x}(0) + \boldsymbol{d}}_{base} + \underbrace{C\sum_{d}^{D_m}\sum_{t'=1}^{t} A^{t-t'}\boldsymbol{b}_{m,d}u_{m,d}^k(t')}_{motion} + \underbrace{C\sum_{d}^{D_c}\sum_{t'=1}^{t} A^{t-t'}\boldsymbol{b}_{c,d}u_{c,d}^k(t')}_{color}$$

In this way, we can decompose the firing rates of the neurons at a given time in terms of the different input contributions, plus a baseline component which is condition independent. The parameters $u_{m/c}^k(t)$ are the motion/color input influence at time t, under condition k, and there are $D_{m/c}$ possible different ones. We consider $D_{m/c} = 1$ in the following, for simplicity. In the task design, color and motion coherence values are repeated across conditions. Therefore, we can use this information to infer the coherence strengths by learning inputs from trials where a particular coherence value was presented. We can learn a whole input time series $u_{m/c}^{coh}(1...T)$ for each coherence value. Alternatively, we can further constrain the inputs by considering a common time series which is scaled by a parameter indicating the coherence value. In this way, we can infer the input coherence values and separately, the overall time course of the inputs.

$$u_m^k(t) = T_m(t)\, m^k$$
$$u_c^k(t) = T_c(t)\, c^k$$

where c^k and m^k indicate the color and motion coherences presented at condition k, which can take one out of 6 possible values

$$m = [m^1, m^2, m^3, m^4, m^5, m^6]$$
$$c = [c^1, c^2, c^3, c^4, c^5, c^6]$$

We find that this assumption is sensible given the way we believe inputs are modulating the neurons. Learning different coherence levels is reasonable as the strength of the evidence clearly affects the magnitude of the FRs. However, it is unlikely that the pattern of temporal modulation changes with coherence level. Nevertheless, we will consider both possibilities, either constraining the inputs as we explained, or allowing for full flexibility in the input time courses across coherence levels.

Finally, we allow the model to learn different input time series for positive and negative coherences (evidence in or out of the neurons' RF)

$$\begin{cases} u_m^k(t) = T_m^{in}(t)\, m^k & if \quad \text{trial } k \text{ motion coh} > 0 \\ u_m^k(t) = T_m^{out}(t)\, m^k & if \quad \text{trial } k \text{ motion coh} < 0 \end{cases}$$

and similarly for color. We found that having this extra flexibility improved the performance of the model, as different time lags of input modulation can be then taken into account. For instance, during T^{out} conditions, it is likely that some inhibitory-mediated suppression of the neurons is taking place, which would in principle have a different latency of modulation than a direct input bias. The equations derived to learn the input time courses and coherence values can be found in the Appendix (A.1.2).

Incorporating the input constraints, we obtain the expression

$$\boldsymbol{y}^k(t) = \underbrace{CA^t\boldsymbol{x}(0) + \boldsymbol{d}}_{base} + \underbrace{C\sum_{d}^{D_m}\sum_{t'=1}^{t} A^{t-t'}\boldsymbol{b}_{m,d}^{\pm}(t')m_d^k}_{motion} + \underbrace{C\sum_{d}^{D_c}\sum_{t'=1}^{t} A^{t-t'}\boldsymbol{b}_{c,d}^{\pm}(t')c_d^k}_{color} \tag{3.2}$$

where

$$\begin{cases} \boldsymbol{b}_m^{+}(t) = \boldsymbol{b}_m T_m^{in}(t) & if \quad \text{trial } k \text{ motion coh} > 0 \\ \boldsymbol{b}_m^{-}(t) = \boldsymbol{b}_m T_m^{out}(t) & if \quad \text{trial } k \text{ motion coh} < 0 \end{cases}$$

and analogously for color.

3.2.2 JF model

Matrices $\{Y^k\}^{(cx)}$ containing the the trial-averaged z-scored PSTHs for each condition k in each context cx, are jointly factored so that the scores for Y^k depend linearly on the inputs u_d^k (see section 2.2)

$$\underset{N\times KT}{\mathcal{Y}} \approx \underset{N\times H}{C} \underset{H\times(D+1)T}{\mathcal{B}} \underset{(D+1)T\times KT}{\mathcal{U}}$$

where $\mathcal{Y} = [Y^1, ..., Y^K]$ contains the data matrices concatenated horizontally. The matrices C and $\mathcal{B}$ can be learned using reduced-rank regression and the inputs can be estimated together via alternating least squares.

We consider a condition independent "base" component and $D = D_m + D_c$ inputs for color and motion influences. The input design matrix $\mathcal{U}$ is defined as

$$\underset{(D+1)\times K}{U} = \begin{bmatrix} 1 & 1 & ... & 1 \\ \boldsymbol{u}^1 & \boldsymbol{u}^2 & ... & \boldsymbol{u}^K \end{bmatrix} = \begin{bmatrix} 1 & 1 & ... & 1 \\ m_1^1 & m_1^2 & ... & m_1^K \\ c_1^1 & c_1^2 & ... & c_1^K \\ & & \vdots & \\ m_{D_m}^1 & m_{D_m}^2 & ... & m_{D_m}^K \\ c_{D_c}^1 & c_{D_c}^2 & ... & c_{D_c}^K \end{bmatrix}$$

$$\underset{(D+1)T\times KT}{\mathcal{U}} = U \otimes I_{T\times T}$$

Same as for the LDS, we do not learn an input value for each condition k, since the coherence values are repeated across conditions. In order to incorporate this constraint, we write U as a product of a matrix P (listing all possible motion and color coherence values for each input dimension d, in our case$M_1 = 6, ..., M_D = 6$) and indicator matrix Q (selecting the values for each condition)

$$P = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & m_{11} & \dots & m_{1M_1} & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & c_{11} & \dots & c_{1M_2} & \dots & 0 & \dots & 0 \\ & & & & & \vdots & & & & & \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & c_{D_c1} & \dots & c_{D_cM_D} \end{bmatrix}$$

$$\underset{(D+1)\times K}{U} = \underset{(D+1)\times(1+\sum M_d)}{P} \underset{(1+\sum M_d)\times K}{Q}$$

The model was fit concatenating the matrices $\mathcal{Y}^{cx}$ from the two contexts $[\mathcal{Y}^1, \mathcal{Y}^2]$ so that we could recover input biases $[U^1, U^2]$ and low dimensional representations $[\mathcal{B}^1, \mathcal{B}^2]$ for each context, but sharing the loading matrix C across contexts, as done for the LDS.

Rewriting the equations for a single condition

$$\underset{N\times T}{Y^k} \approx \underset{N\times H}{C} \underset{H\times(D+1)T}{\mathcal{B}} \underset{(D+1)T\times T}{\mathcal{U}^k}$$

and making explicit the time indexing, we can write this model in a similar fashion as the LDS

$$\boldsymbol{y}^k(t) = \underbrace{C\boldsymbol{b}_0(t)}_{base} + \underbrace{C\sum_{d}^{D_m} \boldsymbol{b}^{\pm}_{m,d}(t) m^k_d}_{motion} + \underbrace{C\sum_{d}^{D_c} \boldsymbol{b}^{\pm}_{c,d}(t) c^k_d}_{color} \tag{3.3}$$

Note that the low-rank representation has been extended, so that different mappings can be used for positive and negative coherences. In this way, the two models are put in exactly the same grounds and differences in performance can uniquely be attributed to the dynamical constraints that the LDS incorporates.

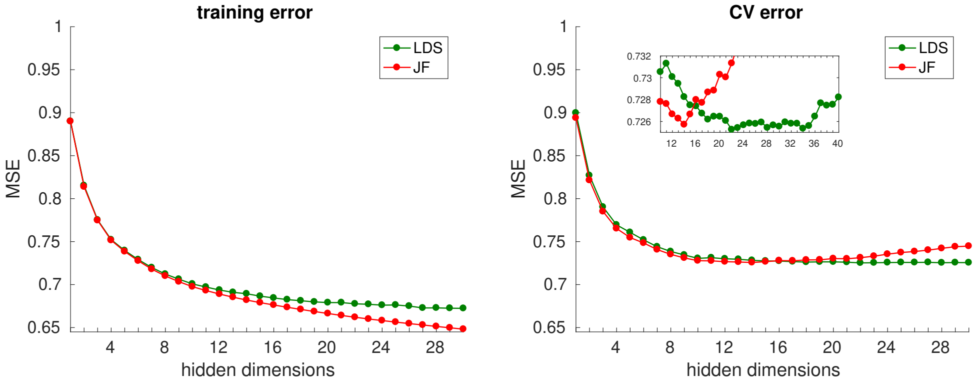
Comparing equation 3.3 with the equivalent derived for the LDS model (eq. 3.2)

$$\boldsymbol{y}^k(t) = \underbrace{CA^t\boldsymbol{x}(0) + \boldsymbol{d}}_{base} + \underbrace{C\sum_{d}^{D_m}\sum_{t'=1}^{t} A^{t-t'}\boldsymbol{b}^{\pm}_{m,d}(t') m^k_d}_{motion} + \underbrace{C\sum_{d}^{D_c}\sum_{t'=1}^{t} A^{t-t'}\boldsymbol{b}^{\pm}_{c,d}(t') c^k_d}_{color}$$

one can explicitly see the constraint the dynamical prior imposes on the parametrization. For the LDS, the inputs and base components learned must be of the form

$$\begin{aligned}
\boldsymbol{b}_0(t) &= A^t\boldsymbol{x}(0) + C^{\intercal}\boldsymbol{d} \\
\boldsymbol{b}^{\pm}_{m,d}(t) &= \sum_{t'=1}^{t} A^{t-t'}\boldsymbol{b}_{m,d}T^{in/out}_{m,d}(t') \\
\boldsymbol{b}^{\pm}_{c,d}(t) &= \sum_{t'=1}^{t} A^{t-t'}\boldsymbol{b}_{c,d}T^{in/out}_{c,d}(t')
\end{aligned}$$

where the dynamics matrix A shapes the direction of the integrated inputs in time, restricting the space of possible mappings that the inputs can have along the trial. Note that both the dynamics and the input time courses contribute to the integrated inputs scaling.



(a) (b)

Figure 3.5: LDS and JF models performance as a function of hidden dimensionality (rank degree for the case of the JF model) a) Training mean squared error b) cross-validation mean squared error. Squared errors are in units of variance, given that the FR responses were z-scored.

3.3 Results: Monkey A

3.3.1 LDS and JF models performance and solution

3.3.1.1 LDS and JF models performance

Once the models were fit to the data, model-generated FR responses were estimated –in the case of the LDS, using the noiseless LDS prior. The models were fit with 4 input dimensions, 2 for color and 2 for motion, as we found that for this dimensionality performance saturated (see 3.3.3). Mean squared errors averaged over conditions, neurons, contexts and time were computed between the neurons PSTHs and the model generated PSTHs. We estimated cross-validation performance by fitting the models leaving one condition out, and then testing on the left-out condition. Training and cross-validation performance are shown in figure 3.5 for a wide range of hidden dimensionalities (or rank degree in the case of the JF model).

The JF model clearly outperforms the LDS in training, as it has more parameters. In cross-validation, however, the two models achieve almost identical performance (see table 3.1). The difference in errors is $\delta MSE = 0.0004$, which accounts to 0.04% of the total variance (as the squared errors are in units of variance, since the FR responses were z-scored). It is quite remarkable that in training, the two models perform very similarly for low dimensionalities. For higher dimensionalities, when the JF and the LDS curves start to diverge, is indeed when the JF model starts to suffer strongly from overfitting. The LDS needs more dimensions to reach a minimum, but once it is reached the error curve flattens out. It is not until dimensionalities close to H=40 that the model seems to overfit. In fact we found that the model suffers from numerical instabilities in this region, so the increase in error could be, at least in part, due to poor convergence issues (see B.1.1.5). This suggests that the LDS constraint on the dynamics does indeed help generalize. Finally, we observed that when computing the LDS likelihood, the curve

Model	min CV error	H
LDS	0.7253	22
JF	0.7257	14

Table 3.1: LDS and JF models minimum cross-validation error and corresponding hidden dimensionality H (or rank degree) for which it is achieved.

does not saturate and we are still getting a likelihood gain for high dimensionalities (see B.1.1.5). Considering that there is indeed a saturation in the MSE curve, which indicates how well the "mean" dynamics is captured, this suggests that the gain in the likelihood must be for capturing additional variance in the firing rates, attributed to residual noise or model mismatch.

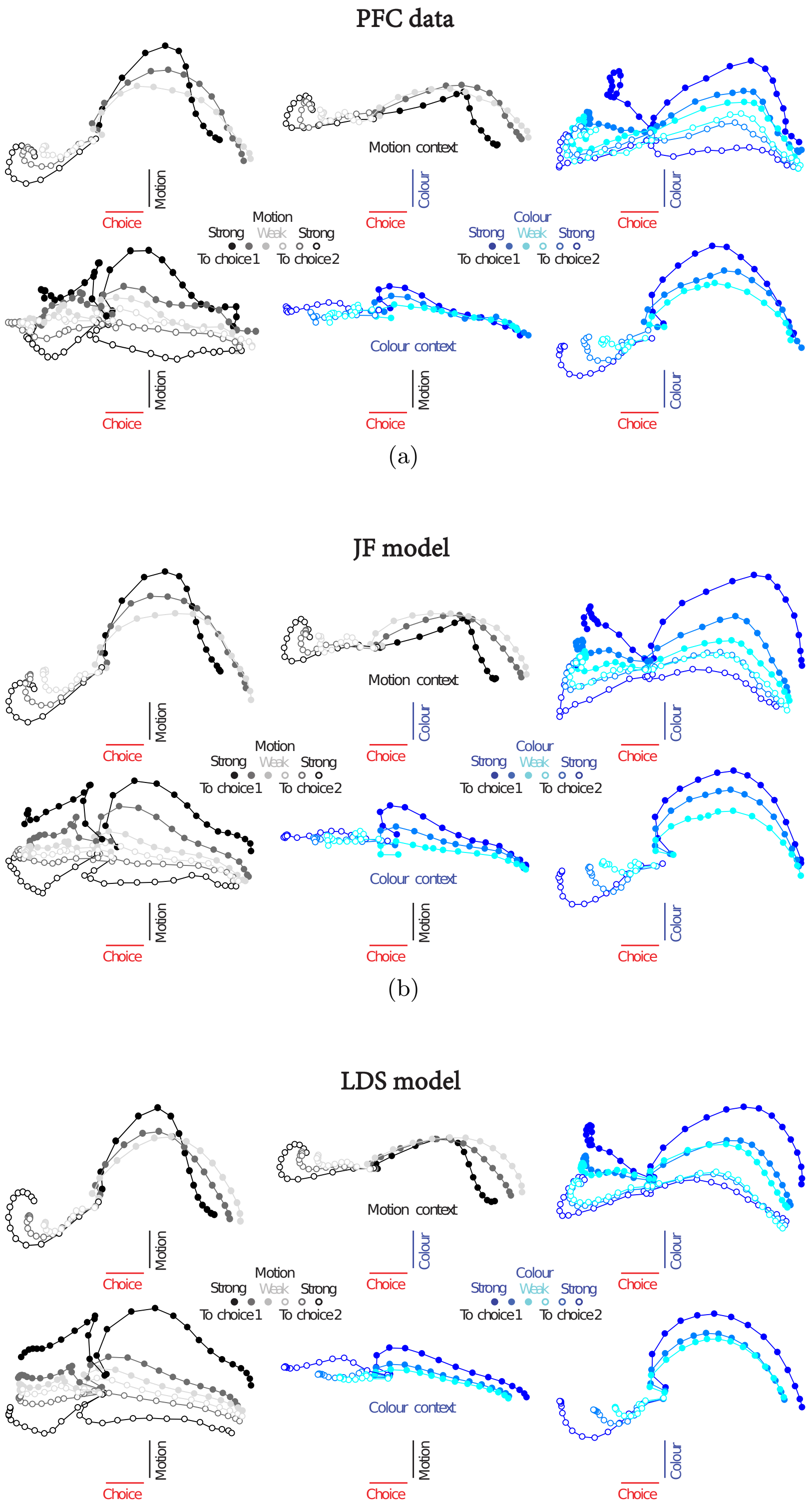
In the supplementary information, we report the performance of a JF model which does not have extra parameters to account for Tin and Tout differences. Relieving the model from such extra parametrization does in fact improve generalization slightly (see figure B.1 and table B.1). Why this feature is not important for the JF model, but helps the performance of the LDS, is something that we have not looked into but we think it would be worth exploring.

Given this results, we conclude that the LDS dynamical prior is a good assumption of the temporal structure in the data, as fitting a model with extra temporal flexibility does not improve the generalization performance. This is remarkable given the fact that the JF model has the potential to capture arbitrary patterns in time, including dynamics specified by non-linear rules.

Next, we compared how well the models are able to capture the trajectories in the task-relevant subspace found in Mante et al. (2013) via targeted dimensionality reduction (TDR). For that, we plotted model-generated trajectories on the exact same subspace identified from the data. The results are shown in figure 3.6. In order to identify the TDR subspace and to generate the plots, we applied the same code used in the paper, which was kindly provided by the authors[1]. Note that the generated trajectories have been cross-validated. Nevertheless, both models seem to accurately capture all their main features, including the ones that could not be reproduced in the previous study with the trained RNN. Notably, the deflections along the input dimension and the convergence of some of the trajectories to the same end point.

The trajectories generated from the model fits directly, without cross-validation, are almost identical B.2. Therefore, the bulk of the CV error is not reflected in this subspace. To investigate this further, we broke down the CV error per conditions, time and contexts (fig. 3.7). The first striking observation is that the error follows the exact same trends in the two models. Generalization performance is the poorest on left-out conditions for the weakest coherences. The biggest error results from conditions where strong irrelevant

[1] The code generated plots for the trajectories that are mirrored imaged with respect to the original figures in the paper 3.2. Also note that the input axes were not re-scaled with respect to the decision axis.



(a)

(b)

(c)

Figure 3.6: Data and model generated cross-validated trajectories, leaving one condition out, in the task-relevant subspace identified by regression Mante et al. (2013) a) PFC data b) JF model c) LDS model. Same plotting conventions as in figure 3.2.

evidence is paired with weak relevant evidence, that is, were there is strong incongruent information. This same effect is in fact reflected in the performance of the monkey, as it is unable to ignore strong irrelevant information perfectly (3.1d,e). The error pattern in time, however, seems to simply reflect Poisson variability, as it matches the general trend of the FRs initial dip at the beginning of the trial plus a gradual building up of activity towards the end –typically observed in PFC neurons during DM tasks. The asymmetry in the per-condition error curve seems to be due to the same reason: the error is higher for positive coherences than for negative coherences, which can be explained from the fact that the FRs are in general higher in trials where the relevant evidence was pointing towards the RF of the neurons. The central peak of the curve, however, cannot be, as the weakest coherences trigger the lowest FRs. Therefore, the source of this error must be different. The training error (fig. B.3) follows exactly the same trends. Therefore, the mismatch does not arise due to a poorer generalization performance when predicting conditions with weak coherences. It seems to be harder for the models in general to capture those conditions. The LDS cross-validated likelihood also reflects the same error pattern, despite the fact that it is computed on the whole model, considering noise terms and inferred posterior (see figure B.8). A possibility could be that the PSTHs for weak coherences are noisier, perhaps, reflecting a noisier integration process. Figure B.4 shows that this is not the case, as FRs are not more variable under those circumstances. We also performed a quantification of the error along the task-relevant dimensions for the two models, per-time and per-condition. We found no particular condition-related structure of the error along those important dimensions, even though they carry a large percentage of the variance in the data (Mante et al., 2013). Finally, we found that the per-condition averaged error pattern follows the distribution of the error across the whole population (B.6). We are still uncertain about its source, but given that the two models are impaired in exactly the same way, this suggests that the limitation does not come from the dynamical constraint, but rather from the assumption that inputs are linearly combined across conditions.

One final observation is that the models, in general, seem to be missing a large percentage of the variance, as much as 70% (see fig 3.5 and supplementary information B.1.1.4). However, the models are capturing the trajectories fairly well and also the individual PSTHs (fig 3.35). This is due to several reasons: first, we did not smooth the data. Second, the MSE is estimated on all conditions, while the trajectories in 3.6 are from color/motion conditions averaged out (in fact, the model-generated PSTHs for all conditions present more clear deviations from the real data B.43). Finally, we did not exclude neurons that were markedly noisy (see fig. B.44).

We performed one last test to convince ourselves that the temporal structure present in the PSTHs is consistent with the assumptions of the dynamical prior. If true, when the temporal correlations in the FRs are destroyed, the LDS should not be able to capture the data well. The JF model, however, should be left unaffected, as it can model any arbitrary pattern in time. This is in fact what we obtain (fig B.10).

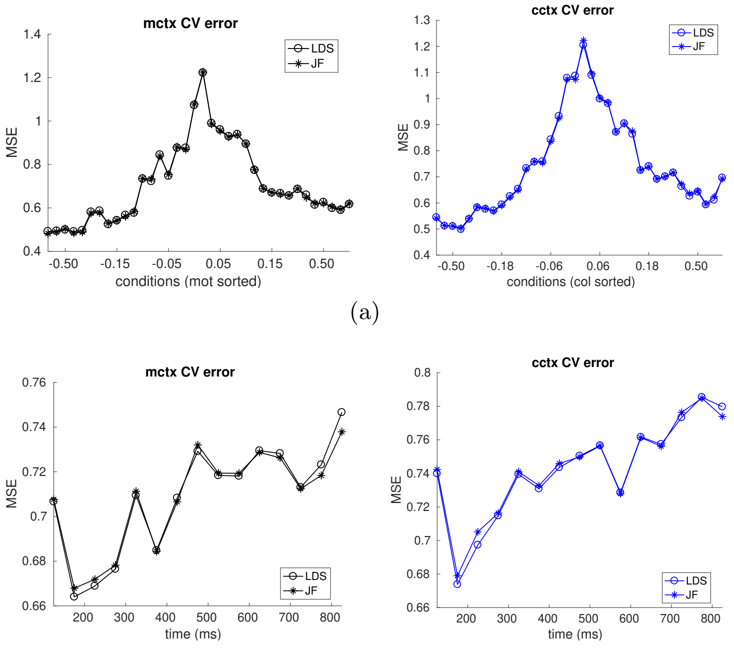
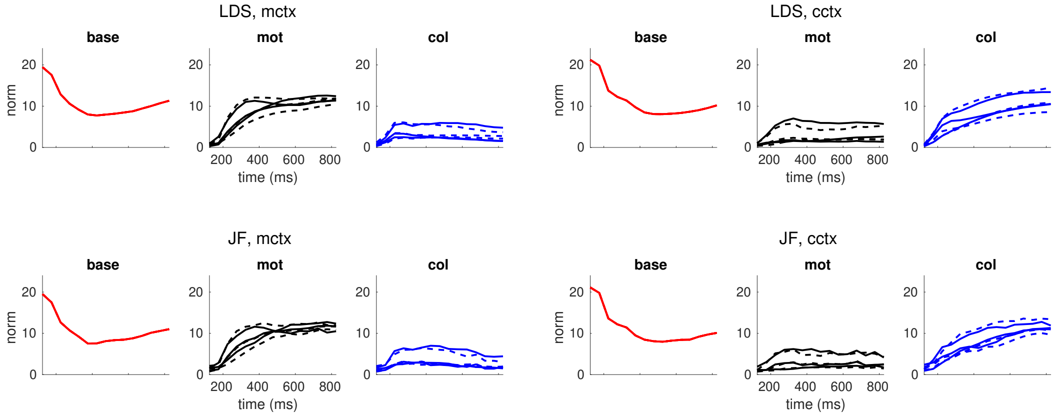


Figure 3.7: LDS and JF models minimum CV error (see table 3.1) a) CV error for the 36 different conditions, grouped by the coherence value of the relevant inputs and sorted in ascending order. Each group contains 6 conditions corresponding to the 6 possible irrelevant coherence values. These are also sorted in ascending order b) CV error across time. Motion context (left), color context (right)

3.3.1.2 Comparing the LDS and the JF models solutions

Using equations 3.2 and 3.3, we can compare the solutions inferred by each model. In particular, we were interested in assessing how each model captures the motion and color related influences in the data. We analyze the models with the dimensionality corresponding to the minimum CV error (table 3.1). The results are summarized in figures 3.8 and B.12. The first figure shows the base, motion and color components norms in time (given by the sum of the squares of the decomposed FRs, squared rooted). We generated model responses using the inferred inputs for the six possible coherence values. The motion and color component norms for positive coherences are shown using filled lines and dashed lines are used for negative coherences. We must clarify that the decomposition we are showing in 3.8 is not exactly the one obtained directly from the model fits. This is because there are degeneracies in the models, which allow for condition independent effects to be allocated to the input components. To remove this effects, we subtracted the across conditions input mean to each input component and incorporated this condition independent term into the base component. The raw solution can be seen in B.11. After this baseline correction, it is apparent that the two models learn almost identical decompositions. Furthermore, when computing the alignment of the relevant input components across models, we find that they are learned to point in very similar directions (figure B.12a) This alignment is very strong across base and



(a) (b)

Figure 3.8: Vector norms of the base, motion and color components at each point in time. For the LDS model (top) and the JF model (bottom), in the motion context (a) and the color context (b) The input components are computed for all the learned motion and color input values (in black and blue, 6 in each case), corresponding to the positive coherences (solid lines) and negative coherences (dashed lines) The base component (red) does not depend on the inputs –so there is only a single trace– and captures condition-independent variance.

relevant input components. The correspondence is a bit weaker for the irrelevant inputs. Therefore, the two models find similar neural mappings for the evolution of the different input patterns in time. A final observation is that during the first time steps the dot products are close to zero. This can be explained by the fact that input influences are very small at the beginning of the trial, so the input directions learned are rather arbitrary.

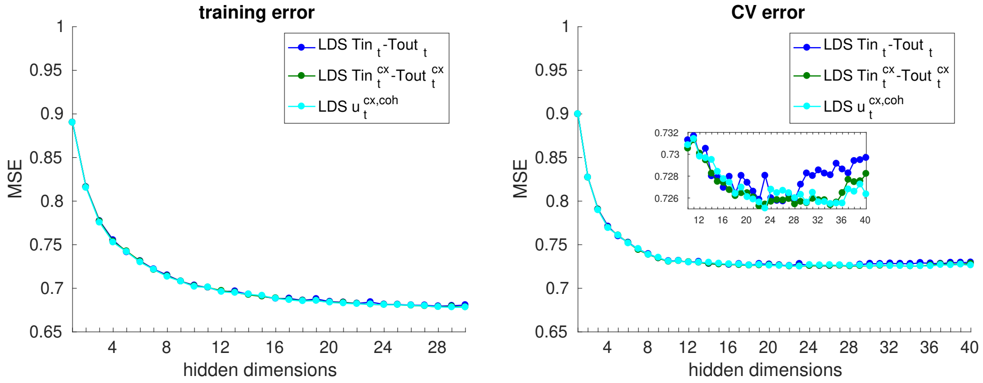
Finally, we find that consistently across the two models, the color and motion input components are fairly orthogonal along the whole trial (figure B.12b) This suggests that the integration process keeps the color and motion inputs in orthogonal dimensions. However, as we will see in the next sections, this lack of alignment is in fact expected by chance.

We conclude that the LDS and the JF models capture the data in an equivalent fashion, which reinforces the idea that the LDS prior is a good assumption of the temporal structure in the FR responses.

Two main observations are to be made from the time courses of the input component norms. First, relevant inputs are strongly amplified and build up until the end of the trial. The strongest motion signal, however, seems to reach a saturation point early in the trial. Second, the irrelevant inputs signals mildly increase in strength but saturate early in the trial. Importantly, their effect does not seem to decay back to zero.

3.3.2 LDS with different input constraints

We explained in the methods than the LDS model can incorporate different constraints in the input biases. The model we focused on assumes a common time course for each input



(a) (b)

Figure 3.9: LDS model performance for different input specifications. Shared time course across coherences and contexts (blue), shared time course across coherences but not contexts (green) and full flexibility per coherence value (cian) a) Training mean squared error b) CV mean squared error.

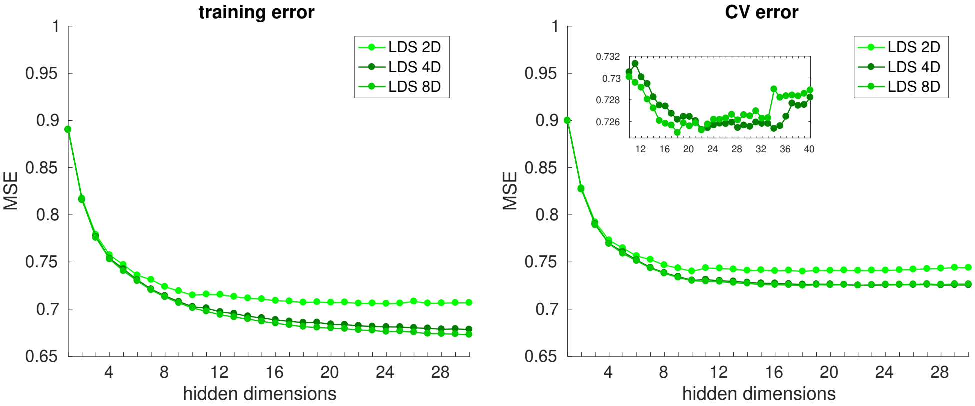
	Model	min CV error	H
1.	LDS $u_t^{coh,cx}$	0.7250	23
2.	LDS $Tin_t^{cx} - Tout_t^{cx}$	0.7253	22
3.	LDS $Tin_t - Tout_t$	0.7257	26

Table 3.2: LDS minimum cross-validation error, for different input specifications, and corresponding hidden dimensionality H for which it is achieved.

which is scaled by a different coherence value and considers different time courses for positive and negative coherences (model 2 in fig 3.9 and table 3.2). We found that this model had almost identical performance to a much more flexible model where the whole input time series was learned differently for each coherence value (model 1). Furthermore, as we will see later, the two models recovered similar input patterns (figures 3.14 and B.21). Therefore, the constraints we incorporated in the input structure seem to be a reasonable assumption. Models 1 and 2 had the flexibility to learn different input biases for each of the contexts. We also tried to constrain the time structure so that it was shared across contexts (model 3). By doing this the performance was impaired slightly, but the trajectories in the task-relevant subspace did not present any evident deficit (see figure B.16b and compare with 3.6c) This is not surprising given that the input patterns recovered in the two contexts independently are indeed very similar –as we will see later– (figures 3.14 and B.21).

3.3.3 LDS input dimensionality

One of the questions that was not addressed in the previous study is the dimensionality of the input subspace. Considering that the effects of color and motion live in only two dimensions reduces the picture of possible input-related patterns that may modulate the FR responses. This point has also been raised by others (see 2.1.3). Therefore,



(a) (b)

Figure 3.10: LDS model performance for different input dimensionalities a) Training mean squared error b) CV mean squared error.

Model	min CV error	H
LDS 8D	0.7250	18
LDS 4D	0.7253	22
LDS 2D	0.74	18

Table 3.3: LDS minimum cross-validation error, for different input dimensionalities, and corresponding hidden dimensionality H for which it is achieved.

we fitted the LDS model varying the number of input dimensions, going from 2D up to 8D. The results are summarized in figure 3.10 and table 3.3. We first observe that increasing the dimensionality from 2D to 4D substantially improves model performance. This was the case even for models with full flexibility in the input time courses (model 1). Furthermore, we observed that the 2D model would often suffer from local optima convergence issues. Finally, we found that increasing the dimensionality as much as by 8 dimensions barely decreased the CV error. The performance seems in fact to saturate at four dimensions. In later sections, we will further look into the dimensionality of the different input subspaces, as it could well be that this is different in each context and also, across color and motion input modalities.

3.3.4 LDS under the same contextual dynamics

In the previous section we proved that a model with a four dimensional input subspace is sufficient to capture the data well. The model is free to learn an arbitrary input sequence

Model	min CV error	H
LDS $A^{(c)} \neq A^{(m)}$	0.7253	22
LDS $A^{(c)} = A^{(m)}$	0.7271	29

Table 3.4: LDS minimum cross-validation error, for unconstrained (top) and constrained (bottom) dynamics across contexts, and corresponding hidden dimensionality H for which it is achieved.

for each of the dimensions, including setting them to zero if there is no residual variance left to be explained. The question is then, given all the flexibility placed in the inputs, do we really need the dynamics? or in other words, can we explain the population FRs in the two contexts uniquely via a change of direct input influences?

In order to address this question, we fitted models to each context constraining them to have the same dynamics, that is, learning a unique transition matrix across contexts. Table 3.4 shows that the dynamically constrained model performs worse in cross-validation compared to the unconstrained one. It could be though, that qualitatively speaking, this moderate increase in error does not impact much the fits. In fact, we found that, surprisingly, the model is able to capture the average trajectories equally well (see fig. B.15). However, additional analysis quantifying the CV error along the task-relevant dimensions –for all conditions, not averaging out color or motion effects– indicates that substantial variance is being missed along the decision dimension (see figure B.14 and compare with B.13). Furthermore, the model is highly unstable. This is due to the fact that the transition matrix eigenspectrum is always learned with a large negative eigenvalue (fig B.19a), which often cause strong oscillations in the output space. More details on the solution found for this dynamically constrained model will be discussed in the following sections.

In conclusion, the data supports a model in which a change in dynamics takes place during each context. This change seems to be necessary to reproduce, in detail, the selective integration computation. However, considering that the differences in performance are very subtle, we do not completely rule out the possibility that other type of models, perhaps non-linear in nature, could capture the data accurately under a single dynamics.

3.3.5 LDS inputs and dynamics

Now that we have justified that the LDS is a suitable model to explain the data, we proceed to analyze the properties of the solution learned. We will consider a model with input dimensionality $D = 4$ and a hidden dimensionality of $H = 22$, for which the best generalization performance was obtained.

3.3.5.1 Transition matrix

We first looked at the eigenspectrum of the transition matrices learned for each context (3.11a) The first observation is that in both contexts, a multitude of slow modes are learned. This can be seen by looking at the stability of the modes, given by the magnitude of the associated eigenvalues $|\lambda_h|$. A magnitude equal to 1 indicates a perfectly stable mode, that is, a direction in state space where the activity of the system neither decays nor diverges. Most of the modes have a magnitude close to 1 (3.11b), indicating that there are several dimensions in neural space along which activity decays, but with a long time constant. Note that one of the eigenvalues has a magnitude larger than 1, meaning

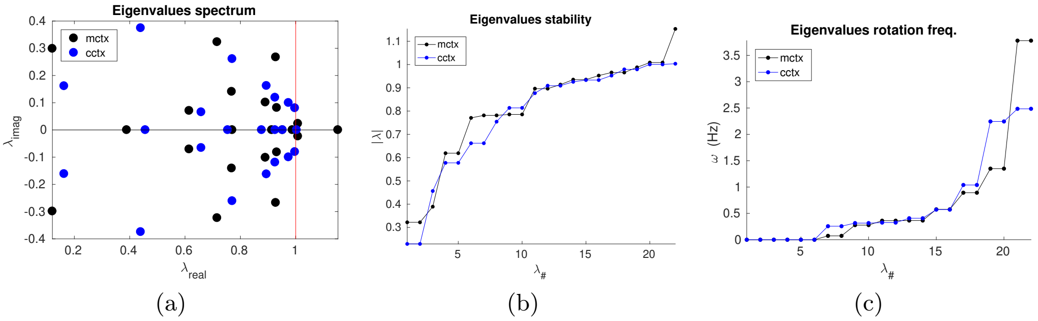


Figure 3.11: Properties of the learned LDS transition matrix a) Eigenspectrum of the transition matrices inferred in each context (black for motion context and blue for color context) b) Stability of the eigenmodes of the dynamics in each context, given by the absolute value of the eigenvalues (sorted in ascending order) c) Rotation frequency in the planes spanned by pairs of complex conjugate eigenvectors (sorted in ascending order).

that the system is globally unstable, as the stability criteria requires

$$|\lambda_h| \leq 1 \qquad \forall \lambda_h \; real$$

$$|\lambda_h| = \sqrt{\lambda_{h,im}^2 + \lambda_{h,real}^2} \leq 1 \qquad \forall \lambda_h, \lambda_h^{\dagger} \; complex$$

This is a limitation of fitting a model on finite trial lengths, which allows to learn this type of solutions, as the system will not diverge during the trial's short time period. Incorporating constraints of the type applied in Buesing et al. (2012) to learn stable dynamical systems is something we would like to explore in future work. A second observation is that the model learns a multitude of complex eigenvalues. Complex modes define planes of rotation with amplitude given by $|\lambda_h|$ and rotation frequency specified by

$$\omega_h = \arctan(\frac{\lambda_{h,im}^2}{\lambda_{h,real}^2}) \qquad \forall \lambda_h, \lambda_h^{\dagger} \; complex$$

The reason we believe the system finds this type of solution is to help capture features of the data such as the deflections observed in the trajectories, via rotations, and to create complex patterns that can, in general, capture the heterogeneity in the PSTHs.

The main properties of the solution discussed in here, namely that a large proportion of the eigenmodes learned are slow and complex, is found for all the models we have explored. The exact eigenspectrum configuration and stability of the modes is not unique and changes for different model specifications and hidden dimensionalities.

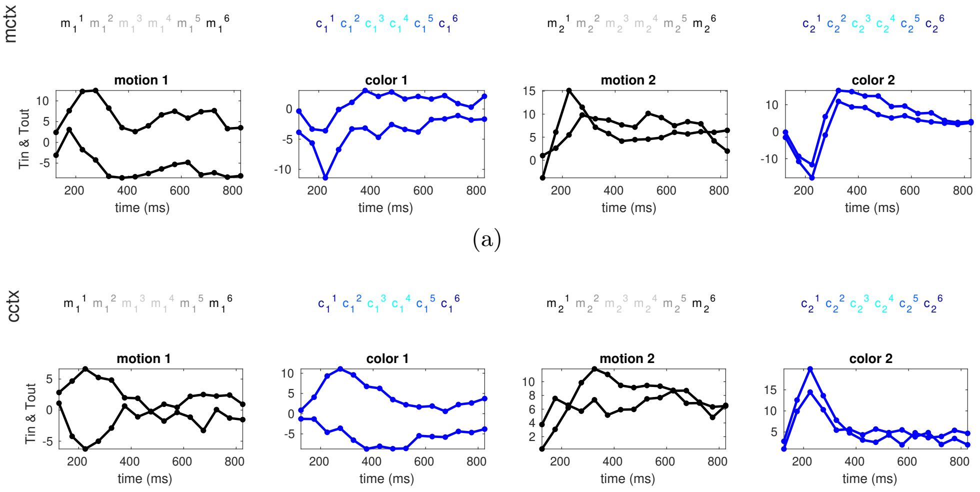
3.3.5.2 Input biases and input directions

The learned input parameters are shown in figures 3.12 and 3.13. Figure 3.12 illustrates the time courses associated to each input in each of the contexts (scaled by the corresponding input vector norm). To obtain the total input influences for each condition, the temporal input traces are then multiplied by the associated coherences-related scalings, whose learned values are shown in 3.13. Two of the inputs clearly capture a coherence related signal, which is scaled differently for each coherence level and has a different sign for Tin and $Tout$ types of evidence. The two other inputs also seem to reflect evidence strength, but surprisingly, not direction. Therefore, for each input modality, color and motion, two type of signals are learned: one that captures coherence magnitude and sign and a second one that only carries coherence magnitude information. These type of signals, however, could simply reflect a particular decomposition along a given basis in the two-dimensional input subspace learned. Set, for instance, by our initialization and model input constraints. Furthermore, the input bases B_c^{cx}, B_m^{cx} are not orthonormal. We therefore specifically looked for directions within the 2D input subspaces that carried coherence information, via simple regression on the input patterns, and found the orthogonal component to them. In this way, we can construct an orthonormal basis in which to represent the inputs in a more intuitive way. Using this procedure, the second dimension –now orthogonal to the dimension carrying coherence information– still reflected coherence magnitude. Such representation is shown in 3.14 –for individual dimensions– and in 3.16a –plotted in 2D. Note that we also subtracted in here the across-condition mean of the inputs, to remove condition independent effects.

In the case of the JF model, we did also recover the same type of coherence-related input scalings B.20, which were similar to the values learned by the LDS 3.13.

Two additional observations are to be made: first, the input time courses have a transient nature, presenting an initial peak that decays towards zero and in some cases is followed by a rebound. In the case of the magnitude signals for motion, inputs seem to be more sustained, in particular during the color context. Second, input strengths are comparable across contexts. We emphasize in here that exact comparison of inputs influences must be made with caution, as the influence of the inputs is always coupled to the dynamics. This gives rise to degeneracies in the model, which means that the same results can be obtained using different parametrizations. Nevertheless, we found that we recovered similar input temporal patterns for different hidden dimensionalities and model specifications. For instance, compare 3.14 with what is found for model 1 (see B.21), which had no constraints in the input time courses. In all the solutions we obtained, input strengths across contexts were similar, although the exact values varied slightly across models. Therefore, we will restrain ourselves of making any claims regarding the absolute strength in which the inputs might be biasing the real system in PFC. When discussing our results, we will focus only on the invariant features we obtain throughout all the different models tested.

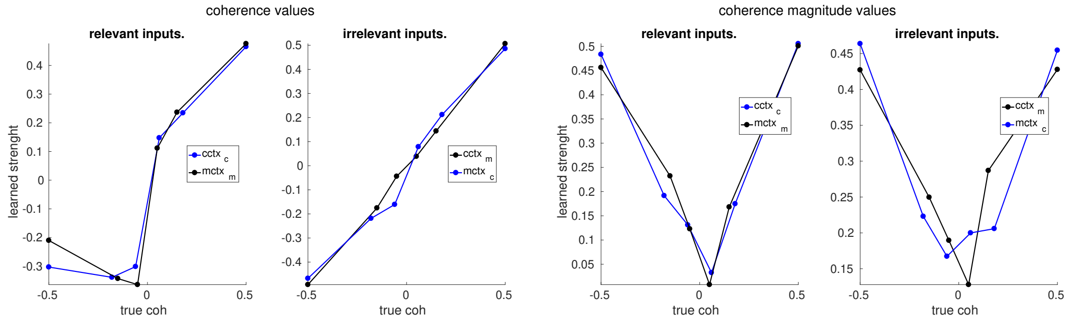
Finally, we compared the input directions obtained within the input subspaces in each context with the input vectors defining the task-relevant subspace in Mante et al. (2013), which were computed on the whole data set via TDR (see 2.1.3). The coherence related input directions estimated using the two methods share a clear correspondence,



(a)

(b)

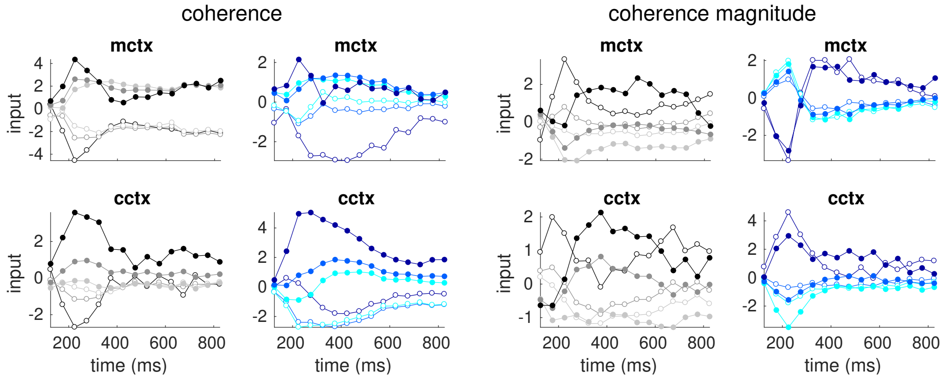
Figure 3.12: LDS learned input time courses in the motion (top) and color (bottom) contexts. Each column shows the input signal associated to each input dimension. Two different time courses are inferred per input, one for positive (Tin) coherences and the other for negative ($Tout$) coherences. To obtain the total input signals for each condition, the temporal input traces are scaled by the corresponding coherence levels (letters on top of the figures, with evidence strengths indicated by the blue and gray scales). The exact learned coherence values are shown in 3.13. To illustrate the two different types of coherence signals found –signed and not signed– we have multiplied the time courses by the sign of the associated coherence values in 3.13. We have also scaled the whole input time series by the norm of the associated input vectors.



(a)

(b)

Figure 3.13: LDS inferred coherence values for each of the four motion and color input dimensions in each context a) associated to the color and motion dimensions that carry coherence information b) associated to the color and motion dimensions that carry coherence magnitude information. The first set of learned values match the true coherence values set in the experiment. The second type of signals, however, are not signed and reflect only the coherence level or strength. In both a) and b) the plots on the left/right show the inferred coherence values when they are relevant/irrelevant.



(a) (b)

Figure 3.14: LDS input signals in the orthogonalised 2D input subspaces. Independent 2D bases are computed for each pair of color and motion input vectors within each context. a) The first dimension is a regression-identified direction that carries coherence information (e.g. top left plot, for motion in the motion context) b) The second dimension is orthogonalised with respect to the first and reflects coherence magnitude (e.g. top left plot, for motion in the motion context). We use the same convention as with the trajectories to indicate coherence level (gray and blue scales) and sign (filled and hollow circles).

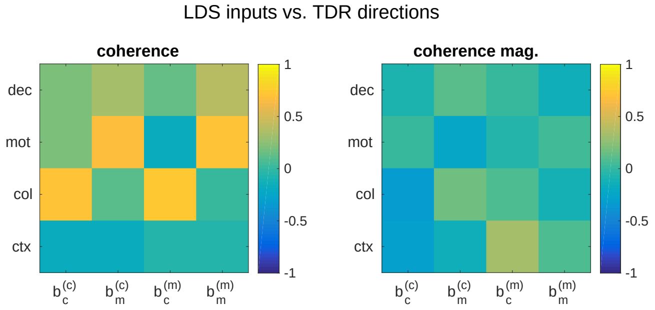


Figure 3.15: Dot products between the TDR task-relevant dimensions (color, motion, decision and context) and the input directions inferred by the LDS in the two contexts. Left: coherence input directions. Right: coherence magnitude input directions. We use subscripts to designate modality color/motion and superscripts to indicate the context (color)/(motion)

but are not perfectly aligned (see fig 3.15). The LDS orthogonal dimensions, which reflect coherence magnitude, do not share the correspondence. Similar picture is obtained when using the TDR estimated non-orthogonalised input dimensions. This results suggest that a dimension exists within the identified input subspaces which carries coherence information, and that this dimension is largely invariant across contexts. Later on we will perform a more detailed analysis on the exact correspondence between all the input subspaces.

Before that, we will briefly discuss the type of input patterns and directions obtained when constraining the models to have the same dynamics across contexts (see section 3.3.4). We wanted to understand how contextual changes were implemented in this case, given that population trajectories were successfully captured by the model. We were expecting that, given the lack of flexibility in the dynamics, the model would learn very different input subspaces across contexts. We found, however, that this was not the case (see B.1.2.3). Surprisingly, all the coherence input vectors were learned to point in the same direction, regardless of context and modality. This direction laid in between the TDR estimated color and motion inputs and the decision dimension (fig. B.19). Furthermore, the input temporal patterns were very different from the model with contextual dynamics. Instead of learning transient biases, the dynamically constrained model placed an integration pattern in the input time courses. In this case, the relevant inputs were scaled with a much larger gain (see B.17).

This model manages to differentially rotate and amplify the input vectors in each context in a complicated way, via the change in gains and relying on the oscillatory dynamics along a mode associated with a large negative eigenvalue (see figure B.19). As we mentioned before, even though selective integration is achieved via this mechanism, it is not enough to capture well the variance along the decision dimension (figure B.14). Furthermore, the solution is highly unstable and often leads to heavy oscillations in the output space.

Finally, we fitted a model in which the input time courses were also constrained to be the same across contexts (same as for model 3 in table 3.2). Only the gains (coherence values) were allowed to change. This setting is interesting because in order to capture contextual changes, as the dynamics is fixed, the model can only rely on the gains, the initial conditions and the input directions. We were therefore expecting that under this situation, the model would this time learn different input subspaces in each context. This, again, was not the case and the model converged to the same solution as found before. The trajectories were still well captured. However, in this model it is a bit more apparent that the influence of the irrelevant input is not as successfully ignored (see irrelevant input trajectories in figure B.16c).

We next asked, how is the pattern of integrated inputs reflected in the input subspaces? Does the dynamics change the input representation? With integrated inputs we refer to the color and motion components in the LDS state equation 3.1, which we will call now $\boldsymbol{c}_{int}(t)$ and $\boldsymbol{m}_{int}(t)$ (for simplicity, in the equation below, we do not make explicit the constraints in the input biases $\boldsymbol{u}$). As we did before when comparing the LDS and the JF models, we subtract condition averages from the input components and incorporate them to the base component.

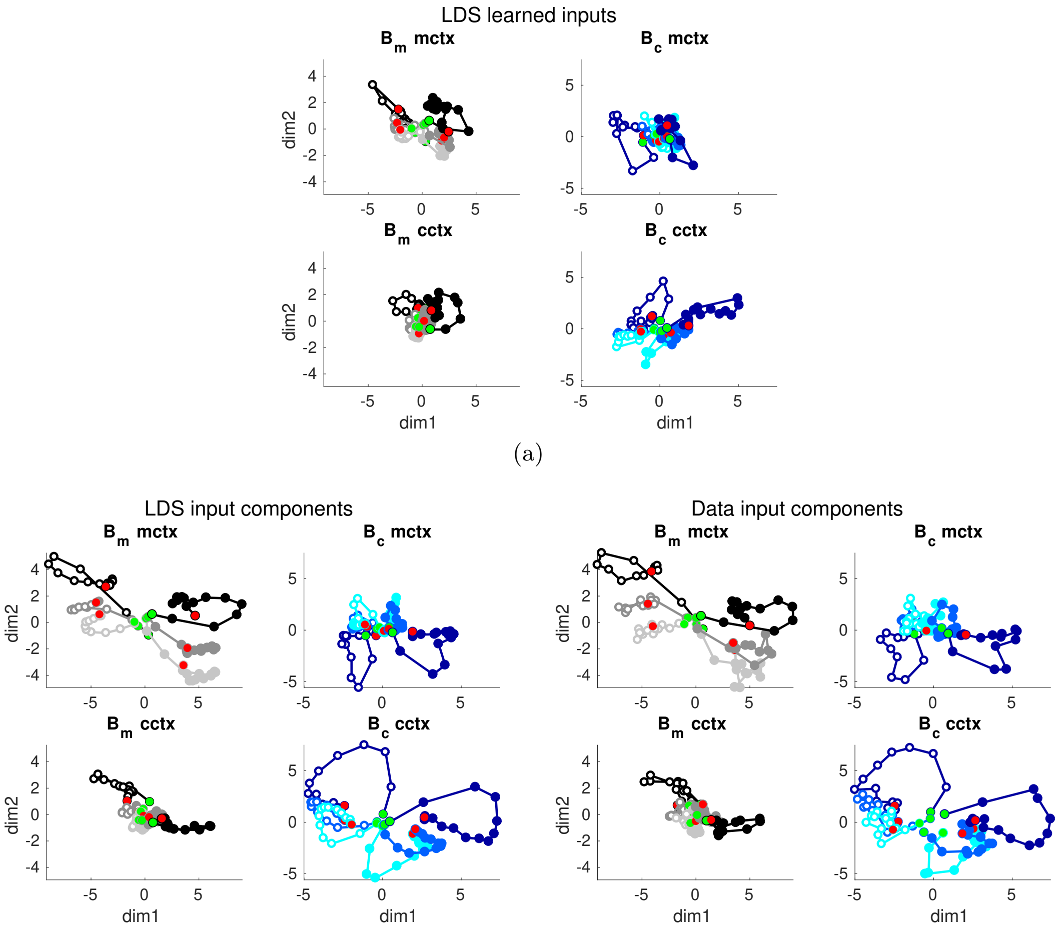
$$\boldsymbol{x}(t) = \underbrace{A^t\boldsymbol{x}(0)}_{base(t)} + \underbrace{\sum_{t'=1}^{t} A^{t-t'} B_c \boldsymbol{u}_c(t')}_{\boldsymbol{c}_{int}(t)} + \underbrace{\sum_{t'=1}^{t} A^{t-t'} B_m \boldsymbol{u}_m(t')}_{\boldsymbol{m}_{int}(t)}$$

At each point in time, these components define a vector whose direction and norm has been shaped both by the dynamics and by the input biases. In figure 3.16 we show that the input dimensions largely preserve the pattern of inputs upon the action of the dynamics. There is a mild scaling – mild compared to the scaling that, as we will see, integrated signals experience in other dimensions, such as the decision dimension– and an smoothing effect, but the overall structure and transient nature of the inputs is not altered by the dynamics. The exception is the pattern of integrated motion in the color context subspace, which seems closer to be 1D rather than 2D. This suggests that either the variance along the second dimension is very small or that the model is simply capturing noise. In fact, as we will show later, adding a second dimension for motion in the color contexts does not seem to substantially change the model's performance.

In the supplementary material we show the alignment, in time, of the motion and color integrated components $\boldsymbol{c}_{int}(t)$ and $\boldsymbol{m}_{int}(t)$ with respect to the coherence inputs (see B.28a). This plot illustrates that the projection of the integrated inputs onto the coherence input dimensions is strong at the beginning of the trial, but goes quickly towards zero as the trial progresses (B.28a). This explains why the action of the dynamics is not strongly reflected along the input dimensions. Interestingly, for motion, the projection reaches its lowest value, close to zero, and immediately afterwards an upwards rebound follows. This is stereotypical of both the integrated relevant and irrelevant inputs and as we will see, it is also observed in the second monkey (B.56a). The final integrated motion dimensions, in both cases, are rendered almost orthogonal to the input dimensions.

Finally, we wanted to find out how aligned the identified input subspaces are, both within and across contexts. For that, we first computed the amount of variance shared between the different subspaces (see B.1.3.2). We did it so for a wide range of hidden dimensionalities (see figure 3.17). When the dimensionality of the hidden state is small, the input subspaces, not surprisingly, are aligned and share a lot of variance. This is the case both within contexts and across contexts. As the hidden dimensionality increases, the shared variance drops quickly close to zero. This, however, occurs only for the color-motion input subspaces within contexts. Across contexts, color-color and motion-motion subspaces share a substantial amount of variance (around 80% for color and 40% for motion, for H=22).

We next asked, is the overlap between the subspaces –or the lack of it– expected for any pair of random subspaces? or more specifically, what is the expected null distribution of overlaps for pairs of randomly drawn subspaces? Considering the size of the observational space ($N = 727$), a fair control would be to draw random samples constrained to the region of state space where the data lives. Following the approach in Elsayed et al. (2016), we drew random subspaces biased to the data covariance structure, which in our case had been projected onto the hidden subspace defined by the columns of the loading matrix C. We then measured minimum and maximum subspace angles and compared it to what is expected from this null (see B.1.3.2). In figure B.22 we showed that within contexts, motion and color subspaces are close to orthogonal. However, this is in fact expected by chance (the curves lie close below, but not above the null 95th



(a)

(b)

(c)

Figure 3.16: a) LDS input signals in the orthogonalised 2D input subspaces. Independent 2D bases are computed for each pair of color and motion input vectors within each context. The first dimension (x axis) defines a direction within the 2D subspace that carries coherence information. The second dimension (y axis) is orthogonalised with respect to the first and reflects coherence magnitude. b) Color and motion integrated input components projected onto the same input subspaces as in a) c) Data input components, projected onto the same subspaces. To reveal the pattern of integrated inputs in the real data: first, we project the observations onto the low-dimensional hidden space using the loading matrix C. Second, we subtract to the data the LDS condition independent (base) component. Finally, we show color/motion trajectories where the motion/color contribution has been averaged out. Green/red dots indicate the beginning/end of the trial.

percentile for random angles). Across contexts, color subspaces are significantly more aligned than expected by chance (the curve lies well below the null 5th percentile for random angles). In the case of the motion subspaces, this is also true for the estimated minimum subspace angle, but not for the maximum one.

Given that the subspaces across contexts seem to be aligned along some dimensions more than others, we looked at the correspondence between individual directions. In particular, we were interested in comparing the coherence and coherence magnitude directions identified by regression within each of the input subspaces. The results are shown in figure 3.18.

We first note that for a small number of hidden dimensions the directions estimated seem to be quite variable, but they later stabilize. This happen around the range of hidden dimensionalities where the best cross-validation performance is achieved (figure 3.5). The pattern of the error can therefore explain some of the noise in the curves shown in 3.18. Furthermore, for a small number of hidden dimensions, the model lacks flexibility to capture multiple dimensions in the data, so the input parameters are used to explain any type of variance. There is an additional feature in the curves worth discussing, namely the general downwards trend showing a decrease in the alignment as a function of the hidden dimensionality. This can be seen both in figures 3.18 and 3.17. This can be explained by the fact that with increasing dimensionalities vectors in general become more and more orthogonal.

Coming back to figure 3.18, we obtain that for the coherence related inputs, the alignment across contexts is larger than expected by chance (the curves lie well above the null 95th percentile for random angles). Note that the 5th percentile is zero, indicating that randomly sampled vectors are likely to be orthogonal. For the coherence magnitude related inputs, this is true only for color, which is expected given the results on the whole input subspaces. We now know though, that within those independently estimated input subspaces, a direction carrying coherence related variance exists which is largely invariant across contexts, and more, for the case of the color, there is an additional dimension carrying coherence magnitude information which is also invariant.

We then proceed to compare the same dimensions against the task-relevant input directions identified in Mante et al. (2013), as we did in 3.15. We will do it this time for a wide range of hidden dimensionalities and comparing the results to chance distributions. The results obtained in figure 3.19 indicate a clear correspondence between the LDS identified input dimensions and the TDR directions, higher than expected by chance. The directions estimated by the two methods, however, are not identical.

In the supplementary material we show alignment, in time, of the LDS motion and color integrated components $\boldsymbol{c}_{int}(t)$ and $\boldsymbol{m}_{int}(t)$ with respect to the TDR inputs (see B.28b). This plot suggests that the TDR procedure may be in fact detecting the early amplification of the inputs by the dynamics before they start getting rotated away from the input subspaces and towards the decision axis.

To end this section, we wanted to test one final thing. The results we just discussed

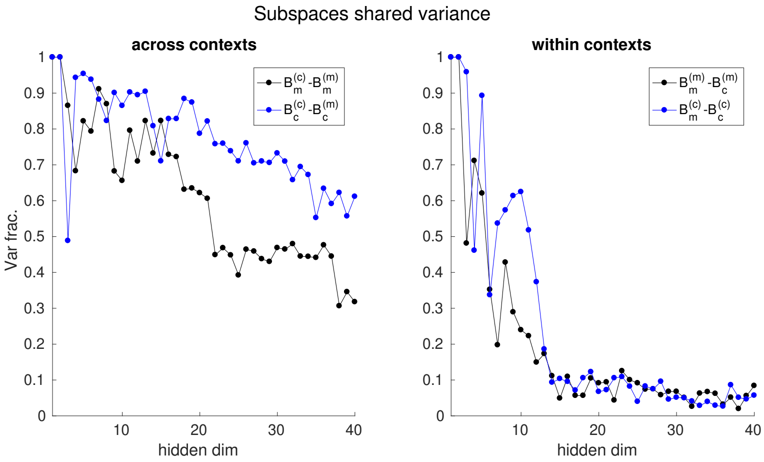


Figure 3.17: Shared variance across the different LDS input subspaces estimated for a wide range of hidden dimensionalities. Left plot, fraction of variance shared between the color/motion subspaces across contexts. Right plot, fraction of variance shared between the color and the motion subspaces within a given context. We use subscripts to designate modality color/motion and superscripts to indicate the context (color)/(motion)

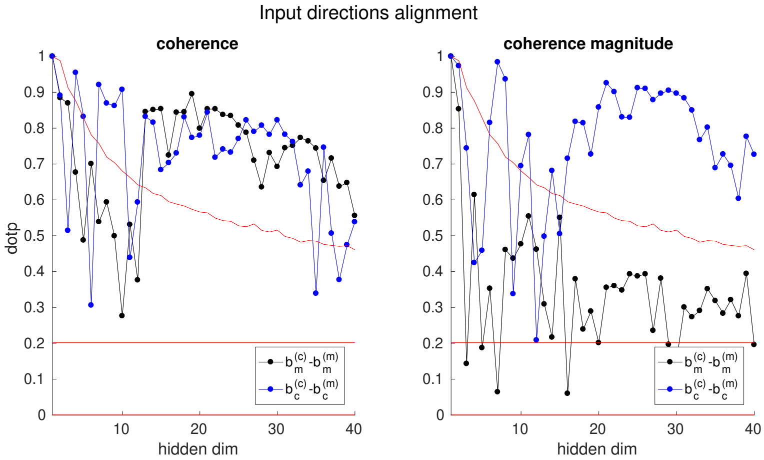


Figure 3.18: Across contexts alignments (dot products) for the LDS motion and color input directions. Left plot, coherence dimensions. Right plot, coherence magnitude dimensions. Alignments are computed for different hidden state dimensionalities. Red lines correspond to the 5th (found close to zero) and 95th percentiles of the null distribution for random alignments. The null is restricted to follow the data covariance structure –once projected onto the hidden subspaces, for the different dimensionalities. The constant lines correspond to the null estimated on the whole data space, without projecting the covariance onto the hidden subspaces.

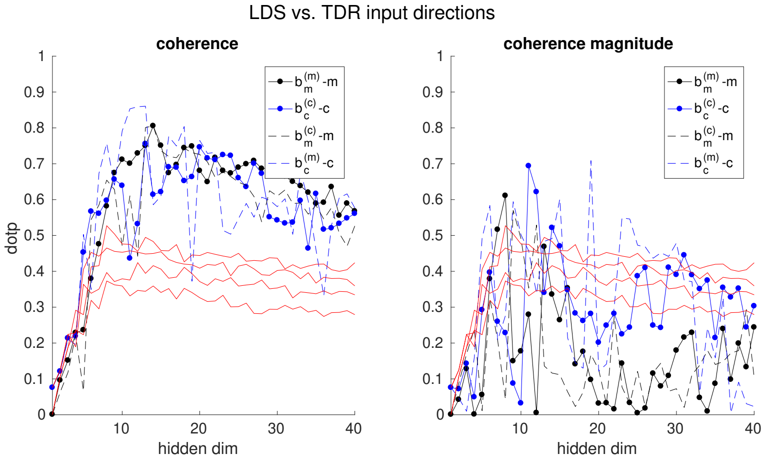


Figure 3.19: LDS input dimensions and TDR task-relevant input directions (m,c) alignment for different hidden dimensionalities. Left plot, coherence dimensions. Right plot, coherence magnitude dimensions. Solid/dashed lines are used for the estimated relevant/irrelevant inputs. Red lines indicate 95th percentiles of the null distribution for expected random alignments with respect to the TDR input vectors (4 lines, corresponding to the color and motion nulls in each contexts). Note that the two nulls within each context, the one for motion and the one for color, are not identical, indicating that the color and motion input dimensions reflect different proportions of the variance.

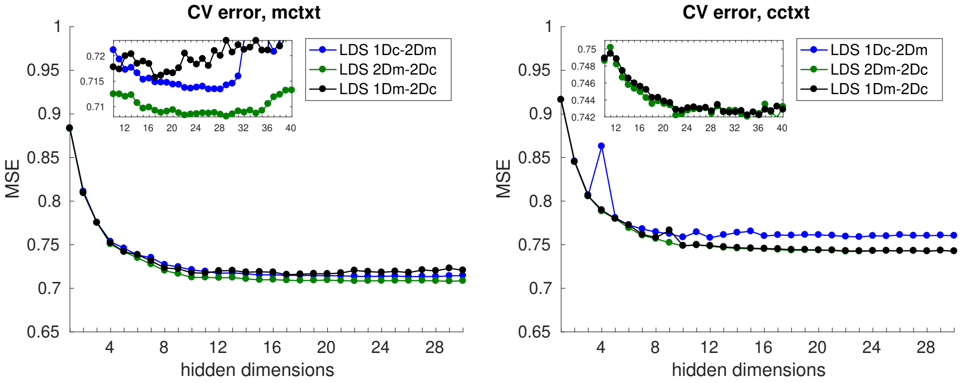
suggests that the dimensionality of at least one of the input subspaces, namely the motion subspace in the color context, is lower than 2D. We therefore fitted models varying the dimensionality of the color and motion subspaces independently and looked for the impact on performance for each individual context (see figure 3.20). The results show that, indeed, adding a second input dimension for motion in the color context barely improves performance, indicating that this secondary dimension is either capturing noise or accounts for very little variance in the data.

3.3.5.3 Dynamics

In order to understand exactly how the LDS is achieving the contextual integration, we will take a closer look at the equations describing the dynamics of the system. Let's rewrite the noiseless state evolution equation so that it is expressed in its eigenmodes basis (see Appendix for detailed derivations A.2). For that, we take the eigendecomposition of the transition matrix A, where R is a matrix containing the right eigenvectors in its columns, $L = R^{-1}$ contains the left eigenvectors in its rows and Λ is a diagonal matrix with the eigenvalues in its diagonal

$$\begin{aligned}\boldsymbol{x}(t) &= A\boldsymbol{x}(t-1) + B\boldsymbol{u}(t)\\ \boldsymbol{x}(t) &= (R\Lambda L)\boldsymbol{x}(t-1) + B\boldsymbol{u}(t)\end{aligned}$$

Defining



(a) (b)

Figure 3.20: LDS cross-validation performance as a function of hidden dimensionality in the motion (a) and color (b) contexts and for different dimensions of the color and motion input subspaces.

$$\boldsymbol{\alpha}(t) = L\boldsymbol{x}(t)$$

$$\boldsymbol{\alpha}(t) = \Lambda^t L\boldsymbol{x}(0) + \sum_{t'=1}^{t} \Lambda^{t-t'} LB\boldsymbol{u}(t')$$

The new state vector $\boldsymbol{\alpha}(t)$, expressed in the left eigenvector's basis, contains in its entries the independent components of the dynamics

$$\alpha_h(t) = \lambda_h^t \boldsymbol{l}_h^\intercal \boldsymbol{x}(0) + \sum_{t'=1}^{t} \lambda_h^{t-t'} \boldsymbol{l}_h^\intercal B\boldsymbol{u}(t') \tag{3.4}$$

which evolve independently from each other with a time constant determined by the eigenvalues λ_h. The extent to which the inputs are carried along a given mode h, depends on the projection of the input vectors onto the associated left eigenvector $\boldsymbol{l}_h^\intercal B$ (See Mante et al. (2013) supplementary material for a similar analysis on the RNN). If the mode is associated to an eigenvalue with magnitude $\lambda_h \sim 1$, that is, it is a slow mode, the input will be accumulated and persists dynamically. To map the low dimensional state vector onto the observational space, in our case the neurons' PSTHs, we use

$$\begin{aligned} \boldsymbol{y}(t) &= CR\boldsymbol{\alpha}(t) + \boldsymbol{d} \\ &= C\sum_h \alpha_h(t)\boldsymbol{r}_h + \boldsymbol{d} \end{aligned} \tag{3.5}$$

Note that the firing rates of the neurons are reconstructed via a linear combination of the components $\alpha_h(t)$, which evolve independently along directions in state space $C\boldsymbol{r}_h$ specified by the right eigenvectors of the system. We looked at the decomposition of the dynamics in this basis, but we did not find it particularly insightful. Furthermore, the individual components are hard to interpret. This is because, even though they describe independent components of the dynamics, the dimensions along which they evolve are not orthogonal, as the dynamics matrix is non-normal.

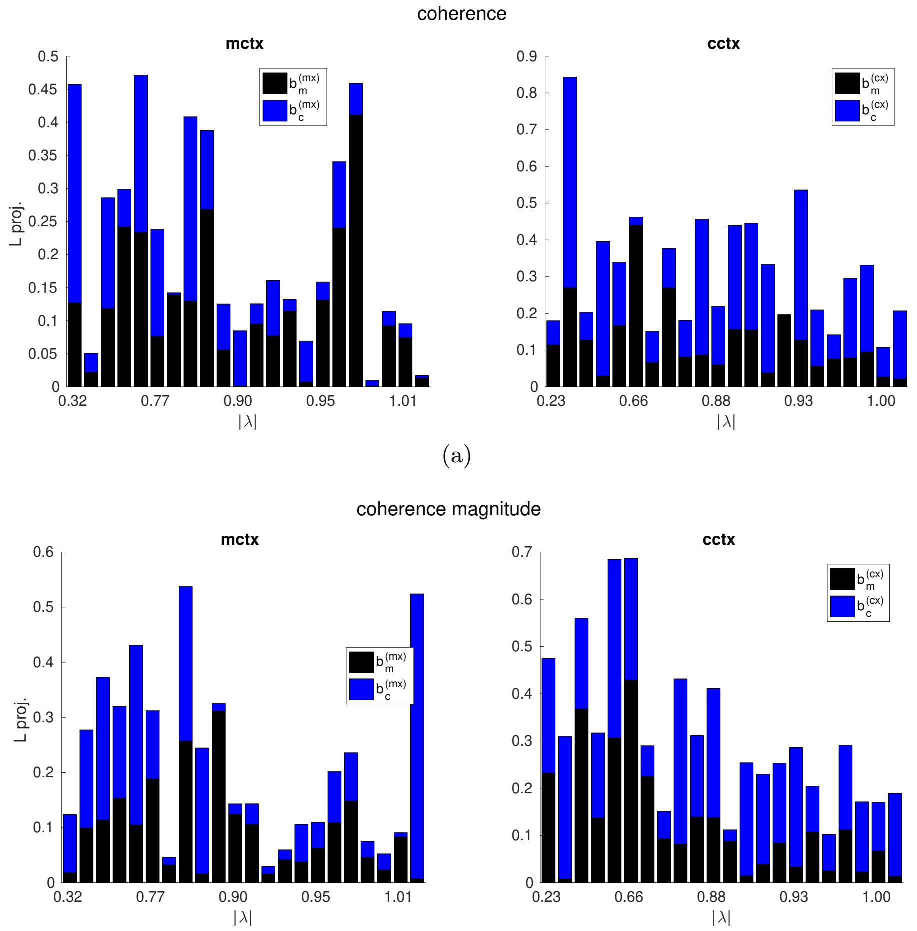
Contextual selection mechanism

Given that in our model,

1. Color and motion inputs strengths are comparable across context (figures 3.14 and B.21), which suggests that no gating of the irrelevant inputs takes place.
2. Input directions are largely invariant across contexts, in particular for the coherence related inputs (figure 3.18).

This suggests that the dynamics plays a role in the selective integration of relevant inputs. As we explained before, the extent to which the inputs are carried along a given mode h depends on the projection of the input vectors onto the associated left eigenvector $\boldsymbol{l}_h^\intercal \boldsymbol{b}_c$, $\boldsymbol{l}_h^\intercal \boldsymbol{b}_m$. For the modes associated with a complex conjugate pair, inputs are mapped into a plane, as determined by the projections onto the real and imaginary components of the complex vectors $\boldsymbol{l}_h - \boldsymbol{l}_{h^\dagger}$ (see Appendix A.2). If the directions $\boldsymbol{b}_c$ and $\boldsymbol{b}_m$ are the same across contexts, the eigenvectors of the system must change to implement a different mapping, i.e. perform selection. In figure 3.21 we show the projection of the inputs onto the left eigenvectors of the system inferred in each context. We do this for the two input directions identified within the 2D input subspaces that reflect coherence and coherence magnitude information. Note that due to the non-normality of the dynamics matrix, the eigenvectors are not orthonormal, so we have normalized them to compute the dot products. We observe that in fact the projection pattern changes so that, in the case of the coherence related inputs, the relevant inputs in each context have a larger projection onto the slowest modes of the dynamics. Note that the total load from each input at a given time is not only given by this projection, as the scaling from the input time courses must be taken into account. The learned input patterns have comparable strengths across modalities and contexts, so the picture implied by 3.21 holds. Interestingly, a similar picture is obtained for the second input pattern, with the difference that the color magnitude signal seems to be mapped so that it persists in the motion context too (we will comment on this later). Finally, note that this projection is largely distributed across all of the modes. The inputs do not target any specific mode, as had been suggested in the previous study (Mante et al., 2013). As we mentioned before, the eigenspectrum and projection patterns are not unique and change slightly depending on model specifications and hidden dimensionality. However, the main feature discussed in here, namely, the change in motion and color projection patterns across contexts, is always found.

In the supplementary material we attach a scatter plot version of figure 3.21, in case the reader finds this depiction of the data more intuitive (see fig. B.26)



(a)

(b)

Figure 3.21: LDS input directions projection onto the left eigenvectors of the dynamics –which were normalized to be unit norm– a) for the inferred motion and color coherence inputs b) for the motion and color coherence magnitude inputs. Each left eigenvector is associated with an eigenvalue, which is specified in the x axis (sorted by magnitude). For complex-conjugate pairs we consider the projection onto the real and the imaginary components, which define the complex planes. (a,b) left figure, motion context; right figure, color context.

Model	min CV error	H
LDS $B^{(c)}_{c,m} \neq B^{(m)}_{c,m}$	0.7253	22
LDS $B^{(c)}_{c,m} = B^{(m)}_{c,m}$	0.7271	27

Table 3.5: LDS minimum cross-validation error, for unconstrained (top) and constrained (bottom) input directions across contexts, and corresponding hidden dimensionality H for which it is achieved.

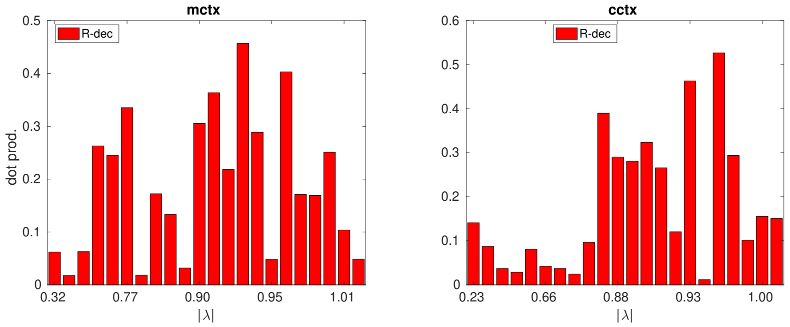
Finally, we decided to perform one last control. We found that color and motion input subspaces are largely aligned across contexts, but do not perfectly overlap –without considering here the second motion dimension in the color context, which seems to be superfluous. We wanted to know whether this small change in orientations is indeed necessary to capture the data well. Therefore, we decided to fit an LDS constrained to learn the same input subspaces across contexts. For this model, the trajectories in the task-relevant subspace were accurately captured, suggesting that selective integration was properly achieved (figure B.24). Furthermore, this model implemented the same solution as the model with flexible input directions, obtaining similar input directions and input biases. The pattern of integrated inputs along these identified input dimensions were also very similar. However, we found that some fraction of the variance was being missed along the decision dimension (see figure B.23). Careful inspection of the trajectories revealed that some features were, in fact, not as well reproduced when compared to the flexible model –a few other features, nevertheless, were arguably better captured. Therefore, incorporating this slight change in directions across contexts seems to be important to capture some aspects of the FRs accurately. In fact, adding this type of input constraint impairs the performance as much as constraining the dynamics (see table 3.4 and compare with table 3.5). In this case though, the model does not suffer from instabilities. Therefore, we conclude that for monkey A, inputs to PFC bias the network in a similar fashion across contexts, but not via the exact same pattern of modulation.

Integration mechanism

Having understood how the model transforms the inputs under each context, we now want to address the following questions

1. Why does the model learn this particular dynamical system?
2. Why does it have several slow modes?

The fact that the inferred dynamical system contains multiple slow modes suggest that evidence integration does not occur uniquely along a single direction. Furthermore, as those modes are associated with a wide range of eigenvalues, it indicates that there is more than one time constant underlying the dynamics of integration. This can be seen by combining equations 3.4 and 3.5, so that the neurons FRs are expressed as



(a)

Figure 3.22: Projection of the TDR identified decision axis onto the right eigenvectors of the dynamics in each context –which were normalized to be unit norm. We use the loading matrix C to bring the decision vector into the hidden space. Each right eigenvector is associated with an eigenvalue, which is specified in the x axis (sorted by magnitude). For complex-conjugate pairs we consider the projection onto the real and the imaginary components, which define the complex planes.

$$\boldsymbol{y}(t) = \underbrace{\sum_h \left(\sum_{t'=1}^{t} \lambda_h^{t-t'} \boldsymbol{l}_h^\intercal \boldsymbol{b}^1 u^1(t') \right) C\boldsymbol{r}_h}_{\boldsymbol{b}^1_{int}(t)} + ...$$

which illustrates that the integrated input vector $\boldsymbol{b}^1_{int}$ at time t points in a direction given by a linear combination of all the right eigenvectors ($C\boldsymbol{r}_h$ in the output space). The weights of this linear combination are specified by the projection of the input onto each of the left eigenvectors ($\boldsymbol{l}_h^\intercal \boldsymbol{b}^1$). Furthermore, in time, the eigenvalues determine which directions $C\boldsymbol{r}_h$ dominate. In the case that there was a single slow dimension, with $\lambda_h \sim 1$, and the rest of the dimensions were fast decaying $\lambda_{\neg h} \sim 0$, the total integrated vector would trivially point into the direction specified by the right eigenvector $\boldsymbol{r}_h$ associated with this unique slow mode. In the case that several slow modes exist, the integrated vector will be a linear combination of the slowest directions. In agreement with this observation, we find that the decision vector estimated via linear regression (Mante et al. (2013)), which reflects the pattern of integrated inputs, does not preferentially point into any of the slowest modes (see figure 3.22). Note that the dot products are not significantly higher than chance, given a null distribution of random alignments (see figure 3.18, the 95th percentile for hidden dimensionality H=22 is about 0.55) Similar projection patterns are obtained when considering the decision vectors we estimated independently in each context, both for the one that separated motion sign in the motion context and for the one separating color sign in the color context (see below).

We also verified that the right eigenvectors are not aligned among themselves –remember that, as the transition matrix is non-normal, the eigenvectors are not orthogonal– (see figure B.25). The picture that can be derived from this analysis is that the integrated input, in time, gets stretched and rotated by the dynamics and that the subspace in which it lives is spanned by a multitude of slow dimensions.

Patterns of inputs and integrated inputs

We will now analyze how the dynamics in each context transforms the color and motion inputs. From equation 3.1, we can define the integrated inputs in the observational space as

$$\boldsymbol{y}(t) = \underbrace{CA^t\boldsymbol{x}(0) + \boldsymbol{d}}_{base(t)} + \underbrace{C\sum_{t'=1}^{t} A^{t-t'}B_c\boldsymbol{u}_c(t')}_{\boldsymbol{c}_{int}(t)} + \underbrace{C\sum_{t'=1}^{t} A^{t-t'}B_m\boldsymbol{u}_m(t')}_{\boldsymbol{m}_{int}(t)}$$

As we did before when comparing the LDS and the JF models, we subtract condition averages from the input components and incorporate them to the base component.

We will focus first in understanding what our model does with each of the inputs along the TDR-identified decision dimension. In 3.23, we show the pattern of integrated inputs $\boldsymbol{c}_{int}(t)$ and $\boldsymbol{m}_{int}(t)$ along the decision axis. We observe that the input components display selective amplification along this dimension. The condition independent $base(t)$ component also projects into the decision dimension. Its projection pattern along this dimension captures the initial downwards dip in FRs –typically observed in PFC neurons during DM tasks–, followed by a sustained FR increase. This latter effect can be interpreted as a form of urgency signal (Hanks et al., 2011; Mante et al., 2013).

In the supplementary material we repeat the analysis but computing the alignments (dot products), not the full projection (fig. B.29). This illustrates that the integrated input components are in fact rotated by the dynamics so that, in each context, the relevant integrated vectors are gradually brought towards the decision axis, but the irrelevant inputs are mapped into orthogonal directions.

In 3.24, we plot model-generated trajectories for all 36 possible input conditions, projected onto the decision axis. Remember that model trajectories are generated using the previous equation, with the FRs explained as the sum of all the components. In the same figure we show the real data trajectories, projected onto the same direction. The model captures the data well in this dimension, as we have seen before when plotting the trajectories in the whole TDR task-relevant subspace.

Next we asked, if we look for a direction that carries decision information within the relevant integrated input subspaces, will we find the same decision dimension? For that, we further expanded the previous equation, so that we explicitly see each individual input dimension (four in our case).

$$\boldsymbol{y}(t) = \underbrace{C\sum_{t'=1}^{t} A^{t-t'}\boldsymbol{b}_c^1 u_c^1(t')}_{\boldsymbol{c}_{int}^1(t)} \underbrace{+C\sum_{t'=1}^{t} A^{t-t'}\boldsymbol{b}_c^2 u_c^2(t')}_{\boldsymbol{c}_{int}^2(t)}$$

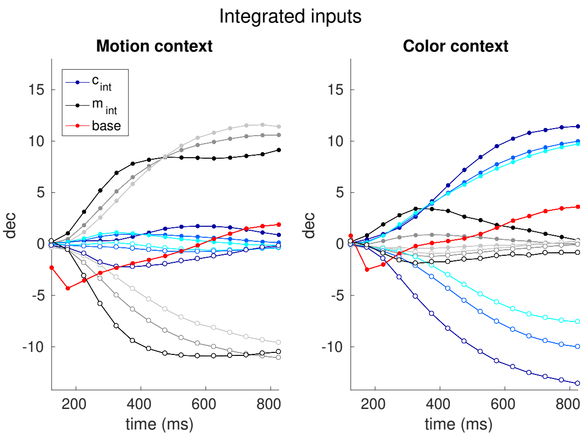
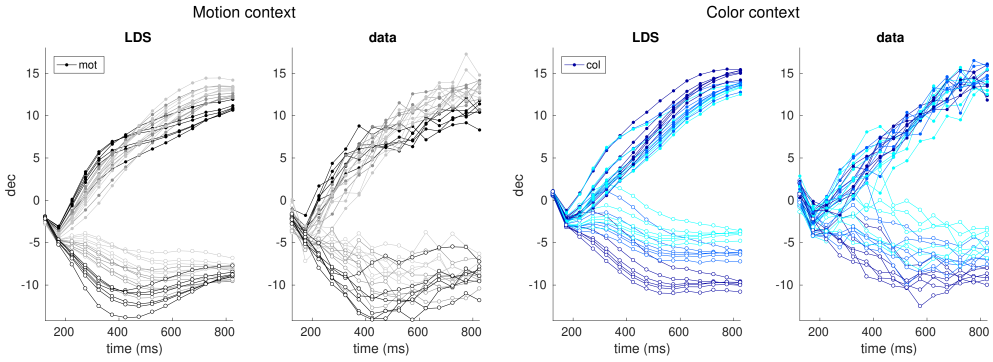


Figure 3.23: LDS integrated inputs and base components projected onto the TDR decision axis, for all possible input values (6 each for color and motion). Same conventions as in previous plots applies.



(a)

(b)

Figure 3.24: LDS model and data trajectories, for the 36 possible conditions, projected onto the TDR decision axis. Each condition is color-coded according to the relevant modality for each context. The data has not been smoothed.

$$+ \underbrace{C \sum_{t'=1}^{t} A^{t-t'} \boldsymbol{b}_m^1 u_m^1(t')}_{\boldsymbol{m}_{int}^1(t)} + \underbrace{C \sum_{t'=1}^{t} A^{t-t'} \boldsymbol{b}_m^2 u_m^2(t')}_{\boldsymbol{m}_{int}^2(t)} + ...$$

At any given moment in time, the integrated color and motion input subspaces are defined by two planes, specified by the vectors $[\boldsymbol{c}_{int}^1(t), \boldsymbol{c}_{int}^2(t)]$ and $[\boldsymbol{m}_{int}^1(t), \boldsymbol{m}_{int}^2(t)]$. Using linear regression, we looked for a dimension within those planes that separated conditions according to decision, or equivalently, by relevant input coherence sign –as we use only correct trials. We considered the integrated inputs at the last time step, right after stimulus offset, when the decision signal is found to be the strongest (Mante et al., 2013). For instance, in the color context, we took the integrated input vectors at the last time step $[\boldsymbol{c}_{int}^1(T), \boldsymbol{c}_{int}^2(T)]$, for all conditions, and looked for a dimension that separated color coherence sign. We did the same in the motion context, but looking for dimensions within the motion subspace that separated motion coherence sign.

What we found is that, indeed, the decision-like dimensions identified in each context, which we name $d_c^{(c)}$ and $d_m^{(m)}$, point in the same direction (figure 3.25a) Furthermore, these dimensions do in fact correspond to the decision axis identified via TDR on the two contexts (figure 3.25b). Trajectories along these two dimensions (figure 3.26), do offer the same picture as in 3.24, with a single decision axis.

We also looked for dimensions, within the irrelevant input subspaces this time, that separated conditions by the irrelevant input sign. We refer to them as $d_m^{(c)}$ and $d_c^{(m)}$. Given that the total color and motion irrelevant input signals does not seem to decay to zero at the end of the trial (figure 3.8), it could well be that the dynamics is also doing something with those inputs, but the transformation occurs along a different direction from the decision axis. In figures 3.27 and 3.28 we show that the irrelevant inputs are in fact integrated, but in a much lower extent. The dimensions where they are eventually mapped at the end of the trial are found to be orthogonal to the decision dimension (figure 3.25). The fact that they are orthogonal, however, is expected by chance as we explained before. Note that the pattern of the trajectories along these dimensions does not follow from the pattern of learned inputs, indicating that the dynamics is indeed transforming these inputs too.

Finally, we looked into the dimensions orthogonal to $(d_m^{(m)} - d_c^{(c)})$ and $(d_m^{(c)} - d_c^{(m)})$, which we call $(d2_m^{(m)} - d2_c^{(c)})$ and $(d2_m^{(c)} - d2_c^{(m)})$. The orthogonal dimensions reflected coherence strength, but not sign (figures B.35 and B.36). We also computed the alignment between these dimensions (fig B.33). Unlike the decision-related dimensions $d_m^{(m)} - d_c^{(c)}$, the orthogonal dimensions $d2_m^{(m)} - d2_c^{(c)}$ were not the same across contexts. Therefore, the integrated relevant subspaces do not completely overlap across contexts, but they do in fact share a single dimension of integration.

Continuing with the picture sketched in the previous section, we explained that the integrated inputs are amplified and rotated by the dynamics within a subspace spanned by a multitude of slow dimensions. What we can add now is that, in each

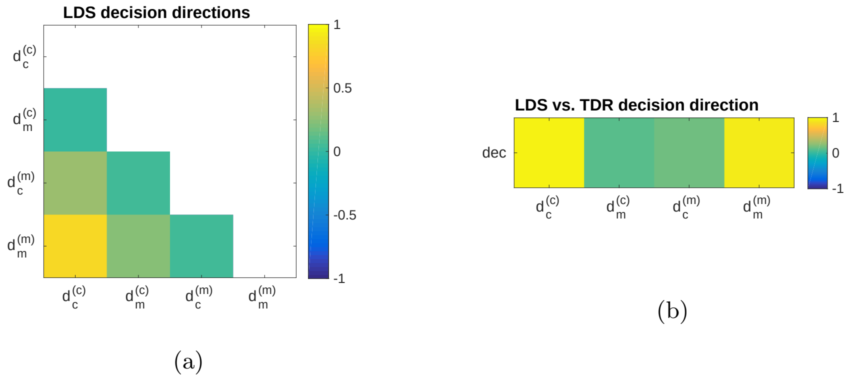


Figure 3.25: Decision-related dimensions within the color and motion integrated 2D subspaces at the last time step. The dimensions are found based on the sign of the relevant inputs, within the relevant inputs subspaces ($d_m^{(m)} - d_c^{(c)}$), or based on the sign of the irrelevant inputs, within the irrelevant inputs subspaces ($d_m^{(c)} - d_c^{(m)}$). The decision dimensions $d_m^{(m)} - d_c^{(c)}$ are found to be the same across contexts. a) decision-related dimensions alignment across and within contexts b) decision-related dimensions and TDR decision axis correspondence.

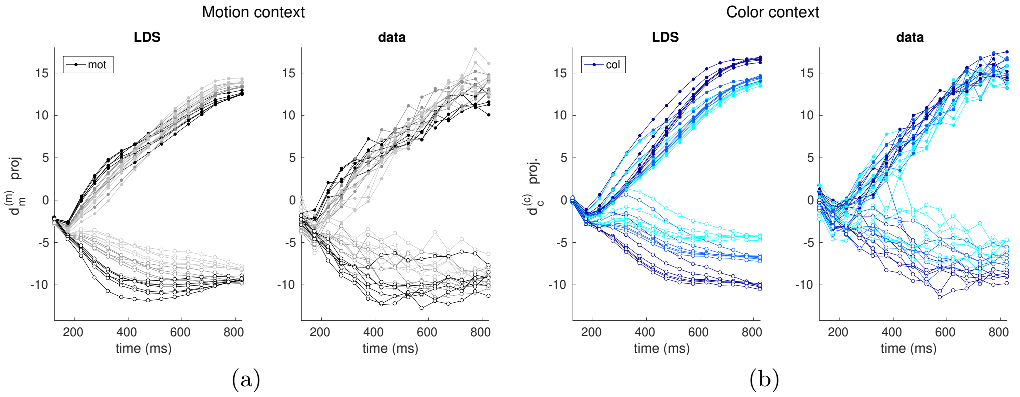


Figure 3.26: LDS model and data trajectories, for the 36 possible conditions, projected onto the "relevant" decision dimension. The y-axis corresponds to a regression-identified direction, lying within the integrated relevant-input 2D subspace, that reflects the sign of the relevant information. Each condition is color-coded according to the relevant modality for each context.

context, the dynamics brings the relevant integrated vectors to the same place in the two contexts, namely, the decision axis. The irrelevant inputs, however, are mapped onto orthogonal directions and the mild integration that they experience is not reflected along the decision dimension.

Finally, we show the pattern of integrated inputs along the coherence input dimensions identified within the input subspaces (figure 3.29, as we did in 3.16b). We use $b1_m^{(m)} - b1_c^{(c)}$ to denote the relevant dimensions and $b1_m^{(c)} - b1_c^{(m)}$ for the irrelevant. Trajectories are shown in figures 3.30 and 3.31. The first observation is that, as we discussed before, the integrated inputs mirror the pattern of the input biases, which are learned to be transient. A second observation is that the integrated irrelevant inputs along this particular dimensions have smaller magnitudes than the relevant inputs. The input biases learned (3.12, 3.14 and 3.16) do not vary substantially across contexts, therefore,

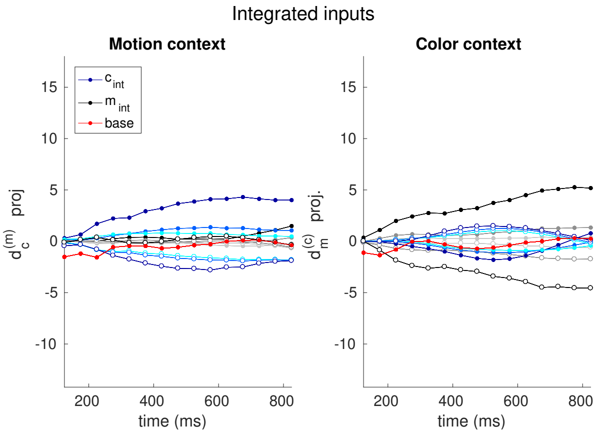
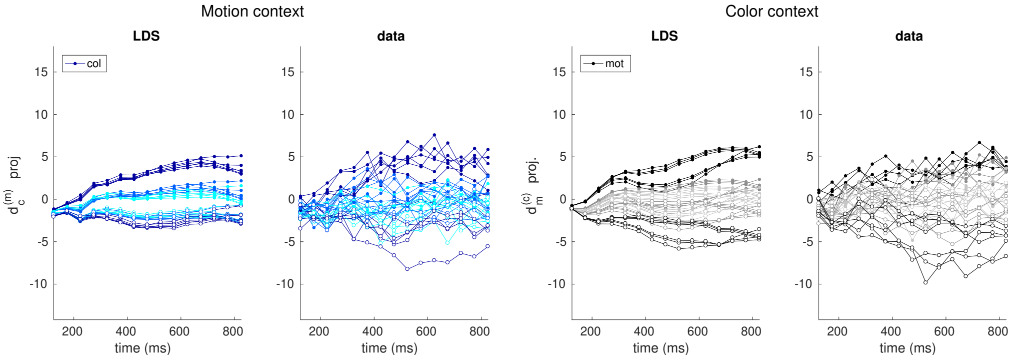


Figure 3.27: LDS integrated input and base components projection onto the “irrelevant” decision dimension, for all possible input values (6 each for color and motion). The y-axis corresponds to a regression-identified direction, lying within the integrated irrelevant-input 2D subspace, that reflects the sign of the irrelevant information.



(a)

(b)

Figure 3.28: LDS model and data trajectories, for the 36 possible conditions, projected onto the “irrelevant” decision dimension. The y-axis corresponds to a regression-identified direction, lying within the integrated irrelevant-input 2D subspace, that reflects the sign of the irrelevant information. Each condition is color-coded according to the irrelevant modality for each context.

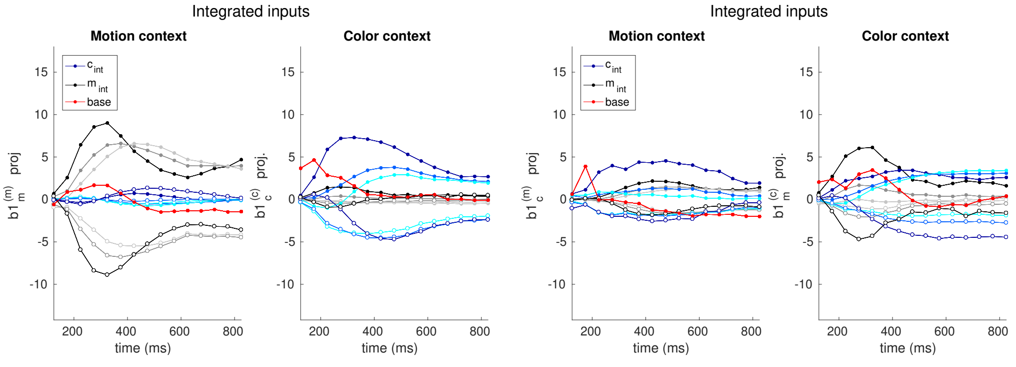


Figure 3.29: LDS integrated inputs and base components projected onto the LDS coherence input directions, for all possible input values (6 each for color and motion) a) relevant dimensions b) irrelevant dimensions.

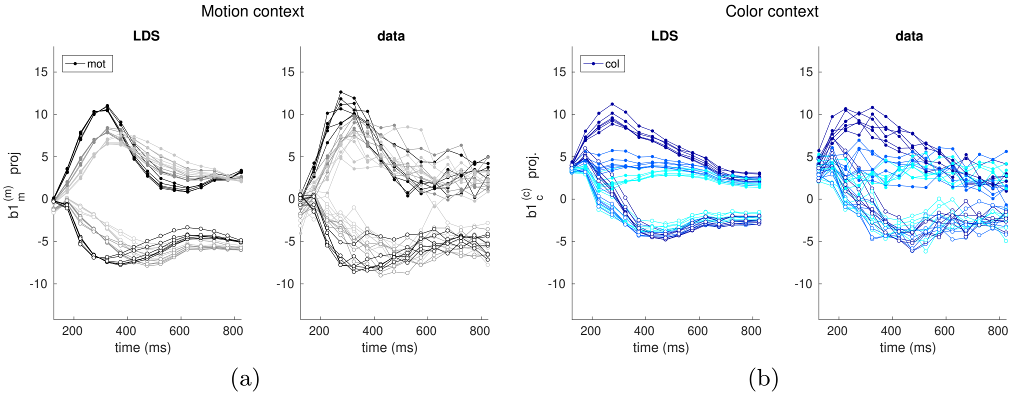


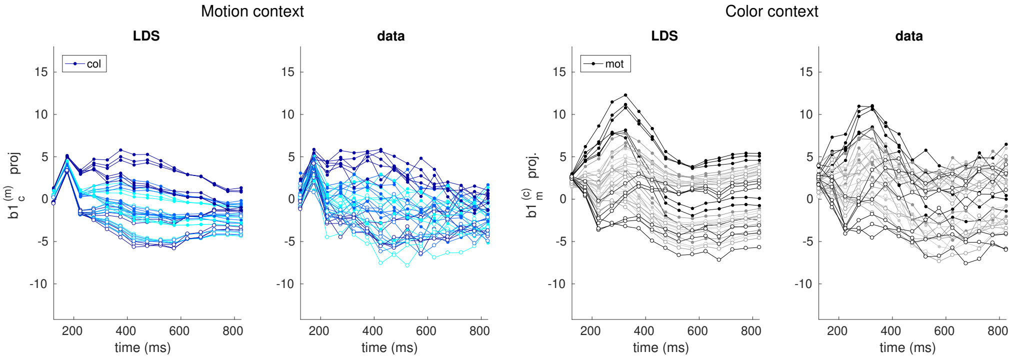
Figure 3.30: LDS model and data trajectories, for the 36 possible conditions, projected onto the relevant coherence input dimensions. The y-axis corresponds to a regression-identified direction, lying within the relevant-input 2D subspace, that reflects the relevant coherence value. Each condition is color-coded according to the relevant modality for each context.

the difference in strengths seems to originate from the projection pattern of the integrated inputs along the input dimensions and not from differences in the input drives.

The corresponding plots for the second input dimensions, which separate by coherence strength, but not sign, can be found in the supplementary material. Coherence magnitude signals are present in the data along the coherence magnitude input dimensions. For the case of motion in the color context, this signal is either very weak or the model is capturing noise, as figures 3.20b, B.30b and B.32b suggest.

Population activity patterns

The dimensions that we have identified allow us to decompose the dynamics of the whole population as a sum of different task-related activity patterns that underlay the tuning of individual neurons. We then wanted to know, how distributed are these patterns of activity in the whole population? How many neurons participate in each of them?



(a)

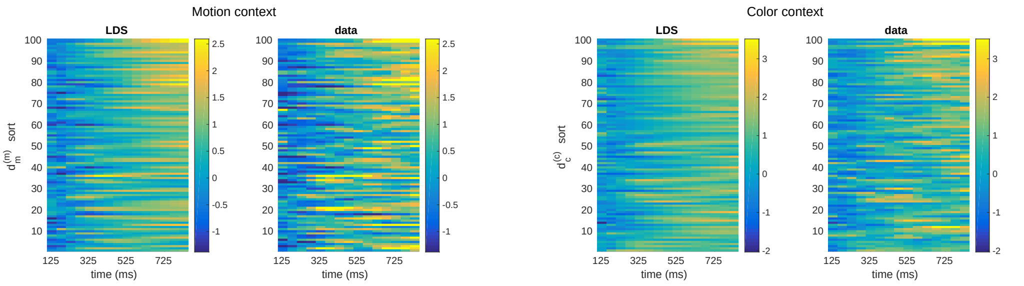
(b)

Figure 3.31: LDS model and data trajectories, for the 36 possible conditions, projected onto the irrelevant coherence input dimensions. The y-axis corresponds to a regression-identified direction, lying within the irrelevant-input 2D subspace, that reflects the irrelevant coherence value. Each condition is color-coded according to the irrelevant modality for each context.

As it was already shown in the study by Mante et al. (2013), the whole PFC population is highly heterogeneous and most cells are characterized for having "mixed" selectivity. We find that, indeed, the dimensions identified are highly uncorraleted (see figure B.39) and that both the inputs and the decision loads into the neurons are broadly distributed. However, a small percentage of the units are strongly tuned to specific parameters and follow closely the activity patterns implied by the input and decision dimensions. To see this, in figures 3.32, 3.33 and 3.34 we show the FRs of the neurons with the strongest loads from the decision, the relevant coherence inputs and the irrelevant coherence input dimensions. The loads are defined by the vectors $Cd^{(m/c)}_{m/c}$, $Cb1^{(c/m)}_{c/m}$, $Cb1^{(c/m)}_{m/c}$, after normalizing them. The PSTHs correspond to the condition with the highest positive color and motion coherence. With this figure we wanted to illustrate that the patterns of activity along the dimensions identified by the LDS are indeed reflected in the population FRs, so the model seems to be extracting meaningful structure from the data. Furthermore, the model seems to approximate well the different patterns at the level of individual neurons. As we will see in the next section, single unit PSTHs are in fact accurately captured by the model.

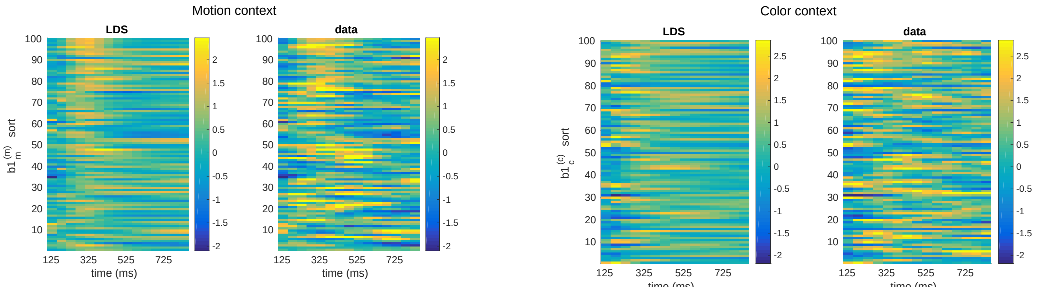
3.3.6 LDS single unit PSTHs

In this section we illustrate how the LDS model performs at the level of individual FR responses. In figure 3.35 we show single unit model PSTHs for different conditions and compare them to the real data. In this particular plot we show conditions where the irrelevant input influence has been averaged out. In the supplementary material one can find the model generated responses for the 36 different conditions B.43. The LDS model is able to capture the whole complexity of individual firing patterns, despite the high degree of heterogeneity in the population responses. These include neurons that strongly and transiently respond to motion and to color (first and third rows); that



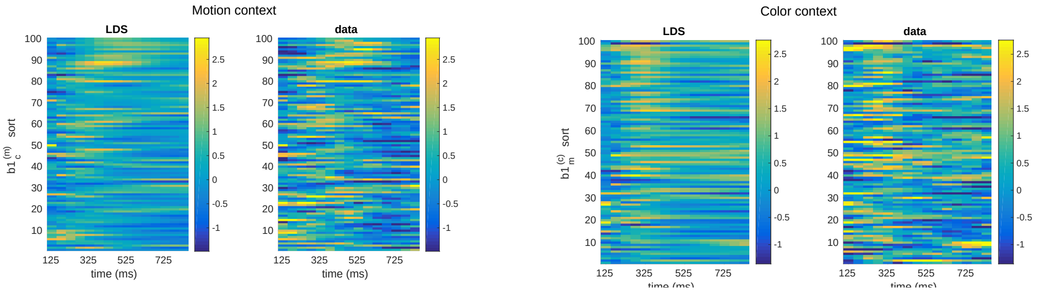
(a) (b)

Figure 3.32: Firing rates of 100 units participating in the activity patterns along the LDS "relevant" decision dimensions (see 3.26). Neurons were selected based on the top 100 loads (by magnitude) and are shown sorted in ascending order. The PSTHs correspond to the condition with the highest positive color and motion coherence. Note that a few neurons have high FRs that are not well captured by the LDS model. To aid visualization, the color scale in the data plots has been saturated to the maximum FR value in the LDS model plots. For the data, responses have been smoothed using a squared window filter, as in Mante et al. (2013).



(a) (b)

Figure 3.33: PSTHs of 100 units participating in the activity pattern along the LDS relevant coherence input dimensions. Same convention as in previous figure.



(a) (b)

Figure 3.34: PSTHs of 100 units participating in the activity pattern along the LDS irrelevant coherence input dimensions. Same convention as in previous figure.

are suppressed and respond to color and motion strengths (second and fourth rows) or that strongly represent decision signals (fifth row) Note, however, that these neurons are not driven only by a single activity pattern. It is evident that a mixture of them is reflected in the FRs. For instance, a clear example can be seen in figure B.45. The unit in the fourth row during the color context starts separating its responses strongly by color strength (and not sign), but then switches to reflect decision, as it does during the motion context.

The fact that we are able to capture the complexity in the FR responses gets us back to the question of why we learned this particular system. As we explained before, the firing rates of the neurons are reconstructed via a linear combination of the components $\alpha_h(t)$, which evolve independently along directions in state space $\boldsymbol{C}\boldsymbol{r}_h$ specified by the right eigenvectors of the system. The fact that the inferred dynamical system contains multiple slow modes suggest that evidence integration does not occur uniquely along a single dimension and that there is more than one time constant underlying the dynamics of integration. In fact, if the system had only one integration mode and the rest of the dimensions were fast decaying, all the neurons would present the same stereotypical ramping-like response profile (except some perhaps reflecting only the input drive). But this is not what we observe. As we have already seen, neurons in PFC have complex, heterogeneous responses. Therefore, a richer basis set, with several components $\alpha_h(t)$ evolving under multiple time constants, is needed to explain the diversity of PFC's firing rates. This can be seen if we express the individual neuron's responses directly as a function of the independent components of the dynamics $\alpha_h(t)$ (see Appendix).

$$y_n(t) = \sum_{h,\,real} w_{nh}\alpha_h(t) + \sum_{h-h^\dagger,\,img} 2(w_{nh^+}\alpha_{h^+}(t) - w_{nh^-}\alpha_{h^-}(t)) + d_n \tag{3.6}$$

where $\alpha_{h^\pm}(t)$ are the sum and difference modes, which are two complementary real solutions from each pair of complex conjugate roots $\alpha_h(t) - \alpha_h^\dagger(t)$

$$\begin{aligned}
\alpha_{h^+}(t) &= \frac{1}{2}(\alpha_h(t) + \alpha_h^\dagger(t)) \\
&= \Re\{\alpha_h(t)\} \\
\alpha_{h^-}(t) &= \frac{1}{2i}(\alpha_h(t) - \alpha_h^\dagger(t)) \\
&= \Im\{\alpha_h(t)\}
\end{aligned}$$

that reflects the contribution from each complex conjugate eigenmode pair to the dynamics of each neuron n . This influence is weighted by the "sum and difference coefficients"

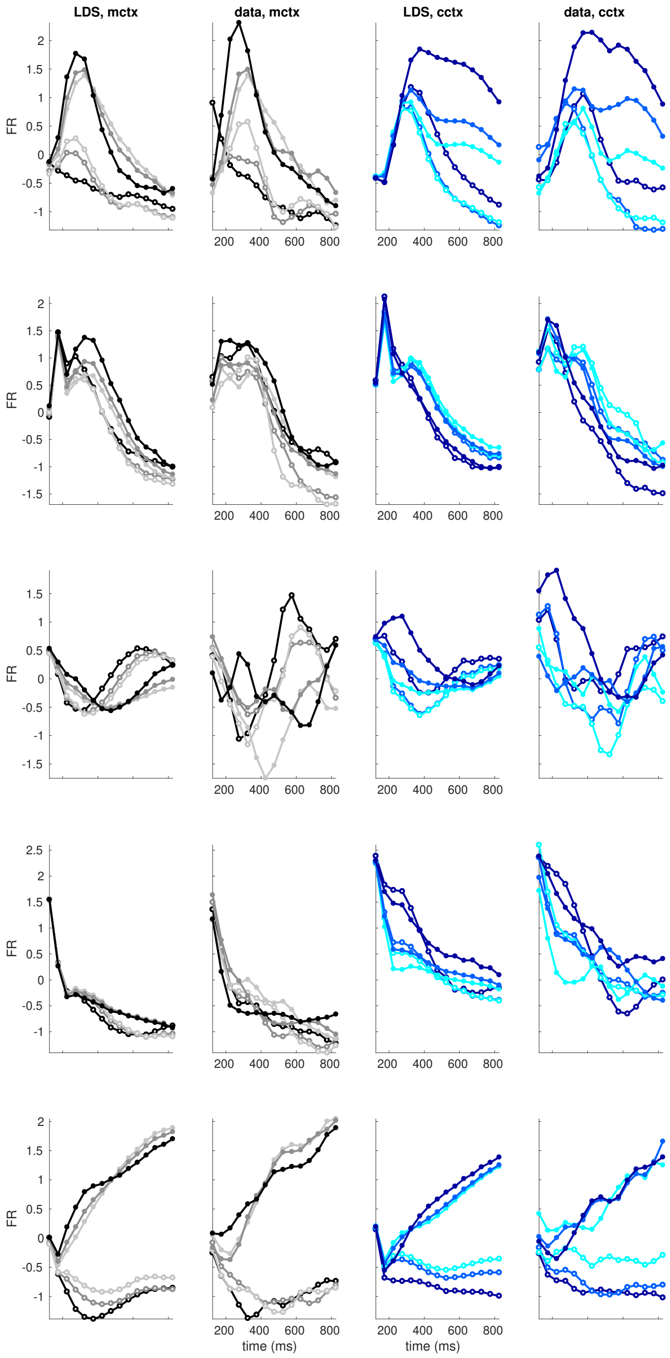


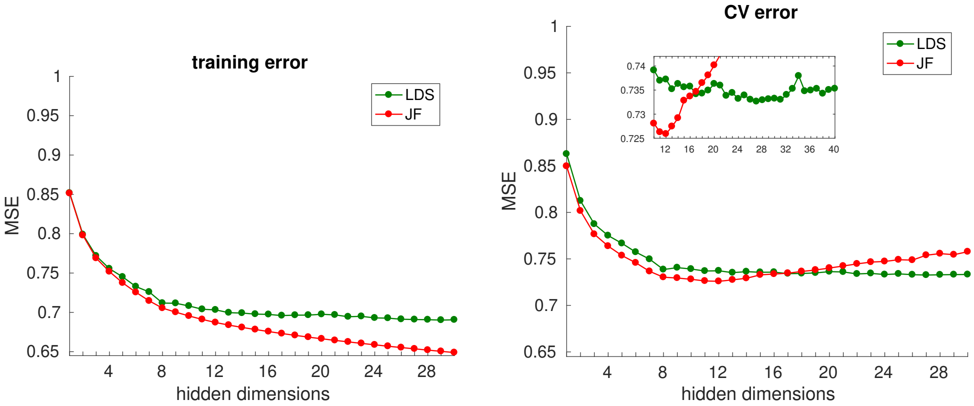
Figure 3.35: Single unit PSTHs sorted by conditions with the irrelevant input influence averaged out (units n=243, 582, 307, 692, 716). First and third column, LDS model. Second and fourth column, PFC data. In each row, from top to bottom, we show units with the largest load from: the motion coherence input (in the motion context), the motion coherence magnitude input (in the motion context), the color coherence input (in the color context), the color coherence magnitude input (in the color context) and the decision axis. For the data, responses have been smoothed using a squared window filter, as in Mante et al. (2013).

$$\begin{aligned} w_{nh^+} &= C_{n,:}\,\Re\{\boldsymbol{r}_h\} \\ w_{nh^-} &= C_{n,:}\,\Im\{\boldsymbol{r}_h\} \end{aligned}$$

and the contribution from each real eigenmode h is weighted by the coefficients

$$w_{nh} = C_{n,:}\,\boldsymbol{r}_h$$

Through these weights we can compute how much each mode contributes to the dynamics of each neuron, which gives us an estimate of the characteristic time constants driving the individual FR responses.



(a) (b)

Figure 3.36: LDS and JF models performance as a function of hidden dimensionality (rank degree for the case of the JF model) a) Training mean squared error b) cross-validation mean squared error. Squared errors are in units of variance, given that the FR responses were z-scored.

3.4 Results: Monkey F

In this section, we perform the analysis on the data recorded from the second monkey. We thought that the results could be very revealing, given the differences observed across the two monkeys –both in the behavior and in the population data (fig. B.46 and B.47). In particular, the representation of the color signal was notably dissimilar. In monkey F, there was some color modulation along the color input axis, but the effect was very weak in the two contexts. Furthermore, color and decision axes were highly correlated. Therefore, the possibility of a gating effect for color in this monkey could not be ruled out. Finally, the bias in the firing rates characteristic of a form of "urgency signal" was also found to be much stronger in this monkey.

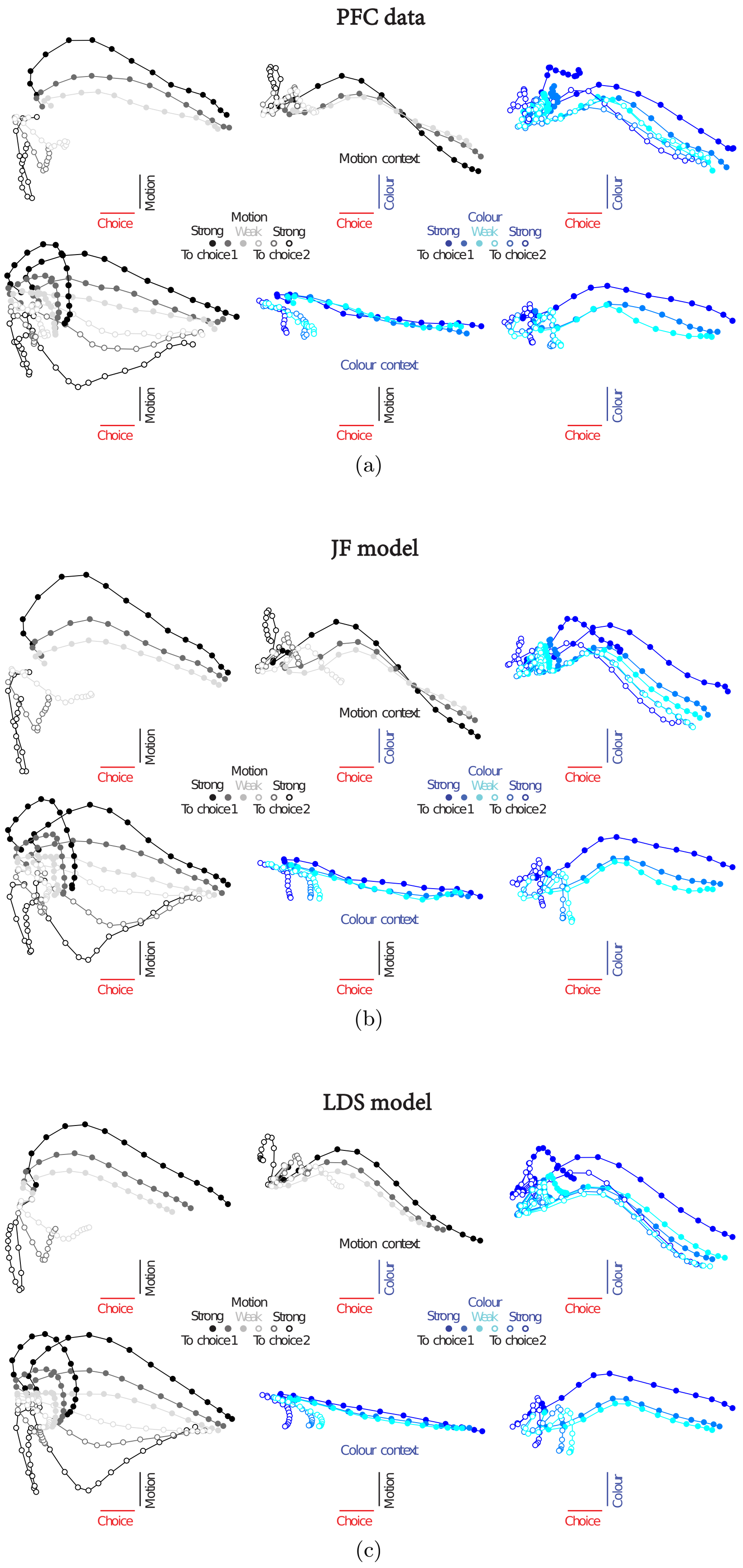
When explaining the results, we will focus on the differences obtained across monkeys, as the commonalities have already been discussed in length during the previous section.

3.4.1 LDS and JF models performance and solution

3.4.1.1 LDS and JF models performance

The two models perform similarly, although for this monkey, the JF model seems to slightly outperform the LDS. In this case, the difference in errors is $\delta MSE = 0.007$, which accounts to 0.7% of the total variance. The two models achieve similar performance across monkeys.

Next, we computed cross-validated trajectories projected into the TDR task-relevant subspace. We find that also for this monkey, the two models accurately capture all the features in the data. This is very interesting given how differently the FR responses seem to be modulated in this monkey, with a very weak color signal and a strong "urgency" effect (see fig. B.47, extended data figure 7 in Mante et al. (2013)).



(a)

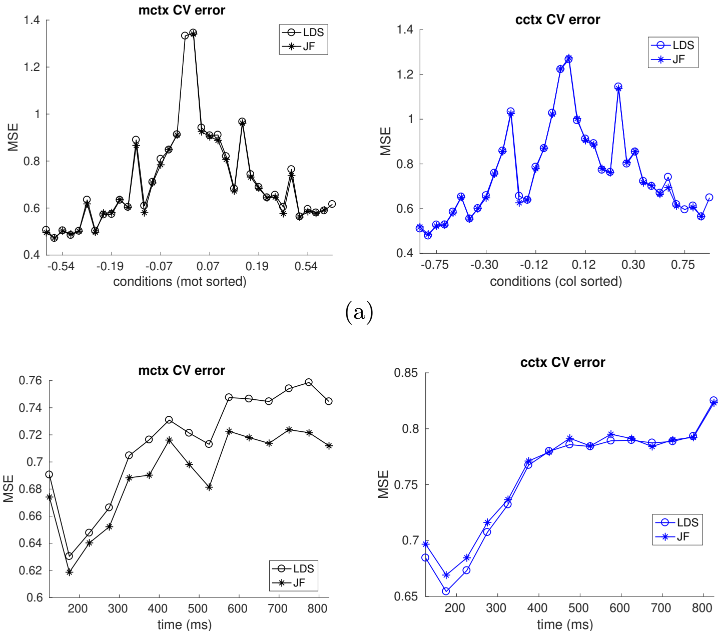
(b)

(c)

Figure 3.37: Data and model generated cross-validated trajectories, leaving one condition out, in the task-relevant subspace identified by regression Mante et al. (2013) a) PFC data b) JF model c) LDS model. Same plotting conventions as in figure 3.2.

Model	min CV error	H
LDS	0.733	27
JF	0.726	12

Table 3.6: LDS and JF models minimum cross-validation error and corresponding hidden dimensionality H (or rank degree) for which it is achieved.



(a)

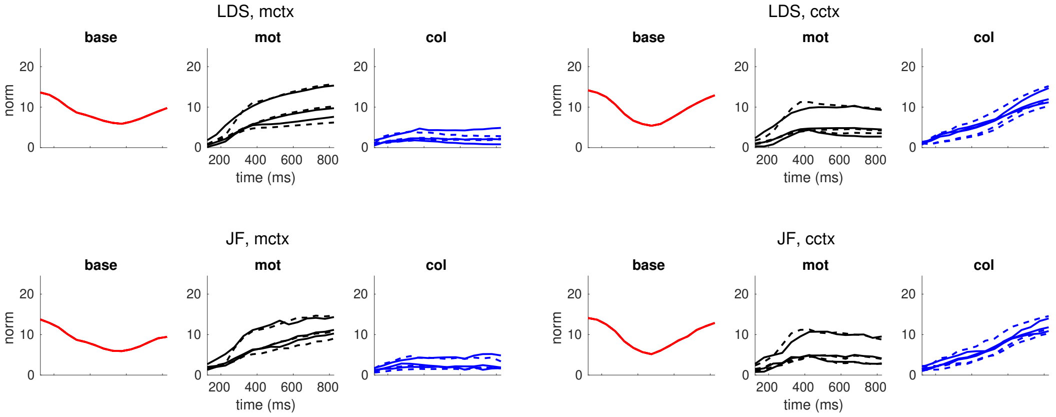
(b)

Figure 3.38: LDS and JF models minimum CV error (see table 3.6) a) CV error for the 36 different conditions, grouped by the coherence value of the relevant inputs and sorted in ascending order. Each group contains 6 conditions corresponding to the 6 possible irrelevant coherence values. These are also sorted in ascending order b) CV error across time. Motion context (left), color context (right).

The CV error pattern across conditions and time presents the same characteristics as for monkey A. The largest errors, which were found for conditions with strong incongruent information in monkey A, are even more severe in this case. This is interesting because this monkey had a poorer performance, and indeed, it seemed to be less successful at ignoring the irrelevant inputs. This is apparent in the pattern of choices of the monkey, as there is a prominent bias towards the irrelevant coherence information when this is strong (see fig. B.46 (a-d), extended data figure 2 in Mante et al. (2013)).

3.4.1.2 Comparing the LDS and the JF models solutions

We analyze now the input decomposition found by the two models, in terms of the condition independent base and input components. The relevant inputs are strongly amplified and their effect builds up until the end of the trial. Note that for the motion signal, the rate of change seems to gradually decrease in time. This is also true in monkey A, for both color and motion relevant inputs. The color signal in monkey F,



(a) (b)

Figure 3.39: Vector norms of the base, motion and color components at each point in time. For the LDS model (top) and the JF model (bottom), in the motion context (a) and the color context (b) The input components are computed for all the learned motion and color input values (in black and blue, 6 in each case), corresponding to the positive coherences (solid lines) and negative coherences (dashed lines) The base component (red) does not depend on the inputs –so there is only a single trace– and captures condition-independent variance.

however, increases steadily without a change in slope. The irrelevant input component norms are also different across monkeys. For monkey F, the effect of color in the motion context is very weak, but present nonetheless. The motion signal in the color context, however, is quite strong compared to monkey A. Motion signals seem to be stronger in general for monkey F. As we found for monkey A, irrelevant input signals are present until the very end of the trial and their effect does not decay back to zero. Finally, the condition independent base component differs across contexts for this monkey, unlike what is found for monkey A. We observe that in the color context, its norm substantially increases towards the end of the trial.

The different components alignment across models are also large, except for the color component in the motion context. Given how weak the color signal is in this case, it is likely that directions are poorly estimated or learned rather arbitrarily. Finally, color and motion components are also fairly orthogonal throughout the trial. This is different for the LDS in the motion context, because the color component as we just explained.

The results of the two models for this monkey reinforce the idea that the LDS prior is a good assumption of PFC's population data.

3.4.2 LDS with different input constraints

In this section we analyze the performance of the LDS model when incorporating different types of constraints in the input biases. We find that, in agreement with monkey A, all models perform similarly. Therefore, the constraints we incorporated on the input's temporal structure seem to be a reasonable assumption of monkey F data as well.

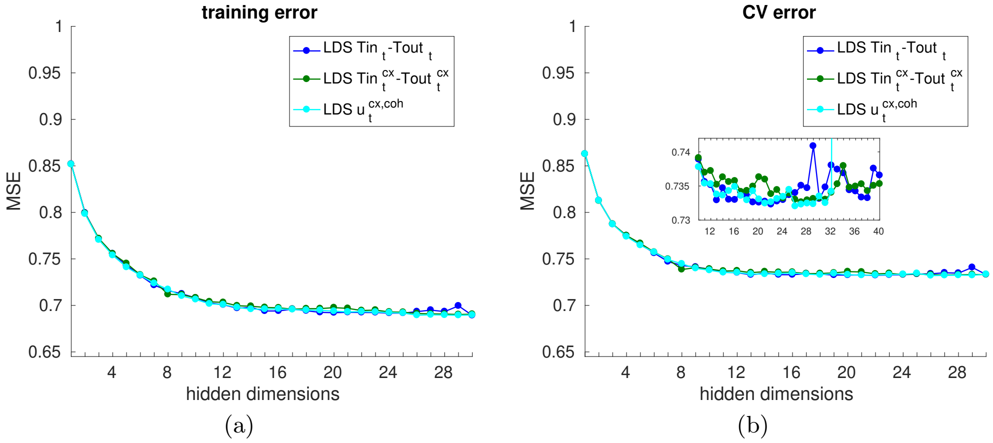


Figure 3.40: LDS model performance for different input specifications. Shared time course across coherences and contexts (blue), shared time course across coherences but not contexts (green) and full flexibility per coherence value (cian) a) Training mean squared error b) CV mean squared error.

	Model	min CV error	H
1.	LDS $u_t^{coh,cx}$	0.7320	26
2.	LDS $Tin_t^{cx} - Tout_t^{cx}$	0.7326	27
3.	LDS $Tin_t - Tout_t$	0.7323	22

Table 3.7: LDS minimum cross-validation error, for different input specifications, and corresponding hidden dimensionality H for which it is achieved.

3.4.3 LDS input dimensionality

We consistently find, as for monkey A, that performance saturates at input dimensionalities larger than four. The dimensionality of each the input subspaces also seems to be 2D. However, as we will see later, a model with a single dimension for color in the motion context performed almost as well as a model where a two dimensional color subspace was considered (see figure 3.51).

3.4.4 LDS under the same contextual dynamics

We fitted the model constrained to have the same dynamics across contexts to the data from this monkey. We found that, surprisingly, this model slightly outperformed the dynamically flexible model. The type of solution implemented, however, was of the same nature as the one found for monkey A, which was highly unstable. The dynamics was

Model	min CV error	H
LDS 8D	0.732	14
LDS 4D	0.733	27
LDS 2D	0.748	27

Table 3.8: LDS minimum cross-validation error, for different input dimensionalities, and corresponding hidden dimensionality H for which it is achieved.

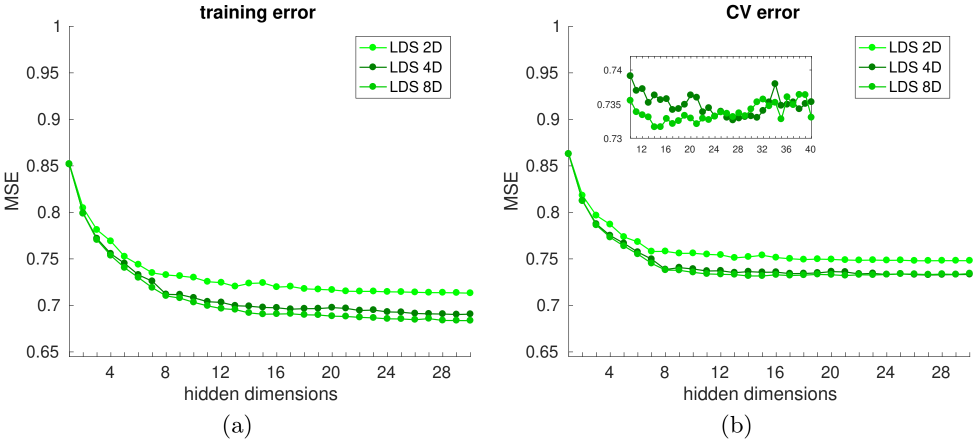


Figure 3.41: LDS model performance for different input dimensionalities a) Training mean squared error b) CV mean squared error.

Model	min CV error	H
LDS $A^{(c)} \neq A^{(m)}$	0.7326	27
LDS $A^{(c)} = A^{(m)}$	0.7324	15

Table 3.9: LDS minimum cross-validation error, for unconstrained (top) and constrained (bottom) dynamics across contexts, and corresponding hidden dimensionality H for which it is achieved.

also learned with a strong negative eigenvalue, which would sometimes lead to strong oscillations in the output space of FRs. Nevertheless, considering that the performance of this model is indeed better, we cannot rule out the possibility that the population of PFC neurons, for this monkey, evolves under a single dynamics in the two contexts. Alternative models than the one explored in this work should be considered, which offer a stable solution and are more likely to be implemented in a neural circuit.

3.4.5 LDS inputs and dynamics

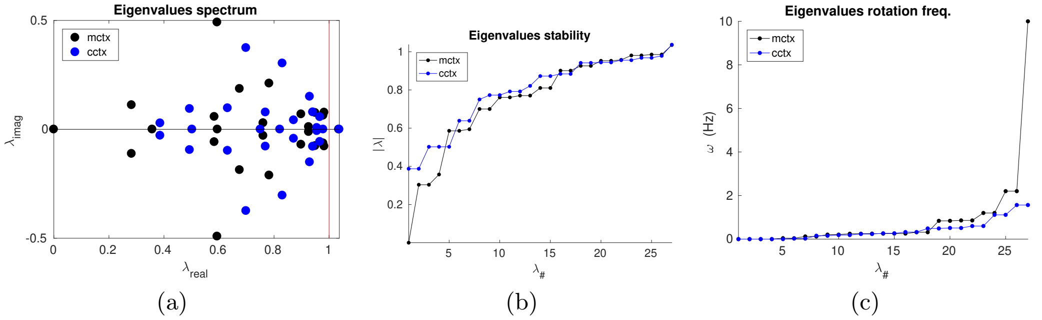
In the previous sections we have justified that, also for monkey F, the LDS is a suitable model to explain the data. To analyze the properties of the solution learned, we will consider an input dimensionality D = 4 and a hidden dimensionality of H = 27, for which the best generalization performance was obtained.

3.4.5.1 Transition matrix

We also observe that in both contexts, a multitude of slow modes are learned and that large proportion of them are complex.

3.4.5.2 Input biases and input directions

The input parameters obtained for this monkey present all the characteristics of the solution found for monkey A, with one exception. The pattern of the color coherence



(a)

(b)

(c)

Figure 3.42: Properties of the learned LDS transition matrix a) Eigenspectrum of the transition matrices inferred in each context (black for motion context and blue for color context) b) Stability of the eigenmodes of the dynamics in each context, given by the absolute value of the eigenvalues (sorted in ascending order) c) Rotation frequency in the planes spanned by pairs of complex conjugate eigenvectors (sorted in ascending order).

input during the color context is not found to be transient, as it is the case for motion, but resembles and integrated signal (figures 3.43 and 3.44). The model had the flexibility to generate an integration pattern using the dynamics matrix, however, it places the integration in the inputs, suggesting that no other type of color coherence related signal is present in the data. Note that the strength of this ramping color signal learned is comparable to the motion input signal. It is when the input is run through the dynamics that a selective amplification is obtained, which is reflected along this same input dimension (see 3.47). In the face of degeneracies in the model, we cannot distinguish between this two possibilities: that the color signal is purely input driven and the local circuit does not transform it, or that the input signal is weak and it gets locally amplified by the dynamics –and it does so along the same input dimension.

Finally, we found that the color coherence signal in the motion context is either very weak or non-existent (see also figures 3.51a, 3.60b and 3.62a).

Magnitude related signals were also found in this monkey, for the two input modalities and in the two contexts. Interestingly, in the motion context, the magnitude-related color signal was much stronger than the signed color coherence signal (see figure 3.47), which as we explained before, was very weak or inexistent.

The JF model recovered similar coherence-related input scalings B.50 to the values learned by the LDS (fig. 3.44).

When fitting a model with constrained dynamics across contexts, as found for monkey A, both color and inputs were non-transient and reflected a pattern of integration, with relevant inputs learned to be much stronger. In this case, unlike what was found for monkey A, color and motion input directions pointed in orthogonal dimensions within each of the contexts. Across contexts, the input subspaces largely overlapped. This solution, however, as we have pointed out, is unstable in nature and unlikely to be implemented in a neural substrate.

As discussed above, the integrated input patterns along the color coherence dimension

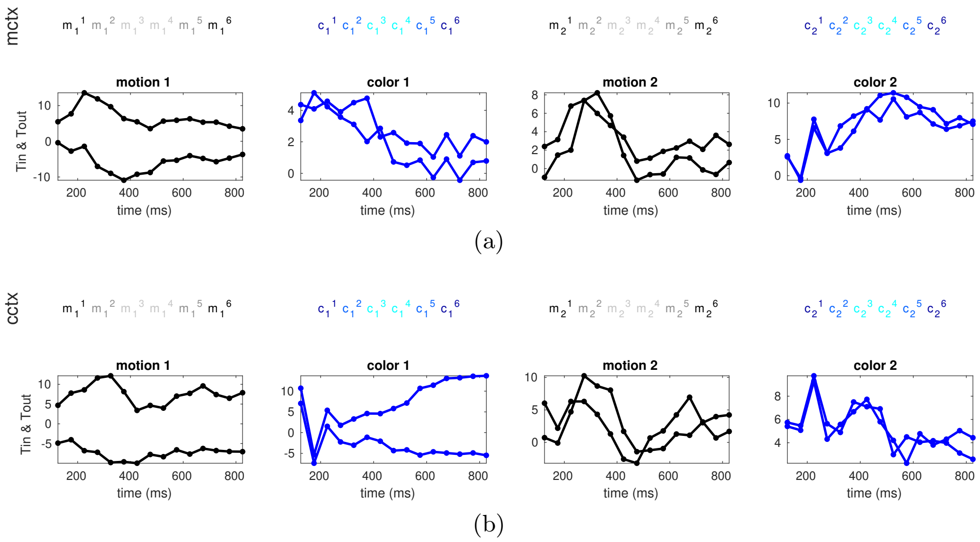
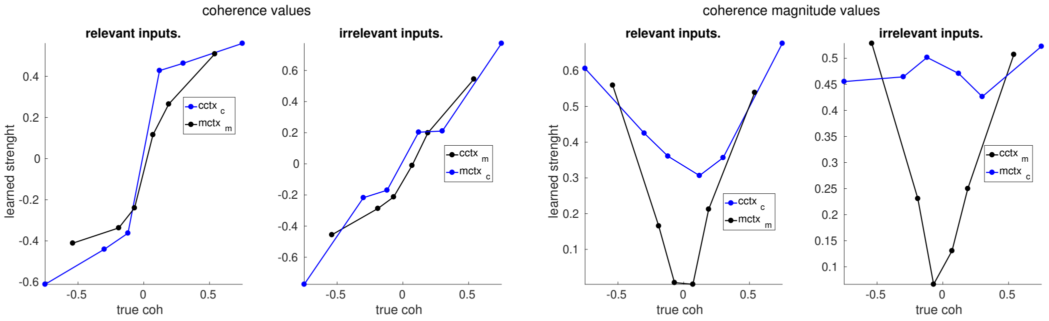


Figure 3.43: LDS learned input time courses in the motion (top) and color (bottom) contexts. Each column shows the input signal associated to each input dimension. Two different time courses are inferred per input, one for positive (Tin) coherences and the other for negative ($Tout$) coherences. To obtain the total input signals for each condition, the temporal input traces are scaled by the corresponding coherence levels (letters on top of the figures, with evidence strengths indicated by the blue and gray scales). The exact learned coherence values are shown in 3.44. To illustrate the two different types of coherence signals found –signed and not signed– we have multiplied the time courses by the sign of the associated coherence values in 3.44. We have also scaled the whole input time series by the norm of the associated input vectors.

change upon the action of the dynamics, reflecting a strong amplification (figure 3.47). For the rest of the input dimensions, the action of the dynamics is weakly reflected in the input subspaces, as found for monkey A.

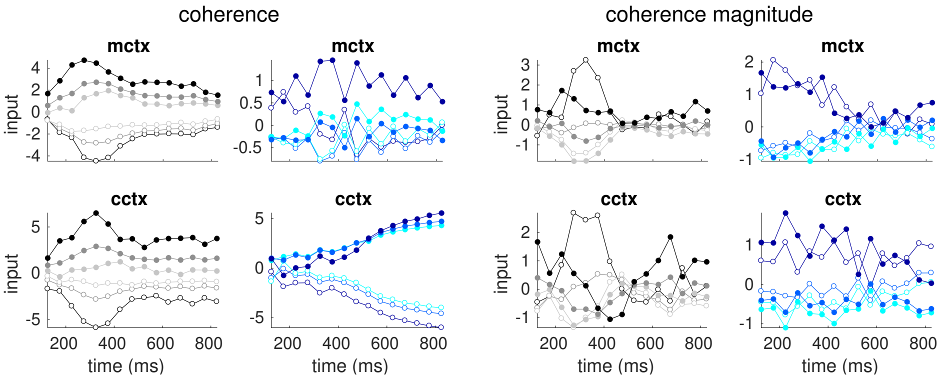
We also compared the input directions inferred using the LDS with the input directions found via TDR in Mante et al. (2013). The coherence related input directions estimated using the two methods share a mild correspondence, but the alignment is consistently greater across inputs of the same modality. The coherence magnitude related dimensions for color also share some correspondence with the TDR color dimension. Finally, the LDS estimated color coherence dimension, not surprisingly, strongly aligns with the decision axis.

In the supplementary material we show the alignment, in time, of the LDS motion and color integrated components $\boldsymbol{c}_{int}(t)$ and $\boldsymbol{m}_{int}(t)$ with respect to the TDR and LDS coherence inputs (see B.56). This plot clearly illustrates that, in the case of motion, the projection of the integrated inputs onto the coherence input dimensions goes quickly towards zero as the trial progresses (fig. B.56a). Interestingly, for this monkey, this happen earlier in the trial –compared to monkey A (fig. B.28a)– This explains why, in this monkey, the pattern of integrated motion inputs along the coherence input dimensions does not appear amplified –as we will also see later. In the plots, as soon as the zero projection is reached, an upwards rebound follows. This is stereotypical of both the integrated relevant and irrelevant inputs and is found consistently across monkeys.



(a) (b)

Figure 3.44: LDS inferred coherence values for each of the four motion and color input dimensions in each context a) associated to the color and motion dimensions that carry coherence information b) associated to the color and motion dimensions that carry coherence magnitude information. The first set of learned values match the true coherence values set in the experiment. The second type of signals, however, are not signed and reflect only the coherence level or strength. In both a) and b) the plots on the left/right show the inferred coherence values when they are relevant/irrelevant.



(a) (b)

Figure 3.45: LDS input signals in the orthogonalised 2D input subspaces. Independent 2D bases are computed for each pair of color and motion input vectors within each context. a) The first dimension is a regression-identified direction that carries coherence information (e.g. top left plot, for motion in the motion context) b) The second dimension is orthogonalised with respect to the first and reflects coherence magnitude (e.g. top left plot, for motion in the motion context). We use the same convention as with the trajectories to indicate coherence level (gray and blue scales) and sign (filled and hollow circles).

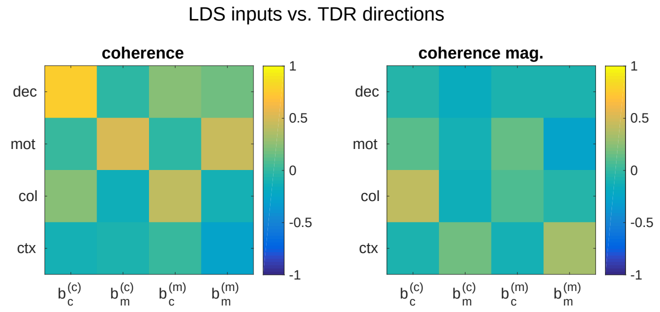


Figure 3.46: Dot products between the TDR task-relevant dimensions (color, motion, decision and context) and the input directions inferred by the LDS in the two contexts. Left: coherence input directions. Right: coherence magnitude input directions. We use subscripts to designate modality color/motion and superscripts to indicate the context (color)/(motion)

The final integrated motion dimensions, in both cases, are almost orthogonal to the input dimensions. In the case of the integrated color input in the color context, the picture is completely different. The integrated input is orthogonal to the coherence input dimension at the beginning of the trial and ends up aligning to it strongly at the end. This is because, as we will see, the final direction of color integration corresponds to this input dimension. At the beginning of the trial the integrated color input lays within the 2D input subspaces, but aligns strongly to the second input dimension, which reflects magnitude –in particular for the highest input coherences– (see fig. 3.47b). Finally, the dot product between the TDR and the LDS integrated vectors for motion picks sharply right before t=400ms, which is consistent across contexts, and reaches a value of 0.8-0.9 (B.56b). This suggests that the TDR procedure may be detecting the early amplification of the motion inputs by the dynamics before they start getting rotated away from the input subspaces and towards the decision axis.

Next, we tested the alignment of the different input subspaces. Within contexts, color and motion subspaces share very little variance (3.48) and are close to orthogonal (B.51). This lack of alignment, however, as we found for monkey A, is expected by chance. Across contexts, we find that for this monkey, the motion subspaces are more aligned than expected by chance, but not the color subspaces.

When looking at individual dimensions, in the case of the motion input, both the coherence related and the coherence magnitude related directions are more aligned than chance across contexts (figure 3.49). This is not the case for any of the color input dimensions.

Finally, we found that the mild alignment of the LDS input dimensions with respect to the TDR identified input directions is well above chance only for the motion dimension in the motion context.

Given this results, which reflect a large variability in some of the estimated dimensions, we assessed the dimensionality of the color and motion subspaces separately. We found that a model with a single dimension for color in the motion context performed almost

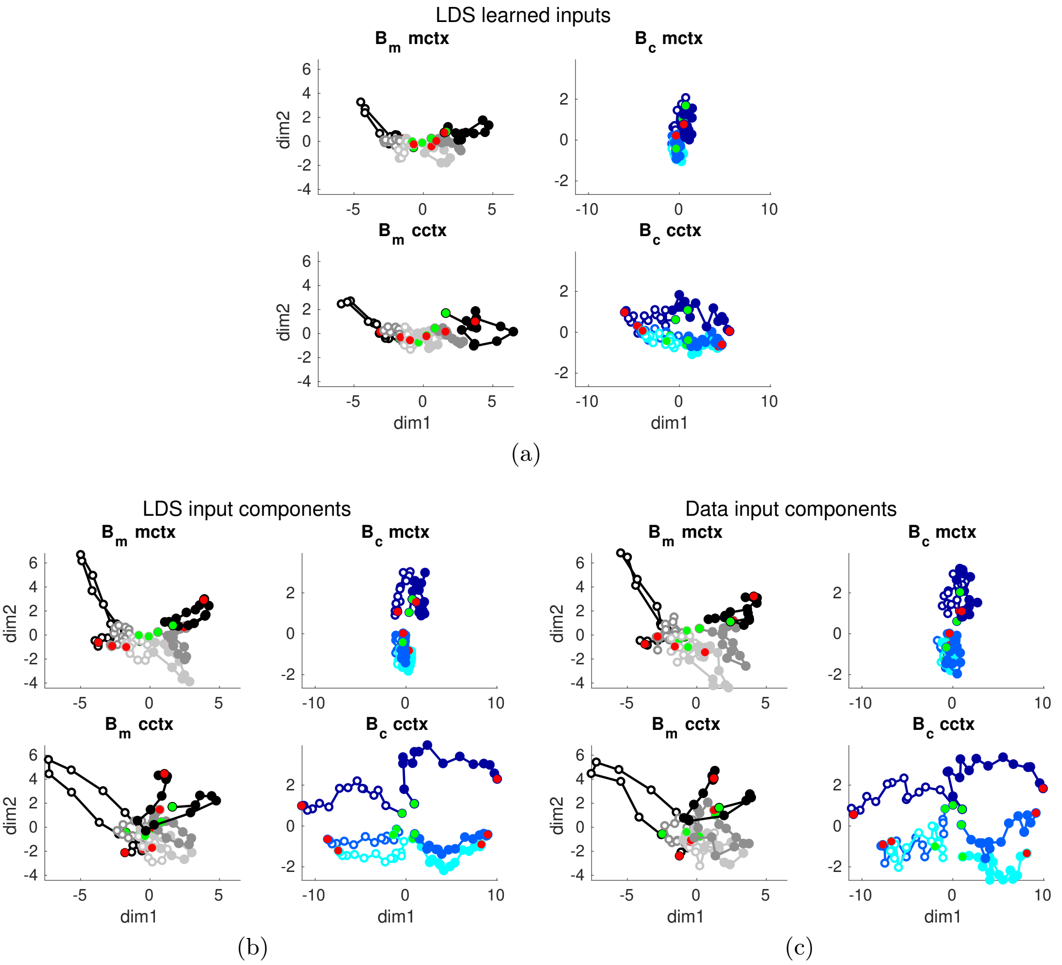


Figure 3.47: a) LDS input signals in the orthogonalised 2D input subspaces. Independent 2D bases are computed for each pair of color and motion input vectors within each context. The first dimension (x axis) defines a direction within the 2D subspace that carries coherence information. The second dimension (y axis) is orthogonalised with respect to the first and reflects coherence magnitude. b) Color and motion integrated input components projected onto the same input subspaces as in a) c) Data input components, projected onto the same subspaces. To reveal the pattern of integrated inputs in the real data: first, we project the observations onto the low-dimensional hidden space using the loading matrix C. Second, we subtract to the data the LDS condition independent (base) component. Finally, we show color/motion trajectories where the motion/color contribution has been averaged out. Green/red dots indicate the beginning/end of the trial.

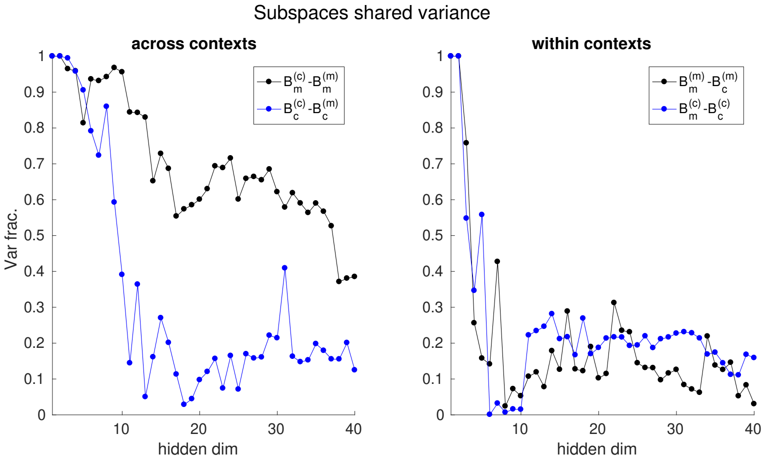


Figure 3.48: Shared variance across the different LDS input subspaces estimated for a wide range of hidden dimensionalities. Left plot, fraction of variance shared between the color/motion subspaces across contexts. Right plot, fraction of variance shared between the color and the motion subspaces within a given context. We use subscripts to designate modality color/motion and superscripts to indicate the context (color)/(motion)

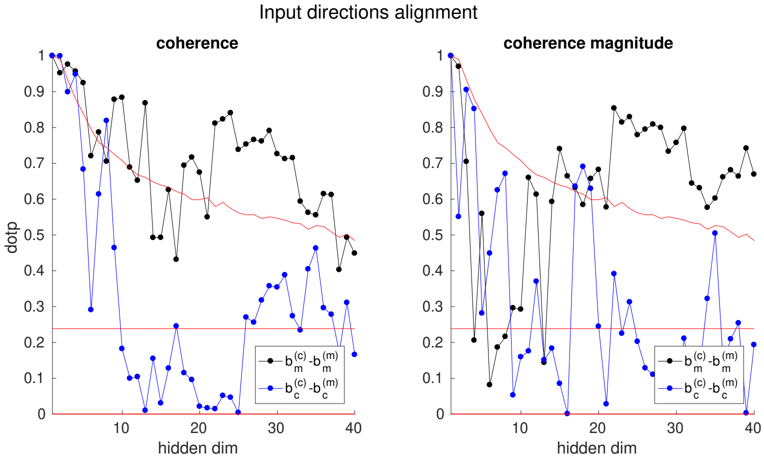


Figure 3.49: Across contexts alignments (dot products) for the LDS motion and color input directions. Left plot, coherence dimensions. Right plot, coherence magnitude dimensions. Alignments are computed for different hidden state dimensionalities. Red lines correspond to the 5th (found close to zero) and 95th percentiles of the null distribution for random alignments. The null is restricted to follow the data covariance structure –once projected onto the hidden subspaces, for the different dimensionalities. The constant lines correspond to the null estimated on the whole data space, without projecting the covariance onto the hidden subspaces.

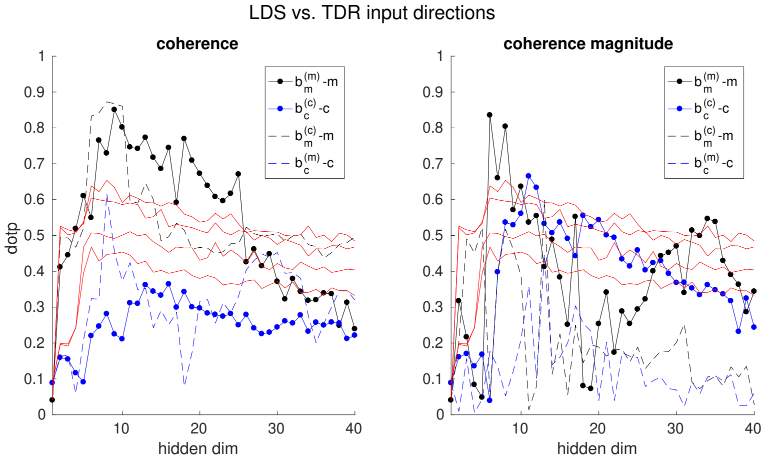


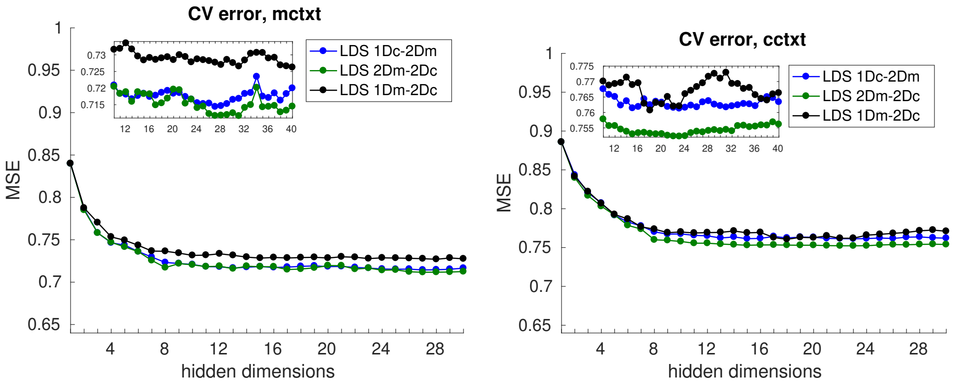
Figure 3.50: LDS input dimensions and TDR task-relevant input directions (m,c) alignment for different hidden dimensionalities. Left plot, coherence dimensions. Right plot, coherence magnitude dimensions. Solid/dashed lines are used for the estimated relevant/irrelevant inputs. Red lines indicate 95th percentiles of the null distribution for expected random alignments with respect to the TDR input vectors (4 lines, corresponding to the color and motion nulls in each contexts). Note that the two nulls within each context, the one for motion and the one for color, are not identical, indicating that the color and motion input dimensions reflect different proportions of the variance.

as well as a model where a two dimensional color subspace was considered. However, the performance does in fact improve slightly when adding an extra color dimension. This indicates that the color coherence signal in the motion context may indeed be present, but if so it is very weak and reflects very little variance.

3.4.5.3 Dynamics

Contextual selection mechanism

The pattern of projections of the inputs onto the left eigenvectors of the dynamics in this monkey did not present an evident contextual change. This, however, can be explained



(a)

(b)

Figure 3.51: LDS cross-validation performance as a function of hidden dimensionality in the motion (a) and color (b) contexts and for different dimensions of the color and motion input subspaces.

Model	min CV error	H
LDS $B^{(c)}_{c,m} \neq B^{(m)}_{c,m}$	0.733	27
LDS $B^{(c)}_{c,m} = B^{(m)}_{c,m}$	0.731	20

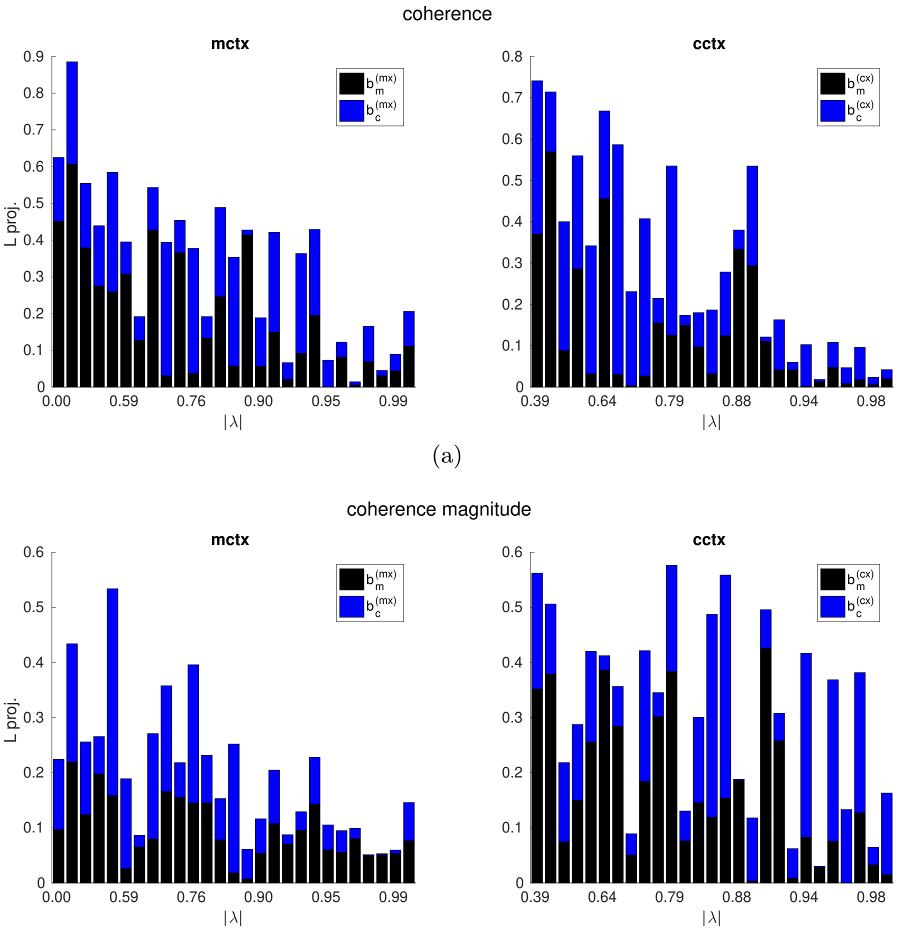
Table 3.10: LDS minimum cross-validation error, for unconstrained (top) and constrained (bottom) input directions across contexts, and corresponding hidden dimensionality H for which it is achieved.

by taking into account the pattern of inputs learned. In the color context, it is the case that the color coherence input projects more strongly into the first few slow modes relative to motion (figure 3.52). Furthermore, the color coherence input signal in this monkey is ramping, not transient, so this must also be taken into account. In the motion context, the motion coherence input vector projects more strongly onto the first slow modes, as compared to what is found in the color context (remember that the motion coherence input signals are similar in strength across contexts, unlike the color signals 3.45). The color coherence input in this case also targets the same modes, projecting into them as strongly as the motion input. However, the color input signal learned in this context is very weak, so the resulting amplification will be smaller compared to motion. Therefore, taking together these differences in both the projection patterns and in the input strengths, it can be explained how the model can selectively amplify the inputs in each context. This selective amplification is clearly reflected in the pattern of integrated inputs along the decision dimension (3.54)

In the supplementary material we attach a scatter plot version of figure 3.52, in case the reader finds this depiction of the data more intuitive (see fig. B.54)

We also fitted a model where we constrained the input subspaces to be the same across contexts. Unlike what was found for monkey A, there was no impairment in the performance under this constraint. In fact, this model performed slightly better in cross-validation (see table 3.10). The pattern of inputs and integrated inputs along the identified coherence and coherence magnitude input dimensions were very similar to the ones recovered under the model with flexibility in the directions. Finally, given the constraint in the input directions, the color coherence information dimension in motion context also aligned with the decision dimension. The color signal in this case was also found to be very weak.

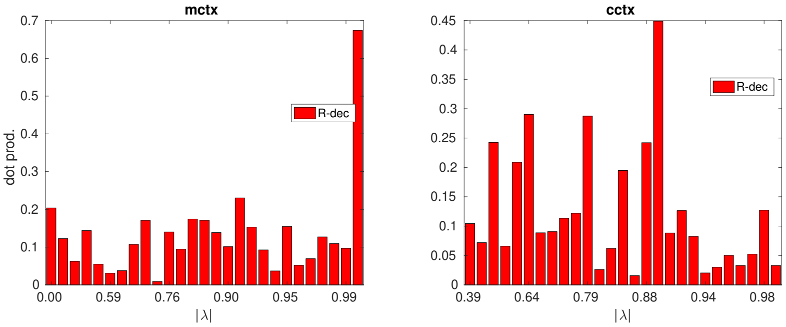
For monkey F, the data suggests that inputs to PFC bias the network along the same dimensions in the two contexts. However, as we have seen before, the strength of the color coherence signal experiences a substantial change across contexts, becoming very weak when it is irrelevant. This is suggestive of a model where a partial gating of the color signals mediates the selection of relevant information. Alternatively, as we will explain in the discussion, the type of solution found could be a consequence of having only partially observed the dynamics.



(a)

(b)

Figure 3.52: LDS input directions projection onto the left eigenvectors of the dynamics –which were normalized to be unit norm– a) for the inferred motion and color coherence inputs b) for the motion and color coherence magnitude inputs. Each left eigenvector is associated with an eigenvalue, which is specified in the x axis (sorted by magnitude). For complex-conjugate pairs we consider the projection onto the real and the imaginary components, which define the complex planes. (a,b) left figure, motion context; right figure, color context.



(a)

Figure 3.53: Projection of the TDR identified decision axis onto the right eigenvectors of the dynamics in each context –which were normalized to be unit norm. We use the loading matrix C to bring the decision vector into the hidden space. Each right eigenvector is associated with an eigenvalue, which is specified in the x axis (sorted by magnitude). For complex-conjugate pairs we consider the projection onto the real and the imaginary components, which define the complex planes.

Integration mechanism

We found that for this monkey, surprisingly, the decision axis strongly aligns to one of the slow modes in the motion context, but not to the rest. This alignment is higher than expected by chance (see figure 3.49, the 95th percentile for H=27 is about 0.55). This is found for different models, across different hidden dimensionalities and model specifications, so it does not seem to be a particular solution. The projection pattern is never identical, but it is the case that the load onto the last mode is substantially higher than for the rest of the dimensions. We are not certain of why this is the case. As we mentioned before, when the dynamics has several slow modes, we expect that the integrated input vector will point into a direction given by a linear combination of the slowest directions. The decision axis, which reflects the pattern of integrated inputs, should therefore point into this direction. From the spectrum of input projections onto the dynamics, one can see that the strongest load from the motion coherence input vector goes into this last slow mode, which furthermore, it is slightly unstable. However, it is not the only slow mode the motion input projects into. Therefore, we were expecting the integrated motion vector to point somewhere in between this largest mode and the rest of the slow modes. Similar projection patterns are obtained when considering the decision vectors we estimated independently in each context, both for the one that separated motion sign in the motion context and for the one separating color sign in the color context.

In future work we would like to provide better means to analyze the complicated dynamical portrait implied by the input projections onto the left eigenvectors, so that we can develop an intuition of the exact mechanism the model is using to solve the task.

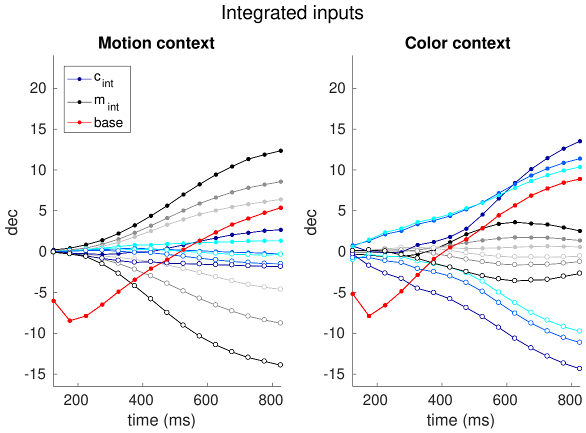


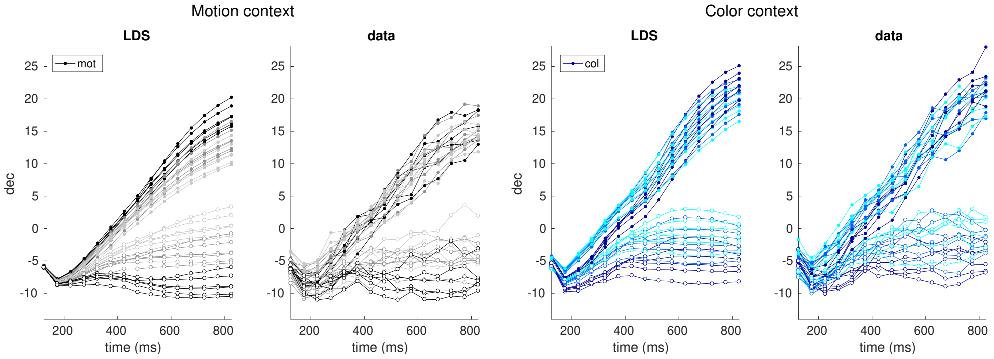
Figure 3.54: LDS integrated inputs and base components projected onto the TDR decision axis, for all possible input values (6 each for color and motion). Same conventions as in previous plots applies.

Patterns of inputs and integrated inputs

The pattern of integrated inputs along the decision axis display selective amplification of the relevant inputs (3.54). Two main observations are to be made. The difference in the strengths between relevant and irrelevant inputs along this decision dimension is smaller for this monkey (compare with 3.23). Furthermore, the influence of the base component is also stronger. This seems to be learned in order to capture the strong condition-independent drift observed in monkey F trajectories.

In the supplementary material we repeat the analysis but computing the alignments (dot products), not the full projection (fig B.57). This illustrates that the integrated input components are in fact rotated by the dynamics so that, in each context, the relevant integrated vectors are gradually brought towards the decision axis. The irrelevant inputs, unlike what is found for monkey A, also experience the same gradual rotation towards the decision axis and present a moderate alignment with it by the end of the trial. Note that in the case of the integrated color input in the color context, the coherence related component points into the decision axis. However, the second component, which carries coherence magnitude information, is orthogonal to it. At the beginning of the trial the integrated input vector aligns mostly into this second component, but then the coherence related input signal starts growing and pulls the integrated input vector towards it (see 2D representations 3.47b). This explains the projection pattern of the integrated color inputs along the decision axis found in B.57.

We then estimated decision-related dimensions within each of the integrated subspaces. More specifically, we looked dimensions within the integrated subspaces at the last time step that separated the sign of the color and motion inputs ($d_m^{(m)} - d_c^{(c)}$ and $d_m^{(c)} - d_c^{(m)}$, see 3.56). The decision dimensions found in the relevant subspaces across contexts were highly aligned and corresponded to the TDR estimated decision axis. The alignment was particularly high for the LDS decision dimension in the color context. Model and data trajectories along all these dimensions are shown in figures 3.55, 3.57.



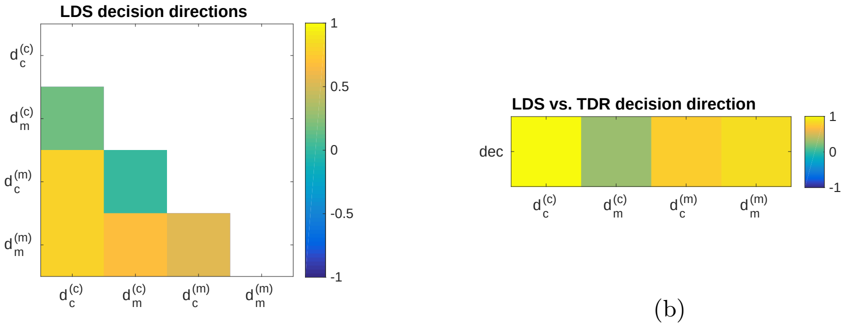
(a) (b)

Figure 3.55: LDS model and data trajectories, for the 36 possible conditions, projected onto the TDR decision axis. Each condition is color-coded according to the relevant modality for each context. The data has not been smoothed.

Regarding the dimensions identified in the irrelevant subspaces, we found that the direction separating color sign in the motion context did also align with the decision axis. This suggests that an integrated color coherence signal is also present in the motion context, but being confined to the decision dimension, its effects are masked by the integrated motion signal (see 3.58a). In fact, for the two monkeys, it is the case that along the decision dimension the integrated irrelevant inputs are weakly represented. However, their effect on the trajectories is not apparent given the influence of the strong integrated relevant signal. In the color context, conversely, a dimension existed along which motion signals were strongly amplified (see 3.58b). This dimension was orthogonal to the decision axis (3.56). Color coherence information was also mildly reflected along this dimension.

In the supplementary material we show the results on the dimensions orthogonal to $d_m^{(m)} - d_c^{(c)}$ and $d_m^{(c)} - d_c^{(m)}$, which we called $d2_m^{(m)} - d2_c^{(c)}$ and $d2_m^{(c)} - d2_c^{(m)}$. These dimensions are all orthogonal to each other and carry very little variance. Unlike what is found for the first relevant integrated dimensions, the corresponding orthogonal dimensions $d2_m^{(m)} - d2_c^{(c)}$ are not the same across contexts. Therefore, the integrated relevant subspaces do not completely overlap across contexts, but they do in fact share a single dimension of integration.

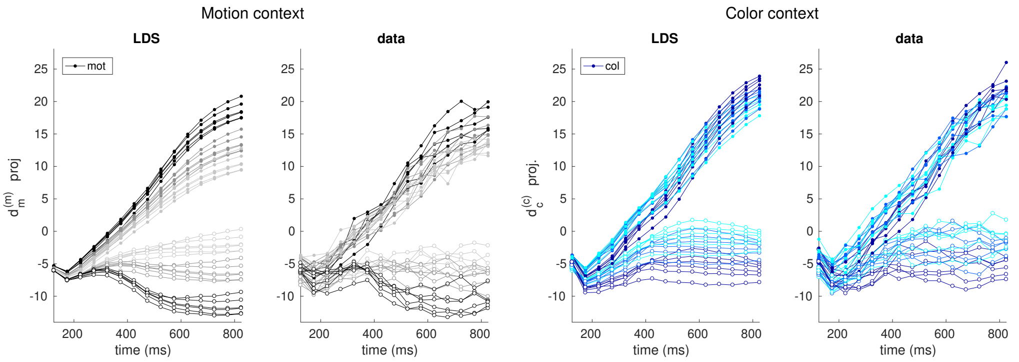
Finally, we show the pattern of integrated inputs along the coherence input dimensions identified within the input subspaces (figure 3.60). Trajectories are shown in figures 3.61 and 3.62. As we discussed before, the integrated inputs largely mirror the pattern of the input biases. The exception is the integrated color signal in the color context, whose amplification is strongly reflected along the input dimension, given that it is aligned with the decision axis. Interestingly, the strengths of the integrated motion signals along the motion input dimensions are comparable across contexts, unlike what it is found for monkey A. It seems like the pattern of motion inputs along the motion dimensions is largely unaffected by the dynamics. However, we note that when motion is irrelevant,



(a)

(b)

Figure 3.56: Decision-related dimensions within the color and motion integrated 2D subspaces at the last time step. The dimensions are found based on the sign of the relevant inputs, within the relevant inputs subspaces ($d_m^{(m)} - d_c^{(c)}$), or based on the sign of the irrelevant inputs, within the irrelevant inputs subspaces ($d_m^{(c)} - d_c^{(m)}$). The decision dimensions $d_m^{(m)} - d_c^{(c)}$ are found to be the same across contexts. a) decision-related dimensions alignment across and within contexts b) decision-related dimensions and TDR decision axis correspondence.



(a)

(b)

Figure 3.57: LDS model and data trajectories, for the 36 possible conditions, projected onto the “relevant” decision dimension. The y-axis corresponds to a regression-identified direction, lying within the integrated relevant-input 2D subspace, that reflects the sign of the relevant information. Each condition is color-coded according to the relevant modality for each context.

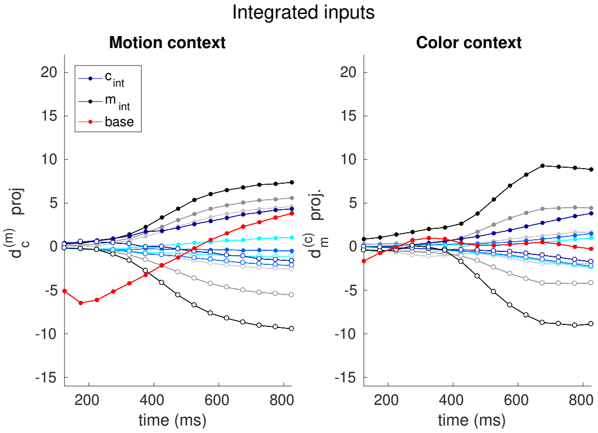
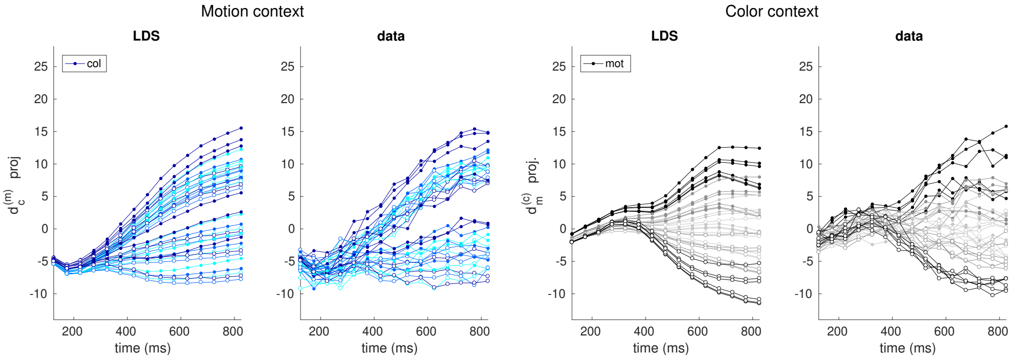
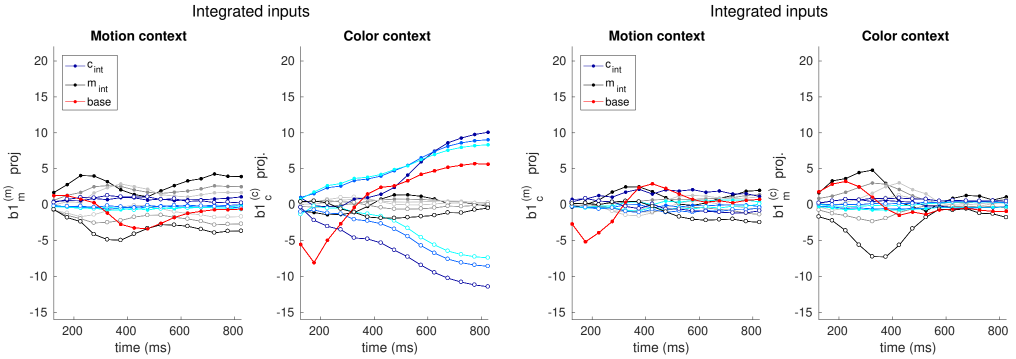


Figure 3.58: LDS integrated input and base components projection onto the "irrelevant" decision dimension, for all possible input values (6 each for color and motion). The y-axis corresponds to a regression-identified direction, lying within the integrated irrelevant-input 2D subspace, that reflects the sign of the irrelevant information.



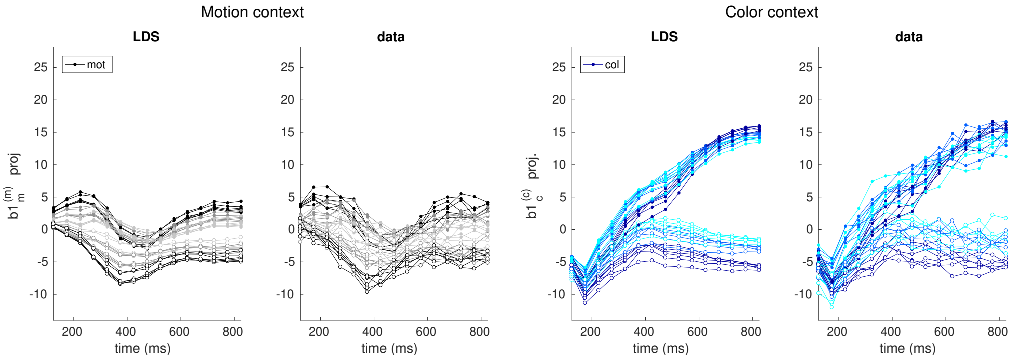
(a) (b)

Figure 3.59: LDS model and data trajectories, for the 36 possible conditions, projected onto the "irrelevant" decision dimension. The y-axis corresponds to a regression-identified direction, lying within the integrated irrelevant-input 2D subspace, that reflects the sign of the irrelevant information. Each condition is color-coded according to the irrelevant modality for each context.



(a) (b)

Figure 3.60: LDS integrated inputs and base components projected onto the LDS coherence input directions, for all possible input values (6 each for color and motion) a) relevant dimensions b) irrelevant dimensions.



(a) (b)

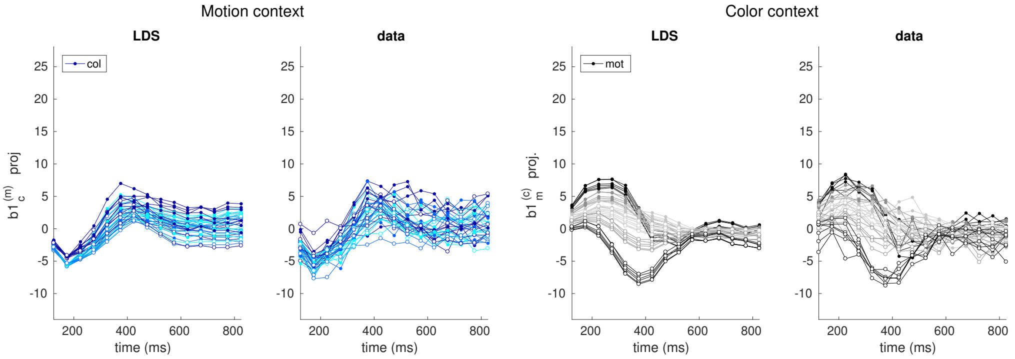
Figure 3.61: LDS model and data trajectories, for the 36 possible conditions, projected onto the relevant coherence input dimensions. The y-axis corresponds to a regression-identified direction, lying within the relevant-input 2D subspace, that reflects the relevant coherence value. Each condition is color-coded according to the relevant modality for each context.

the integrated motion signal gets more strongly suppressed towards the end of the trial.

The corresponding plots for the second input dimensions, which separate by coherence strength, but not sign, can be found in the supplementary material. Coherence magnitude signals along the coherence magnitude input dimensions are present in the data, for both color and motion modalities and during the two contexts.

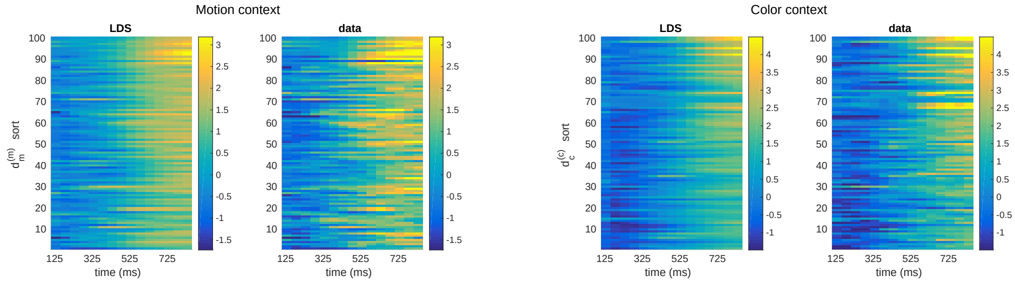
Population activity patterns

It is also the case for this monkey that the whole PFC population is highly heterogeneous and most cells are characterized for having “mixed” selectivity. The dimensions identified in this monkey are also highly uncorraleted and both the inputs and the decision loads into the neurons are broadly distributed. The exception, clearly, is for the color input in the color context, which is correlated with the decision dimension (see figure B.66). This is consistent with what was found in the study by Mante et al. (2013).



(a) (b)

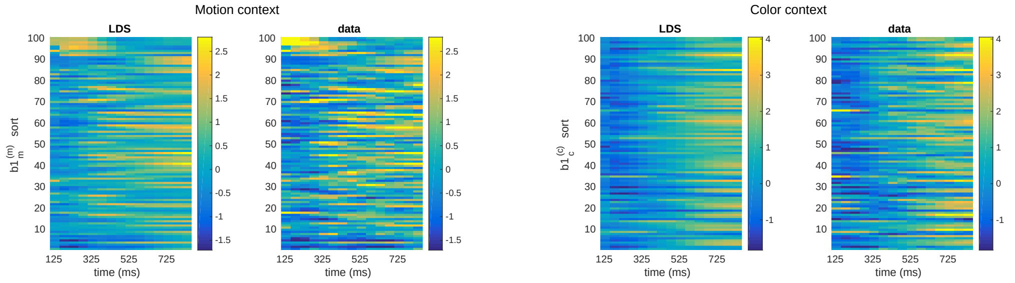
Figure 3.62: LDS model and data trajectories, for the 36 possible conditions, projected onto the irrelevant coherence input dimensions. The y-axis corresponds to a regression-identified direction, lying within the irrelevant-input 2D subspace, that reflects the irrelevant coherence value. Each condition is color-coded according to the irrelevant modality for each context.



(a) (b)

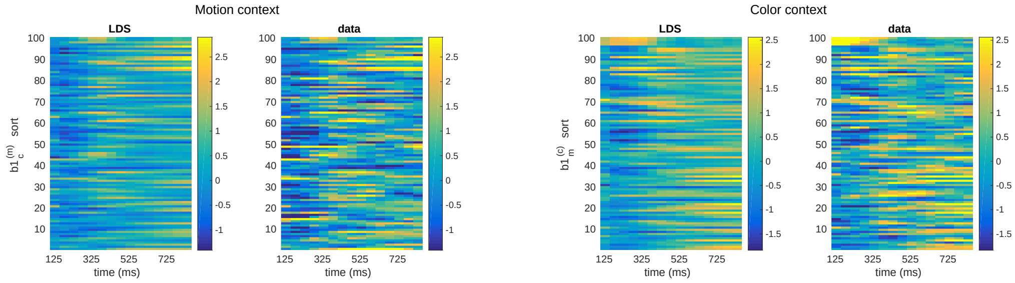
Figure 3.63: PSTHs of 100 units participating in the activity patterns along the LDS “relevant” decision dimensions (see 3.26). Neurons were selected based on the top 100 loads (by magnitude) and are shown sorted in ascending order. The PSTHs correspond to the condition with the highest positive color and motion coherence. Note that a few neurons have high FRs that are not well captured by the LDS model. To aid visualization, the color scale in the data plots has been saturated to the maximum FR value in the LDS model plots. For the data, responses have been smoothed using a squared window filter, as in Mante et al. (2013).

Next, we looked at population of neurons being highly driven by the different activity patterns along the directions we identified. In figures 3.63, 3.64 and 3.65 we show the FRs of the neurons with the strongest loads from the decision, the relevant input and the irrelevant input dimensions respectively. The plotted PSTHs correspond to the condition with the highest positive color and motion coherence. The patterns of activity along the dimensions identified by the LDS are indeed reflected in the population FRs. Furthermore, the model seems to capture well the different patterns at the level of individual neurons.



(a) (b)

Figure 3.64: PSTHs of 100 units participating in the activity pattern along the LDS relevant input dimensions. Same convention as in previous figure.



(a) (b)

Figure 3.65: PSTHs of 100 units participating in the activity pattern along the LDS irrelevant input dimensions. Same convention as in previous figure.

3.4.6 LDS single unit PSTHs

The LDS model is also able to capture the whole complexity of individual firing patterns in monkey F (fig. 3.66), despite the high degree of heterogeneity in the population responses. These include neurons that strongly and transiently respond to motion (first row); to color (third row. Note that the color input pattern in this monkey is ramping-like, but the initial transient in the PSTHs can be explained because this neuron also receives a large load from the color magnitude vector); that respond to color and motion strengths (second and fourth rows. Weak for color in this particular neuron) or that strongly represent decision signals (fifth row).

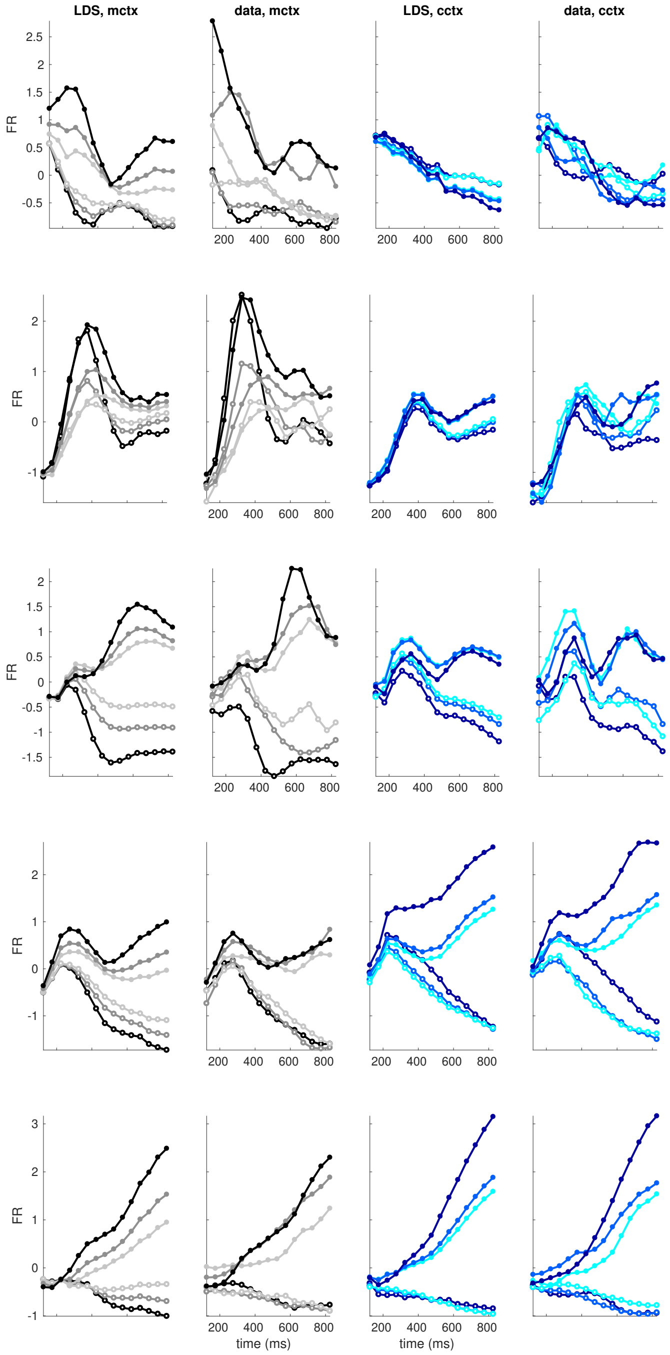


Figure 3.66: Single unit PSTHs sorted by conditions with the irrelevant input influence averaged out (units n=5, 98, 422, 277, 410). First and third column, LDS model. Second and fourth column, PFC data. In each row, from top to bottom, we show units with the largest load from: the motion coherence input (in the motion context), the motion coherence magnitude input (in the motion context), the color coherence input (in the color context), the color coherence magnitude input (in the color context) and the decision axis. For the data, responses have been smoothed using a squared window filter, as in Mante et al. (2013).

Chapter 4

Discussion and Future Work

4.1 Discussion

A central quest in the history of neuroscience has been to identify the integral parts of the brain that give rise to our cognition. At the beginning of the 20th century, Cajal, following his neuron doctrine (Waldeyer, 1891; Cajal, 1954), postulated that structural changes at the level of individual neurons mediate abstract mental processes (DeFelipe, 2006). The neuron doctrine motivated the vast majority of the research in neuroscience until the end of the last century. It also triggered the inception of a new promising field, theoretical neuroscience, which established the view that single neurons constitute the computational building blocks giving rise to sensation and perception (Barlow, 1972). During the 80s and 90s, in the advent of an expansion in methodological techniques, several discoveries started to shift the neuro-centric view towards a circuit-level understanding of the brain. This included seminal studies such as the ones conducted by Bill Newsome's group at Stanford University (Newsome et al., 1989; Shadlen et al., 1996). The core question under scrutiny was that of how simple perceptual decisions could be executed by the brain. Drawing inspiration from Sherrington's research in spinal cord circuits, who postulated the existence of reflex-like integration centers (Glimcher, 2003), this and other groups looked for areas in the brain where sensory-to-motor transformations could be implemented. These ideas have lead to the discovery that many cortical and subcortical areas are part of a complicated and highly distributed decision-making circuitry (Siegel et al., 2015). However, a region stands out given its key role in mediating such transformations and for its direct implication in the generation of complex behaviors: The Prefrontal Cortex (PFC) (Fuster, 2015; Miller and Cohen, 2001).

The goal of this study was to understand the contribution of PFC to higher order cognitive functions. The main focus was to gain intuitions about how this brain area may generate complex and flexible behaviors, such as contextual decision-making. In particular, we addressed the question of how populations of neurons in PFC can flexibly select relevant sensory information and transform it towards a choice. For that, we built on the work by Valerio Mante and colleagues (Mante et al., 2013), by developing a new methodological approach that provided new insights into the nature of the computations implemented in this PFC circuit. We considered the problem within a dynamical systems framework, providing a mechanistic account that allowed us to explicitly study the dynamics of the decision formation. However, instead of working with RNN models of the type used in (Wang, 2002; Machens et al., 2005; Mante et al., 2013) –which were designed and/or trained to solve a specific task– we considered a model fitting procedure, which allowed us to infer the dynamics directly from the data.

We found that the PFC data presented systematic dynamics consistent with the assumptions of a linear dynamical system model. The model identified high-dimensional motion and color input signals spanning independent subspaces, which were largely invariant across contexts. Within each subspace, we were able to identify directions that separated coherence input information and others that separated coherence magnitude, but not sign. To mediate the selection and integration of relevant input information, the dynamics of the circuit changed in each context so that relevant stimuli were strongly amplified. The integration subspace was also high-dimensional and spanned multiple directions, corresponding to the slowest modes of the dynamics. Some of the inputs were integrated towards a final dimension which was consistent across contexts and separated decision information. Finally, in spite of the nature of PFC's responses, which are complex and highly heterogeneous, the model was able to accurately capture individual PSTHs and reproduced the main dynamical features of the population trajectories. In this study we uncovered a whole space of sensory-related input signals invariantly modulating PFC responses across different contexts and proposed a novel mechanism by which such signals are selected and integrated for contextual computations.

4.1.1 Summary of the findings

In this section we address, one by one, the questions that were raised at the beginning of chapter 3 (see 3.1.2 and 3.1) and provide a summary of our findings.

1. We compared the performance of the LDS with that of a linear model which incorporates no dynamical constraints, the JF model. Both models perform similarly in cross-validation, meaning that the extra flexibility offered by the JF model is not necessary to capture the data well (fig. 3.5 , 3.36). Therefore, we conclude that the temporal structure present in PFC's PSTHs is consistent with the assumptions of the dynamical prior. This indicates that a dynamical systems framework is in fact adequate for understanding the activity patterns of population of neurons in PFC under complex computations.

2. The LDS is able to reproduce the temporal structure in the population trajectories just as well as the JF model does (fig. 3.6, 3.37), even on cross-validated data. This is quite remarkable given that the JF model has the potential to capture arbitrary patterns in time, including dynamics specified by non-linear rules. Furthermore,

the LDS accurately captures the data at the level of individual PSTHs, despite the complex and highly heterogeneous nature of the population firing patterns (fig. 3.35, 3.66). The fact that the dynamical trends of the whole trajectories can be generated by this model, suggests that the linear account holds globally within the region of neural space explored by the data during each context.

3. We introduce an alternative way of estimating input directions from the method proposed in Mante et al. (2013) (see section 2.1.3). The LDS model allows to incorporate input influences, which bias the system along specific dimensions and control the evolution of the dynamical system. We could learn both the input biases and the input directions. Therefore, this setting allows to explicitly model *input signals*, separating the network inputs from the *integrated input* patterns. Furthermore, the estimation procedure does not require performing any ad hoc decisions of the type performed in the previous study, where inputs had to be chosen among all the estimates at each point in the trial. Finally, our method does not require making any assumptions about the dimensionality of the inputs and allows to estimate whole input subspaces.

4. The cross-validation performance on the LDS for different input dimensionalities suggests that the input subspace is no larger than 4D (fig. 3.10, 3.41 and tables 3.3, 3.8). The color and motion signals live in 2D subspaces. Within these subspaces, we were able to identify a dimension that separates *coherence* information and a second dimension that reflects *coherence magnitude*, but not sign –the first reflecting both evidence strength and direction and the second only strength– (fig. 3.14, 3.45).

   (a) For monkey A, the dimensionality of the motion subspace in the color context was close to 1D, as the secondary motion dimension, reflecting a coherence magnitude related signal, captured very little variance (fig. 3.20).

   (b) For monkey F, the dimensionality of the color subspace in the motion context was close to 1D, given that the color dimension reflecting coherence information captured very little variance (fig. 3.51).

5. Input subspaces of the same modality are largely invariant across contexts, being much more aligned than expected by chance. Color and motion input subspaces within each context are close to orthogonal, although this is in fact expected by chance.

   (a) In monkey A, the color subpaces were highly aligned across contexts. The motion subspaces overlapped only along one of the dimensions, the one that reflected coherence information (fig. 3.17, B.22, 3.18). However, the second dimension captured very little variance in the color context, so this direction was either poorly estimated or simply captured noise. We also found a correspondence between the coherence input dimensions identified within the input subspaces and the input vectors found via TDR in the study by Mante et al. (fig. 3.15, 3.19). A model constrained to learn the same input subspaces across contexts captured the trajectories well, but performed slightly worse (fig. B.24 and table 3.5). This suggests that in order to achieve selective integration, a large reorientation of the input dimensions across contexts is not required. However, a small change in directions seem to be necessary to accurately capture the data. Therefore, for monkey A, inputs to PFC seem to bias the network in a similar fashion across contexts, but not via the exact same pattern of modulation.

(b) In monkey F, the motion subpaces were highly aligned across contexts. The color subspaces, however, were close to orthogonal (fig. 3.48, B.51, 3.49). We also found a correspondence between the coherence input dimensions identified within the input subspaces and the input vectors found via TDR in the study by Mante et al. (fig. 3.46, 3.50). For this monkey, however, this was weak and was found only for the motion vectors. In this case, we found that both color dimensions in the motion context reflected little variance, so the directions are likely to be poorly estimated or to capture mere noise. A model constrained to learn the same input subspaces across contexts slightly outperformed the model with flexible directions (table 3.10). This suggests that a reorientation of the input dimensions across contexts is not required in order to achieve selective integration. Therefore, for monkey F, inputs to PFC seem to bias the network along the same dimensions in the two contexts.

6. We find no evidence for a strong gating of input signals, based on what we could infer from the solution found by the LDS –with the exception of what occurs to color signals for monkey F.

   (a) For monkey A, the model-identified input strengths are comparable across contexts, for both the coherence and the coherence magnitude types of *input signals* –along the coherence and the coherence magnitude input dimensions (fig. 3.14). When inputs are run through the dynamics, however, the *integrated input* components *projected* onto the coherence input dimensions appear slightly amplified. This amplification is bigger when they are relevant than when they are irrelevant (fig. 3.29 and 3.16b), which is consistent with what was observed in the data (Mante et al., 2013). Therefore, in this model, the difference in strengths mainly originates from the projection pattern of the integrated inputs along the input dimensions and not from actual differences in the input drives. Finally, the *norm* of the integrated input components, which represents the strength of the total color and motion related signals driving the population FRs (fig. 3.8), steadily increases during the trial for the relevant inputs. For the irrelevant inputs, it saturates early in the trial and stays at the same level until the very end. This is found consistently in both the LDS and the JF models.

(b) For monkey F, the model-identified input strengths are comparable across contexts, for both the coherence and the coherence magnitude types of input signals –except for the color coherence ones– (fig. 3.45). The pattern of integrated motion inputs along the motion coherence input dimensions presents a very mild amplification and is similar in amplitude across contexts (3.60 and 3.47b) –unlike to what is found for monkey A. This indicates that the motion integrated input vectors are largely orthogonal to the motion coherence input dimensions throughout the whole trial. The color inputs, however, differ in several ways. First, coherence inputs, when irrelevant, are very weak. If relevant, they are much stronger and have the form of an integrated input signal. When inputs are run through the dynamics, the pattern of integrated color inputs displays a strong amplification along the color coherence input dimension. This is because, as we will explain below, the color coherence input direction in the color context is strongly correlated with the final direction of integration (fig. B.66b). The norm of the integrated input components, which represent the strength of the total color and motion related signals in the population FRs (fig. 3.39), steadily increases during the trial for the relevant inputs. For the irrelevant inputs, it saturates early in the trial and stays at the same level until the very end. In the case of the color, the signal is weak, but present throughout the whole trial nonetheless. This is found consistently in both the LDS and the JF models. The results, therefore, suggests that a partial –but not complete– gating of the color signal takes place across contexts. However, as we will comment on later, this findings could as well indicate that the dynamics was only partially observed.

7. In the model, the input signals within the input subspaces have a transient nature –except for color in monkey F–, meaning that they are attenuated towards the end of the trial. The temporal input patterns typically present an initial peak followed by a gradual decay, which however, does not go back to zero (fig. 3.14, 3.45). Therefore, there is a substantial residual input signal at the end of the trial. This effect is also present in the pattern of integrated inputs. The results in 3.16, 3.47 suggests that the input subspaces are largely *invariant* to the action of the dynamics, as the pattern of integrated inputs within these subspaces largely follows the temporal structure present in the input biases –although they display a mild amplification, as explained before. This means that the integrated input vectors must be rotated away from the input subspaces, so that they project weakly into them. If the input subspaces and the integration subspaces overlapped, the pattern of integrated inputs along the input subspaces would no longer be transient, as the effect of the integration would be strongly reflected in there. Alternatively, it could have been that the input signals learned are constant and that the transient

patterns that we observe in the data –along the input dimensions– arise entirely due to the fact that the integrated inputs are being rotated away. However, this is not the solution that this model infers, as the input biases themselves are transient. Therefore, the presence of arcs in the trajectories is explained –at least in this model– by the existence of an incoming input signal along the input dimensions which is transient, and not uniquely via a projection pattern from inputs being rotated by the dynamics. The integrated inputs, however, project strongly into the input subspaces at the beginning of the trial (B.28a, B.56a) –with the exception of color for monkey F– and therefore, the integration pattern is reflected in there. As we explained before, this gives rise to the differences in strength between the relevant and the irrelevant integrated input patterns observed along the input dimensions in monkey A (fig. 3.29 and 3.16). The projection, however, quickly becomes small and gets close to zero towards the end of the trial. This happens earlier for the motion vectors in monkey F (B.28a, B.56a), which explains the fact that for this monkey, the motion integrated pattern is not reflected in the motion coherence dimensions. The final integrated relevant inputs, which as we will see correspond to the decision axis, become then orthogonal to the coherence input dimensions –again, with the exception of what occurs to color for monkey F, which on the contrary, strongly aligns to it.

(a) In monkey A, the patterns of input signals found across both contexts and modalities are similar.

(b) In monkey F, the patterns of input signals found across contexts for motion are similar and consistent with the patterns found in monkey A. For color, however, coherence signals are very different. In the color context, the model learns a ramping-like pattern, which resembles an integrated signal. In the motion context, the color signal is very weak and mildly separates coherence information.

8. The evolution of the data trajectories in the task-relevant subspace is suggestive of a process slowly accumulating sensory evidence towards a choice along a single decision dimension (Mante et al., 2013). The LDS model implements a solution in which relevant inputs are indeed integrated by a dynamical process (fig. 3.23 and 3.24, 3.54 and 3.55). However, the integration does not seem to occur along a unique direction, given that the dynamical system inferred contains multiple slow modes (fig. 3.11, 3.42), which are in fact distinct and point into different directions in state space (fig. B.25, B.53). Therefore, a whole integration subspace exists in each context. The solution that the model found is not characteristic of that of a line attractor, as proposed in Mante et al. (2013), which presents a unique mode of integration and fast decaying dynamics on the rest of the dimensions. The mechanism for contextual input selection seems indeed to be largely mediated by

a reorientation of the dynamics, so that relevant inputs project more strongly onto the slowest modes in each context and are hence amplified (fig. 3.21, 3.52). Finally, we found that a model with dynamics constrained to be the same in both contexts performed surprisingly well (fig. B.15), either comparably or slightly worse than the model with flexible dynamics (tables 3.4, 3.9). The nature of its solution, however, made us rule it out. This is because the model implemented a highly unstable system involving strong oscillatory dynamics along a single negative eigenmode (see section B.1.2.3), which we believe is unlikely to be implemented in a neural substrate. Therefore, the differences in the population activities across contexts are consistent with a model that involves a change in the dynamics of the underlying circuit and are unlikely to arise uniquely from direct input modulations[1] –with the exception of the color signal for monkey F, as we discuss below.

(a) In monkey A the same mechanisms for contextual selection and integration of inputs seem to apply for both color and motion signals.

(b) In monkey F, the mechanism for contextual selection and integration of inputs as we have described it applies to the motion inputs. For the color inputs, however, an additional mechanism seems to be taking place, which involves the partial gating of the color signals when they are irrelevant. The integration mechanism seems also to be different, as the model learns a pattern of integrated color inputs that is not internally generated by the circuit, but that arises via direct input modulation (fig. 3.45). However, as we will see later, this findings could as well indicate that the dynamics were only partially observed.

9. As discussed above, integration does not seem to occur along a single dimension. However, we found that the integrated relevant coherence inputs are indeed mapped to the same direction in state space, regardless of the context. This was the case in both the two monkeys. In particular, on the model-generated data, we were able to identify directions within the integrated relevant 2D subspaces –defined at the end of the trial– that separated the sign of the relevant coherence inputs. These directions were consistent with the decision axis identified on the PFC data via TDR (Mante et al., 2013) (fig. 3.25, 3.56). Both the model and the data trajectories displayed selective amplification along these model-identified decision dimensions

[1] At the time of the deposition of this thesis --a couple of months after the defense– we found that, given a particular initialization, another optimum can be found for the model with constrained dynamics across contexts which does not converge to an unstable LDS solution. We found this different "initialization" incidentally, when swapping the order of the M-step updates for matrices A and B –which are in this case learned independently. In the stable case, the input matrix B is updated first, using the value of A found in the previous iteration –and the other way around for the unstable case. We decided to keep the conclusions of the thesis unchanged in order to be consistent with the work presented in the Viva. In future publications the conclusions will have to be revised. We found that the CV performance of this model was slightly worse than that of the model with different dynamics in each context. However, the drop in performance was very small. Therefore, given that this new solution is stable, the evidence against the model with single dynamics is currently too weak to reject it.

(fig. 3.26, 3.57). The integration pattern that is reflected in them originates from the projection of the integrated relevant input vectors, which are both amplified and rotated by the dynamics until they converge to these particular directions (fig. B.29, B.57). Given that the dynamics presents multiple slow modes –which are distinct– the rotation experienced by the integrated input vectors occurs on a high-dimensional integration subspace, spanned by the multiple slow directions. Within the final integrated relevant 2D subspaces, only the decision-like dimensions were consistent across contexts (fig. B.33, B.61). Therefore, the integrated relevant subspaces do not completely overlap across contexts, but they do in fact share this final dimension of integration. The integrated color and motion input components are kept orthogonal throughout the whole trial (B.12, B.48), although this is in fact expected by chance. Finally, the fact that the inferred linear dynamical system contains multiple slow modes suggests that there is more than one time constant underlying the dynamics of integration. This seems to be key in order to capture the complexity in the PSTHs. This can be seen from the fact that the firing rates in the model are reconstructed via a linear mixture of the modes of the dynamics –each of them evolving independently with a characteristic time constant and under different degrees of input modulation (equations 3.4 and 3.5).

(a) In monkey A, interestingly, the irrelevant inputs are also amplified, although in a much lower extent. The dynamics maps them into directions orthogonal to the decision axis (fig. 3.25), so the mild integration that they experience is not reflected in there (in 3.27 but not reflected in 3.23). The second integrated relevant dimensions, orthogonal to the decision-like dimensions B.33, reflect coherence magnitude (fig. B.35 and also when irrelevant B.36). These signals, however, are much weaker than the strongly amplified relevant coherence input patterns along the decision axis 3.26.

(b) In monkey F, the motion signal in the color context is also amplified along a dimension orthogonal to the decision axis (fig. 3.56), but in this case, quite strongly (in 3.58b but not reflected in 3.54b; see also 3.39 and compare with monkey A 3.8). Regarding color, we also found a moderate integrated signal in the motion context. This, however, was localized to the decision dimension (fig. 3.56), so its effect on the trajectories is almost entirely masked by the strong integrated motion signal (fig. 3.58a). The second integrated relevant dimensions, orthogonal to the decision-like dimensions B.61, also reflect coherence magnitude variance (fig. B.63, B.64).

10. For monkey F, color coherence information in the color context seems to enter the system directly along the decision axis as a form of an integrated color signal. In the motion context, however, the color influence is very weak. These findings suggest that, either a mechanism of the type implied by an early selection model takes place in this monkey, or that the amount of neurons recorded were not

enough to characterize the full dynamics. Furthermore, we found that compared to monkey A, the difference in the strengths between the relevant and the irrelevant integrated inputs along the decision dimension was smaller. Regarding the nature of the strong "urgency signal" present in the data from this monkey (fig. B.47), our model unfortunately does not allow us to determine its origin and function. We could simply independently verify that condition independent effects were stronger in this monkey than in monkey A. This can be seen from the condition independent component inferred by the model, which is stronger for monkey F, in particular along the decision axis (compare fig. 3.23 and 3.54).

4.1.2 Further considerations and limitations

We expand on the previous points, making additional comments and discussing multiple caveats

1. We remind the reader that to be able to compare the LDS and the JF models, we restricted our analysis to the noiseless, deterministic component of the LDS prior. However, to fit the LDS, we applied the standard EM optimization procedure on the full probabilistic model[2]. The JF model, on the contrary, does not incorporate a noise model and the optimization is based on a low-rank reconstruction of the data, in a least-squares sense. In order to be completely rigorous, the the two optimization procedures should be matched. We could do this in two different ways, by either constructing a deterministic version of an LDS or by incorporating a noise distribution into the JF model. In spite of the discrepancy in the fitting procedures, we found that the two models learned almost identical FR decompositions, as given by the color and motion input components and the base components (fig. 3.8, 3.39 and B.12, B.48). Therefore, the two models find similar neural mappings for the different input patterns in time.

2. The results obtained when comparing the LDS model and the JF model solutions strongly support the linearity assumption on the dynamics unfolding within each context. The underlying system, however, is likely to be non-linear if a switching of the dynamics is required to occur in each context. For instance, the RNN model used in the study by Mante et al. was non-linear, a key feature that allowed the network to effectively implement a different dynamics in each context. Furthermore, the solution implemented in each context was also characteristic of non-linear systems, given the existence of multiple dynamical attractors. However, the dynamics could be largely approximated by a linear system, at least in the vicinity of the fixed points. What our model indicates is that the dynamics observed in each context can be globally supported by a simple linear dynamical system.

[2] Note that we are fitting the LDS to trial *averaged* data, so the "noise" terms capture *residual* trial-to-trial noise that was not fully averaged away and also reflect model *mismatch*.

3. When fitting the LDS, input directions and input signals are estimated based on the residuals of the dynamics (equations A.1, A.2 and section A.1.2). That is, the inputs capture a component that is being missed by the dynamics at each state transition. This component, importantly, is constrained to live in the same subspace throughout the whole trial, defining what we call the input subspace. A valid concern could be that the model is incorporating arbitrary input influences and that the dynamics is learned accordingly to compensate for that –as mentioned several times, the same problem of having coupled parameters that can lead to degeneracies. We believe that the inputs learned in this case are not arbitrary, given that we converge to similar solutions for a wide range of different models and hidden dimensionalities, but it is true that different solutions, in principle, could have been inferred[3]. Furthermore, the input dimensions discovered by the model are associated with groups of neurons which clearly follow the activity patterns along them (fig. 3.33, 3.34 and 3.64, 3.65). This indicates that the model is indeed capturing lawful structure in the population activity.

4. The fact that the color and motion input subspaces B_c, B_m have more than one dimension indicates that input influences are not simply collinear. This also means that the LDS, effectively, can implement inputs that change direction, both in time and across conditions. This can be understood when looking at the pattern of inputs in the 2D planes (fig. 3.16a, 3.47a). The points along each of the six possible input trajectories define the input vectors, which present different orientation at each time step and also across trajectories. The changes in orientation, however, are constrained to be within the 2D plane defining the input subspace. With this perspective in mind, lets introduce again the JF model. We mentioned before that this model has the flexibility to capture the influence of inputs biasing the system via *arbitrary* directions at each point in time. We understand now that the LDS, by considering higher dimensional input subspaces, can effectively achieve this too. What the LDS allows us to do then, is to impose *constraints* on the dimensionality of such input subspaces, so that each input vector is restricted to live in the same subspace throughout the whole trial. Therefore, at each point in time, the color and motion input directions are defined by the vectors $B_c\boldsymbol{u}_c(t)$ and $B_m\boldsymbol{u}_m(t)$, which live in the planes spanned by B_c and B_m. These subtle changes in orientation allow the system to process the input conditions differently, as in reality, each input vector –across conditions and times– is associated with a different projection pattern into the left eigenvectors of the dynamics. This can be seen from equation 3.4, as at each point in time the inputs enter the dynamics via the projection $\boldsymbol{l}_h^\intercal B\boldsymbol{u}(t)$.

[3] It is always the case that some form of input must be incorporated to the system in order to specify the conditions and provide coherence information. The question is knowing with certainty which exact form this bias has in reality.

- Finally, we comment on the fact that the LDS input subspaces contained a dimension reflecting coherence magnitude information (the JF model did also recover such type of input influences B.20, B.50). We are not certain about the nature of this signal. It appears early in the trial, in the form of a sharp peak in the case of color in monkey A, or as a smooth transient in the case of motion in monkey F. The effect seems to persists throughout the trial, in particular in the case of the color in monkey F and motion in monkey A. Ding and Gold (2012) reported similar type of signals in FEF. They found neurons that were modulated by motion strength with the same sign for the two choices. They also found that, because motion coherence strength largely determines trial difficulty, the activity correlated with reward probability. Therefore, these signals could represent additional behavioral variables, such as expected reward values or confidence (Kiani and Shadlen, 2009; Grimaldi et al., 2015). Alternatively, they could reflect something very simple, like the overall saliency of the motion and color patterns. A more speculative idea would be that, given that the magnitude or absolute value is a non-linear function, the system is performing some type of non-linear transformation of the inputs. Finally, this particular 2D input representation could arise when the system is operating within certain normative frameworks. In particular, the existence of dimensions representing the absolute value of a signal is broadly consistent with the predictions of efficient coding schemes (Machens et al., 2005; Barrett et al., 2016).

5. We found that input subspaces across contexts were highly aligned along certain dimensions and conversely, that color and motion subspaces within a given context were very close to orthogonal[4]. In the study by Mante et al. the fact that motion and color input directions were orthogonal was of key importance in order to implement the selection of relevant inputs. In the same way, mapping the patterns of integrated inputs to a component perpendicular to the irrelevant inputs was crucial to be able to correctly decode decision. Similar mechanisms have been reported in Rossi-Pool et al. (2017), where orthogonal activity profiles in dorsal premotor cortex disentangle input transients from a sustained working memory and choice related component. In another recent study (Elsayed et al., 2016), the computational advantages of allocating preparatory and movement computations into orthogonal subspaces were also emphasized. This allowed the system to perform separate but linked computations. Therefore, both motor cortex and PFC

[4] We would like to comment on the quality of the evidence we gathered to support these conclusions. We are aware that the plots illustrating the alignment of the different input subspaces and of some dimensions within them are fairly noisy, specially for low dimensionalities. In future work, we would like to explore different types of input initializations during the fitting procedure in order to improve this. We could also consider models with different input dimensionality across contexts, so that input dimensions reflecting very little variance are not taken into account. This might improve the estimation of the rest of the dimensions.

exploit separate subspaces when performing a series of *subsequent* computations, such as movement preparation and movement execution –in motor cortex– or sensory input processing and choice formation –in PFC. Additionally, when dealing with the processing of several sensory inputs, PFC also seems to exploit this feature by allocating *simultaneous* computations into orthogonal subspaces. Importantly, in motor cortex, the orthogonality results were not expected to arise by chance. Therefore, it seemed like such subspaces were being actively misaligned. What we found in PFC, however, is that orthogonality is indeed expected by chance, given the dimensionality of the state space. The system, therefore, does not explicitly need to construct the subspaces to be that way. The clear advantage is that multiple input computations can easily be supported simultaneously. What it is not expected, however, is the fact that the motion and color subspaces are aligned across contexts. The view that we take is not that these subspaces are being actively aligned by the circuit, but rather that contextual changes –mediated potentially by contextual inputs– do not orthogonalise them, so the subspaces are *invariant* or *preserved* across contexts. This brings us back to the neuro-centric vs. circuit-wide views for coding. It is widely accepted that early sensory areas hold largely invariant response properties at the level of individual neurons, meaning that their representations are not much influenced by context. Coding principles can then be understood at the level of individual neurons (Jazayeri, 2017). Single neurons in higher order areas, however, do not present such invariant properties, as they exhibit highly heterogeneous response profiles, or "mixed" selectivity, even in simple tasks involving basic stimuli (Rigotti et al., 2013; Jazayeri, 2017). Our study suggests that sensory-related representations in PFC are in fact preserved at the population level, but the invariance of the representations is hidden in the complexity of the responses.

6. For the LDS models considered, the learned input strengths are comparable across contexts. We observe this consistently for different model specifications and hidden dimensionalities (for example, see fig. B.21). However, we emphasize that direct comparison of the learned input biases must be made with caution, as their influence is always coupled to the dynamics. This give rise to degeneracies in the model, which means that the same results can be obtained using different parametrizations. Therefore, we cannot completely rule out the alternative that, in the real PFC network, inputs arriving into the system have different strengths and that the dynamics creates dimensions where the differences are not apparent. What we can say, is that the data is consistent with the solution obtained by this particular model. We will restrain ourselves of making any strong claims regarding the absolute strength in which the inputs might be biasing the real system in PFC. Finally, for Felix, we experience a similar issue regarding the inferred intensity of the color coherence ramping-like signal in the color context, which seems to

directly target the decision dimension. This input signal is weaker than the final integrated pattern. We cannot tell whether in reality, as the model suggests, this ramping input signal is further amplified by the internal dynamics or that the integrated pattern arises entirely from the input –arguably, a more parsimonious scenario. Alternatively, as we have explained, the type of solution found for this monkey could be a consequence of having only partially observed the dynamics (Seely et al., 2016). This issue is very different from the degeneracy problem in the model fitting procedure, as it is a limitation from the data collection stage, if not enough neurons are sampled. We will elaborate on this point later in the discussion.

7. We emphasize here that in order to reproduce the transient responses observed in the trajectories along the input directions, the model does not require learning transient input signals. As we explained, this effect could arise from projection patterns of the integrated inputs, or via the action of non-normal dynamics producing transient amplifications along certain dimensions (Murphy and Miller, 2009; Hennequin et al., 2012) (in fact, the dynamics that we learn is non-normal). For instance, the LDS model with constrained dynamics across contexts does indeed learn a pattern of inputs that is sustained, but it is still able to capture the input transients. This is evident form the model generated trajectories in the task-relevant subspace (see section B.1.2). However, we have reasons to favor the solution implemented by the model with flexible dynamics, as it has been argumented before. One last consideration is that when fitting the model, we in fact initialized the input biases to be constant along the trial. If that was an equally valid solution, the model could have in principle found it[5]. Finally, we clarify again that the behavior of the trajectories along the input dimensions is not uniquely driven by the input biases. This is because these dimensions do reflect to some extent the pattern of integrated inputs. Similar patterns of activation have been observed in other associative areas. For instance in dorsal premotor cortex, Rossi-Pool et al. (2017) found that orthogonal activity profiles disentangle sensory input transients from a sustained working memory and choice related component. Importantly, in another study (Wimmer et al., 2015), the time course of choice probabilities in MT was shown to arise from an early component reflecting the *"transient integration"* of sensory inputs and a separate late contribution reflecting decision build-up. The early component captured the bottom-up effect of sensory inputs on the activity of the decision network as it approached an attractor. Later on, the attractor dynamics would decrease the impact of the sensory signal fluctuations on the network activity –and hence, on the upcoming decision–, generating the transient effect. In this case, the sensory input biases to the network were, in fact, modeled

[5] We did in the past fit a model that learned constant input biases, but we assumed only two dimensions, so the performance was poorer than the models we considered in here. In future work, as a control, we will certainly repeat the analysis for higher input dimensionalities.

to be *constant* and the dynamics created the *transient* effect. Therefore, the highly recurrent nature of such networks makes it almost impossible to infer from the data the *true* inputs arriving to the area. It is possible that the input signals that our model infers correspond to such bottom-up *transient integration* effects of sensory inputs. In any case, we are still able to separate such early *input effects* from the later *integrated input* components that give rise to the decision.

- Finally, we comment on the fact that the input signals that we learn are present until the end of the trial. It could be that our linear model requires this form of sustained inputs in order to keep them in memory. A non-linear model would be able to implement a more complicated portrait of attractor dynamics which allows to preserve the input effects in different regions of state space, as in the non-linear RNN proposed in Mante et al. (2013). In this network, two point attractors are implemented at the extremes of the approximated line attractor, which allows the system to preserve a memory of the decision. However, as soon as the trajectories converge to them, the input information is destroyed. We believe that it is unlikely that this solution arises in the real data, as it is the case that, more often than not, the data trajectories present a clear coherence effect at the end of the trial (see end points of the trajectories, for the two monkeys). Stimulus-related persistent activity has also been observed in FEF following decision commitment (Ding and Gold, 2012). This "memory trace" of task information could be used for decision reevaluation purposes.

8. The mechanism for contextual input selection seems to be largely mediated by a reorientation of the dynamics, so that relevant inputs project more strongly onto the slowest modes in each context. However, the projection patterns of the inputs onto the left eigenvectors of the dynamics that we recovered are not straightforward to interpret (fig. 3.21, 3.52), even when looking in isolation at the input component carrying coherence information, as we did. It is the case that overall, the relevant inputs seem to project more strongly into the slowest modes, but for all inputs the projection pattern is highly distributed across all the modes in non-trivial way. In future work we would like to provide better means to analyze this complicated dynamical portrait, so that we can develop an intuition of the exact mechanism the model is using to selectively integrate the inputs.

    - Finally, as we concluded, the data in the two monkeys supports a model with different dynamics in each context and fixed inputs over a model with the same dynamics but with different input gains across contexts –on the grounds that, given the unstable nature of this last model, it is unlikely to be operative in a real circuit. However, considering that the differences in performance between the two models are very subtle, we do not completely rule out the possibility that other type of models, perhaps non-linear in nature, could capture the data accurately under a single dynamics.

9. The integration pattern that is reflected along the decision dimension originates from the projection of the integrated input vectors, which are both amplified and rotated by the dynamics, until they converge to this particular direction. The subspace of integration, where this vector is being rotated, is high dimensional given that the dynamics presents multiple slow modes –which are distinct and slowly decay along different dimensions, with time constants specified by the eigenvalues λ as $|\lambda|^t$. An important distinction between the picture implied by our model and the one implied by a line attractor is in the dimensionality of this integration subspace. In the case of the line attractor, the integration subspace is confined to a plane, which is spanned by a single direction of integration, given by the line attractor, and the direction specified by the input bias –note that we are considering here a single input, for simplicity, which we assume constant thorough the trial as was set for the nlRNN in Mante et al. (2013). The integrated input vector, then, has two components, the component that is being amplified in the direction of the line attractor and the input bias component, which in the setting discovered by the nlRNN, is orthogonal to the line attractor. As the component along the line attractor grows, the integrated input vector gets rotated away from the input dimension and aligns more and more strongly into the line attractor dimension. In our model the picture is similar conceptually, but involves thinking in higher dimensions for both the input subspaces and the integration subspaces. The integrated input in our case gets rotated away from the 2D input subspaces, as it is amplified by the dynamics, and changes direction within a whole integration subspace spanned by multiple slow modes. This subspace is largely orthogonal to the input dimensions, as the amplification is weakly reflected in them. Eventually, the component of the integrated vector that reflects relevant coherence information gets mapped into a particular direction in state space, consistently across contexts, which defines the decision dimension.

   - We proceed now to discuss a related issue. We mentioned that the firing rates of the neurons are reconstructed via a linear combination of the different components of the dynamics (equation 3.4). These components $\alpha_h(t)$ evolve independently along directions in state space specified by the right eigenvectors of the system $\boldsymbol{C}\boldsymbol{r}_h$ (equation 3.5). The fact that the inferred dynamical system contains multiple slow modes, as we just discussed, suggests that evidence integration does not occur uniquely along a single dimension. It also suggests that there is more than one time constant underlying the dynamics of integration. In fact, if the system had only one integration mode and the rest of the dimensions were fast decaying –as implied by a line attractor solution–, all the neurons would present the same stereotypical ramping-like response profile (except the ones reflecting only the input drive). But this is not what we observe. As we have already seen, neurons in PFC have complex, heterogeneous responses. Therefore, a richer basis set, with several components $\alpha_h(t)$ evolving under multiple time constants, is needed to explain the diversity of PFC's firing rates. This can be seen if we express the individual neuron responses directly as a function of the independent components of the dynamics $\alpha_h(t)$ (equation 3.6).

- Finally, we comment on our finding regarding the integration of irrelevant inputs. A study by Kumano et al. (2016) found that in LIP single cell ramping patterns also reflect the irrelevant inputs, but the build up is much weaker[6]. This is consistent with what we observe, as our model identifies that the irrelevant input signals are also integrated into some extent (fig. 3.27, 3.58). In our case, however, we found that the integration occurs along a dimension that is orthogonal to the decision dimension.

10. For monkey F, we explained that color coherence information seems to enter the system directly along the decision axis, already in the form of an integrated color signal –given that the color coherence input direction and the decision axis were found to be the same. We clarify that the color signal is in reality 2D, as the color input vector lives within a whole plane, so it is not confined to the decision axis. This can be deduced from the projection pattern of the integrated color input along the decision axis in the color context (fig. B.57). The projection pattern can be explained by taking into account the influence of the magnitude signal. At the beginning of the trial, the integrated input vector aligns mostly with this magnitude component, but then the coherence input signal starts growing and pulls the integrated input vector towards it (see 2D representations, in particular for the strongest coherences 3.47b) So, effectively, the integrated input vector is rotated towards the decision dimension at the end the trial. This not exactly the same picture implied by the early selection model as discussed in Mante et al. (2013). Alternatively, instead of implying early selection, the solution found –requiring a direct integrated input influence– could be a consequence of having only partially observed the dynamics (Seely et al., 2016). For instance, we might have missed key circuit components that mediate the transformation of color inputs towards decision signals. We will further comment on this below.

Having discussed in length the points explained in the previous section, we wanted to make a few additional comments:

In the first couple of points at the beginning of the previous section, we justified

[6] The fact that the irrelevant inputs have an impact in the evidence accumulation process is somewhat expected from the study by Mante et al. as the irrelevant information does weakly influence the pattern of the trajectories along the decision dimension, as well as the behavior.

the claim that PFC data is well described by a dynamical system model. Several research groups have also argued in favor of using a dynamical systems framework to understand the complexity of neural population data (Churchland et al., 2007; Shenoy et al., 2013; Wang et al., 2018). This approach has motivated a lot of work training RNNs to model a variety of tasks (Barak, 2017), involving motor control (Hennequin et al., 2014; Sussillo et al., 2015), decision making (Machens et al., 2005; Mante et al., 2013) or timing computations (Carnevale et al., 2015; Wang et al., 2018; Remington et al., 2018) among others. This framework has been extremely successful in providing intuitions about the computations that might underlie the activity of populations of neurons. However, the justification for this modeling approach is based on rather indirect evidence, motivated by the capability of such models to qualitatively capture important dynamical features of the population responses. In our case, we were concerned that our constrained LDS model, despite being able to capture some general trends in the data, would be missing important structure. By comparing the LDS with the JF model we verified that the data is well supported by a dynamical system, as the additional flexibility offered by the JF model is not necessary to capture the data well. Therefore, the type of structure present in the PSTHs lawfully follows the assumptions of the LDS prior. Similar concerns have been raised before in the literature, although from a different perspective. It is often argued that population structure typically found in neuroscience datasets is a hallmark of an emergent collective behavior, arising from an underlying dynamical process or reflecting a circuit-level gestalt computation. However, this view has been recently challenged (Elsayed and Cunningham, 2017). It could well be that the population structure simply reflects properties of individual neurons, such as single cell tunings, and not a circuit-wide emerging code. In a similar line of argument, it could be that the temporal structure present in the data is not indicative of a dynamical process unfolding, but that it simply reflects correlations among neurons and smoothness properties of the PSTHs. The authors of the study Elsayed and Cunningham (2017) developed a methodological framework to help address these concerns. The method consists of generating surrogate data sets that preserve the first and second moments of the original data, but remove the rest of the structure. In this way, the surrogate data sets are characterized by the set of "primary features" of the original data, which give rise to basic properties such as neural and temporal correlations, but have no higher level structure. These data sets can be used as a control, forming a null distribution which is more conservative than the ones created with standard shuffling tests. This allows to test for significance when a particular population metric is computed, such as the goodness of fit for a dynamical systems model. Using this method, the authors found that the dynamical structure present in recordings from motor cortex in monkeys is not a byproduct of basic features. Even though this approach clearly proves that the dynamical system model is capturing additional structure in the data, it does not tell us how good the dynamical assumption is. In that sense, our analysis is a stronger test for dynamics.

We have shown that the LDS and the JF models capture the data similarly, finding almost identical decompositions in terms of input influences and base components. However, the decomposition found by the LDS was not at first sight similar to the decomposition provided by the JF model, as we had to subtract condition independent effects from the motion and color components (fig. B.11). Therefore, it is evident from this analysis that caution must be taken when attempting to interpret results in the presence of degeneracies in the models. The invariants of the fits are given by the sum of all the components, which constitute the total FRs, but in principle different decompositions can be found. Once we understood that condition independent effects could be allocated to all the components, and accounted for that, comparing the different input terms across models was possible.

Studying the pattern of both training and CV errors across individual conditions, we found that performance of both the JF and the LDS models is the poorest for left-out conditions on the weakest coherences. The biggest errors result from conditions where strong irrelevant evidence is paired with weak relevant evidence, that is, were there is strong incongruent information (fig. 3.7, 3.38). This same effect is in fact reflected in the performance of the monkeys, as they are unable to ignore strong irrelevant information perfectly (fig. 3.1 d,e and B.46 b,c). Remember that the RNN network in Mante et al. (2013) did not present such bias (fig. B.46, f,g). The model fits from the two monkeys presented these type of errors, but in the case of monkey F, they were much more severe. This is interesting because monkey F, indeed, performed more poorly and in particular, he seemed to be less successful at ignoring irrelevant information. This is apparent in the pattern of choices of the monkey, as there is a prominent bias towards the irrelevant input information when it is strong (fig. B.46 b,c). We are still uncertain about the exact source of the error pattern, but given that the two models are impaired in exactly the same way, this suggests that the limitation does not come from the dynamical constraint, but rather from the assumption that inputs are linearly combined across conditions. We are tempted to speculate on the possibility that a similar non-linear mechanism underlies the source of errors –of this type– made by the monkeys. For instance, it could be that when the irrelevant input is very strong, it is able to trigger some form of suppression mechanism which partially inhibits the relevant inputs, if these are weak.

In the LDS model we focused on, the input biases are constrained to share the temporal structure across different coherence strengths. This common input time series is then scaled by a different number, one for each possible coherence value (fig. 3.12, 3.43). We also allowed the model to learn different temporal patterns for Tin and Tout types of evidence (positive and negative coherences), as we found that this helped improve performance. For instance, we found that in some cases, the Tin and Tout types of evidence pick at slightly different times in the trial, with Tout evidence arriving with a greater latency. The LDS model with the above specified input constraints performed as well as an LDS with full flexibility to learn arbitrary temporal patterns per coherence (tables 3.2, 3.7). Furthermore, the flexible model learned similar input patterns to the constraint one (fig. B.21). Therefore, our results support a model in which the input signals, regardless of their intensity, bias the system throughout the trial via the same pattern of modulation across time. The strength of such modulation is scaled by a multiplicative gain factor, which is related to the coherence value presented in the experiment. In the case of the irrelevant coherence inputs, this relationship is learned to be linear. In the case of the relevant coherence inputs, the weakest and medium coherences are often learned to be stronger than the true coherence strength presented (fig. 3.13, 3.44 and the same is found for the JF model B.20, B.50). This non-linear scaling of the relevant inputs is also reflected in the trajectories. Given that our model is linear, such effect can only be captured via the inputs. Therefore, we cannot distinguish whether this is a true feature of the inputs arriving to PFC or whether this effect is generated via non-linearities in the local circuit when the relevant inputs are processed.

We have demonstrated that the LDS model is able to capture the whole complexity of individual firing patterns, despite the high degree of heterogeneity in the population. In the main text we commented on the fact that some units seem to be highly sensitive to certain features of the inputs, such as color strength, but later in the trial they loose that information to become strongly tuned to the decision (fig. B.45, unit in fourth row). This behavior is what we would expect from neurons that receive direct input influences, but that are embedded in a larger recurrently connected circuit which is performing a transformation of the incoming signals. The alternative, which cannot be distinguished, is that all activity patterns are externally driven and that the dynamics that we locally observe is being *fully* inherited from other circuits upstream. In either case, further subtleties apply depending on whether we assume that we are *fully* or *partially* observing the *local* system. It has been argued that these two scenarios result in very different types of tensor data structures (Seely et al., 2016). Hence, in a population that exhibits recurrent dynamics –either produced locally or inherited from elsewhere– the expected data structure would be different depending whether we are able to *fully* or *partially* observe the dynamics. Given that in reality we do not have access to the whole system, this depends on whether we have recorded from enough neurons to fully characterize the dynamics. If the system is only partially observed, according to Seely et al. (2016), the type of structure that would be recovered would be equivalent to that of an input driven system (see footnote in 2.2.2). Therefore, coming back to Mante's dataset –and assuming that our LDS model is somehow sensitive to the stated differences in population tensor structure–, the fact that the model infers a dynamical solution in some cases (motion and color for monkey A, motion for monkey F) and an input driven system in other cases (color for monkey F) suggests the following scenarios. In the first case, the model would be identifying the *full* dynamics, which could be either produced *locally* or be *fully* inherited from elsewhere. In the second case, the structure could arise via direct inputs, as suggested by the model, with the local system *partially* inheriting the dynamics from elsewhere. The alternative, however, would be that the dynamics are in fact produced *locally* –or are *fully* inherited from elsewhere– but we are only *partially* observing the system. Further analysis of the type performed in Seely et al. (2016) could be made in order to test for differences in the tensor structures implied by the data from the two monkeys.

Having said this, it is almost certain that the dynamics that we observe results from the orchestration of many different brain regions jointly contributing to the decision formation. The current view, in fact, is that decision signals emerge from a highly interconnected and distributed network of different cortical and subcortical areas (Siegel et al., 2015). It is likely that different elements of the computation are distributed and supported by different brain regions, but without the existence of sharp functional boundaries. For instance, in rats, evidence accumulation signatures are found in both parietal (PPC) and prefrontal cortices (FOF), but while PPC encodes accumulating evidence in a graded manner, the FOF does so in a more binary-like fashion, closer to a action triggering signal (Hanks et al., 2015). The extent of this distributed network is likely to include sensory areas, at least the ones up in the hierarchy. In a recent study, Tajima et al. (2017) found evidence for context-dependent attractor dynamics in the higher visual cortex of macaque monkeys, demonstrating the emergence of discrete attractors in IT during categorical perceptions. This suggests that sensory areas may be involved in the contextual modulation of the dynamics not only as mere relays, but also as a part of the same recurrent circuit that carries out the input transformation.

Finally, in order to understand the complexity of individual PSTHs, we summarized the activity of different neurons as participating into different population-wide activity patterns (fig. 3.32, 3.33, 3.34 and 3.63, 3.64, 3.65). With this analysis we were able to provide a summary of the population tuning to the different task variables. This highlights the advantages of the modeling approach we took. First, we could incorporate task information in the fitting procedure by labeling conditions according to the input categories presented –the color and motion coherence levels and direction. We could then use this information to infer different types of input signals. Second, our fitting approach allowed us to capture as much variance as possible from the data, as we could increase the flexibility of the model by expanding both the hidden and the input dimensionalities. Importantly, this approach combines the strengths of both supervised and unsupervised methods, as we tailored the model to the task, providing it with some form of labeled data, but we also required it to capture as much structure in the data as possible. This approach allowed us to find dimensions that captured the coherence-related tunings expected to arise from the task, but after expanding the input dimensionality, we were also able to identify other coherence-related signals in the data. In this way, we believe that we have overcome the limitation that the LDS models have been argued to present, namely, that they do not take task parameters into account and that they do not provide a summary statistic of the population tuning (see discussion and Table1 in Kobak et al. (2016)). Furthermore, we were able to determine the dimensionality of the task-related subspaces –via cross-validation in this case. Other methods such as dPCA can also consider multiple components per task parameter (unlike TDR). However, the current method does not incorporate any criteria to select the rank of the subspaces[7]. Furthermore, input influences cannot be explicitly modeled. Finally, this method suffers from similar limitations as TDR when attempting to distinguish variance from highly correlated parameters, in this case, the inputs and the integrated inputs. Finally, factorization-based methods of the type of dPCA, TDR –and possible extensions– and also our own JF method, present a clear limitation over dynamical systems models. Even though they do no need to commit to heavy assumptions about the underlying structure in the data, they cannot provide the mechanistic insights that dynamical models offer.

In this work we have developed a novel theoretical framework that provides new insights into the computations implemented in PFC during complex tasks, such as contextual decision making. We found that input signals bias this network in a similar fashion across contexts, with the dynamics of the circuit changing so that the relevant information is selected and amplified. This mechanism seems to underlie the processing of motion inputs in the two monkeys under study. In the case of the color inputs, however, a different mechanism seems to be taking place in one of the monkeys. The color integrated signal arises in this case via direct input modulation and it is partially gated when irrelevant. This supports a model in which a form of early selection of inputs mediates the integration of relevant information, or alternatively, could indicate that the dynamics of the decision process was not fully observed. Our study introduces the view that independent high-dimensional input subspaces exist within PFC which are invariant across contexts. Decision signals emerge as the inputs are transformed by the dynamics of the circuit, which changes in each context. Our model provides a new perspective on the type of input signals that might be arriving into PFC and implies a novel mechanism by which sensory-related information is selected and integrated for contextual computations.

4.2 Open questions and future work

In this section we point out additional limitations of our method and discuss possible extensions.

We start with a very important point. The fact that the PFC data was not simulta-

[7] although this can be accounted for post-hoc, based on the cumulative variance explained by a certain number of components.

neously recorded, so we had to work with averaged PSTHs and not with single-trial spike trains. The obvious limitation of this approach is that averaging can destroy valuable information about the correlational structure in the population, so the dynamics that we are inferring may in fact not be representative of what the circuit is doing on a single trial basis. It seems to be the case that, typically, the dynamics recovered on single trial data are consistent with the dynamics recovered on averaged responses[8]. However, recent studies have argued that individual responses during single trials involve sharp transitions or steps (Latimer et al., 2015; Morcos and Harvey, 2016). This is very different from the smooth temporal structure implied by the averages, and hence, not lawfully characterized by a dynamical systems model. Our method can readily be applied to single trial data, so this hypothesis can easily be tested by comparing its performance against alternative methods. Such analysis would be one of our highest priorities for future work, so that we can verify that our claims indeed hold at the single trial level.

Having said that, caution must be taken when attempting to compare models that are very different in nature. In particular, when using standard methods for model selection (Latimer et al., 2015). Favoring one model over another on the grounds of a marginal gain in performance does not justify endorsing the particular mechanism implied by the winning model class. It will almost always be the case that the data is far from any of the model classes, so performing post-hoc analysis to understand why one particular model instance –within a given class– is outperforming another particular model instance –within a different class– is an essential step in model selection. In particular, if one wants to make any claims about the data following the mechanisms implied by the winning model class (Chandrasekaran et al., 2018). We encountered a similar issue when comparing the results obtained from the LDS models with constrained and unconstrained dynamics across contexts. The model with unconstrained dynamics marginally outperformed the one with single dynamics –for monkey A–, but this was not the only reason why we discarded it. This was also based on a qualitative assessment of the solution found, after testing the stability and plausibility of the identified dynamics.

Latent variable models, such as linear dynamical systems, have proven to be extremely successful at describing the statistical structure in populations of spiking networks. These models assume that the shared variability and the temporal correlations in the data arises by a low-dimensional latent process evolving with smooth dynamics (Macke et al., 2011). Regardless of the true underlying processes generating the dynamics observed at the population level, whether continuous or discrete, the intuitions provided by such models could still prove to be very useful. After all, we do still launch satellites based on simple Newtonian physics, without the need to compute extensive relativistic calculations.

An extension to the LDS that we had previously considered incorporates Poisson observations (Macke et al., 2011, 2015). Given our finding that the dynamics within

[8] personal communication, not published data.

each context is well approximated by a linear system, this is the obvious extension as the PLDS model is still linear in the dynamics. Another model worth considering is PfLDS (Gao et al., 2016), which also assumes linear dynamics but features a more general observational model, with arbitrary non-linear mappings and the option to incorporate different noise models.

If these are to be fit to the single trial count data from Mante et al. study, we must consider that neurons had not been simultaneously recorded. Meaning that, when pooling data from all neurons, one must take into account that an unequal number of trials were recorded for each neuron. This can be considered as a missing observations problem[9]. We did in fact derive an expression to fit a PLDS considering this scenario (see section 2.3.2.1). To account for the discrepancy, we weighted each neuron's contribution to the likelihood by the number of trials recorded in each case. We found that this procedure was almost equivalent to fitting the model to the averaged count data.

We assessed the performance of the PLDS model on Mante's data and found that it was slightly worse to that of the LDS. Furthermore, due to the approximate inference step required in this model, it was much slower to fit, so the extensive cross-validation analysis that we performed on the multiple variants of the gaussian models would have not been possible in this case. A final caveat was that the non-linear exponential mapping between latents and observations rendered the dynamics more difficult to interpret.

The results in this thesis can be used to guide the development of new RNN models. In the context of Mante et al. study, it would be very interesting to train their nlRNN with the input specifications we obtained and then analyze the dynamical portrait that it is recovered. Further extensions could incorporate different levels of biological realism, as in Wang (2002); Machens et al. (2005). Other RNN models that are hand-crafted with minimal assumptions about core connectivity and structure could also benefit from our results. For instance in (Mastrogiuseppe and Ostojic, 2017) the network connectivity consisted of a combination of a random part and a minimal, low dimensional structure. The authors were able to implement context-dependent computations by using a simple rank-two structured connectivity. Interestingly, the dynamics of the constructed network formed an approximate ring attractor[10]. Extending the rank of the structured part of the network and incorporating some of the input features we inferred from Mante's data could help bring important theoretical insights into these type of contextual computations.

Finally, it is worth mentioning that other groups are working on combining the expressivity of RNNs –given that they are powerful function approximators– within the framework of latent variable models. For instance, the LFADS model (Pandarinath et al., 2017), which is a sequential extension of a variational auto-encoder (Kingma and

[9] note, however, that this does not circumvent the obvious limitation that, in reality, the neurons did not simultaneously participate in the dynamics.

[10] A similar ring attractor structure emerged when a nlRNN was trained with a contextual input that was transient –and was provided only at the beginning of the trial. This was work by Sepp Kollmorgen, a postdoc at Valerio Mante's group, which was presented at Cosyne 2016.

Welling, 2013). This model seems to be very powerful at extracting data structure and discovering latent factors. However, the recurrent connectivity in the RNN is assumed fixed. It remains an open question whether such type of models can also offer mechanistic insights into the computations implemented by neural populations.

A limitation of our model is that, without specifying any constraints on the type of dynamics that can be inferred, the system can be learned to be unstable. Incorporating prior constraints of the type applied in Buesing et al. (2012) and Liu and Hauskrecht (2016); She et al. (2018) to learn stable dynamical systems is something we would like to explore in future work.

Finally, non-linear extensions of the latent dynamics are also to be considered. This is because the mechanism implied by our model is not very useful in reality, as it cannot keep a memory of the inputs. As we mentioned before, a non-linear model would be able to implement a more complicated portrait of attractor dynamics which allows to preserve the input effects in different regions of state space, as in the non-linear RNN proposed in Mante et al. (2013). Furthermore, we could then capture the data from the two contexts using a single model and importantly, explicitly implement the contextual switch. This could also be achieved using a switching LDS (HSLDS) (Petreska et al., 2011) which approximates non-linear dynamics with a piecewise linear system. It would be very interesting to test how many potential linear systems are required by this model to capture the data within each context and to compare the inferred systems across contexts.

We finish the discussion at this point, hoping that we were able to convey both the strengths and the weaknesses of the approach we took in this thesis. A final note to the readers –and to myself– is to always keep in mind that, as the statistician George Box said back in 1978....

> “All models are wrong but some are useful”

and that...

> “It would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations.”

A quote that brings us to an end and which summarizes what has been our humble goal during this thesis, that of providing a useful and insightful approximation to the neural mechanisms orchestrating the intricacies of our cognition.

Appendix A

Appendix

A.1 LDS equations

A.1.1 LDS M-step updates

The equations provided in this appendix are derived from the notes in Macke et al. (2015) and from other standard sources. We implemented all our code in MATLAB. The backbone of our LDS algorithm is based on some of the scripts provided by Macke et al. (2015) to fit linear dynamical systems with Poisson observations. We are also planning to make our code publicly available.

Observations

$$[C, \boldsymbol{d}]_{new} = \left(\sum_{k=1}^{K} \sum_{t=1}^{T} \begin{bmatrix} \boldsymbol{y}_t \left\langle \boldsymbol{x}_t^{\top} \right\rangle, & \boldsymbol{y}_t \end{bmatrix} \right) \left(\sum_{k=1}^{K} \sum_{t=1}^{T} \begin{bmatrix} \left\langle \boldsymbol{x}_t \boldsymbol{x}_t^{\top} \right\rangle & \left\langle \boldsymbol{x}_t \right\rangle \\ \left\langle \boldsymbol{x}_t^{\intercal} \right\rangle & 1 \end{bmatrix} \right)^{-1}$$

$$R = \frac{1}{KT} \left(\sum_{k=1}^{K} \sum_{t=1}^{T} \boldsymbol{y}_t \boldsymbol{y}_t^{\intercal} - [C, \boldsymbol{d}]_{new} \sum_{k=1}^{K} \sum_{t=1}^{T} [\left\langle \boldsymbol{x}_t \right\rangle ; 1] \boldsymbol{y}_t^{\intercal} \right)$$

We have omitted the trial index k in $\boldsymbol{x}_{k,t}$ and $\boldsymbol{y}_{k,t}$ for clarity.

Dynamics

$$[A, B]_{new} = \left(\sum_{k=1}^{K} \sum_{t=1}^{T} S_{t,t-1} \right) \left(\sum_{k=1}^{K} \sum_{t=1}^{T} S_{t-1,t-1} \right)^{-1}$$

with

$$[A, B]_{new}^{2:T} = \left(\sum_{k=1}^{K} \sum_{t=2}^{T} \begin{bmatrix} \left\langle \boldsymbol{x}_t \boldsymbol{x}_{t-1}^{\top} \right\rangle & \left\langle \boldsymbol{x}_t \right\rangle \boldsymbol{u}_t^{\intercal} \end{bmatrix} \right) \left(\sum_{k=1}^{K} \sum_{t=2}^{T} \begin{bmatrix} \left\langle \boldsymbol{x}_{t-1} \boldsymbol{x}_{t-1}^{\top} \right\rangle & \left\langle \boldsymbol{x}_{t-1} \right\rangle \boldsymbol{u}_t^{\intercal} \\ \boldsymbol{u}_t \left\langle \boldsymbol{x}_{t-1}^{\intercal} \right\rangle & \boldsymbol{u}_t \boldsymbol{u}_t^{\top} \end{bmatrix} \right)^{-1}$$

and

$$[A,B]^1_{new} = \left(\sum_{k=1}^{K} \begin{bmatrix} \langle \boldsymbol{x}_1 \rangle \, \boldsymbol{x}_0^\top & \langle \boldsymbol{x}_1 \rangle \, \boldsymbol{u}_1^\intercal \end{bmatrix}\right) \left(\sum_{k=1}^{K} \begin{bmatrix} \boldsymbol{x}_0 \boldsymbol{x}_0^\top & \boldsymbol{x}_0 \boldsymbol{u}_1^\intercal \\ \boldsymbol{u}_1 \boldsymbol{x}_0^\intercal & \boldsymbol{u}_1 \boldsymbol{u}_1^\top \end{bmatrix}\right)^{-1}$$

$$Q_{new} = \frac{1}{KT} \left(\sum_{k=1}^{K} \sum_{t=1}^{T} \left\langle \boldsymbol{x}_t \boldsymbol{x}_t^\top \right\rangle - [A,B]_{new} \sum_{k=1}^{K} \sum_{t=1}^{T} S_{t,t-1}^\intercal \right)$$

$$\boldsymbol{x}_{0,new} = \frac{1}{K} A_{new}^{-1} \left(\sum_{k=1}^{K} \langle \boldsymbol{x}_1 \rangle - B_{new} \sum_{k=1}^{K} \boldsymbol{u}_1 \right)$$

The state noise covariance at the first step Q_0 has been set to Q for mathematical convenience.

In case we wanted to update A and B independently,

$$A = \left(\sum_{k=1}^{K} \sum_{t=1}^{T} \left\langle \boldsymbol{x}_t \boldsymbol{x}_{t-1}^\top \right\rangle - B \boldsymbol{u}_t \left\langle \boldsymbol{x}_{t-1}^\top \right\rangle \right) \left(\sum_{k=1}^{K} \sum_{t=1}^{T} \left\langle \boldsymbol{x}_{t-1} \boldsymbol{x}_{t-1}^\top \right\rangle \right)^{-1}$$

$$B = \left(\sum_{k=1}^{K} \sum_{t=1}^{T} (\langle \boldsymbol{x}_t \rangle - A \langle \boldsymbol{x}_{t-1} \rangle) \boldsymbol{u}_t^\top \right) \left(\sum_{k=1}^{K} \sum_{t=1}^{T} \boldsymbol{u}_t \boldsymbol{u}_t^\top \right)^{-1} \tag{A.1}$$

This equation shows explicitly that inputs are computed based on the residuals of the expected state transitions. Precisely, B is estimated as a regression problem from the residuals to the inputs.

A.1.2 Learning input signals with constrained time courses.

We constrain the coherence inputs to share a common time series, which is then scaled by a parameter indicating the coherence value

$$\begin{cases} u_m^k(t) = T_m^{in}(t)\, m^k & if \quad \text{trial } k \text{ motion coh} > 0 \\ u_m^k(t) = T_m^{out}(t)\, m^k & if \quad \text{trial } k \text{ motion coh} < 0 \end{cases}$$

where c^k and m^k indicate the color and motion coherences presented at trial k, which can take one out of 6 possible values

$$m = [m^1, m^2, m^3, m^4, m^5, m^6]$$
$$c = [c^1, c^2, c^3, c^4, c^5, c^6]$$

The update equations for these parameters are given by

$$T_t^{c_{in}} = \left(\boldsymbol{b}_c^\intercal Q^{-1} \boldsymbol{b}_c\right)^{-1} \left(\sum_{k \subset c_{in}}^{K_{c_{in}}} (c^k)^2 \right)^{-1} \boldsymbol{b}_c^\intercal Q^{-1} \sum_{k \subset c_{in}}^{K_{c_{in}}} c^k \left([\boldsymbol{x}_{k,t}^k - A \boldsymbol{x}_{t-1}^k] - B_{\neg c} \boldsymbol{u}_{\neg c,t}^k \right)$$

$$c_1 = \left(\boldsymbol{b}_c^\intercal Q^{-1}\boldsymbol{b}_c\right)^{-1}\left(\sum_{t=1}^{T}\sum_{k\subset c_1}^{K_{c_1}}(T_t^{c_{in/out}})^2\right)^{-1}$$
$$\boldsymbol{b}_c^\intercal Q^{-1}\sum_{t=1}^{T}\sum_{k\subset c_1}^{K_{c_1}} T_t^{c_{in/out}}\left([\boldsymbol{x}_{k,t}^k - A\boldsymbol{x}_{t-1}^k] - B_{\neg c}\boldsymbol{u}_{\neg c,t}^k\right) \tag{A.2}$$

and similarly for the rest of the coherence levels and for the motion inputs. Here $\neg c$ indicates that the color dimension c has been excluded.

A.2 Eigenmodes decomposition of the dynamics

We can rewrite the dynamics of the dynamical system so that it is expressed in its eigenmodes basis. For that, we construct a new state vector $\boldsymbol{\alpha}(t)$ which is expressed in the basis of the left eigenvectors of the system, given by the rows of the matrix $L = R^{-1}$

$$A = R\Lambda R^{-1} = R\Lambda L$$

where R contains the right eigenvectors in its columns, so that $LR = I$. In our case, the transition matrix A is non-normal, so the left and right eigenvectors are distinct and non-orthogonal among themselves. We can thus rewrite the dynamics equation as

$$\begin{aligned}\boldsymbol{x}(t) &= A\boldsymbol{x}(t-1) + B\boldsymbol{u}(t)\\ \boldsymbol{x}(t) &= (R\Lambda L)\boldsymbol{x}(t-1) + B\boldsymbol{u}(t)\end{aligned}$$

$$\boldsymbol{\alpha}(t) = L\boldsymbol{x}(t)$$

$$\begin{aligned}L\boldsymbol{x}(t) &= \Lambda(L\boldsymbol{x}(t-1)) + LB\boldsymbol{u}(t)\\ \boldsymbol{\alpha}(t) &= \Lambda\boldsymbol{\alpha}(t-1) + LB\boldsymbol{u}(t)\end{aligned}$$

In this new basis, as Λ is a diagonal matrix, the dynamics of the new state vector components α_h are decoupled from each other. The new components are the eigenmodes of the system. Therefore, we can treat the evolution of each eigenmode independently

$$\alpha_h(t) = \lambda_h\alpha_h(t-1) + \boldsymbol{l}_h^\intercal B\boldsymbol{u}(t)$$

Note that in our formalism the dynamics matrix acts already on the initial state

$$\boldsymbol{x}(1) = A\boldsymbol{x}(0) + B\boldsymbol{u}(1)$$

In this way, we can express the independent modes as a function of the initial state and inputs history

$$\boldsymbol{\alpha}(t) = \Lambda^t L\boldsymbol{x}(0) + \sum_{t'=1}^{t} \Lambda^{t-t'} LB\boldsymbol{u}(l)$$

We can do the same for each independent component $\alpha_h(t)$

$$\alpha_h(t) = \boldsymbol{l}_h^\intercal \boldsymbol{x}(t)$$

$$\begin{aligned}
\alpha_h(t) &= \lambda_h \boldsymbol{l}_h^\intercal \boldsymbol{x}(t-1) + \boldsymbol{l}_h^\intercal B\boldsymbol{u}(t) \\
\alpha_h(t) &= \lambda_h^t \boldsymbol{l}_h^\intercal \boldsymbol{x}(0) + \sum_{t'=1}^{t} \lambda_h^{t-t'} \boldsymbol{l}_h^\intercal B\boldsymbol{u}(l)
\end{aligned}$$

Each component has a time constant specified by the eigenvalue λ_h, as the evolution in time is given by λ_h^t.

Finally, if we want to express the original state vector $\boldsymbol{x}$ as a function of the eigenmodes, we can do it as follows

$$\boldsymbol{x}(t) = R\Lambda(L\boldsymbol{x}(t-1)) + B\boldsymbol{u}(t)$$

$$\boldsymbol{x}(t) = \sum_h \lambda_h \alpha_h(t-1)\boldsymbol{r}_h + B\boldsymbol{u}(t)$$

Therefore,

$$\boldsymbol{x}(t) = \sum_h \lambda_h^t \boldsymbol{l}_h^\intercal \boldsymbol{x}(0)\boldsymbol{r}_h + \sum_{t'=1}^{t-1} \lambda_h^{t-t'} \boldsymbol{l}_h^\intercal B\boldsymbol{u}(l)\boldsymbol{r}_h + B\boldsymbol{u}(t)$$

Finally, in order to bring the dynamics into the output space we use

$$\begin{aligned}
\boldsymbol{y}(t) &= C\boldsymbol{x}(t) + \boldsymbol{d} \\
\boldsymbol{y}(t) &= CL^{-1}L\boldsymbol{x}(t) + \boldsymbol{d} \\
\boldsymbol{y}(t) &= CL^{-1}(L\boldsymbol{x}(t)) + \boldsymbol{d}
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{y}(t) &= CR\boldsymbol{\alpha}(t) + \boldsymbol{d} \\
\boldsymbol{y}(t) &= C\sum_h \alpha_h(t)\boldsymbol{r}_h + \boldsymbol{d}
\end{aligned}$$

Note that when the eigenspectrum of A contains imaginary eigenvalues, as it is in our case, the associated independent components $\alpha_h(t)$ are also complex numbers, which makes the dynamics slightly more complicated to interpret. The final expression in that case is

$$\boldsymbol{y}(t) = C \sum_{h-h^\dagger,\, img} (\alpha_h(t)\boldsymbol{r}^h + \alpha_h^\dagger(t)\boldsymbol{r}^{h\dagger}) + C \sum_{h,\, real} \alpha_h(t)\boldsymbol{r}_h + \boldsymbol{d}$$

One can prove that the way the complex roots contribute to the dynamics is given by the real and imaginary parts of each corresponding conjugate pair. For each pair of complex conjugate roots, two complementary real solutions exist, which are given by the sum and difference modes $\alpha_{h^\pm}(t)$

$$\begin{aligned} \alpha_{h^+}(t) &= \frac{1}{2}(\alpha_h(t) + \alpha_h^\dagger(t)) \\ &= \Re\{\alpha_h(t)\} \\ \alpha_{h^-}(t) &= \frac{1}{2i}(\alpha_h(t) - \alpha_h^\dagger(t)) \\ &= \Im\{\alpha_h(t)\} \end{aligned}$$

For instance, given a two dimensional system with a complex conjugate pair of eigenvalues λ_h and $\lambda_h^\dagger$:

$$\begin{aligned} \boldsymbol{y}(t) &= C(\alpha_h(t)\boldsymbol{r}_h + \alpha_h^\dagger(t)\boldsymbol{r}_{h\dagger}) + \boldsymbol{d} \\ &= 2C(\Re\{\alpha_h(t)\}\Re\{\boldsymbol{r}_h\} - \Im\{\alpha_h(t)\}\Im\{\boldsymbol{r}_h\}) + \boldsymbol{d} \\ &= 2C(\alpha_{h^+}(t)\Re\{\boldsymbol{r}_h\} - \alpha_{h^-}(t)\Im\{\boldsymbol{r}_h\}) + \boldsymbol{d} \end{aligned}$$

as

$$\begin{aligned} \alpha_h(t)\boldsymbol{r}_h + \alpha_h^\dagger(t)\boldsymbol{r}_{h\dagger} &= (\Re\{\alpha_h(t)\} + i\Im\{\alpha_h(t)\})(\Re\{\boldsymbol{r}_h\} + i\Im\{\boldsymbol{r}_h\}) \\ &\quad +(\Re\{\alpha_h(t)\} - i\Im\{\alpha_h(t)\})(\Re\{\boldsymbol{r}_h\} - i\Im\{\boldsymbol{r}_h\}) \\ &= 2\Re\{\alpha_h(t)\}\Re\{\boldsymbol{r}_h\} - 2\Im\{\alpha_h(t)\}\Im\{\boldsymbol{r}_h\}\} \end{aligned}$$

For a given neuron,

$$\begin{aligned} y_n(t) &= 2\alpha_{h^+}(t)\, C_{n,:}\Re\{\boldsymbol{r}_h\} - 2\alpha_{h^-}(t)\, C_{n,:}\Im\{\boldsymbol{r}_h\} + d_n \\ &= 2\alpha_{h^+}(t)w_{nh^+} - 2\alpha_{h^-}(t)w_{nh^-} + d_n \end{aligned}$$

Therefore, the contribution from each complex conjugate pair of eigenmodes to the dynamics of each neuron n is weighted by the “sum and difference coefficients”

$$\begin{aligned} w_{nh^+} &= C_{n,:}\,\Re\{\boldsymbol{r}_h\} \\ w_{nh^-} &= C_{n,:}\,\Im\{\boldsymbol{r}_h\} \end{aligned}$$

and the contribution from each real eigenmode h is weighted by the coefficients

$$w_{nh} = C_{n,:}\,\boldsymbol{r}_h$$

In this way, we can break down the dynamics of each neuron as

$$y_n(t) = \sum_{h,\,real} w_{nh}\alpha_h(t) + \sum_{h-h^\dagger,\,img} 2(w_{nh^+}\alpha_{h^+}(t) - w_{nh^-}\alpha_{h^-}(t)) + d_n$$

Through the coefficients w_{nh}, we can exactly compute how much each mode contributes to the dynamics of each neuron.

We can also understand each neuron's dynamics in terms of the amplitude and phase of the complex conjugate pairs

$$y_n(t) = \sum_{h,\,real} w_{nh}\alpha_h(t) + \sum_{h-h^\dagger,\,img} 2w_{nh^{amp}}|\alpha_h(t)|\,cos\,(\phi_h(t) + w_{nh^{pha}}) + d_n$$

The estimation of the weights is a bit more involved and we decided not to show it in here.

Schur modes decomposition of the dynamics

We can repeat the same analysis using the Schur decomposition $A = UTU^{-1}$, were T is an upper triangular matrix and U, which contains the Schur modes, is unitary $U^{-1} = U^\dagger$. Therefore $A = UTU^\dagger$. In this case, as A is real, we have two possible decompositions, one with complex Schur vectors and another with real ones. We choose a real decomposition, which provides a U with real-valued entries. Therefore, $U^{-1} = U^\dagger = U^\intercal$. The matrix T is quasitriangular and has the real eigenvalues on the diagonal and the complex eigenvalues in 2-by-2 blocks on the diagonal.

$$\begin{aligned} \boldsymbol{x}(t) &= A\boldsymbol{x}(t-1) + B\boldsymbol{u} \\ \boldsymbol{x}(t) &= (UTU^\intercal)\boldsymbol{x}(t-1) + B\boldsymbol{u}(t) \\ (U^\intercal\boldsymbol{x}(t)) &= T(U^\intercal\boldsymbol{x}(t-1)) + U^\intercal B\boldsymbol{u}(t) \end{aligned}$$

$$\boldsymbol{\alpha}^s(t) = T\boldsymbol{\alpha}^s(t) + U^\intercal B\boldsymbol{u}(t)$$

The advantage of this decomposition is that the Schur vectors form an orthogonal basis. In this case, however, we do not obtain a set of uncoupled equations for the evolution of each Schur mode α_h^s, like we did using the eigenmodes decomposition. This is because the evolution of the modes is no longer independent. There's cross-talk among them, some mixing happening at each time step due to the action of the upper triangular matrix T, which is not diagonal like Λ. Nevertheless, we can still look at how each entry of the alpha vector changes in time, which tell us how each Schur mode will evolve in time.

$$\boldsymbol{\alpha}^s(t) = T^t U^\intercal \boldsymbol{x}(0) + \sum_{t'=1}^{t} T^{t-t'} U^\intercal B \boldsymbol{u}(t)$$

Appendix B

Supplementary information

B.1 Monkey A

B.1.1 LDS and JF models performance and solution

B.1.1.1 JF model without Tin-Tout extra flexibility

The JF model has slightly better performance in CV in this case (figure B.1 and table B.1).

B.1.1.2 Non cross-validated model generated trajectories

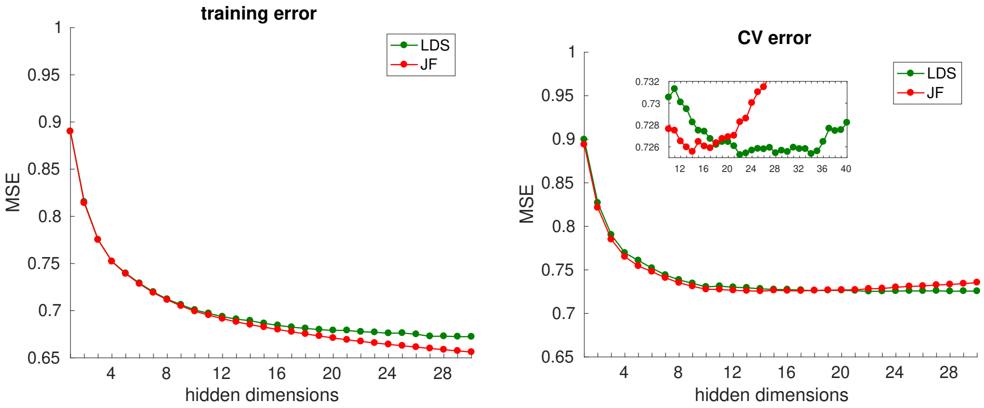
Model generated trajectories on training data are almost identical to the cross-validated trajectories, for both the LDS and the JF models (fig. B.2).

B.1.1.3 Training error patterns, FRs variability and CV error patterns across neurons.

The same error pattern across conditions is obtained for both the training and the CV errors (fig. B.3). We therefore asked, are FRs more variable for weak coherence conditions? (see main text). Figure B.4 shows that this is not the case. The pattern of FR variability in these figures can in fact be entirely attributed to Poisson statistics, as the same trends are found for the FRs (fig. B.5). Finally, we found that the per-condition CV error pattern is distributed across neurons (fig. B.6). This means that the shape of the averaged error pattern arises from the distribution of the error across the whole population.

Model	min CV error	H
LDS	0.7253	22
JF	0.7255	14

Table B.1: LDS and JF models minimum cross-validation error and corresponding hidden dimensionality H (or rank degree) for which it is achieved.



(a) (b)

Figure B.1: LDS and JF models performance as a function of hidden dimensionality (rank degree for the case of the JF model). The JF model in this case does not incorporate extra parameters to account for different Tin and Tout types of evidence a) Training mean squared error b) cross-validation mean squared error. Squared errors are in units of variance, given that the FR responses were z-scored.

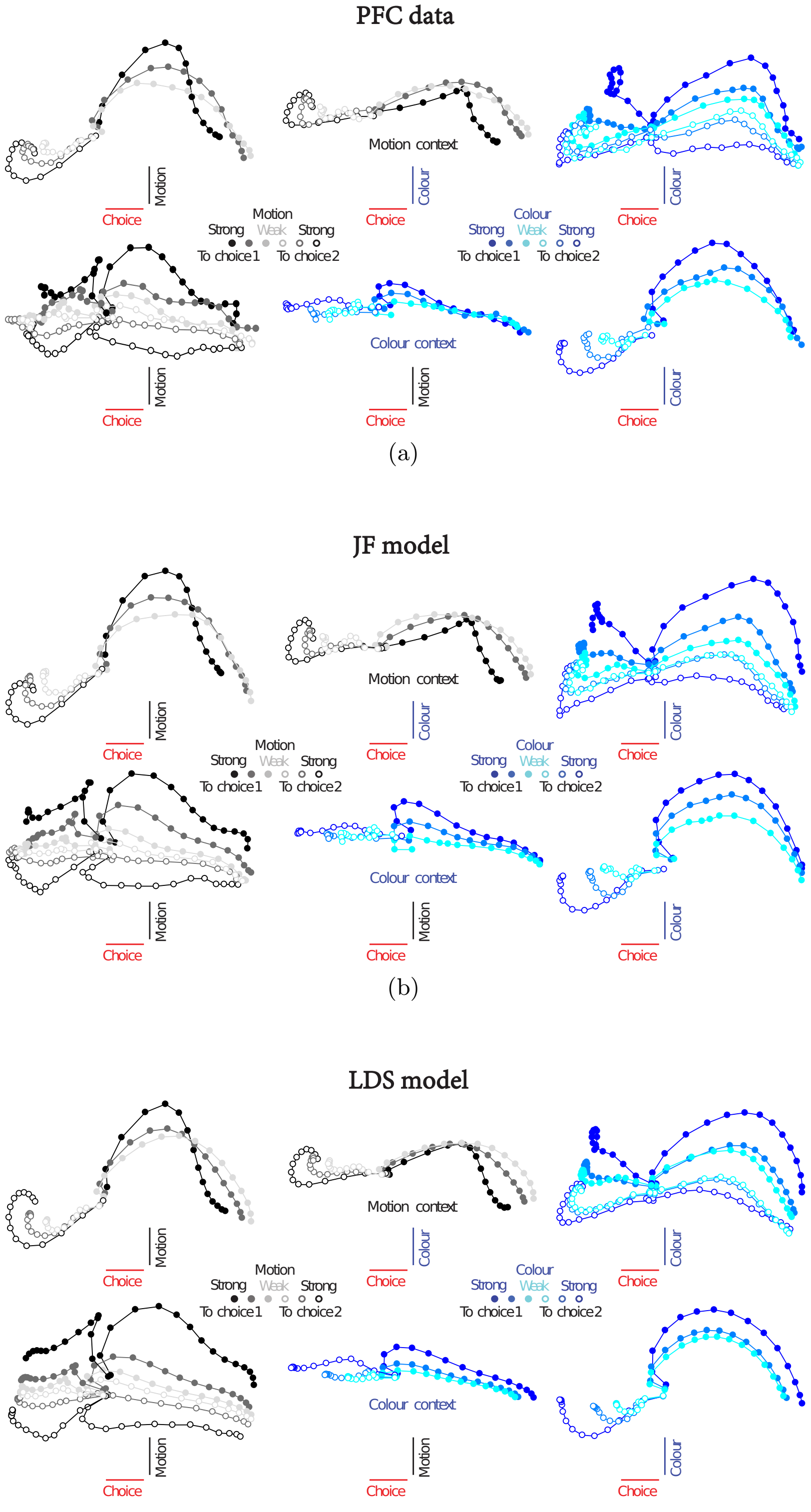
B.1.1.4 Variance explained and PCA

We already noted that little variance is explained by the two models (about 30-35% in training for H=30, fig 3.5), despite being able to capture well the overall structure of the trajectories and individual PSTHs. For comparison, PCA with 30 PCs gets about 45% of the variance (figure B.7). On smoothed data, the amount goes up to 60%. For cross-validated data, the LDS performance flattens out as a function of hidden dimensionalities, explaining less than 30% of the variance. Note that PCA does not incorporate any restriction in the way the variance is allowed to be captured, unlike the JF and the LDS models. This leads to the characteristic monotonic increase in variance explained as dimensions (number of PCs) are added.

It is important to mention all this in case that the meticulous reader notices a discrepancy with respect to the Mante et al. (2013) study. In the paper's supplementary information (extended data figure 4) it is implied that with only 20 PCs about 80% of the variance of the data can be captured. The reason of this discrepancy is that the data being used in this analysis is the same data used to plot the trajectories, which has color and motion conditions averaged out. Therefore, this reduced data set contains 18 "doubly-averaged" conditions, whereas we are fitting the model to all the 36 averaged experimental conditions.

B.1.1.5 LDS likelihood and shuffled data test

In the main text we hinted at some of the differences between the JF and the LDS models. The LDS is a probabilistic model, optimized via the EM algorithm in which latent variables and parameters are jointly estimated in alternating steps until convergence, leading to the maximum likelihood parameter estimates (see 2.3). The parameters



PFC data

JF model

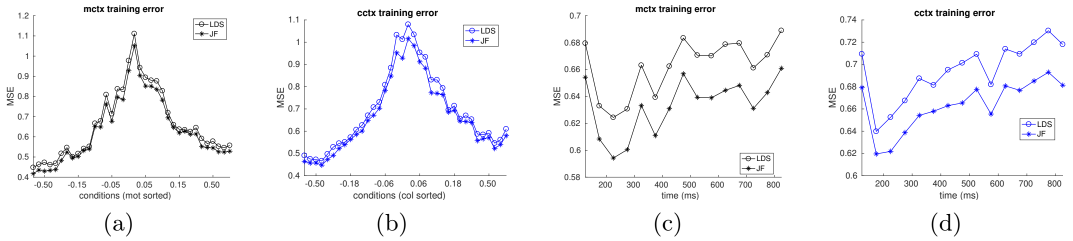
LDS model

(a)

(b)

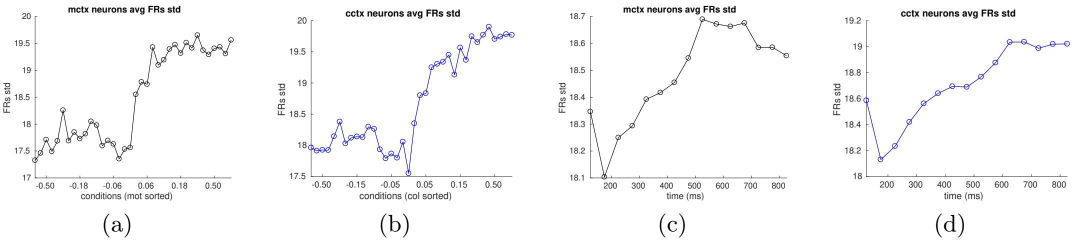
(c)

Figure B.2: Data and model generated trajectories in the task-relevant subspace identified by regression Mante et al. (2013) a) PFC data b) JF model c) LDS model. Same plotting conventions as in figure 3.2.



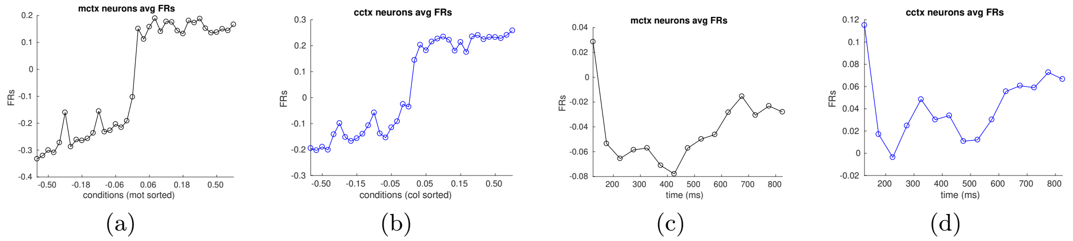
(a) (b) (c) (d)

Figure B.3: LDS and JF models minimum training error (MSE) (a,b) CV error for the 36 different conditions, grouped by the coherence value of the relevant inputs and sorted in ascending order. Each group contains 6 conditions corresponding to the 6 possible irrelevant coherence values. These are also sorted in ascending order (c,d) CV error across time. Motion context (left), color context (right)



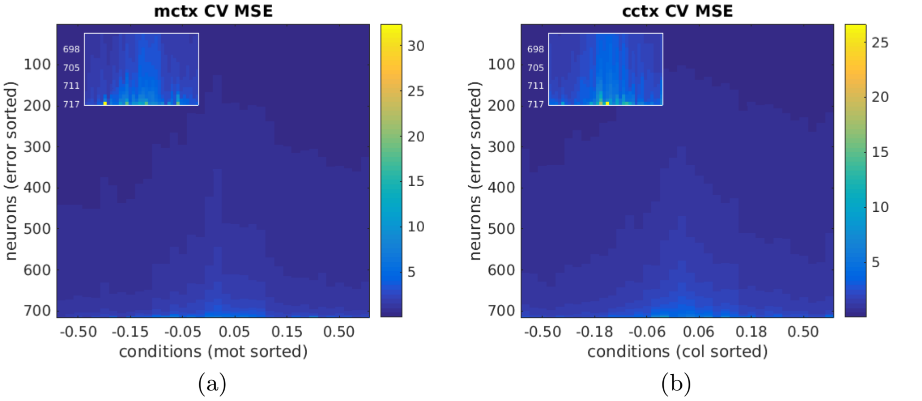
(a) (b) (c) (d)

Figure B.4: FRs standard deviation per time and condition. Same convention as in previous figures.



(a) (b) (c) (d)

Figure B.5: FRs (z-scored) per time and condition. Same convention as in previous figures.



(a) (b)

Figure B.6: LDS model CV minimum error (see table 3.1) per neuron and condition for a) the motion context b) the color context. In the x-axis, the same coherence groupings as in previous figures have been applied. In the y-axis, neurons have been sorted by the magnitude of the error –in ascending order from top to bottom. This is done independently for each condition (i.e. for each vertical slice). The magnitude of the error is indicated by the color bar.

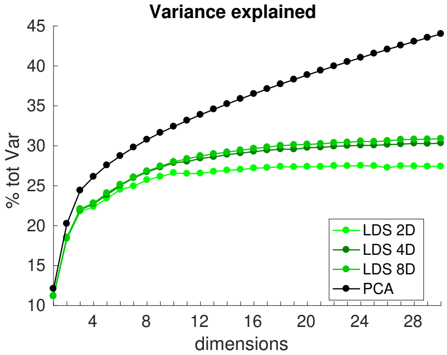


Figure B.7: Percentage of variance explained by the LDS, as a function of hidden dimensions and for different input dimensionalities. As a reference, in black, we show the amount of variance explained using PCA on the same data set.

include the innovations and observations noise covariance matrices. The noiseless LDS prior, therefore, is specified taking only the means of the distributions inferred during the optimization procedure. The JF model, on the contrary, is not endowed with any noise model and the way it is optimized is via least squares minimization on the data reconstruction. Put it simply, the JF model is adjusted to capture the mean of the data, whereas the LDS captures both the mean and the variance. In further work we plan to equalize the two optimization procedures, either by considering a model with dynamical constraints, but without a noise model, or by specifying a noise distribution for the JF model.

The LDS likelihood is computed for the two contexts jointly, as some of the parameters are shared across contexts and are therefore jointly optimized. We found that the per-condition full model cross-validated log-likelihood follows the exact same pattern as the per-condition CV error on the noiseless LDS prior (figure B.8), despite the fact that the likelihood is computed for the whole model, considering the noise terms and the inferred posterior. Note that the shape of the per-condition error curve in this figure is different from the ones shown in figure 3.7. This is because this curve reflects the joint error across contexts, so it has been computed by averaging the errors across contexts, without the sorting of the conditions used in the previous figures.

It is important to mention some minor problems we have encountered during the model optimization stages, in case they may impact the results presented in here. We observe that, sometimes, numerical instabilities arise in the fitting procedure which lead to a decrease in the likelihood, causing the EM iterations to be prematurely aborted. Such instabilities are prominent for high dimensionalities (over H=30, see fig B.9). We do not think that this issue affects the estimation of the minimum CV error, as for the models we fitted, the minimum is reached before, at around H=18-26. However, the fact that the CV error increases rather sharply at around H=35-40 could in fact be caused by such instabilities, and not because of overfitting.

Finally, we perform a control to test for the dynamical assumption in the data by assessing how the models perform on temporally shuffled data. As expected, the LDS

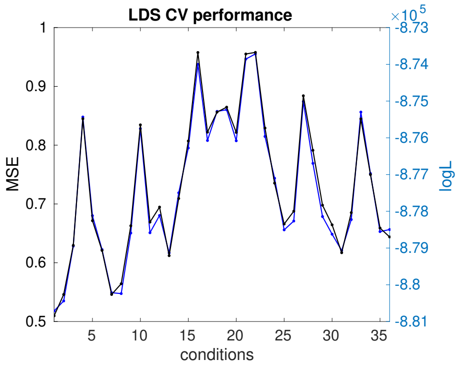


Figure B.8: Cross-validated minimum MSE for the noiseless LDS prior (black) (see table 3.1) and corresponding full model cross-validated log-likelihood (blue) for the 36 different conditions.

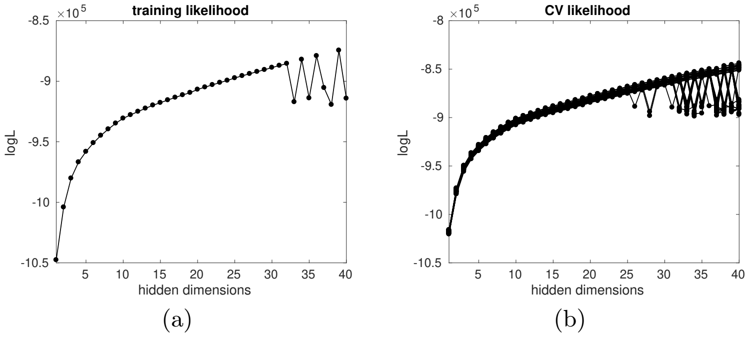
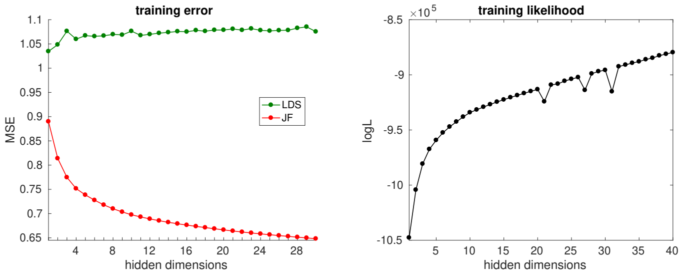


Figure B.9: LDS log-likelihood as a function of hidden dimensionality a) Training log-likelihood averaged across conditions b) Cross-validated log-likelihood, computed for each left-out condition.



(a) (b)

Figure B.10: LDS and JF models performance on time-shuffled data, as a function of hidden dimensionality a) training MSE b) LDS log-likelihood.

performs terribly, but the JF model is left unaffected (fig. B.10a). Note that the training MSE has a slight upwards trend, which obviously should not occur in training data as the number of parameters increase. However, we remind the reader that the objective we are optimizing in the case of the LDS is the log likelihood of the data, which takes into account the inferred posterior and noise models. The log likelihood does indeed steadily increase (see fig. B.10b) as a function of hidden dimensionality. What exactly causes the upward trend in the MSE curve is something we have not investigated further.

B.1.1.6 Comparing the LDS and the JF models solutions

In figure B.11 we show the solution found by the two models, without removing the condition independent effects from the motion and color input components. To further compare the solutions across models we perform the analysis shown in figure B.12.

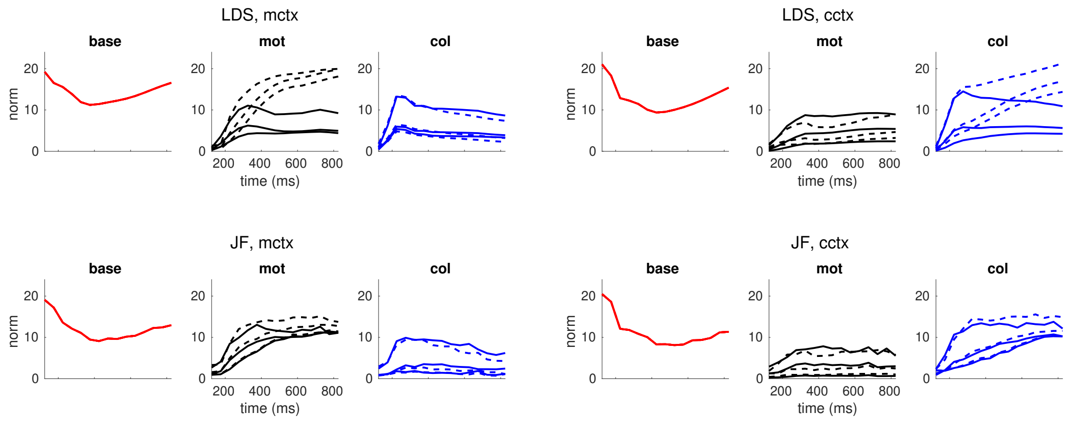
B.1.1.7 CV performance in TDR dimensions

How does each model perform, in CV, along the task-relevant TDR dimensions? We found no particular structure of the error across conditions along such dimensions, but there is trend in time in the case of the decision dimension (fig. B.13). This is because this dimensions captures the characteristic ramping up of FRs towards the end of the trial, so noise increases due to Poisson variability. The two models perform similarly, with the JF model performing slightly worse in the motion context.

B.1.2 LDS under the same contextual dynamics

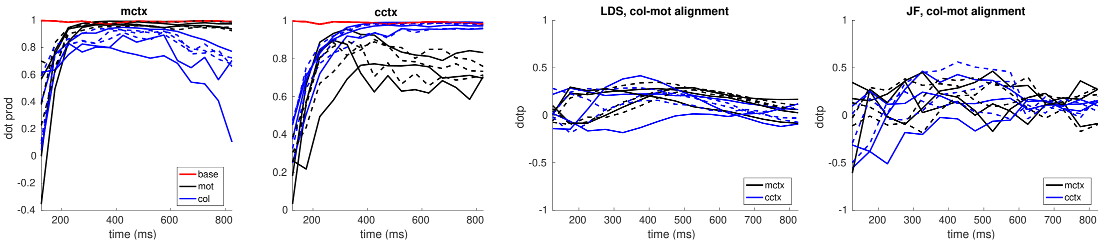
B.1.2.1 CV performance in TDR dimensions

Figure B.14 illustrates that the model with constrained dynamics across contexts is missing substantial variance along the decision dimension. Compare with previous figure.



(a) (b)

Figure B.11: Vector norms of the base, motion and color components at each point in time –without subtracting condition-independent effects to the input components– . For the LDS model (top) and the JF model (bottom), in the motion context (a) and the color context (b) The input components are computed for all the learned motion and color input values (in black and blue, 6 in each case), corresponding to the positive coherences (solid lines) and negative coherences (dashed lines). The base component (red) does not depend on the inputs –so there is only a single trace– and captures condition-independent variance.



(a) (b)

Figure B.12: a) LDS and JF models base, motion and color components directions correspondence in the motion and color contexts b) Alignment between the color and motion components within each context, for both the LDS and the JF models.

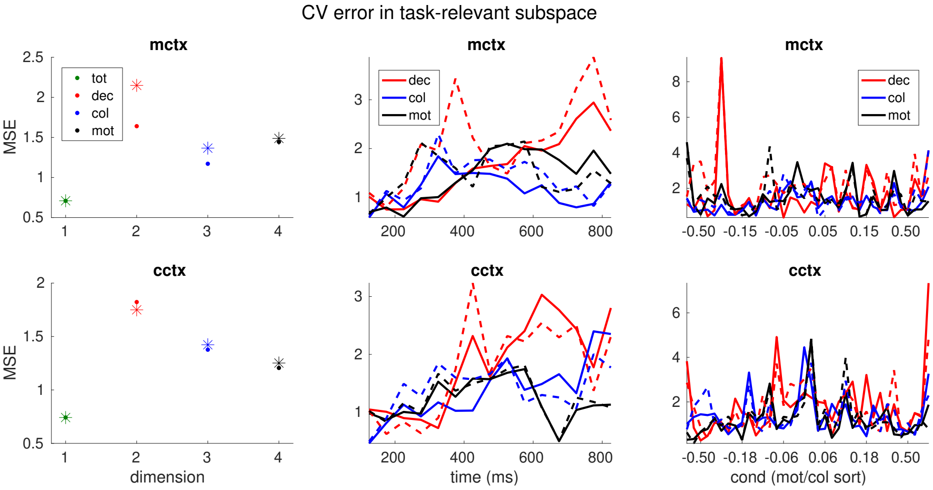


Figure B.13: LDS and JF models CV performance in the task-relevant subspace. First row, motion context. Second row, color context. We use dots/stars and filled/dashed lines for the LDS/JF models a) Total CV MSE (green) and MSE along the TDR decision, color and motion input dimensions b) CV error across time, for each dimension b) CV error across conditions, for each dimension. Same conventions as for previous figures apply.

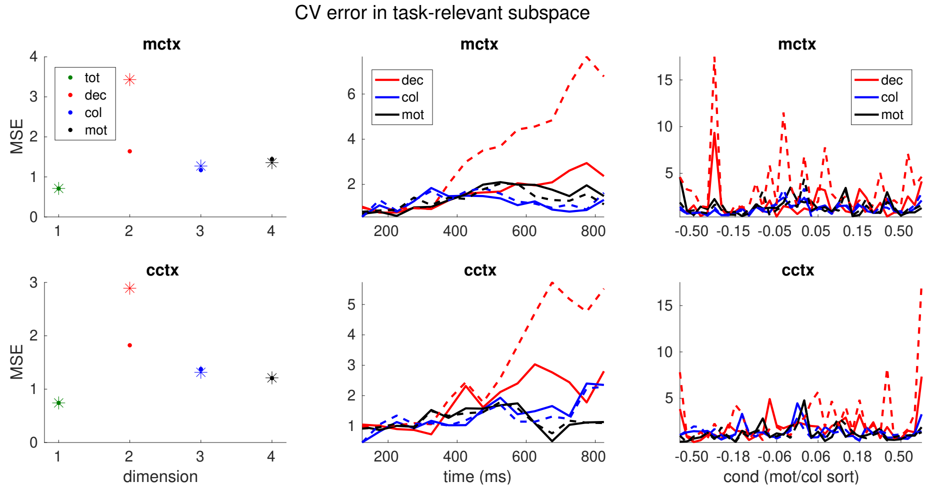


Figure B.14: LDS CV performance in the task-relevant subspace for constrained and unconstrained dynamics across contexts. First row, motion context. Second row, color context. We use dots/stars and filled/dashed lines for the unconstrained/constrained models a) Total CV MSE (green) and MSE along the TDR decision, color and motion input dimensions b) CV error across time, for each dimension b) CV error across conditions, for each dimension. Same conventions as for previous figures apply.

Model	min CV error	H
LDS $A^{(c)} \neq A^{(m)}$	0.7257	26
LDS $A^{(c)} = A^{(m)}$	0.7265	28

Table B.2: LDS minimum cross-validation error, for unconstrained (top) and constrained (bottom) dynamics and time courses across contexts, and corresponding hidden dimensionality H for which it is achieved.

B.1.2.2 Cross-validated trajectories

The model with the dynamics constrained to be the same across contexts, surprisingly, captures the data trajectories equally well (fig. B.15).

We also fitted a model with the input time courses constrained across contexts. We found that under this additional constraint the model was more stable, which explains why it is better in CV than the one with more flexibility in the inputs (see table 3.4). The corresponding trajectories are shown in figure B.16. The model we have used for the trajectories in figure B.16b also incorporates the constraint on the input's time courses (model 3 in table 3.2). Note that this model is able to capture the trajectories well too, and we obtain essentially the same picture as for the model with flexibility in the time courses across contexts (B.15b).

B.1.2.3 Dynamics, input biases and input directions

The LDS model with the transition matrix A constrained to be the same across contexts places the integration pattern in the inputs (figs. B.17 and B.18). Interestingly, when looking at the integrated inputs along the same input dimensions, that is, running the inputs through the dynamics, the model output manages to create a transient pattern, as can be seen in the trajectories along the input dimensions.

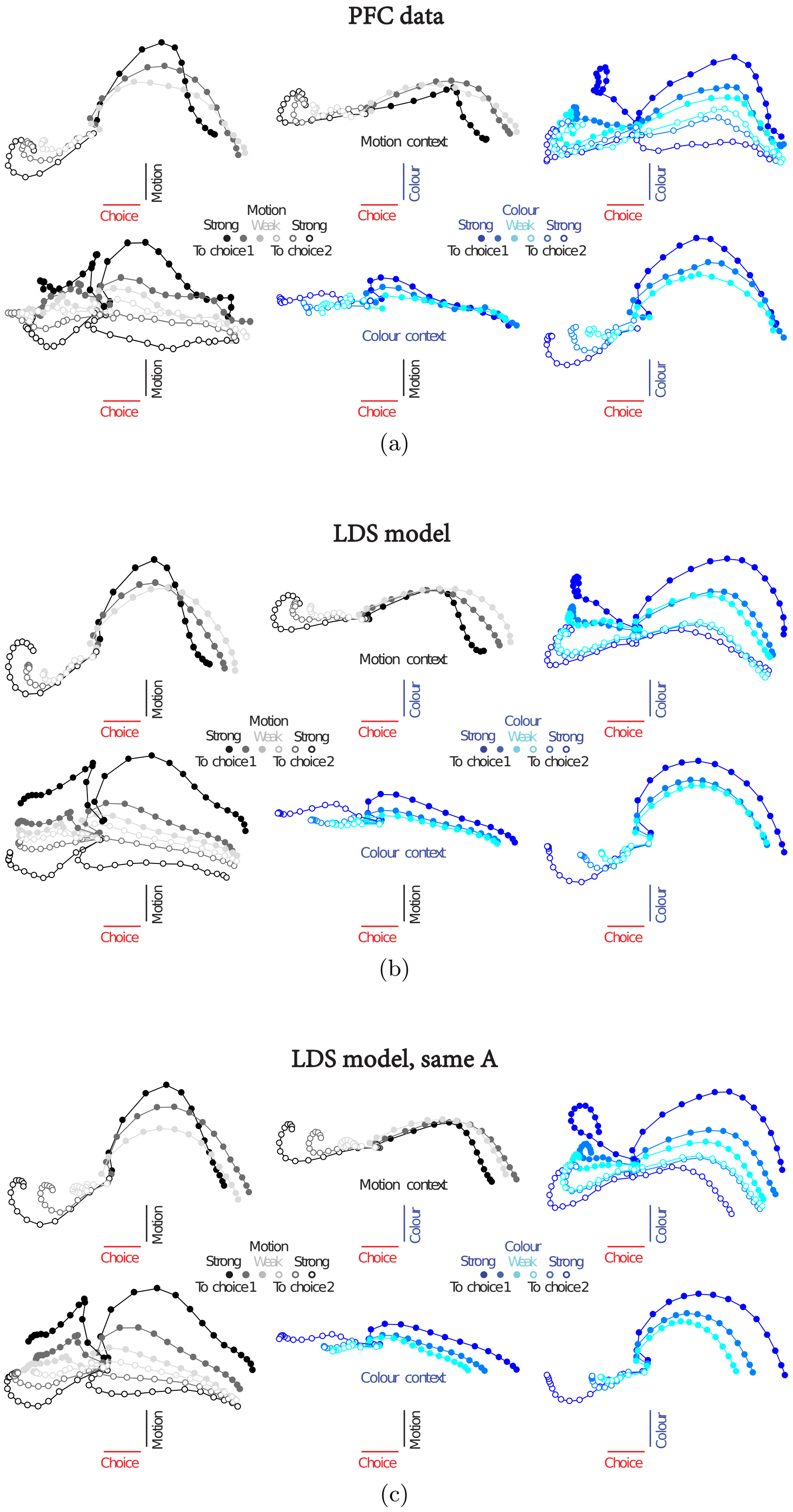
We found that, surprisingly, the inferred input vectors are learned to point in the same direction, regardless of context and the modality (B.19b). This direction lays in-between the TDR input dimensions and the decision axis (B.19c). We also found that this directions sharply aligns with the large negative eigenmode always found for the dynamics and barely projects to the rest of the modes.

B.1.3 LDS inputs and dynamics

B.1.3.1 Input biases

We found that the JF model recovers similar coherence-associated scalings as the LDS (fig. B.20).

A model with unconstrained input signals, for which a whole input time series is learned for each coherence (model 1 in table 3.2), converges to a similar pattern of inputs (see fig. B.21 and compare with fig. 3.14). In the models we fitted, inputs biases were initialized as a constant time series set to the coherence values or the coherence

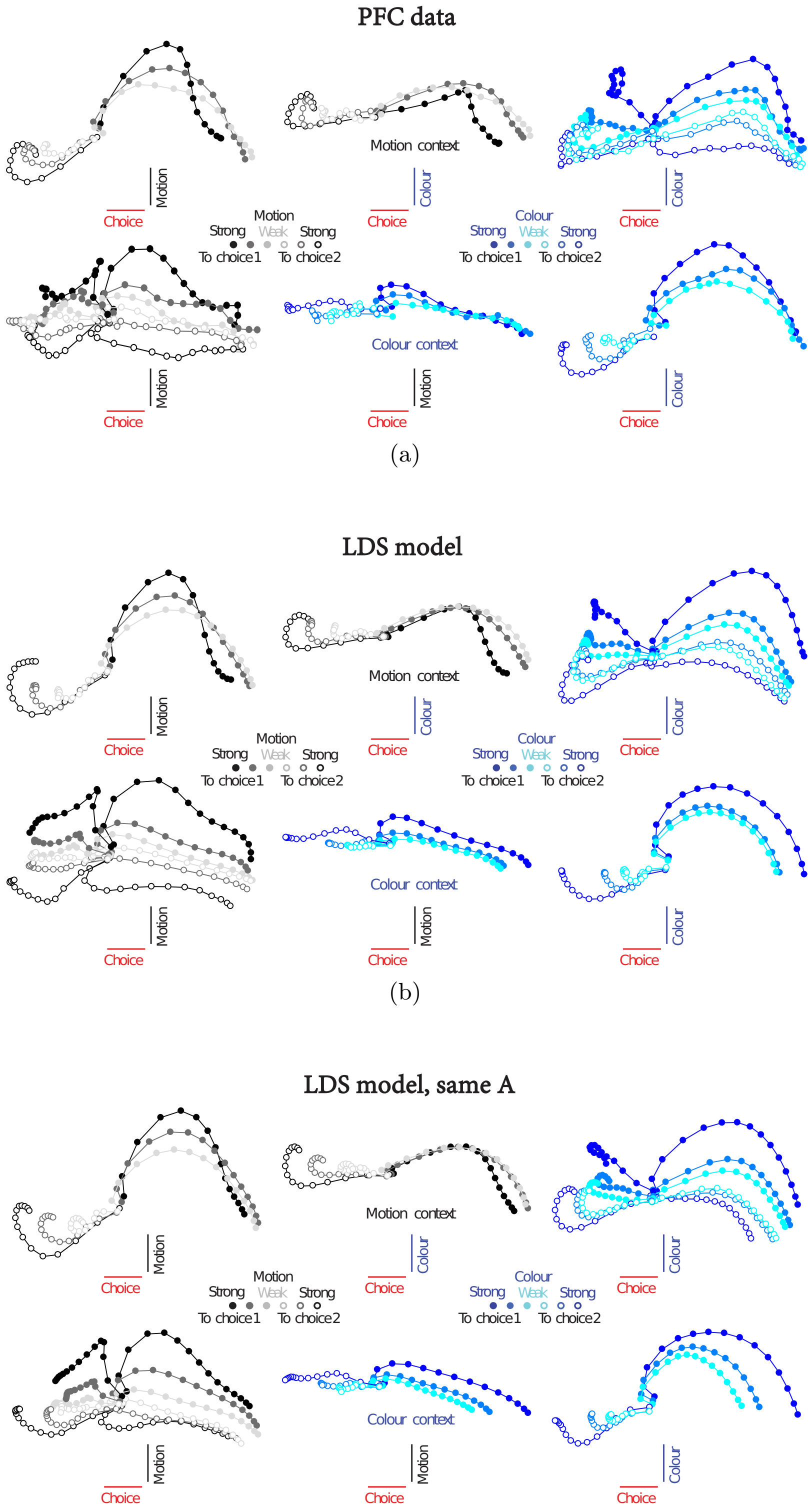


(a)

(b)

(c)

Figure B.15: Data and model generated cross-validated trajectories, leaving one condition out, in the task-relevant subspace a) data b) LDS model with flexible dynamics across contexts c) LDS model constrained with the same dynamics across contexts.



(a)

(b)

(c)

Figure B.16: Data and model generated cross-validated trajectories, leaving one condition out, in the task-relevant subspace a) data b) LDS model with flexible dynamics across contexts. c) LDS model constrained with the same dynamics across contexts. In b) and c) input time courses are also constrained to be the same across contexts, so that only the coherence values and the input directions are allowed to change across context.

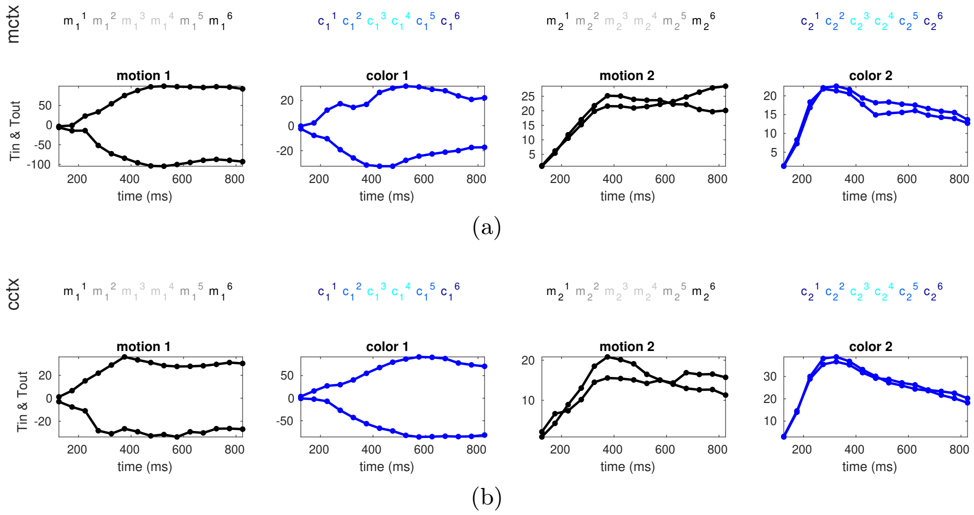


Figure B.17: LDS with constrained dynamics input signals time courses. Same conventions as for the figures in the main text.

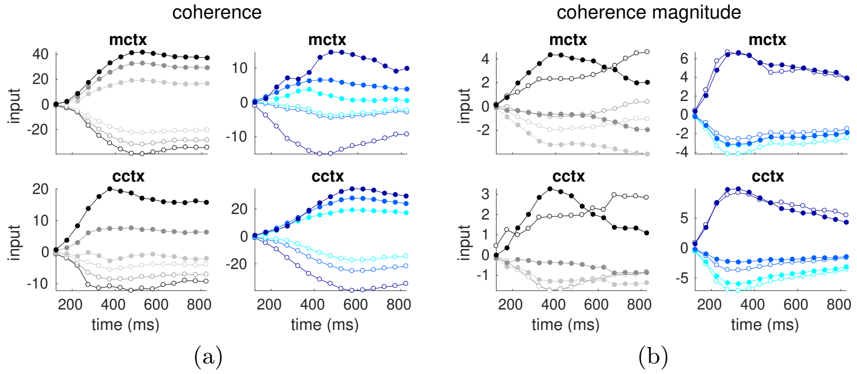
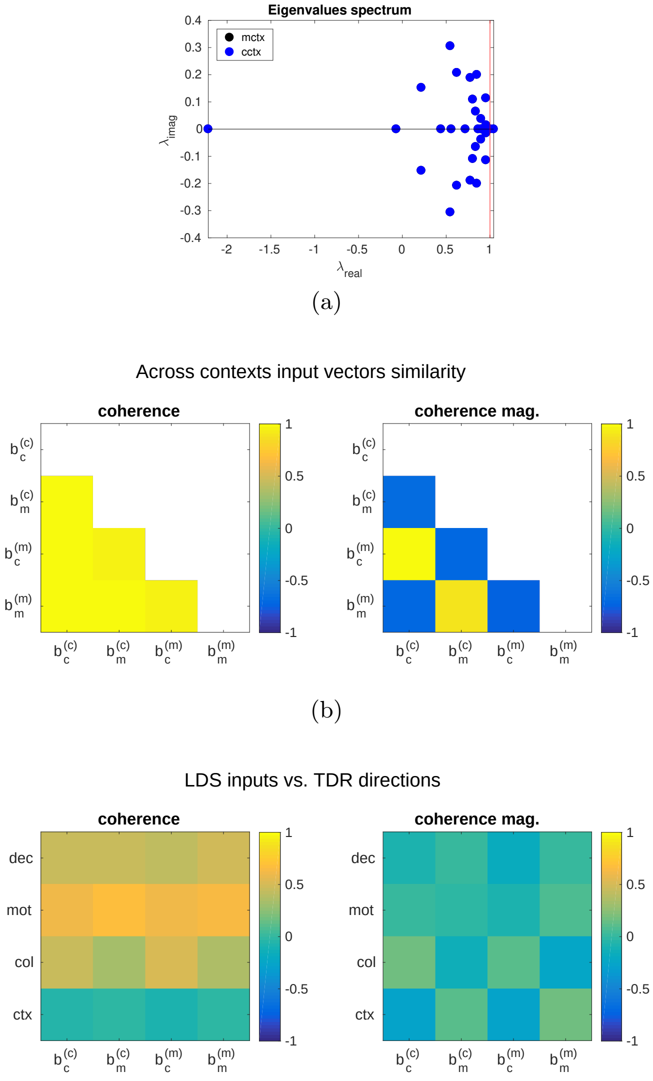


Figure B.18: LDS with constrained dynamics, input signals in the orthogonalised 2D input subspaces. Same conventions as for the figures in the main text.

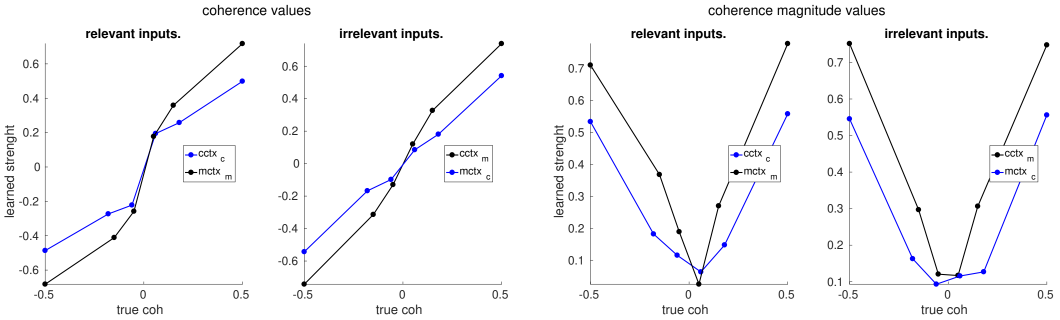


(a)

(b)

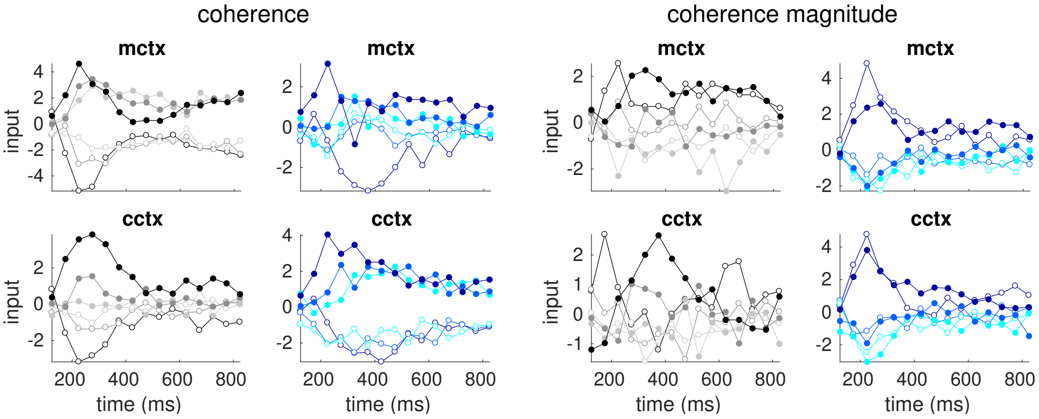
(c)

Figure B.19: a) eigenspectrum b) across contexts input similarity c) correspondence with TDR task-relevant dimensions.



(a) (b)

Figure B.20: JF model inferred coherence values for each of the four motion and color input dimensions in each context. Same conventions as for the figures in the main text.



(a) (b)

Figure B.21: Input signals in the orthogonalised 2D input subspaces for the LDS model with unconstrained input patterns. Same conventions as for the figures in the main text.

magnitude values. However, similar input patterns are learned even when initializing them at random.

B.1.3.2 Subspaces alignment

In order to compute the percentage of variance from one subspace contained in another subspace, we first estimate the input covariance matrix across times and conditions for each of the four subspaces B_c^{cx}, B_m^{cx}. Then, we calculate the amount of variance lost when projecting the full covariance matrix onto a different subspace. For instance, to estimate the amount of variance from the motion input subspace contained in the color input subspace –in a given context cx– we compute

$$\Sigma_m^{cx} = B_m^{cx} \,\langle U_m^{cx} (U_m^{cx})^\intercal \rangle\, B_m^{cx}$$

$$Var\,frac = \frac{Trace\,[B_c^{cx} \Sigma_m^{cx} B_c^{cx}]}{Trace\,[\Sigma_m^{cx}]}$$

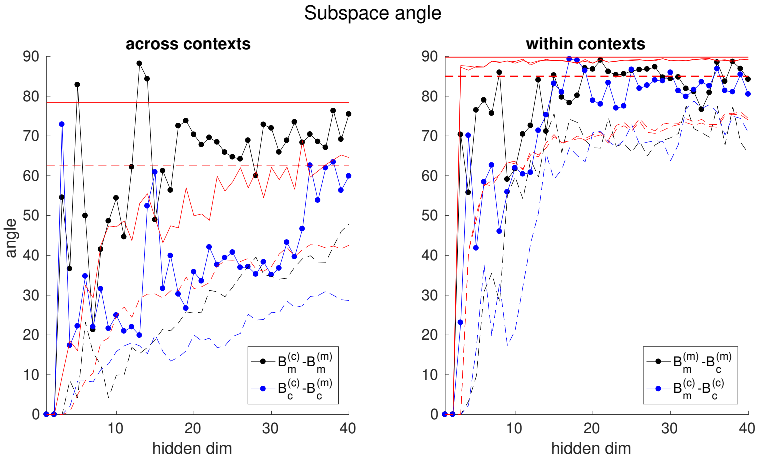


Figure B.22: Subspace angles within and across contexts, for different hidden dimensionalities. Solid lines designates maximum angles and dashed lines minimum angles. Red lines in the left plot indicate the 5th percentile of the null distribution for random angles. In the right plot they indicate the 95th percentile. The constant lines correspond to the null estimated on the whole data space, without projecting the covariance onto the hidden subspaces. In the right plot there are two red lines because a different null has been computed for each context.

where U^{cx}_{m} contains the estimated motion input time courses for all conditions.

In order to measure subspace angles (fig. B.22), we use the same algorithm as in the MATLAB function `subspace`. This function provides the largest angle between two subspaces, as it estimates the angle based on the largest singular value of the projection matrix of the two subspaces. In order to estimate the minimum angle, we instead take the minimum singular value.

B.1.3.3 Input directions constrained across contexts

When constraining the input directions to be the same across context, some variance is lost along the decision dimension (fig. B.23) but the trajectories all well captured (fig. B.24).

B.1.3.4 Dynamics

Coherence magnitude signals are present in the data along the coherence magnitude dimensions. For the case of motion in the color context, this signal is either very weak or the model is capturing noise, as figures 3.20b, B.30b and B.32b suggest.

Within both the relevant and irrelevant integrated subspaces, the orthogonal dimensions to the decision-like dimensions reflect coherence strength, but not sign (figures B.35 and B.36).

In the case of the color in the motion context, the presence of this integrated signal is consistent with the pattern of left eigenvalues projection, as the color magnitude input is strongly mapped into the slowest mode and hence, integrated. In the case of the motion in the color context, it is consistent with the pattern of learned inputs, which are more sustained (figure 3.14). This is an additional example of possible degeneracies in the

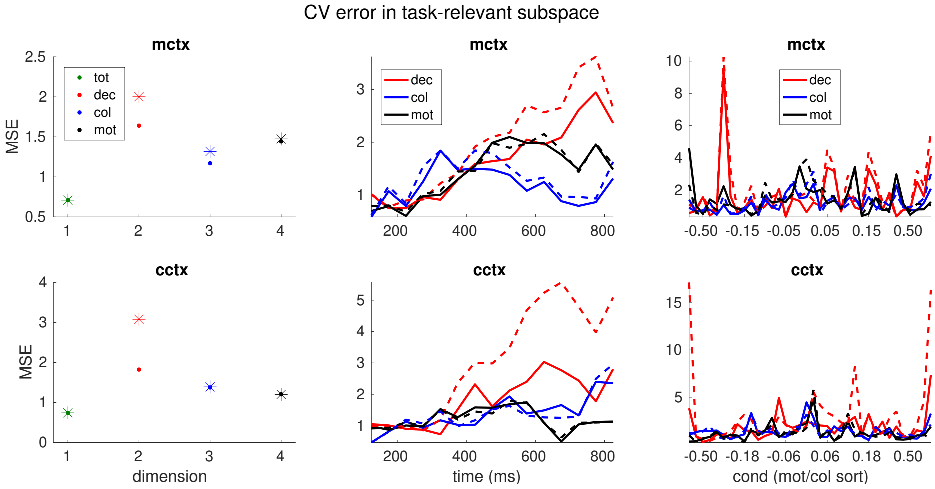
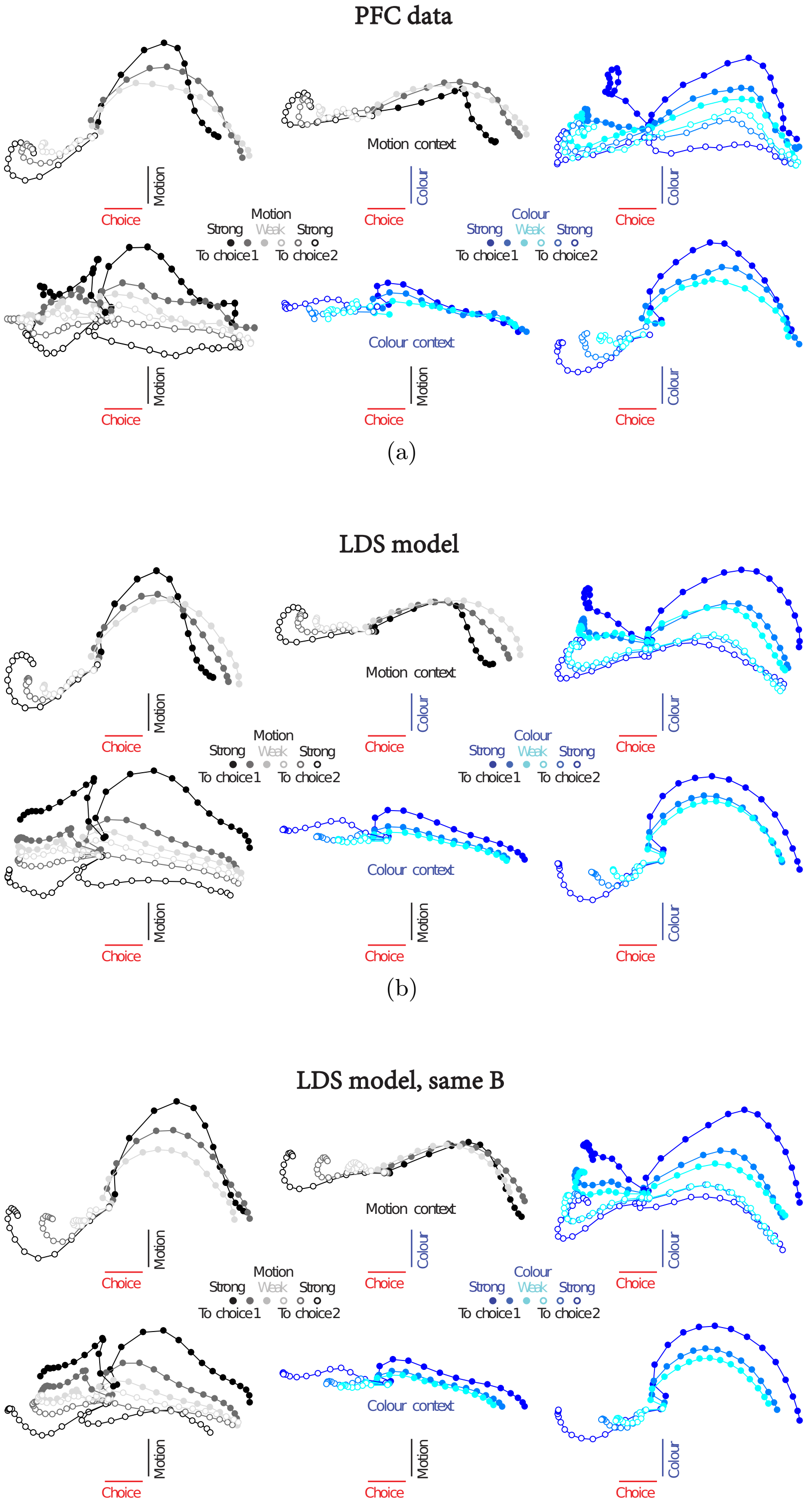


Figure B.23: LDS CV performance in the task-relevant subspace for constrained and unconstrained input directions across contexts. First row, motion context. Second row, color context. We use dots/stars and filled/dashed lines for the unconstrained/constrained models a) Total CV MSE (green) and MSE along the TDR decision, color and motion input dimensions b) CV error across time, for each dimension b) CV error across conditions, for each dimension. Same conventions as for previous figures apply.

model. For instance, we found that for some solutions the irrelevant color magnitude input pattern is learned to be more persistent after the initial pick. The projection pattern on the left eigenvalues is also different and the input is no longer amplified by the dynamics. The integrated color magnitude signal is in this case is then described via direct input modulation.

B.1.4 LDS single unit PSTHs

PSTHs for all conditions, without averaging out color/motion effects (fig. B.43).



(a)

(b)

(c)

Figure B.24: Data and model generated cross-validated trajectories, leaving one condition out, in the task-relevant subspace identified by regression Mante et al. (2013) a) data b) LDS model with flexible input directions across contexts c) LDS model with constrained input directions across contexts.

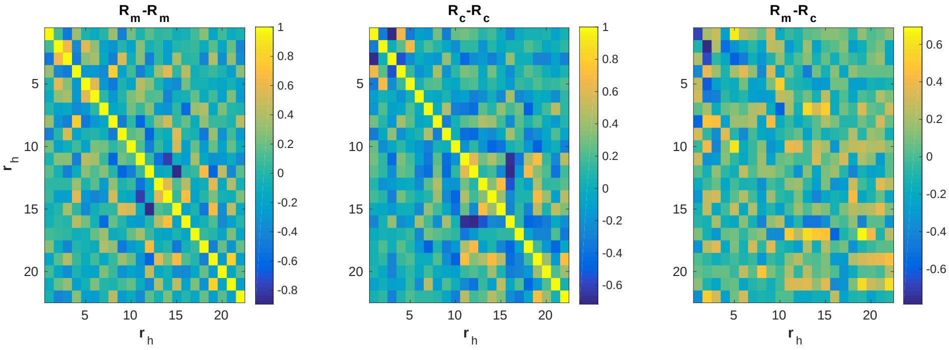


Figure B.25: Right eigenvectors alignment within and across contexts, sorted by increasing magnitude of the associated eigenvalues.

B.2 Monkey F

B.2.1 LDS and JF models performance and solution

B.2.1.1 Comparing the LDS and the JF model solutions

LDS and JF models base, motion and color components directions correspondence (fig. B.48).

B.2.2 LDS under the same contextual dynamics

B.2.2.1 CV performance in TDR dimensions

Unlike what is found for monkey A, in this case not much variance is lost along the decision dimension (fig. B.49).

B.2.3 LDS inputs and dynamics

B.2.3.1 Input biases

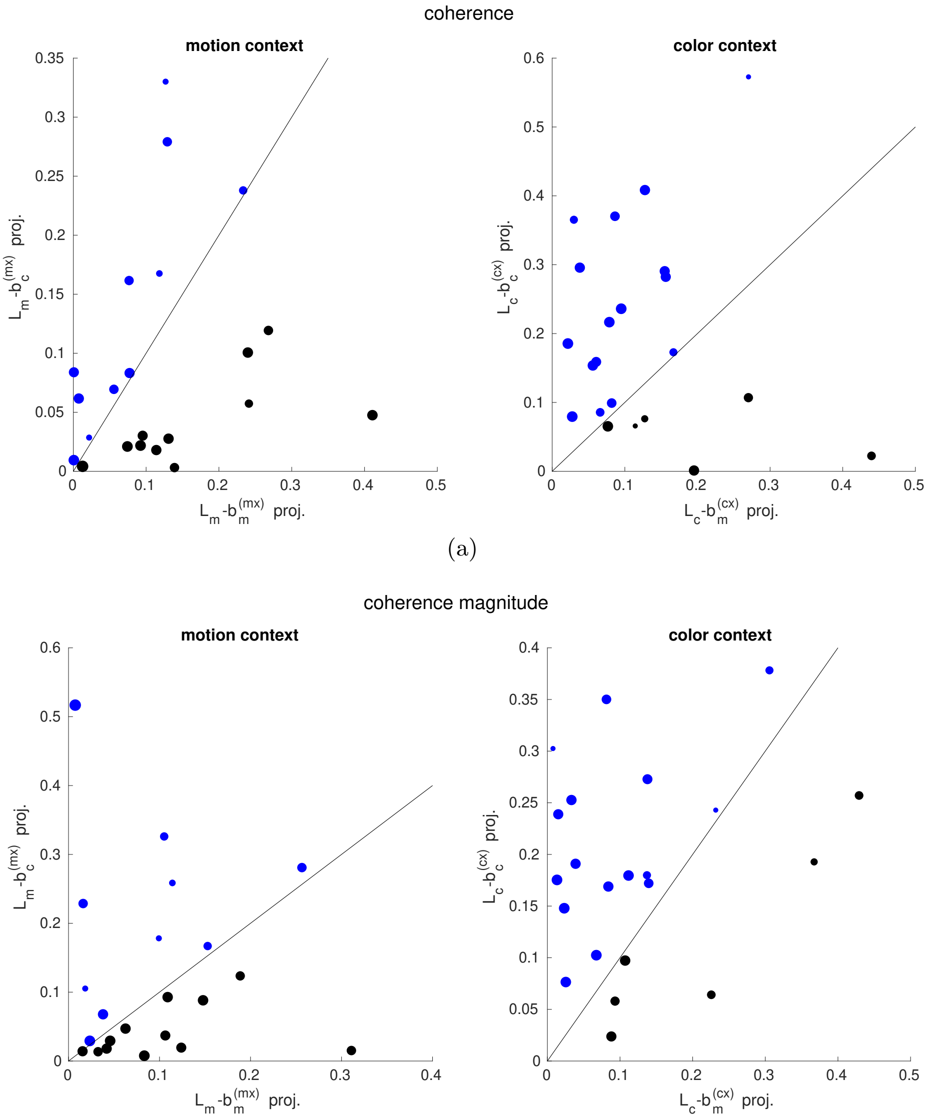
We found that the JF model recovers similar coherence-associated scalings as the LDS (fig. B.50).

B.2.3.2 Subspaces alignment

Subspace angles between the different input subspaces (fig. B.51).

B.2.3.3 Input directions constrained across contexts

In the color context, the flexible model outperforms the constrained one. In the motion context, however, the flexible input directions model misses some variance along the decision dimension, performing worse than the constrained model (fig. B.52).



(a)

(b)

Figure B.26: LDS input directions projection onto the left eigenvectors of the dynamics –which were normalized to be unit norm– a) for the inferred motion and color coherence inputs b) for the motion and color coherence magnitude inputs. Each left eigenvector is associated with an eigenvalue, whose magnitude is given by the size of each dot. The dots have been colored to illustrate which input has the biggest projection onto a given mode, either the color input (blue, dots above the diagonal) or the motion input (black, dots below the diagonal). For complex-conjugate pairs we consider the projection onto the real and the imaginary components, which define the complex planes. (a,b) left figure, motion context; right figure, color context.

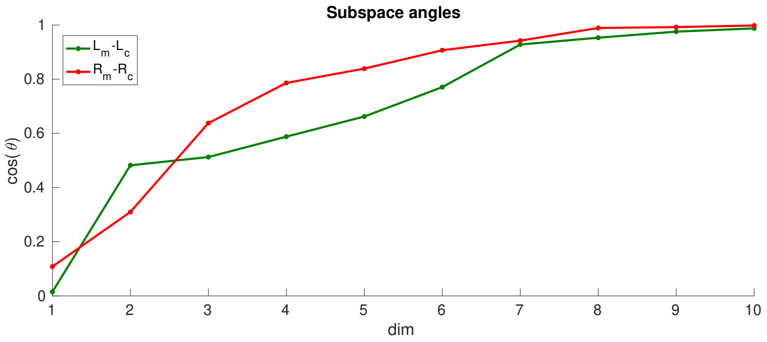
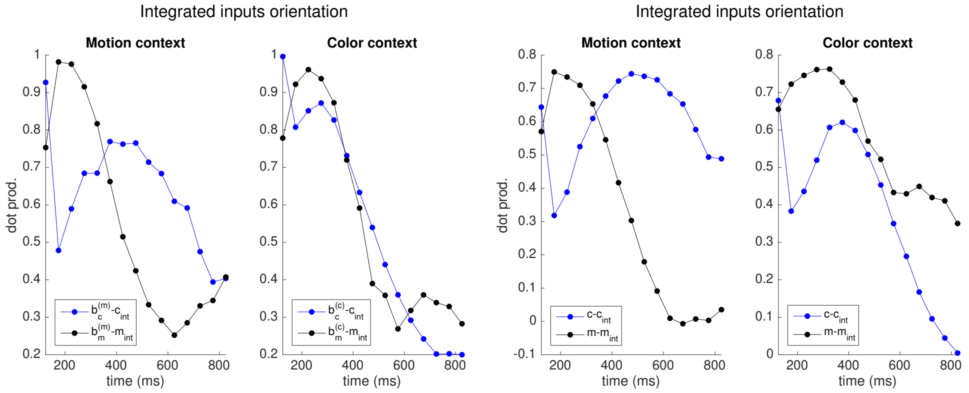


Figure B.27: Across contexts alignment between the subspaces spanned by the left $(L_m - L_c)$ and right $(R_m - R_c)$ 10 slowest modes of the dynamics. Note that both the right and the left subspaces are different across contexts, but they strongly overlap along a few dimensions. Notably, the left (input) subspaces differ much more across contexts than the right (output) subspaces, indicating a dynamical realignment of input-mapping slow modes across contexts.



(a) (b)

Figure B.28: LDS integrated inputs components alignment, in time, with respect to a) the LDS coherence input vectors. b) the TDR input vectors. The components have been estimated for the condition with the largest positive motion and color coherence inputs.

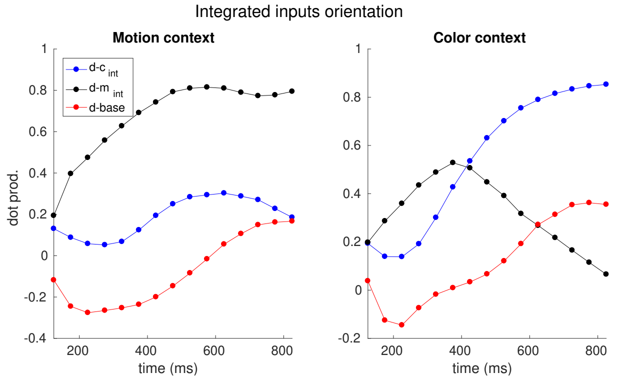
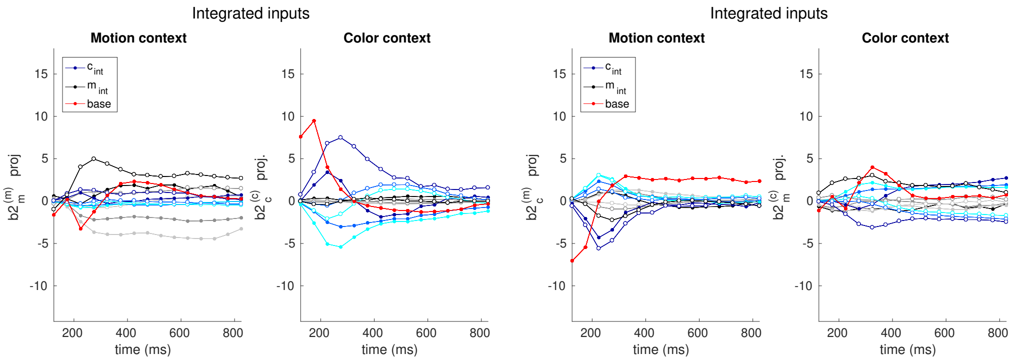
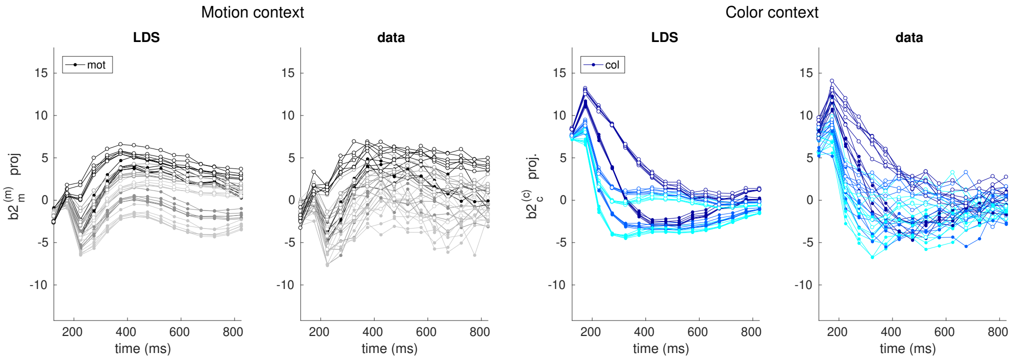


Figure B.29: LDS integrated inputs and base components alignment, in time, with respect to the decision axis. The components have been estimated for the condition with the largest positive motion and color coherence inputs.



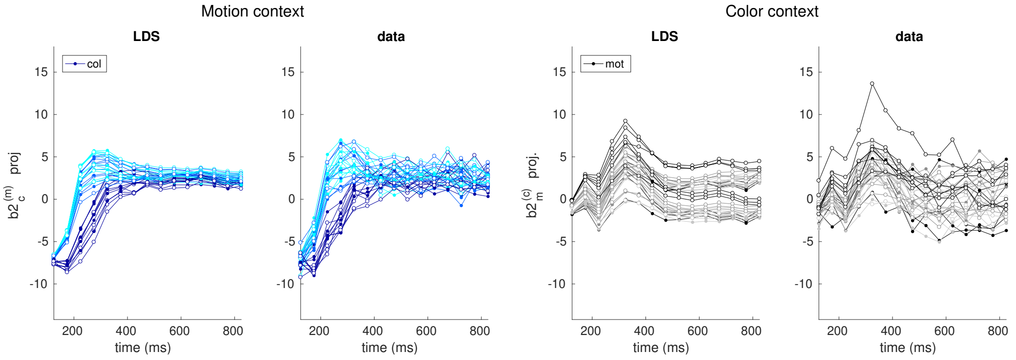
(a) (b)

Figure B.30: LDS integrated inputs and base components projected onto the LDS coherence magnitude input directions, for all possible input values (6 each for color and motion) a) relevant dimensions b) irrelevant dimensions.



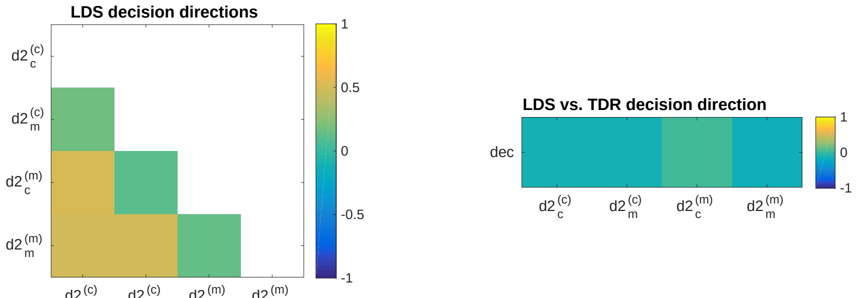
(a) (b)

Figure B.31: LDS model and data trajectories, for the 36 possible conditions, projected onto the relevant coherence input dimensions. The y-axis corresponds to a regression-identified direction, lying within the relevant-input 2D subspace, that reflects the relevant coherence strength. Each condition is color-coded according to the relevant modality for each context.



(a) (b)

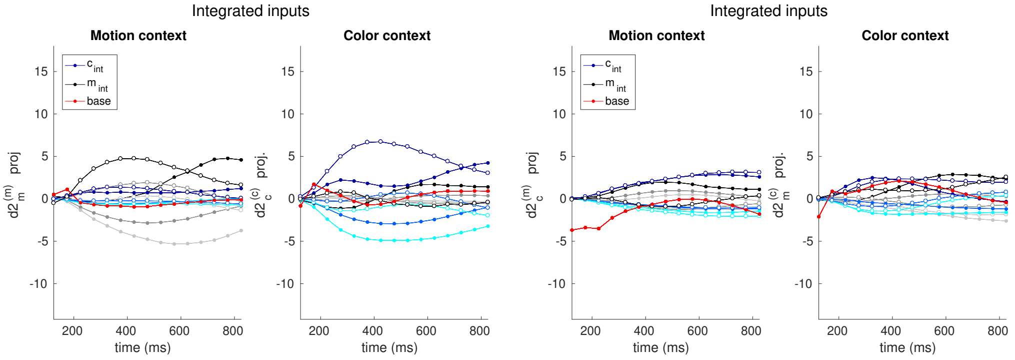
Figure B.32: LDS model and data trajectories, for the 36 possible conditions, projected onto the irrelevant coherence input dimensions. The y-axis corresponds to a regression-identified direction, lying within the irrelevant-input 2D subspace, that reflects the irrelevant coherence strength. Each condition is color-coded according to the irrelevant modality for each context.



(a)

(b)

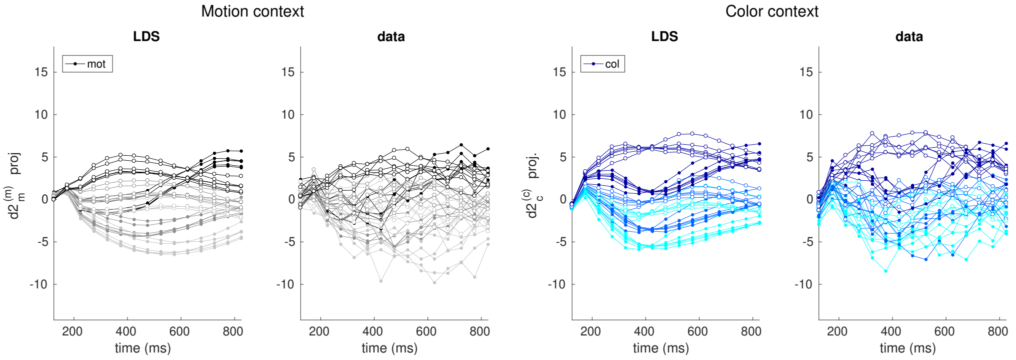
Figure B.33: Orthogonal dimensions to the decision-related dimensions within the color and motion integrated 2D subspaces at the last time step. The dimensions are orthogonal to: directions that separate the sign of the relevant inputs, within the relevant inputs subspaces $(d2_m^{(m)} - d2_c^{(c)})$, and directions that separate the sign of the irrelevant inputs, within the irrelevant inputs subspaces $(d2_m^{(c)} - d2_c^{(m)})$. Unlike the decision dimensions $d_m^{(m)} - d_c^{(c)}$, the orthogonal dimensions $d2_m^{(m)} - d2_c^{(c)}$ are not the same across contexts. a) orthogonal decision dimensions correspondence, across and within contexts b) orthogonal decision dimensions and TDR decision axis correspondence.



(a)

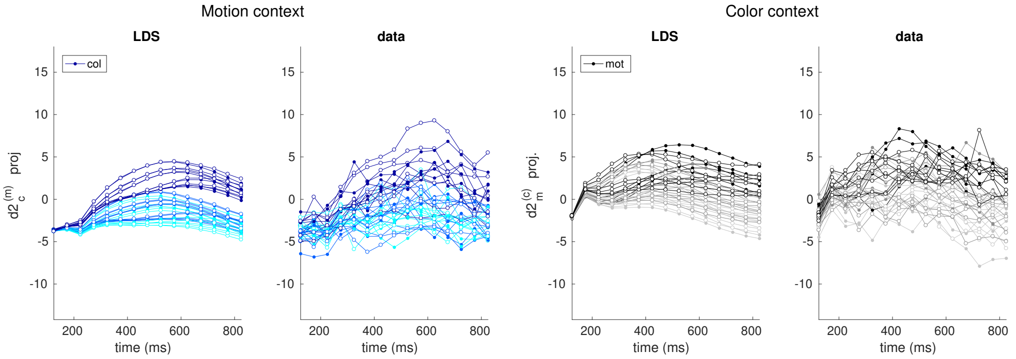
(b)

Figure B.34: LDS integrated inputs and base components projected onto the LDS coherence magnitude integrated input directions, for all possible input values (6 each for color and motion) a) relevant dimensions b) irrelevant dimensions.



(a) (b)

Figure B.35: LDS model and data trajectories, for the 36 possible conditions, projected onto the dimension orthogonal to the “relevant” decision dimension. Within the integrated relevant-input 2D subspaces, we first identified a dimension, via regression, that separates the relevant input sign. The y-axis corresponds to the dimension orthogonal to it. Each condition is color-coded according to the irrelevant modality for each context.



(a) (b)

Figure B.36: LDS model and data trajectories, for the 36 possible conditions, projected onto the dimension orthogonal to the “irrelevant” decision dimension. Within the integrated irrelevant-input 2D subspaces, we first identified a dimension, via regression, that separates the irrelevant input sign. The y-axis corresponds to the dimension orthogonal to it. Each condition is color-coded according to the irrelevant modality for each context.

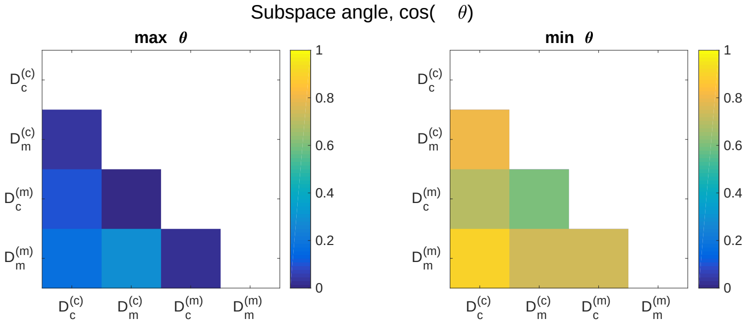


Figure B.37: Alignment within and across contexts between the color and the motion integrated 2D subspaces at the last time step a) maximum subspace angle, b) minimum subspace angle.

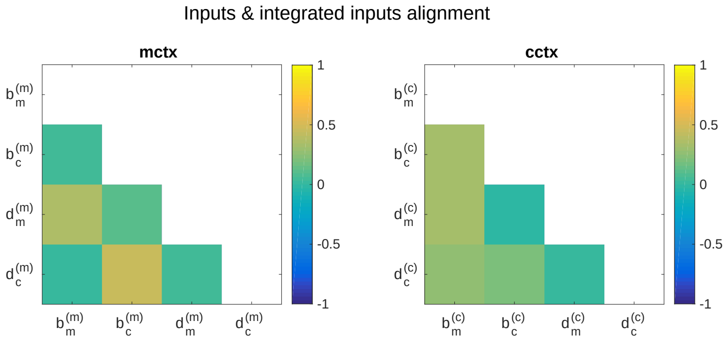
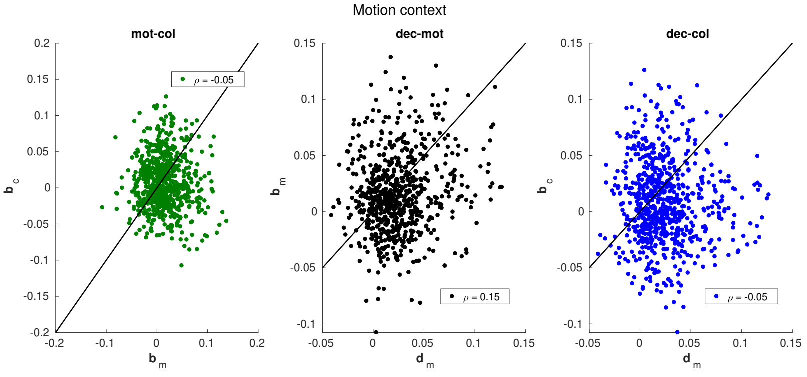
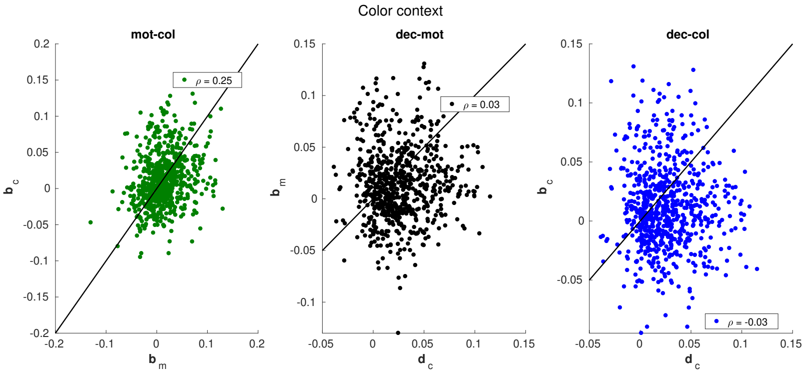


Figure B.38: LDS coherence input dimensions and decision-like dimensions similarity.

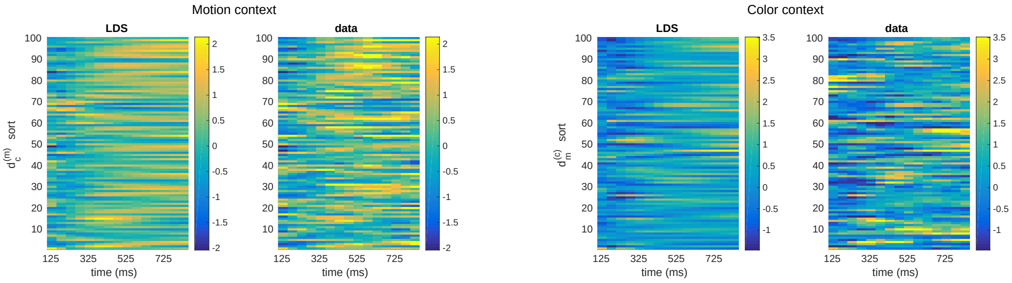


(a)



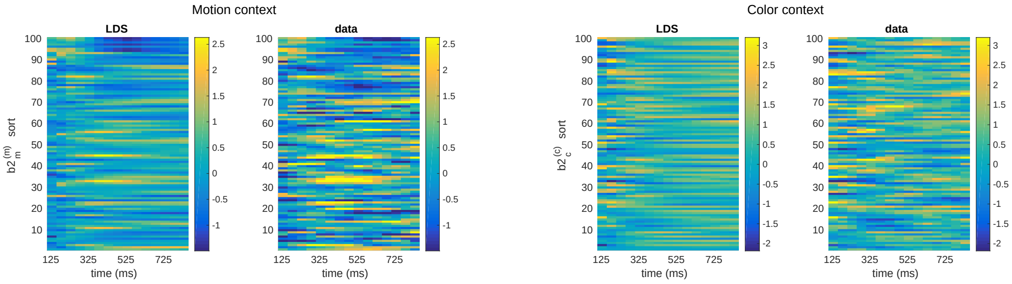
(b)

Figure B.39: LDS coherence inputs and decision-like dimensions loads into the population. The loads are defined by the normalized vectors $Cd_{m/c}^{(m/c)}$, $Cb1_{c/m}^{(c/m)}$, $Cb1_{m/c}^{(c/m)}$.



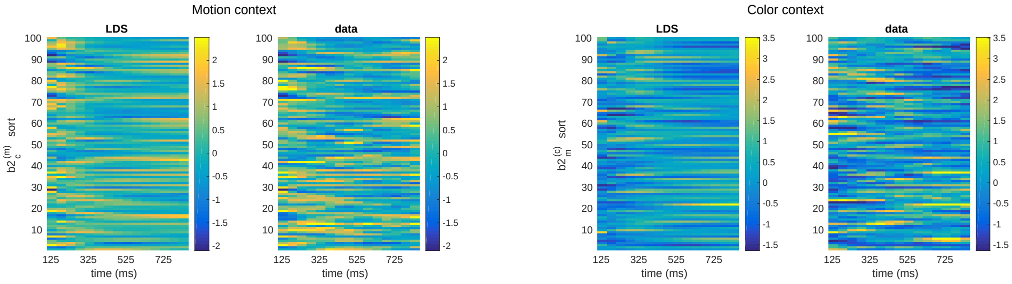
(a) (b)

Figure B.40: PSTHs of 100 units participating in the activity patterns along the LDS “irrelevant” decision dimensions. Same convention as in the main text.



(a) (b)

Figure B.41: PSTHs of 100 units participating in the activity patterns along the LDS relevant coherence magnitude input dimensions. Same convention as in the main text.



(a) (b)

Figure B.42: PSTHs of 100 units participating in the activity patterns along the LDS irrelevant coherence magnitude input dimensions. Same convention as in the main text.

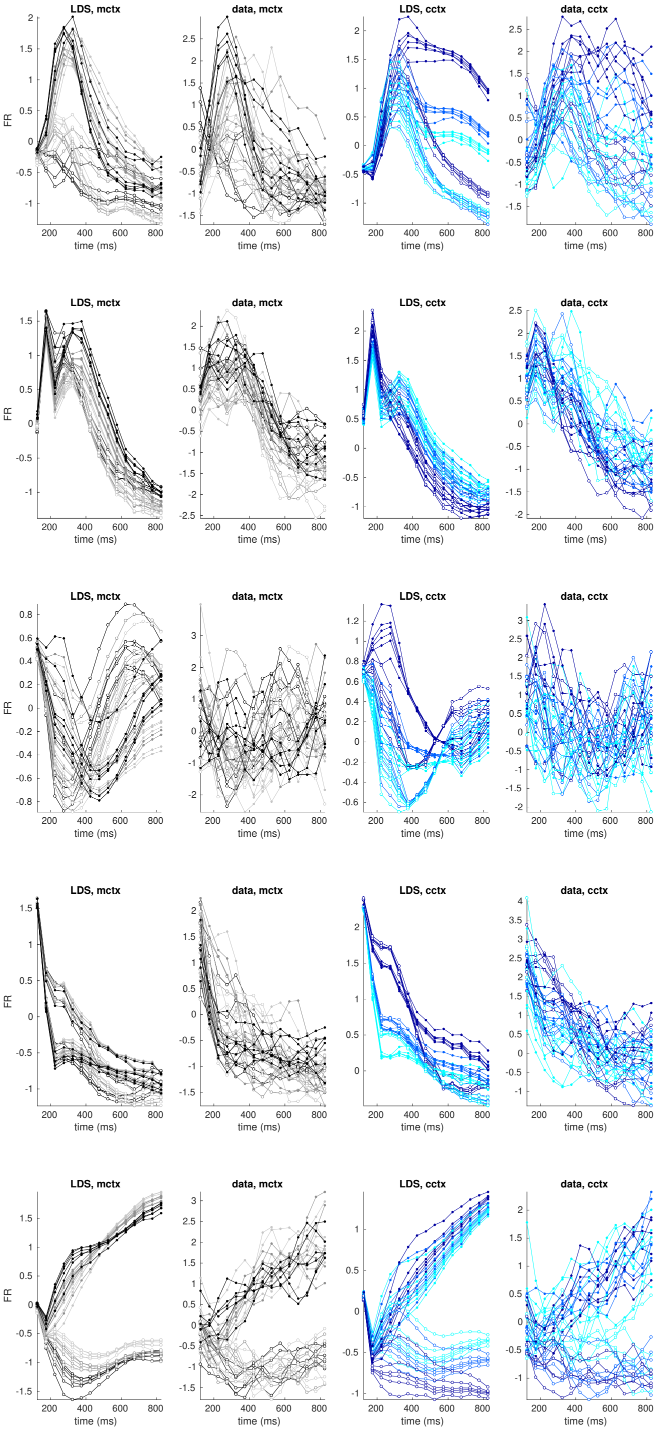


Figure B.43: Single unit PSTHs for the 36 conditions (units n=243, 582, 307, 692, 716). First and third column, LDS model. Second and fourth column, PFC data. In each row, from top to bottom, we show units with the largest load from: the motion coherence input (in the motion context), the motion coherence magnitude input (in the motion context), the color coherence input (in the color context), the color coherence magnitude input (in the color context) and from the decision axis. For the data, responses have been smoothed using a squared window filter, as in Mante et al. (2013).

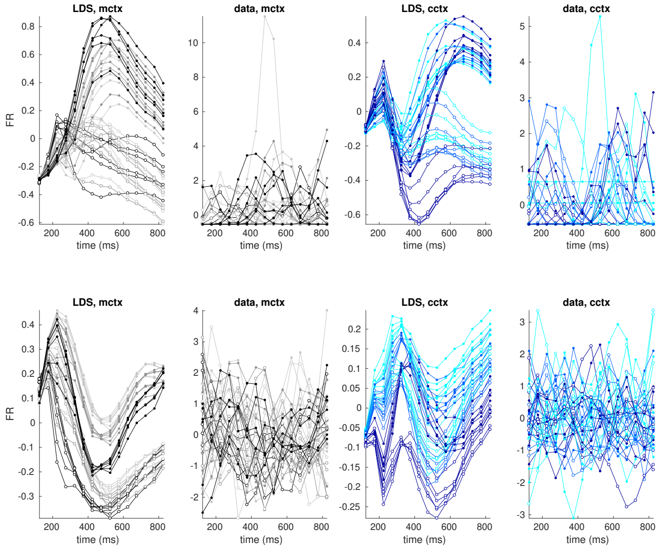


Figure B.44: Two example PSTHs of units with large CV error (units n=418, 4). Same convention as previous figure. First row, the neuron with the maximum mean error in the motion context, $\text{ME}_{max} = 12.8$ (in standard deviation units). Second row, an example neuron with maximum mean error $\text{ME}_{max} = 3$. Note that responses have been smoothed and that the real data for these neurons looks even noisier. Neurons with very large errors like this impact the overall model performance substantially.

B.2.3.4 Dynamics

Coherence magnitude signals are present in the data along the coherence magnitude dimensions (fig. B.58, B.60 and B.60).

The second integrated relevant dimensions, orthogonal to the decision-like dimensions, also reflected a magnitude signal (fig. B.61, B.63 and B.64).

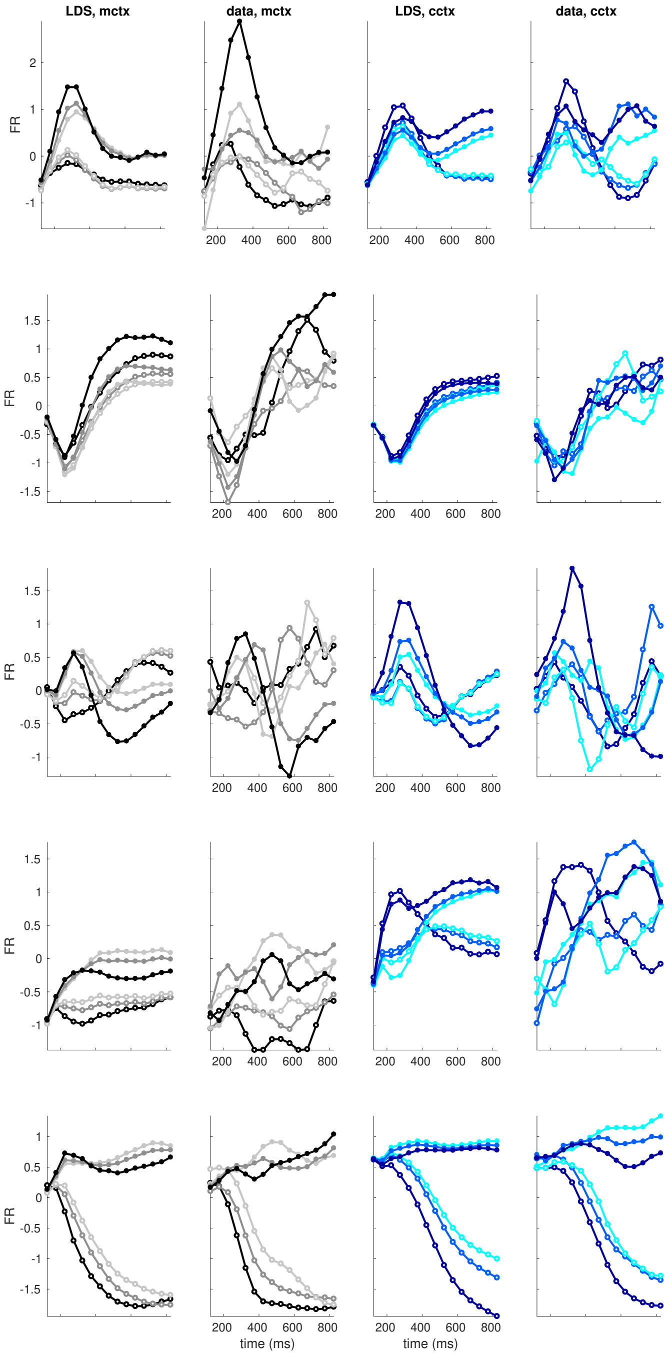


Figure B.45: Single unit PSTHs sorted by conditions with the irrelevant input influence averaged out (units n=17, 650, 308, 640, 16). First and third column, LDS model. Second and fourth column, PFC data. In each row, from top to bottom, we show units with the largest 4th load from: the motion coherence input (in the motion context), the motion coherence magnitude input (in the motion context), the color coherence input (in the color context), the color coherence magnitude input (in the color context) and from the decision axis. For the data, responses have been smoothed using a squared window filter, as in Mante et al. (2013).

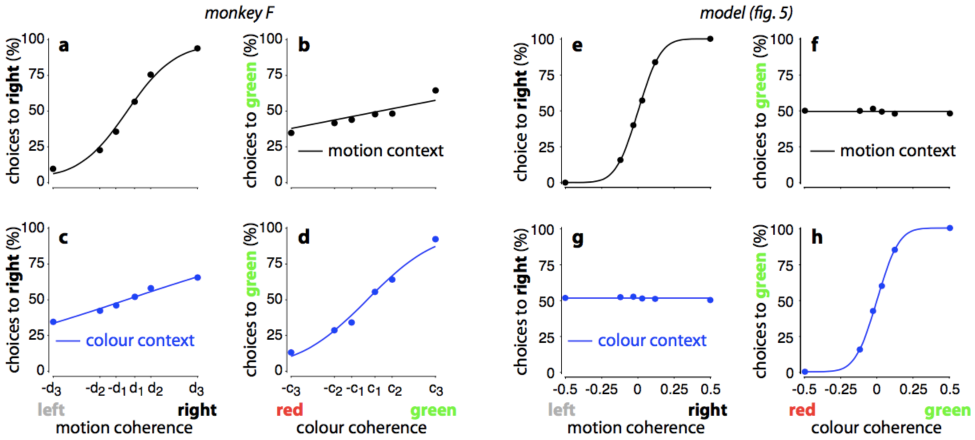


Figure B.46: (a-d) Psychometric curves for monkey F. The monkey was able to perform the contextual DM task, given that the irrelevant inputs had almost no impact on the monkey's choices. (e-h) Psychometric curves for the non-linear RNN. The network successfully learned the task. Note, however, that the irrelevant inputs did not bias the choices, unlike what was observed in the monkeys, meaning that the irrelevant information was completely disregarded by this network. Reprinted by permission from Mante et al. (2013)

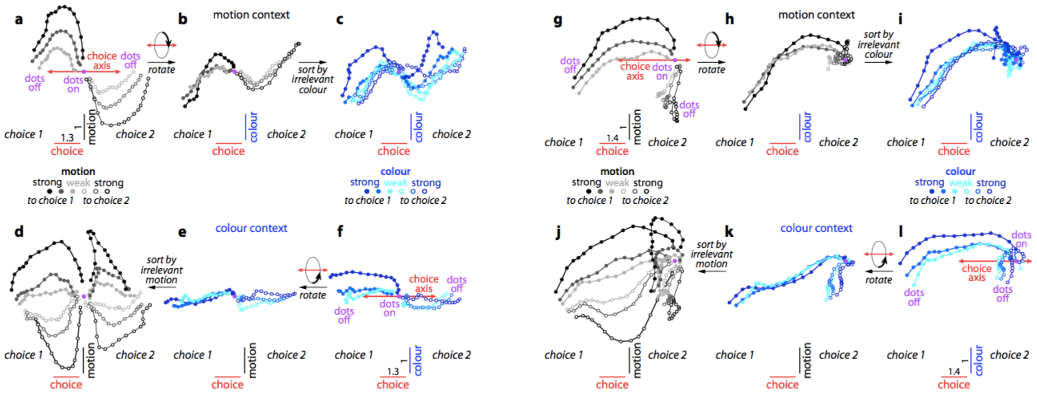
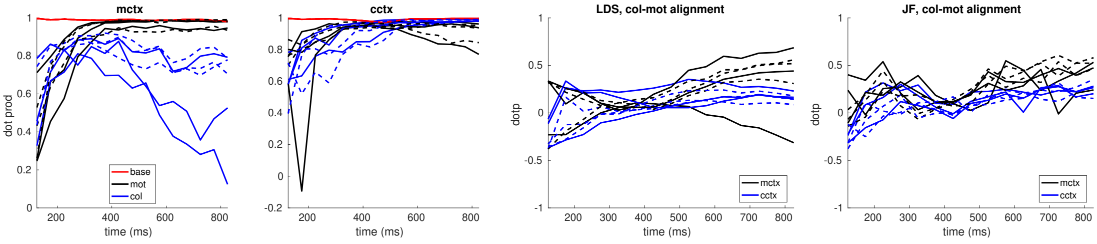


Figure B.47: Population trajectories are shown on the right. On the left the same trajectories are plotted, but condition independent effects have been subtracted out in order to remove the drift effect towards choice1. This drift reflects a tendency of individual firing rates to increase throughout the stimulus presentation time and can be interpreted as an "urgency signal". Trajectories are shown for the period of dots presentation, projected onto the task-relevant subspace spanned by the decision, the color and the motion directions. The first/second row corresponds to the data for the motion/color contexts. Each trajectory reflects the average population activity in a given condition, with the effects of either color (first column) or motion (third column) averaged out. The strength of the evidence in each condition is indicated by the gray and blue scales. Filled/hollow circles designate Tin/Tout types of evidence. Reprinted by permission from Mante et al. (2013)



(a) (b)

Figure B.48: a) LDS and JF models base, motion and color components directions correspondence in the motion and color contexts b) Alignment between the color and motion components within each context, for both the LDS and the JF models.

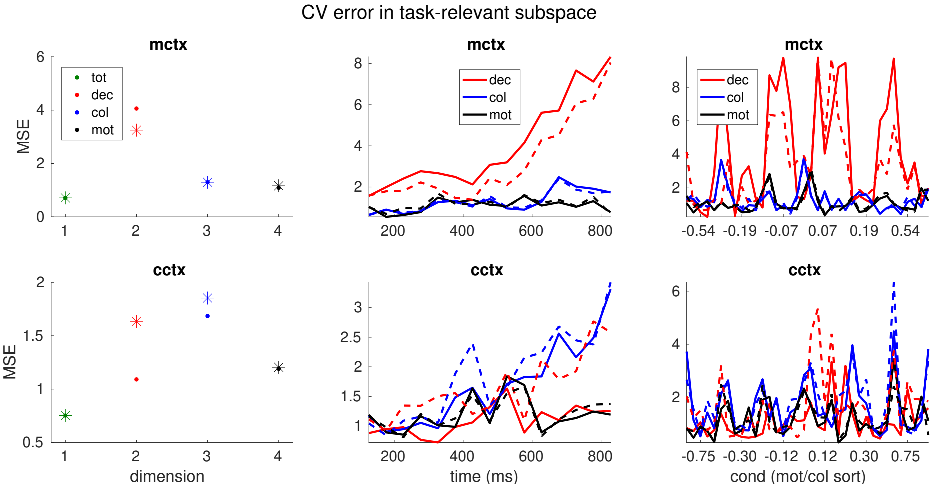
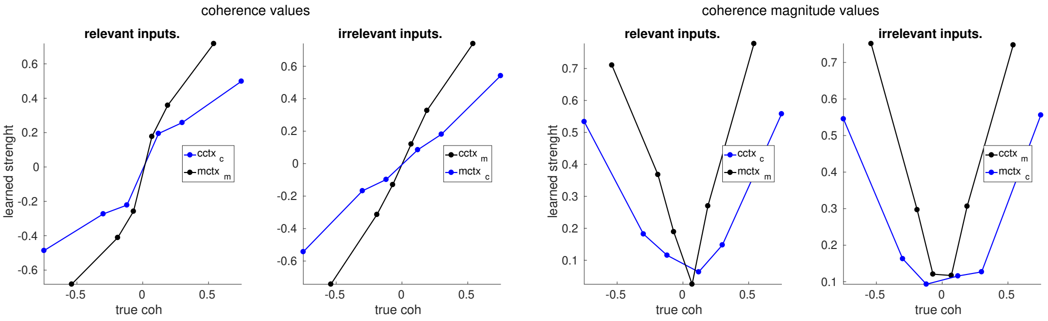


Figure B.49: LDS CV performance in the task-relevant subspace for constrained and unconstrained dynamics across contexts. First row, motion context. Second row, color context. We use dots/stars and filled/dashed lines for the unconstrained/constrained models a) Total CV MSE (green) and MSE along the TDR decision, color and motion input dimensions b) CV error across time, for each dimension b) CV error across conditions, for each dimension. Same conventions as for previous figures apply.



(a) (b)

Figure B.50: JF model inferred coherence values for each of the four motion and color input dimensions in each context. Same conventions as for the figures in the main text.

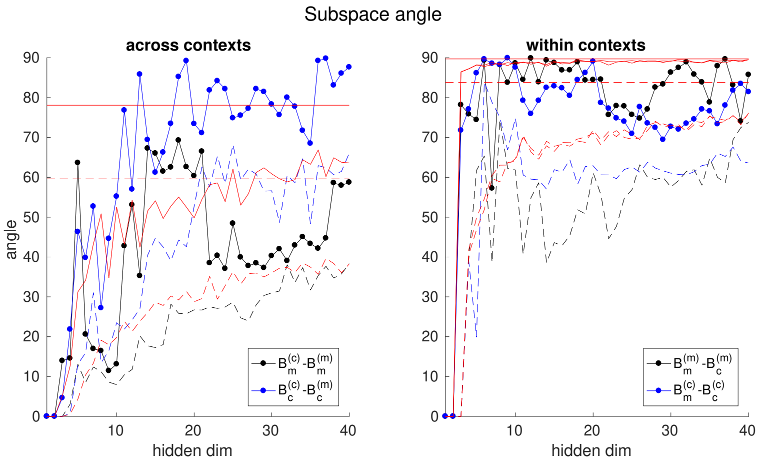


Figure B.51: Subspace angles within and across contexts, for different hidden dimensionalities. Solid lines designates maximum angles and dashed lines minimum angles. Red lines in the left plot indicate the 5th percentile of the null distribution for random angles. In the right plot they indicate the 95th percentile. The constant lines correspond to the null estimated on the whole data space, without projecting the covariance onto the hidden subspaces. In the right plot there are two red lines because a different null has been computed for each context.

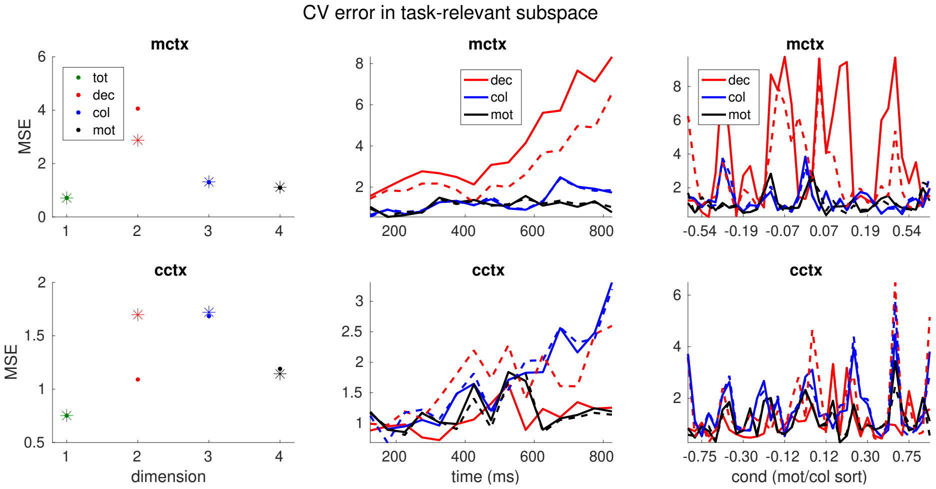
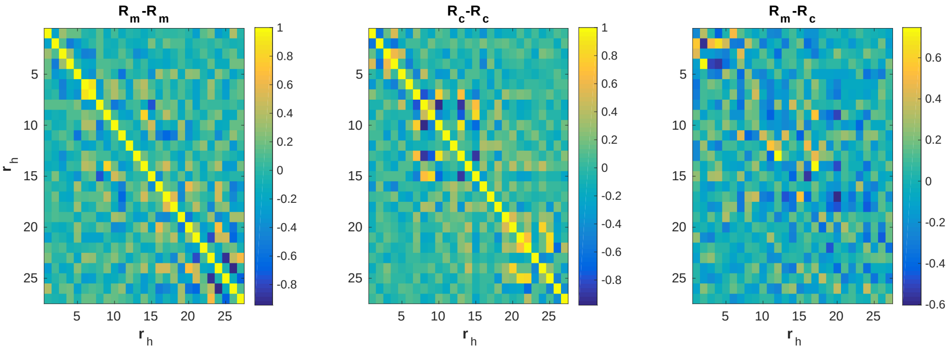
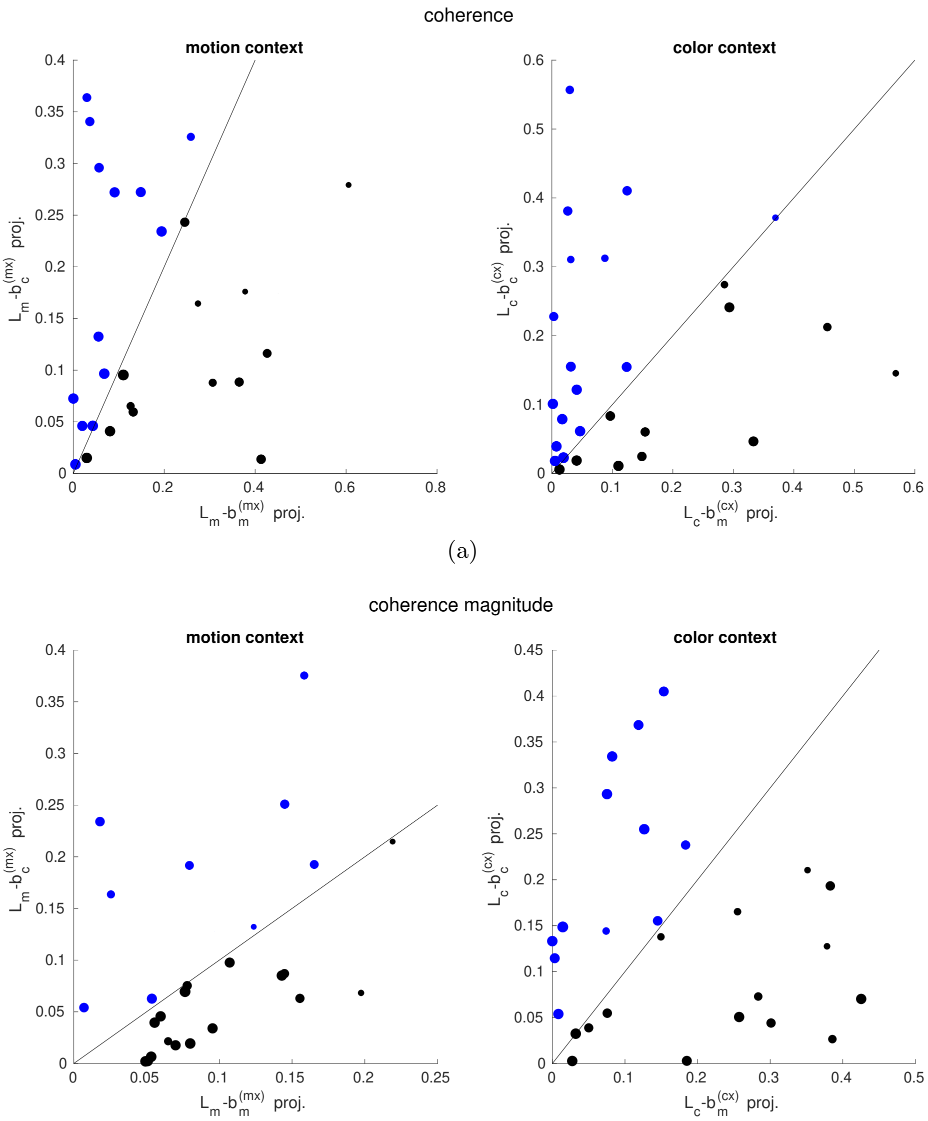


Figure B.52: LDS CV performance in the task-relevant subspace for constrained and unconstrained input directions across contexts. First row, motion context. Second row, color context. We use dots/stars and filled/dashed lines for the unconstrained/constrained models a) Total CV MSE (green) and MSE along the TDR decision, color and motion input dimensions b) CV error across time, for each dimension b) CV error across conditions, for each dimension. Same conventions as for previous figures apply.



(a)

Figure B.53: Right eigenvectors alignment within and across contexts, sorted by increasing magnitude of the associated eigenvalues.



(a)

(b)

Figure B.54: LDS input directions projection onto the left eigenvectors of the dynamics –which were normalized to be unit norm– a) for the inferred motion and color coherence inputs b) for the motion and color coherence magnitude inputs. Each left eigenvector is associated with an eigenvalue, whose magnitude is given by the size of each dot. The dots have been colored to illustrate which input has the biggest projection onto a given mode, either the color input (blue, dots above the diagonal) or the motion input (black, dots below the diagonal). For complex-conjugate pairs we consider the projection onto the real and the imaginary components, which define the complex planes. (a,b) left figure, motion context; right figure, color context.

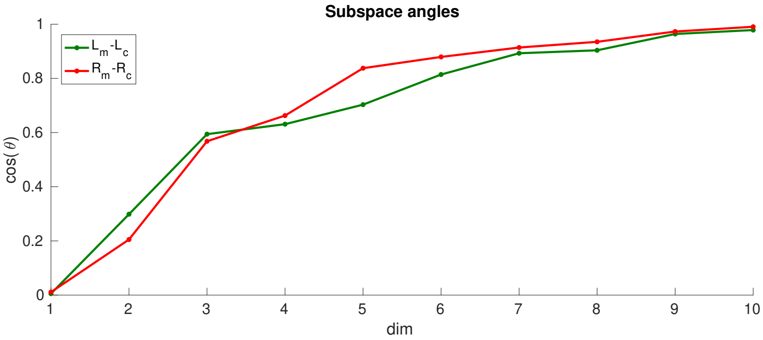


Figure B.55: Across contexts alignment between the subspaces spanned by the left ($L_m - L_c$) and right ($R_m - R_c$) 10 slowest modes of the dynamics. Note that both the right and the left subspaces are different across contexts, but they strongly overlap along a few dimensions. Notably, the left (input) subspaces differ more across contexts than the right (output) subspaces –albeit more subtly than for monkey A–, indicating a dynamical realignment of input-mapping slow modes across contexts.

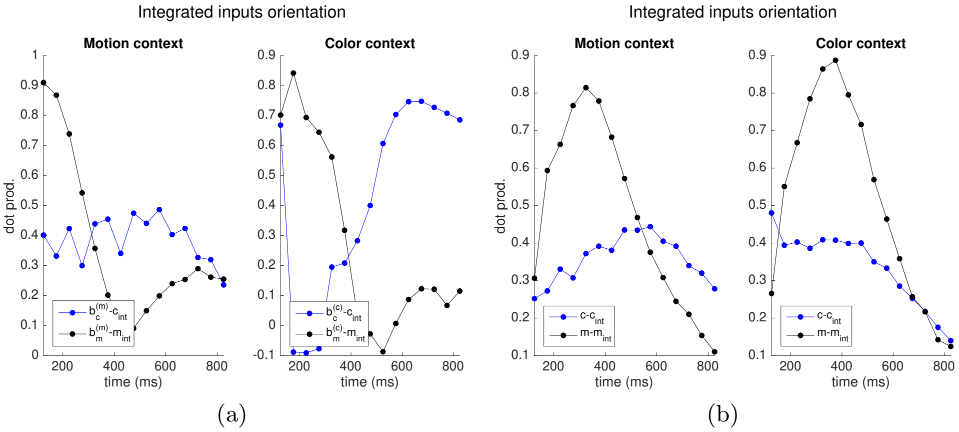


Figure B.56: LDS integrated inputs components alignment, in time, with respect to a) the LDS coherence input vectors. b) the TDR input vectors. The components have been estimated for the condition with the largest positive motion and color coherence inputs.

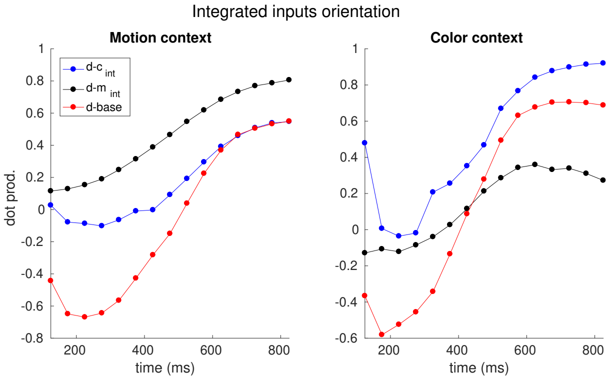
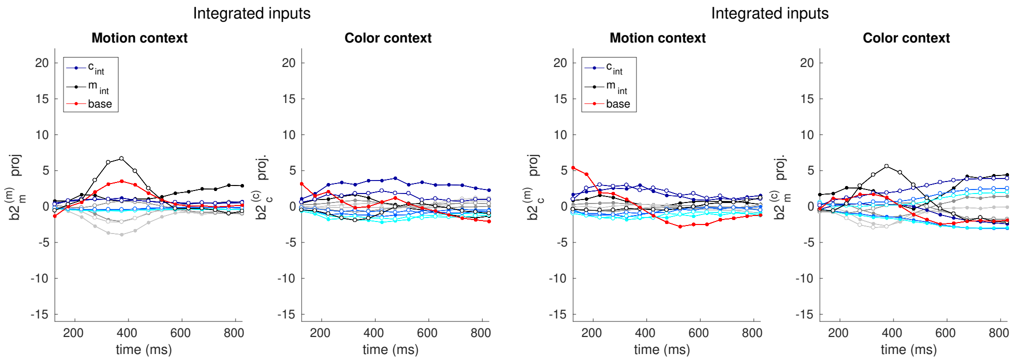
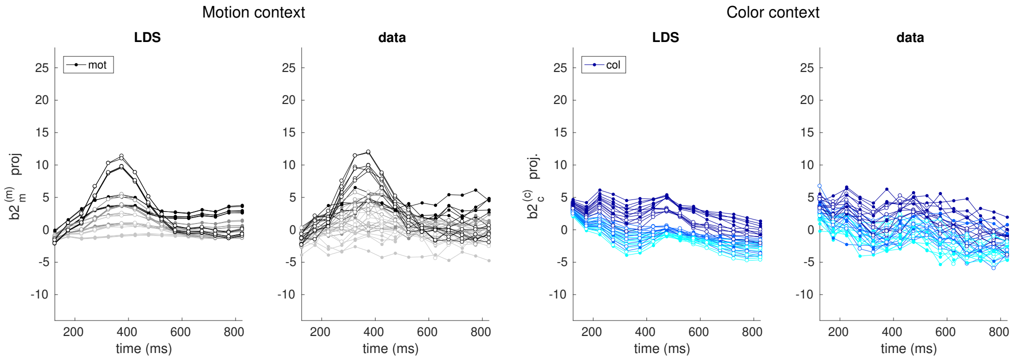


Figure B.57: LDS integrated inputs and base components alignment, in time, with respect to the decision axis. The components have been estimated for the condition with the largest positive motion and color coherence inputs.



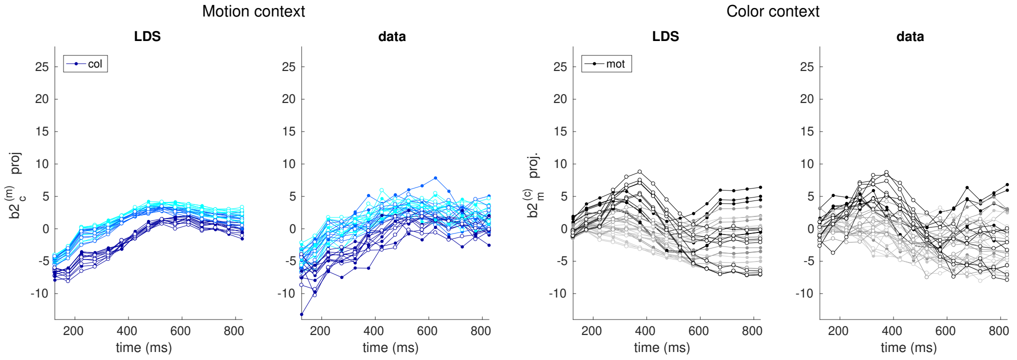
(a) (b)

Figure B.58: LDS integrated inputs and base components projected onto the LDS coherence magnitude input directions, for all possible input values (6 each for color and motion) a) relevant dimensions b) irrelevant dimensions.



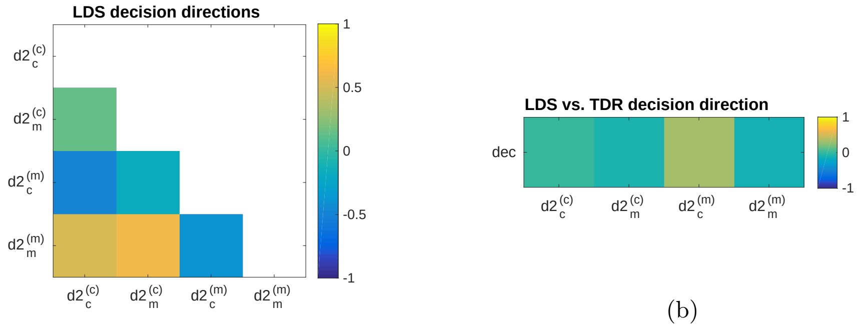
(a) (b)

Figure B.59: LDS model and data trajectories, for the 36 possible conditions, projected onto the relevant coherence input dimensions. The y-axis corresponds to a regression-identified direction, lying within the relevant-input 2D subspace, that reflects the relevant coherence strength. Each condition is color-coded according to the relevant modality for each context.



(a) (b)

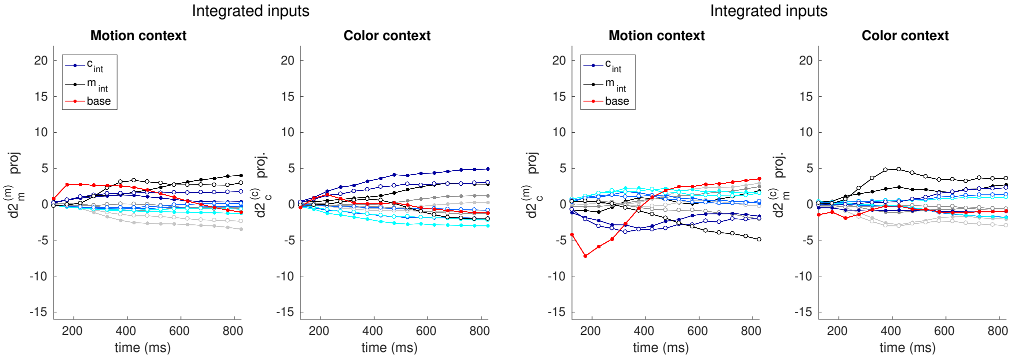
Figure B.60: LDS model and data trajectories, for the 36 possible conditions, projected onto the irrelevant coherence input dimensions. The y-axis corresponds to a regression-identified direction, lying within the irrelevant-input 2D subspace, that reflects the irrelevant coherence strength. Each condition is color-coded according to the irrelevant modality for each context.



(a)

(b)

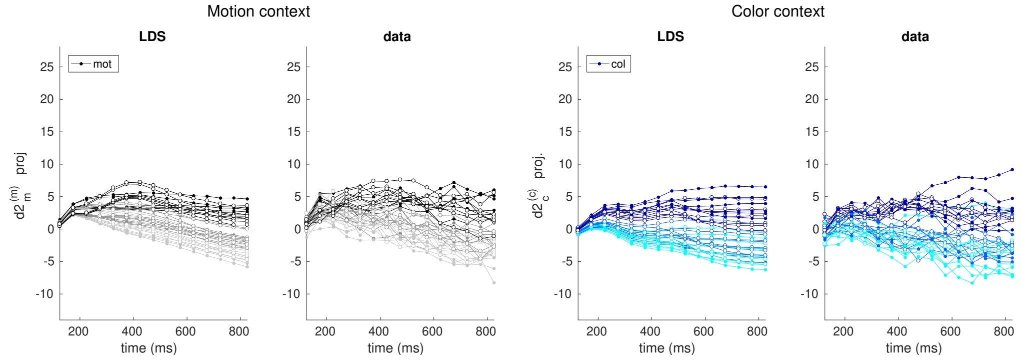
Figure B.61: Orthogonal dimensions to the decision-related dimensions within the color and motion integrated 2D subspaces at the last time step. The dimensions are orthogonal to: directions that separate the sign of the relevant inputs, within the relevant inputs subspaces ($d2_m^{(m)} - d2_c^{(c)}$), and directions that separate the sign of the irrelevant inputs, within the irrelevant inputs subspaces ($d2_m^{(c)} - d2_c^{(m)}$). Unlike the decision dimensions $d_m^{(m)} - d_c^{(c)}$, the orthogonal dimensions $d2_m^{(m)} - d2_c^{(c)}$ are not the same across contexts. a) orthogonal decision dimensions correspondence, across and within contexts b) orthogonal decision dimensions and TDR decision axis correspondence.



(a)

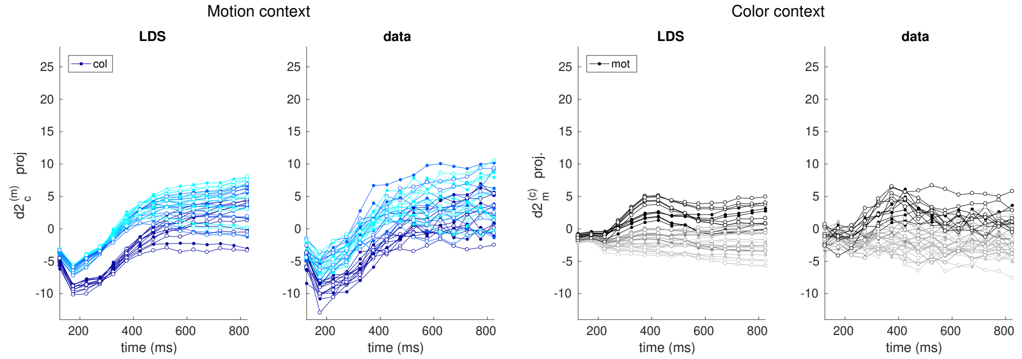
(b)

Figure B.62: LDS integrated inputs and base components projected onto the LDS coherence magnitude integrated input directions, for all possible input values (6 each for color and motion) a) relevant dimensions b) irrelevant dimensions.



(a) (b)

Figure B.63: LDS model and data trajectories, for the 36 possible conditions, projected onto the dimension orthogonal to the "relevant" decision dimension. Within the integrated relevant-input 2D subspaces, we first identified a dimension, via regression, that separates the relevant input sign. The y-axis corresponds to the dimension orthogonal to it. Each condition is color-coded according to the irrelevant modality for each context.



(a) (b)

Figure B.64: LDS model and data trajectories, for the 36 possible conditions, projected onto the dimension orthogonal to the "irrelevant" decision dimension. Within the integrated irrelevant-input 2D subspaces, we first identified a dimension, via regression, that separates the irrelevant input sign. The y-axis corresponds to the dimension orthogonal to it. Each condition is color-coded according to the irrelevant modality for each context.

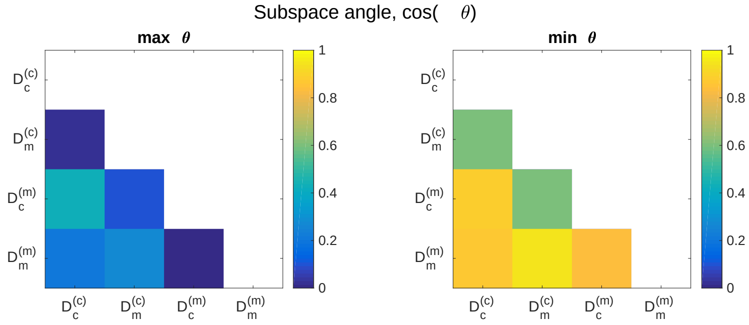
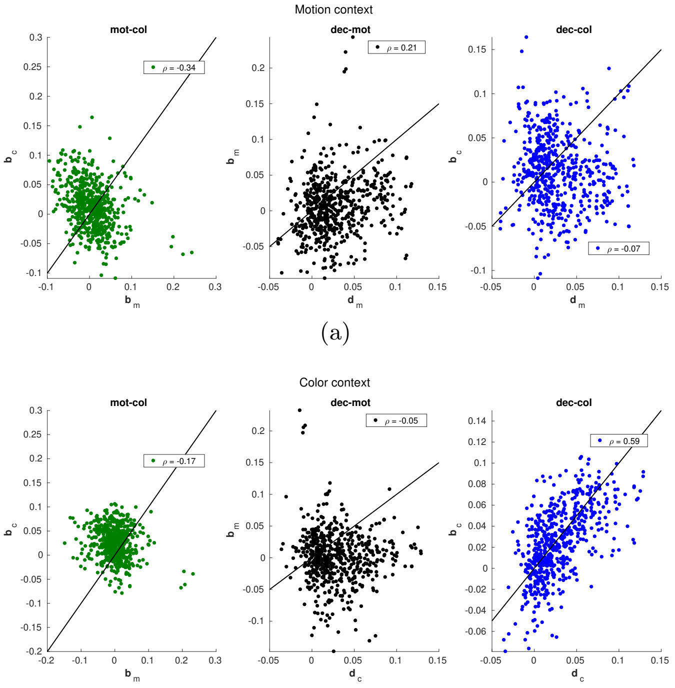


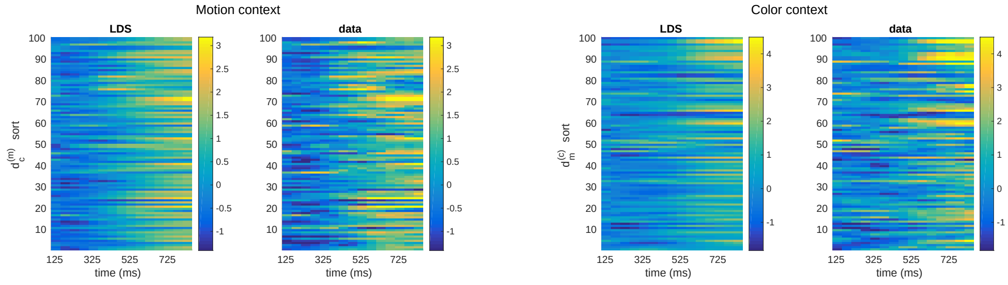
Figure B.65: Alignment within and across contexts between the color and the motion integrated 2D subspaces at the last time step a) maximum subspace angle, b) minimum subspace angle.



(a)

(b)

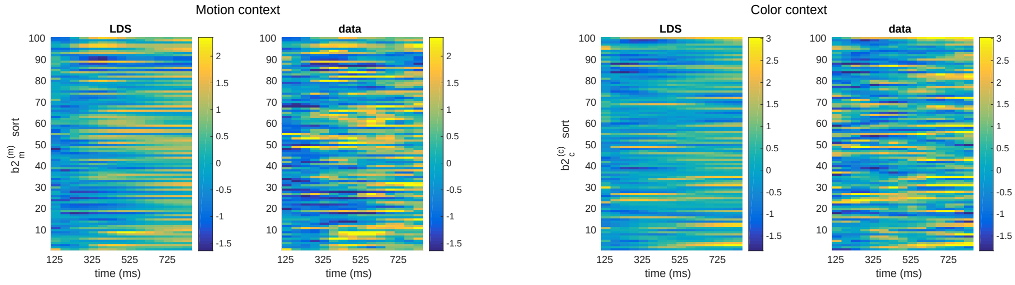
Figure B.66: LDS coherence inputs and decision-like dimensions loads into the population. The loads are defined by the vectors $Cd^{(m/c)}_{m/c}$, $Cb1^{(c/m)}_{c/m}$, $Cb1^{(c/m)}_{m/c}$.



(a)

(b)

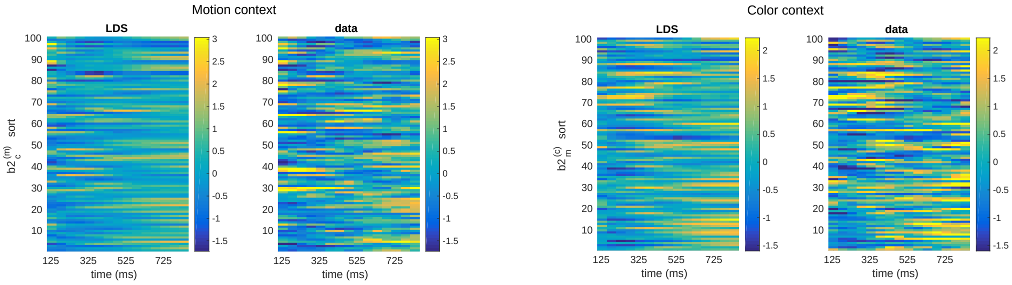
Figure B.67: PSTHs of 100 units participating in the activity patterns along the LDS “irrelevant” decision dimensions. Same convention as in the main text.



(a)

(b)

Figure B.68: PSTHs of 100 units participating in the activity patterns along the LDS relevant coherence magnitude input dimensions. Same convention as in the main text.



(a) (b)

Figure B.69: PSTHs of 100 units participating in the activity patterns along the LDS irrelevant coherence magnitude input dimensions. Same convention as in the main text.

Bibliography

Barak O (2017) Recurrent neural networks as versatile tools of neuroscience research. *Curr. Opin. Neurobiol.* 46:1–6. 141

Barborica A, Ferrera VP (2003) Estimating invisible target speed from neuronal activity in monkey frontal eye field. *Nature Neuroscience* 6:66–74. 50

Barlow HB (1972) Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* 1:371–394. 125

Barrett DG, Deneve S, Machens CK (2016) Optimal compensation for neuron loss. *eLife Sciences* 5:e12454. 135

Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD (2006) The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev* 113:700–765. 27, 29

Britten KH, Newsome WT (1998) Tuning Bandwidths for Near-Threshold Stimuli in Area MT. *Journal of Neurophysiology* 80:762–770. 24

Brooks JL, List A (2006) Searching for the Role of the Frontal Eye Fields in the Visual Attention Network. *J Neurosci* 26:2145–2146. 50

Buckley MJ, Mansouri FA, Hoda H, Mahboubi M, Browning PGF, Kwok SC, Phillips A, Tanaka K (2009) Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. *Science* 325:52–58. 28

Buesing L, Macke JH, Sahani M (2012) Learning stable, regularised latent models of neural population dynamics. *Network* 23:24–47. 74, 148

Cajal SRy (1893) Neue Darstellung vom histologischen Bau den Centralnervensystems In *Arch. Anat. Physiol., Anat. Abth.*, Vol. V & VI, pp. 310–428. 24

Cajal SRy (1954) Neuron theory or reticular theory. *Consejo Superio De Investigaciones Cientificas* . 24, 125

Carnevale F, de Lafuente V, Romo R, Barak O, Parga N (2015) Dynamic Control of Response Criterion in Premotor Cortex during Perceptual Detection under Temporal Uncertainty. *Neuron* 86:1067–1077. 141

Cassanello CR, Nihalani AT, Ferrera VP (2008) Neuronal Responses to Moving Targets in Monkey Frontal Eye Fields. *Journal of Neurophysiology* 100:1544–1556. 50

Chandrasekaran C, Soldado-Magraner J, Peixoto D, Newsome WT, Shenoy KV, Sahani M (2018) Brittleness in model selection analysis of single neuron firing rates. *In preparation* . 146

Churchland AK, Kiani R, Shadlen MN (2008) Decision-making with multiple alternatives. *Nature Neuroscience* 11:693. 54

Churchland MM, Yu BM, Sahani M, Shenoy KV (2007) Techniques for extracting single-trial activity patterns from large-scale neural recordings. *Curr. Opin. Neurobiol.* 17:609–618. 141

Cisek P, Puskas GA, El-Murr S (2009) Decisions in Changing Conditions: The Urgency-Gating Model. *J. Neurosci.* 29:11560–11571. 54

Cunningham JP, Ghahramani Z (2015) Linear Dimensionality Reduction: Survey, Insights, and Generalizations. *Journal of Machine Learning Research* 16:2859–2900. 31

Cunningham JP, Yu BM (2014) Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience* 17:1500–1509. 31

DeFelipe J (2006) Brain plasticity and mental processes: Cajal again. *Nature Reviews Neuroscience* 7:811–817. 125

Ding L, Gold JI (2012) Neural Correlates of Perceptual Decision Making before, during, and after Decision Commitment in Monkey Frontal Eye Field. *Cereb Cortex* 22:1052–1067. 26, 50, 135, 138

Duncker L, Sahani M (2018) Temporal alignment and latent Gaussian process factor inference in population spike trains. *bioRxiv* p. 331751. 38

Elsayed GF, Cunningham JP (2017) Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nature Neuroscience* 20:1310–1318. 141

Elsayed GF, Lara AH, Kaufman MT, Churchland MM, Cunningham JP (2016) Reorganization between preparatory and movement population responses in motor cortex. *Nature Communications* 7:13239. 79, 135

Fuster J (2015) *The Prefrontal Cortex* Elsevier. 28, 125

Gao Y, Archer EW, Paninski L, Cunningham JP (2016) Linear dynamical neural population models through nonlinear embeddings In Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, editors, *Advances in Neural Information Processing Systems 29*, pp. 163–171. Curran Associates, Inc. 147

Glaser JI, Wood DK, Lawlor PN, Segraves MA, Kording KP (2018) From preliminary to definitive plans: two classes of neurons in frontal eye field. *bioRxiv* p. 251835. 50

Glimcher PW (2003) The neurobiology of visual-saccadic decision making. *Annu. Rev. Neurosci.* 26:133–179. 23, 24, 27, 125

Gold JI, Shadlen MN (2000) Representation of a perceptual decision in developing oculomotor commands. *Nature* 404:390–394. 26

Gold JI, Shadlen MN (2001) Neural computations that underlie decisions about sensory stimuli. *Trends Cogn. Sci. (Regul. Ed.)* 5:10–16. 25

Gold JI, Shadlen MN (2007) The Neural Basis of Decision Making. *Annu. Rev. Neurosci.* 30:535–574. 25

Gregoriou GG, Rossi AF, Ungerleider LG, Desimone R (2014) Lesions of prefrontal cortex reduce attentional modulation of neuronal responses and synchrony in V4. *Nat Neurosci* 17:1003–1011. 28, 50

Grimaldi P, Lau H, Basso MA (2015) There are things that we know that we know, and there are things that we do not know we do not know: Confidence in Decision-Making. *Neurosci Biobehav Rev* 55:88–97. 135

Gupta N, Mehra R (1974) Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations. *IEEE Transactions on Automatic Control* 19:774–783. 45

Hamilton J (1986) State-space models Handbook of Econometrics, Elsevier. 45

Hanes DP, Schall JD (1996) Neural control of voluntary movement initiation. *Science* 274:427–430. 26

Hanks TD, Ditterich J, Shadlen MN (2006) Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nat. Neurosci.* 9:682–689. 26, 27

Hanks TD, Kopec CD, Brunton BW, Duan CA, Erlich JC, Brody CD (2015) Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature* 520:220–223. 26, 144

Hanks TD, Mazurek ME, Kiani R, Hopp E, Shadlen MN (2011) Elapsed Decision Time Affects the Weighting of Prior Probability in a Perceptual Decision Task. *J. Neurosci.* 31:6339–6352. 54, 89

Hennequin G, Vogels TP, Gerstner W (2012) Non-normal amplification in random balanced neuronal networks. *Physical Review E* 86. 137

Hennequin G, Vogels T, Gerstner W (2014) Optimal Control of Transient Dynamics in Balanced Networks Supports Generation of Complex Movements. *Neuron* 82:1394–1406. 141

Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. *J Physiol* 148:574–591. 24

Hunt LT, Behrens TE, Hosokawa T, Wallis JD, Kennerley SW (2015) Capturing the temporal evolution of choice across prefrontal cortex. *eLife Sciences* 4:e11945. 27

Jazayeri M (2017) Zooming Out of Single Neurons Reveals Structure in Mnemonic Representations. *Neuron* 96:1210–1212. 136

Katsuki F, Constantinidis C (2012) Unique and shared roles of the posterior parietal and dorsolateral prefrontal cortex in cognitive functions. *Front Integr Neurosci* 6. 26, 28

Katz LN, Yates JL, Pillow JW, Huk AC (2016) Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature* 535:285–288. 27

Kiani R, Shadlen MN (2009) Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science* 324:759–764. 135

Kim JN, Shadlen MN (1999) Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nat. Neurosci.* 2:176–185. 26, 27, 28, 50

Kingma DP, Welling M (2013) Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]* arXiv: 1312.6114. 147

Kobak D, Brendel W, Constantinidis C, Feierstein CE, Kepecs A, Mainen ZF, Qi XL, Romo R, Uchida N, Machens CK (2016) Demixed principal component analysis of neural population data. *eLife Sciences* 5:e10989. 32, 33, 34, 35, 36, 145

Krauzlis RJ (2004) Recasting the Smooth Pursuit Eye Movement System. *Journal of Neurophysiology* 91:591–603. 26

Kumano H, Suda Y, Uka T (2016) Context-Dependent Accumulation of Sensory Evidence in the Parietal Cortex Underlies Flexible Task Switching. *J. Neurosci.* 36:12192–12202. 140

Latimer KW, Yates JL, Meister MLR, Huk AC, Pillow JW (2015) Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science* 349:184–187. 27, 29, 146

Lawrence ND (2004) Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data In Thrun S, Saul LK, SchÃ¶lkopf B, editors, *Advances in Neural Information Processing Systems 16*, pp. 329–336. MIT Press. 38

Liu Z, Hauskrecht M (2016) Learning Linear Dynamical Systems from Multivariate Time Series: A Matrix Factorization Based Framework. *Proc SIAM Int Conf Data Min* 2016:810–818. 45, 148

Machens CK, Gollisch T, Kolesnikova O, Herz AVM (2005) Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron* 47:447–456. 135

Machens CK, Romo R, Brody CD (2005) Flexible Control of Mutual Inhibition: A Neural Model of Two-Interval Discrimination. *Science* 307:1121–1124. 29, 126, 141, 147

Macke JH, Buesing L, Sahani M (2015) Estimating state and parameters in state space models of spike trains In Chen Z, editor, *Advanced State Space Methods for Neural and Clinical Data*, pp. 137–159. Cambridge University Press, Cambridge. 44, 47, 146, 149

Macke JH, Buesing L, Cunningham JP, Yu BM, Shenoy KV, Sahani M (2011) Empirical models of spiking in neural populations In Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, editors, *Advances in Neural Information Processing Systems 24*, pp. 1350–1358. Curran Associates, Inc. 47, 146

Mackevicius EL, Bahle AH, Williams AH, Gu S, Denissenko NI, Goldman MS, Fee MS (2018) Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience. *bioRxiv* p. 273128. 43

Mante V, Sussillo D, Shenoy KV, Newsome WT (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503:78–84. 4, 21, 28, 29, 31, 35, 49, 51, 52, 53, 54, 66, 67, 68, 75, 81, 84, 85, 88, 89, 91, 95, 96, 98, 100, 101, 102, 107, 120, 121, 124, 126, 127, 128, 130, 131, 138, 139, 140, 141, 142, 148, 158, 159, 174, 183, 185, 186

Mastrogiuseppe F, Ostojic S (2017) Linking connectivity, dynamics and computations in recurrent neural networks. *arXiv:1711.09672 [q-bio]* arXiv: 1711.09672. 147

Mazurek ME, Roitman JD, Ditterich J, Shadlen MN (2003) A role for neural integrators in perceptual decision making. *Cereb. Cortex* 13:1257–1269. 26, 29

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24:167–202. 28, 125

Morcos AS, Harvey CD (2016) History-dependent variability in population dynamics during evidence accumulation in cortex. *Nature Neuroscience* 19:1672. 146

Murphy BK, Miller KD (2009) Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron* 61:635–648. 137

Newsome WT, Britten KH, Movshon JA (1989) Neuronal correlates of a perceptual decision. *Nature* 341:52–54. 24, 125

Noudoost B, Chang MH, Steinmetz NA, Moore T (2010) Top-down control of visual attention. *Curr Opin Neurobiol* 20:183–190. 28

Pandarinath C, O'Shea DJ, Collins J, Jozefowicz R, Stavisky SD, Kao JC, Trautmann EM, Kaufman MT, Ryu SI, Hochberg LR, Henderson JM, Shenoy KV, Abbott LF, Sussillo D (2017) Inferring single-trial neural population dynamics using sequential auto-encoders. *bioRxiv* p. 152884. 147

Pang R, Lansdell BJ, Fairhall AL (2016) Dimensionality reduction in neuroscience. *Current Biology* 26:R656–R660. 31

Petreska B, Yu BM, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M (2011) Dynamical segmentation of single trials from population neural data In Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, editors, *Advances in Neural Information Processing Systems 24*, pp. 756–764. Curran Associates, Inc. 38, 148

Rabusseau G, Kadri H (2016) Low-Rank Regression with Tensor Responses In Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, editors, *Advances in Neural Information Processing Systems 29*, pp. 1867–1875. Curran Associates, Inc. 43

Ratcliff R, Rouder JN (1998) Modeling response times for two-choice decisions. *Psychological Science* 9:347–356. 25

Remington ED, Narain D, Hosseini E, Jazayeri M (2018) Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics. *bioRxiv* p. 261214. 141

Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, Miller EK, Fusi S (2013) The importance of mixed selectivity in complex cognitive tasks. *Nature* 497:585–590. 27, 50, 136

Roitman JD, Shadlen MN (2002) Response of Neurons in the Lateral Intraparietal Area during a Combined Visual Discrimination Reaction Time Task. *J. Neurosci.* 22:9475–9489. 25

Rossi-Pool R, Zainos A, Alvarez M, Zizumbo J, Vergara J, Romo R (2017) Decoding a Decision Process in the Neuronal Population of Dorsal Premotor Cortex. *Neuron* 96:1432–1446.e7. 135, 137

Santhanam G, Yu BM, Gilja V, Ryu SI, Afshar A, Sahani M, Shenoy KV (2009) Factor-analysis methods for higher-performance neural prostheses. *J. Neurophysiol.* 102:1315–1330. 37

Seely JS, Kaufman MT, Ryu SI, Shenoy KV, Cunningham JP, Churchland MM (2016) Tensor Analysis Reveals Distinct Population Structure that Parallels the Different Computational Roles of Areas M1 and V1. *PLOS Computational Biology* 12:e1005164. 40, 137, 140, 143, 144

Seung HS (1996) How the brain keeps the eyes still. *PNAS* 93:13339–13344. 29

Shadlen MN, Britten KH, Newsome WT, Movshon JA (1996) A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J. Neurosci.* 16:1486–1510. 25, 29, 125

Shadlen MN, Newsome WT (1996) Motion perception: seeing and deciding. *PNAS* 93:628–633. 25

Shadlen MN, Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol.* 86:1916–1936. 25

She Q, Gao Y, Xu K, Chan RHM (2018) Reduced-Rank Linear Dynamical Systems In *Thirty-Second AAAI Conference on Artificial Intelligence*. 45, 148

Shenoy KV, Sahani M, Churchland MM (2013) Cortical control of arm movements: a dynamical systems perspective. *Annu. Rev. Neurosci.* 36:337–359. 141

Shumway RH, Stoffer DS (1982) An Approach to Time Series Smoothing and Forecasting Using the Em Algorithm. *Journal of Time Series Analysis* 3:253–264. 45

Siegel M, Buschman TJ, Miller EK (2015) Cortical information flow during flexible sensorimotor decisions. *Science* 348:1352–1355. 26, 125, 144

Sussillo D, Churchland MM, Kaufman MT, Shenoy KV (2015) A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience* 18:1025. 141

Tajima S, Koida K, Tajima CI, Suzuki H, Aihara K, Komatsu H (2017) Task-dependent recurrent dynamics in visual cortex. *eLife* 6. 144

Tanji J, Hoshi E (2008) Role of the lateral prefrontal cortex in executive behavioral control. *Physiol. Rev.* 88:37–57. 28

Waldeyer W (1891) Ueber einige neuere Forschungen im Gebiete der Anatomie des Centralnervensystems1). *Dtsch med Wochenschr* 17:1213–1218. 24, 125

Wallis JD, Anderson KC, Miller EK (2001) Single neurons in prefrontal cortex encode abstract rules. *Nature* 411:953–956. 28

Wang J, Narain D, Hosseini EA, Jazayeri M (2018) Flexible timing by temporal scaling of cortical responses. *Nat. Neurosci.* 21:102–110. 141

Wang XJ (2002) Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. *Neuron* 36:955–968. 29, 126, 147

Williams AH, Kim TH, Wang F, Vyas S, Ryu SI, Shenoy KV, Schnitzer M, Kolda TG, Ganguli S (2018) Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron* 98:1099–1115.e8. 43

Williamson RC, Cowley BR, Litwin-Kumar A, Doiron B, Kohn A, Smith MA, Yu BM (2016) Scaling Properties of Dimensionality Reduction for Neural Populations and Network Models. *PLOS Computational Biology* 12:e1005141. 37

Wimmer K, Compte A, Roxin A, Peixoto D, Renart A, de la Rocha J (2015) Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nature Communications* 6:6177. 137

Wu A, Roy NG, Keeley S, Pillow JW (2017) Gaussian process based nonlinear latent structure discovery in multivariate spike train data In Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors, *Advances in Neural Information Processing Systems 30*, pp. 3496–3505. Curran Associates, Inc. 38

Wu CFJ (1983) On the Convergence Properties of the EM Algorithm. *The Annals of Statistics* 11:95–103. 46

Yartsev MM, Hanks TD, Yoon AM, Brody CD (2018) Causal contribution and dynamical encoding in the striatum during evidence accumulation. *bioRxiv* p. 245316. 26

Yu BM, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M (2009) Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *J Neurophysiol* 102:614–635. 37

Zhao Y, Park IM (2017) Variational Latent Gaussian Process for Recovering Single-Trial Dynamics from Population Spike Trains. *Neural Computation* 29:1293–1316. 38

Zhou H, Desimone R (2011) Feature-Based Attention in the Frontal Eye Field and Area V4 during Visual Search. *Neuron* 70:1205–1217. 28, 50