Title

Normative Approaches to the Analysis of Neural Dynamics and Connectivity

Author

Kumar, Ankit

Publication Date

2024

Peer reviewed|Thesis/dissertation

Normative Approaches to the Analysis of Neural Dynamics and Connectivity

by

Ankit Kumar

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Physics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Dr. Kristofer E. Bouchard, Co-chair
Professor Michael DeWeese, Co-chair
Professor Bruno Olshausen
Professor Na Ji

Spring 2024

Normative Approaches to the Analysis of Neural Dynamics and Connectivity

Abstract

Normative Approaches to the Analysis of Neural Dynamics and Connectivity

by

Ankit Kumar

Doctor of Philosophy in Physics

University of California, Berkeley

Dr. Kristofer E. Bouchard, Co-chair

Professor Michael DeWeese, Co-chair

Brain functions, ranging from perception to cognition to action are produced by the collective dynamics of populations of neurons. Our ability to simultaneously record from and map the connectivity between large numbers of neurons across brain areas has increased substantially over the past decade. In contrast, our understanding of the resulting complex and dynamic data in terms of principles of brain computations is lacking. This thesis presents theory and statistical methods that address this gap. I first describe a novel, normative theory of neural population dynamics based on control theory. I introduce novel dimensionality reduction methods that identify subspaces of neural activity that are most amenable to feed-forward (i.e. open-loop) control vs. feedback control (i.e. closed-loop) control. Through new theorems/simulations, I demonstrate that for systems exhibiting non-normal dynamics, generically present in cortex due to Dale's Law, directions most important for feedforward vs. feedback control are geometrically distinct. I then analyze neural recordings from macaque primary motor and somatosensory cortices and show that the dynamics that are most feedback controllable are aligned with those that generate reaching behavior. These feedback controllable dynamics are shown to be mediated by the functional interactions between a population of neurons whose characteristics map to known features of Layer 5 intratellenchephalic neurons. Lastly, I show that feedback controllability provides a normative account for the presence of rotational dynamics in motor cortex. Next, I present an approach an analysis of the *Drosophila* hemibrain connectome using novel maximum entropy models of random graphs inspired from statistical physics. I provide preliminary results indicating that the controllability of *Drosophila* brain networks relies on emergent principles of connectivity between neurons. Finally, I report on work that characterizes the performance of statistical estimation of sparse linear models in the case when model features exhibited correlated variability, a common issue in neural data analysis. The results provide practical guidelines relevant for the estimation of functional connectivity.

Contents

List of Figures

List of Tables

Acknowledgments

This thesis is dedicated first and foremost to my parents, Omlata and Dinesh, who were instrumental in encouraging me to pursue science as a passion first and later as a career, and were unquestioningly supportive of my decision to pursue graduate school. I thank my loving wife, Urmi, who has been an inexhaustible source of respite and companionship during the ups and downs of research. Lastly, this thesis could not have been completed without my loyal dog Lada, who despite not knowing why I was constantly perched in front of my computer these past several years, nonetheless patiently kept watch by my side.

I would like to thank the members of the NSDS Lab that I overlapped with for creating a lively professional environment where scientific discussion and speculation was encouraged. I am particularly appreciative of Jesse Livezey and Sharmodeep Bhattacharyya for being a valuable source of mentorship early on in my time in graduate school. I would like to thank my thesis committee for guiding me through the end of this journey. Finally, I would like to acknowledge the excellent mentorship I received from Kris Bouchard. Kris taught me how to be a scientist through his example and guidance, and encouraged me to pursue a degree of intellectual breadth in my work that few other mentors would have.

Preface

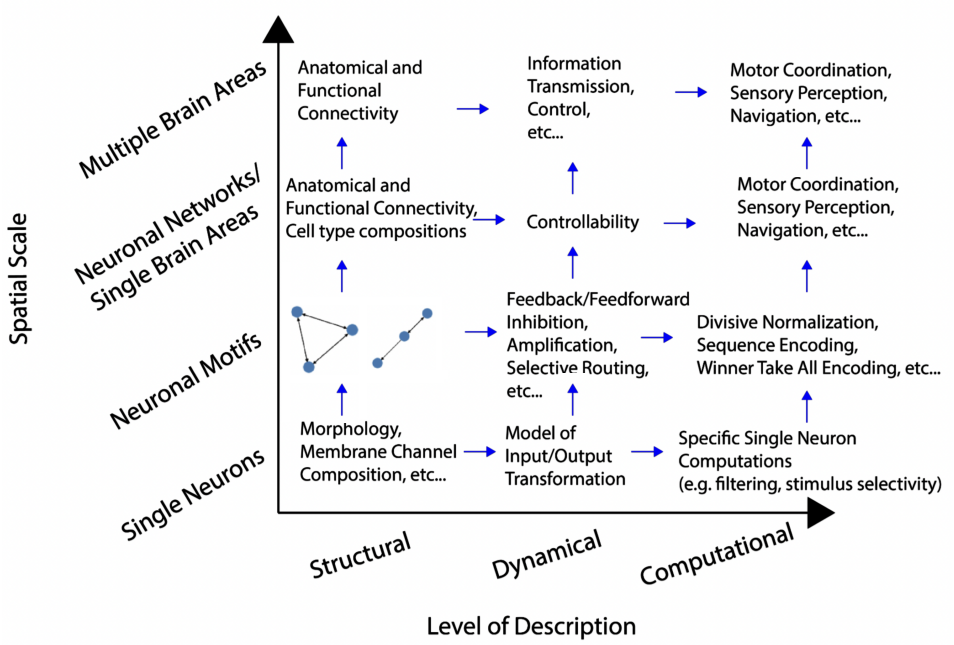
The brain is an enormously complicated system organized hierarchically across spatiotemporal scales. Stubbornly impeding our ability to understand the principles by which the brain functions is that all scales are strongly coupled to each other. Thus, the computations performed by single neurons are inseparable from the biophysical characteristics of channel membranes and synaptic inputs; the dynamics of populations of neurons are shaped by the properties of individual neurons; the coordinated computations performed across distributed brain areas is shaped by the structure of the individual cortical column, and so on. The brain is also replete with emergence, meaning that while phenomena at each scale depends on the details and organization of the lower scales, it is not reducible to these details.

Paralleling this organization across physical levels of description is the fact that phenomena in the brain can be studied at (at least) three distinct analytic levels of description. This framework was most famously expounded by Marr in his seminal work on the vision [2]. Marr differentiated between the computational, algorithmic, and implementational levels of description. Roughly speaking, the computational level of description specifies the function or goal of the system. Taking vision as an example, this level of description would elucidate the features of the external world that the visual system *should* attend to and be able to recognize to facilitate organismal survival. The algorithmic level pertains to *how* the computational goals are solved vis a vis the sequence of transformations and representations of information the system relies upon. In vision, examples include edge detection, object segmentation, and predictive processing. Lastly, the implementational level deals with how the algorthmic level is physically instantiated in the hardware (in this case the circuits of, for example, primary visual cortex).

These three levels of description may be more coarsely partitioned into so-called *normative* theories of brain function and *mechanistic* explanations of brain function. Normative theories pertain to the question of *what* the nervous system is doing and *why* neural circuits or neural activity appears the way it does. Normative theories thus encompass both Marr's computational and algorithmic levels. A central goal of theoretical neuroscience is to construct normative theories which encode the computational goal of brain circuits into a mathematically defined objective function that makes specific, quantitative predictions and postdictions about neural activity. Examples of such normative theories include sparse coding [3], which posits that the goal of the primary visual cortex is to facilitate representations of natural images with sparse activations across neural populations, and the predictive information bottleneck [4], which posits that neural circuits (e.g., in the retina) should encode information about dynamic stimuli only insofar as that information is predictive of the stimuli's future temporal evolution. Mechanistic explanations account for *how* computations are carried out, and therefore encompass the algorithmic and implementational levels. A classical example, again relating to visual perception, is the emergence of spatial tuning in retinal ganglion cells through lateral inhibition [5].

This co-existence of phenomena at these different scales is not only a descriptive fact about the nervous system, but a prescriptive principle by which one may organize scientific

inquiry. In the schematic figure below, I attempt to do this by partitioning topics of study in computational and theoretical neuroscience as they pertain to specific spatial and descriptive levels. I substitute Marr's algorithmic level of description with a dynamical level of description to emphasize that all algorithms implemented by the brain are the result of dynamics that unfold over time. I also emphasize that the arrows indicated in the diagram provide by no means an exhaustive accounting of the conceptual and physical relationships between phenomena, and in fact, a more accurate diagram would likely yield something closer to all-to-all connectivity.



The work of this thesis can be thought of as addressing, in a decidedly modest way, a narrow subset of the scales articulated in this diagram. In particular, in chapter 1, I present a novel, normative theory of neural population dynamics based on the idea of controllability (i.e., the ability of a dynamical system to be controlled). Control theory is fundamentally the study of how one can optimally steer dynamical systems to achieve prescribed functions. Its potential to serve as a core theoretical underpinning for neurobiology goes back to the work of Wiener and the cybernetics movement [6] and ideas surrounding the internal model principle [7]. In the diagram, I indicate controllabillity as a principle at the dynamical level of description that is shaped by the structure of neuronal networks in individual brain areas. Indeed, our results shed light on the role played by Dale's Law (the principle that every neuron in cortex exerts either excitatory or inhibitory effects on its postynaptic targets, but not both) and the ability of neuronal dynamics to be controlled under feedforward/open loop and feedback/closed loop control strategies. We find feedback controllable dynamics to underly the production of reaching behavior, hence bridging the gap to the computational

level. Though not discussed in this thesis, control theory also provides a conceptual framework to understand interactions between brain regions at the dynamical level, as different brain regions may be understood to be trading off the roles of controller and controlled.

The second chapter of this thesis concerns the relationship between local network structure (e.g., neuronal motifs) and the patterns of emergent connectivity across brain regions and indeed, the entire brain. We build models of the *Drosophila* hemibrain connectome using maximum entropy models probability distributions inspired by statistical physics. These probability distributions are the most unstructured distributions over a configuration space (in this case, the possible connectivity patterns between neurons) that are consistent, on average, with a prescribed set of observed statistics. When these statistics are local in nature (e.g., the average pairwise strength in connectivity between different types of neurons), these models provide a means of deciding between the hypothesis that structure at larger scales (e.g. across an entire brain region or neural population) is emergent, or simply a byproduct of structure at the lower scale. We also propose an algorithm to include top-down, global functional constraints on maximum entropy distributions. These constraints in some sense "reverse the arrow" from the dynamical to the structural level of description, allowing one to interrogate how the algorithmic demands placed on neural circuits shape the possible patterns of connectivity they could have exhibited. By identifying where observed connectomes lie within this constrained "network morphospace" [8], this analysis can shed light on the mechanism by which function is achieved and the particular tradeoffs made by biology in doing so. The last chapter of the thesis contains an empirical investigation of the fundamental limits in sparse recovery in linear statistical models, a more practical issue relevant to the estimation of functional connectivity from neural recordings.

Chapter 1

Feedback Controllability as a Normative Theory of Neural Population Dynamics

1.1 Introduction

A purpose of the brain is to produce adaptive behaviors that increase organismal fitness [9]. For complex behaviors, such as finding, identifying, and grasping food, this is accomplished using feedback [10–13]. Feedback (FB) occurs when outputs of a system are routed back as inputs [10], and feedback loops are ubiquitous in the brain. Feedback can be used to correct observed errors in the output of a system relative to a target. For example, the sensory consequences of arm reaches (action) are perceived and used by the brain to control the arm (**Fig. 1.1a**, left). Importantly, multiple aspects of motor coordination underlying reaches are parsimoniously accounted for by optimal feedback control theory [11, 14]. Similar normative accounts based on feedback control (FBC) have been posited for speech production [13], general perception [15], and higher order cognition [12, 16]. Indeed, a plethora of studies have observed impacts of sensory FB on neural recordings [17, 18]. Anatomically, at the brain-systems scale, reciprocal connectivity between brain areas is widespread, providing the requisite anatomical architecture to support FBC [19]. For example, in the context of motor control systems, FB loops between motor cortex and, e.g., the somatosensory cortex [20] and thalamus [21] may support control of activity within motor cortex (**Fig. 1.1a**, center). Similar anatomical considerations hold for high-order perception and cognitive systems as well [22]. Finally, at columnar scales, FB connections permeate canonical cortical microcircuits, with highly recurrent connectivity between cell types as well as asymmetric connectivity between types, giving rise to microcircuit feedback loops [23] (**Fig. 1.1a**, right). Therefore, while there is overwhelming evidence that brains use feedback (FB), the implications of feedback control (FBC) for the dynamics of neural population data has not been directly considered.

Colloquially, the different configurations of activity patterns across a neural population define its state-space [24]. Control of a neural population can be defined as steering the neural population from (any) initial state to a desired state by an external input. Examples of neural population control could include ensuring that motor cortex generates the required neural population dynamics to produce movement [25, 26], updating representations (i.e., neural states) of faces in inferior temporal cortices [27], or updating representations of animal location in the hippocampus/medial prefrontal cortex [28]. In this context, controllability denotes the ability to control a neural population. Controllability is a graded quantity, and control theory considers the cost of control: loosely speaking, how much "energy" is required to control the system [29, 30]. Crucially, controllability is an intrinsic property of the neural population– if the (unobserved) inputs to the neural population are probing the neural state space sufficiently, the controllability of neural populations can be assayed from observations of neural population dynamics itself, without knowing what the inputs or outputs are [31]. Thus, with appropriate methodology, we should be able to assess the controllability of a neural population just from neurophysiological recordings of that population.

Control theory distinguishes between systems that do not use feedback to correct errors (**Fig. 1.1b**, feedforward control, 'FFC', i.e., open-loop control, e.g., eye-blink reflex) and those that do (**Fig. 1.1c**, feedback control, 'FBC', i.e., closed-loop control, e.g., arm reaches). Feedforward control (FFC) of a neural population requires an internal model of that neural population, and deviations from target neural states not predictable by the internal model can not be corrected, i.e., controlled. In contrast, feedback control (FBC) of neural populations would enable correcting errors relative to a target neural state based on observations of those errors [11]. [For ease of exposition, we use FFC/FBC to refer to feedforward and feedback control/controlability interchangeably]. A central benefit of FBC relative to FFC is to enable robust control of a neural population in the presence of unpredicted perturbations and/or noise. FBC is not always possible. For example, in the neural control of reaching there is thought to be an initial FFC phase (proprioceptive FB notwithstanding) before visual processing has time to impact behavior, followed by a FBC phase which utilizes visual FB to guide the arm [32]. Given these differences, systems may be more FFC vs. FBC, and complex systems like the brain can simultaneously be both FFC and FBC to varying degrees.

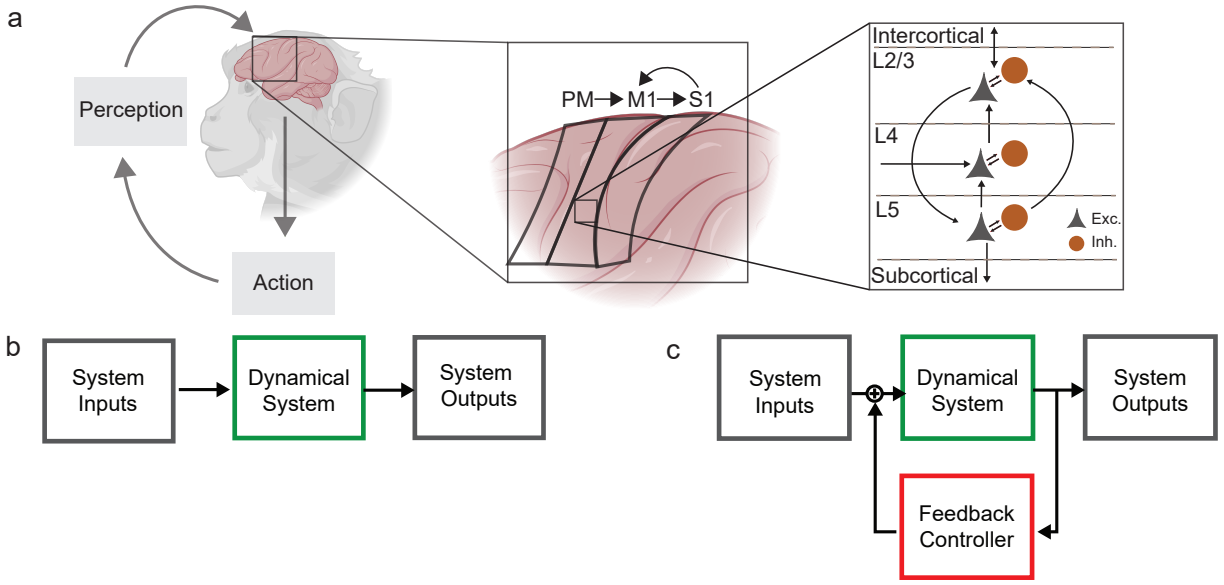


Figure 1.1: Neural population dynamics generating diverse naturalistic behaviors are produced by feedback loops across spatiotemporal scales. (a) Feedback loops are ubiquitous in the nervous system across spatiotemporal scales. At the highest level, there is feedback between perception and action by the organism (left). (Center) The architecture of interactions between distributed brain areas also contains many feedback loops. A prominent example of this architecture is provided by the circuits underlying reaching, which are comprised of both feedforward pathways (premotor (PM) to primary motor (M1) to somatosensory (S1) cortices), but crucially also feedback pathways from sensory to motor areas that enable online error correction of behavior (e.g., S1 $\rightarrow$ M1). (Right) At the scale of individual cortical columns, recurrent connections within canonical microcircuits are prolific, providing an anatomical substrate for feedback control at local scales. (b,c) A dynamical system may be controlled either in a feedforward or open loop sense (b), where system inputs (left grey box) drive a dynamical system (green box) without dependence on system outputs (right grey box), or in a feedback or closed loop sense (c), where system outputs can be used by a feedback controller (red) to modify system inputs.

Here, we test the hypothesis that feedback controllability (FBC) is a normative theory of neural population dynamics. To do so, we developed novel machine learning methods to identify directions (i.e., extract subspaces) in high-dimensional neural population data that are most FBC and compared those to directions (subspaces) that are most feedforward controllable (FFC). We demonstrate that the neuro-anatomical constraint on synaptic connectivity imposed by Dales law generates neural population dynamics for which the FBC and FFC subspaces are distinguishable. In neural population data previously recorded from primary motor (M1) and somatosensory (S1) cortex of monkeys performing a reaching task,

we found that FBC subspaces were substantially better decoders of reach kinematics. Furthermore, we demonstrate that FBC and FFC are mediated by populations of single neurons with distinct neural activity profiles, and that FBC is an emergent property of the neural population that is highly dependent on neuronal interactions. Finally, we show that FBC and FFC engage distinct types of dynamics, and that this maybe a consequence of avoiding dynamic instabilities.

1.2 Results

Controllable Subspaces of Neural Population Dynamics

Our goal was to test the hypothesis that feedback (as opposed to feedforward) controllability is a normative theory of neural population dynamics. A common characteristic of high-dimensional neural population data is that it can be succinctly described by a projection into a lower-dimensional subspace. Such lower dimensional projections can provide compact descriptions of the high-dimensional data that are easier to visualize and understand [33]. As such, to test our hypothesis, we developed dimensionality reduction methods that operate on simultaneously recorded multiple single-neuron neurophysiology data commonly acquired by, e.g., Utah arrays. These dimensionality reduction methods take high-dimensional, neural population firing-rate time-series data (where each dimension is a single-neuron) and extract a lower-dimensional 'subspace'. As described below, the subspaces we extract from neural data maximize the controllability of neural population dynamics under feedforward (**Fig. 1.2**, FFC subspaces, black lines) and feedback (**Fig. 1.2**, FBC subspace, red lines) control schemes. This allowed us to directly compare and contrast the properties of neural population dynamics under these different normative principles.

In its simplest form (i.e., a linear stochastic dynamical system), the dynamics of a neural population can be described as:

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1.1}$$

Here $x(t)$ is a vector describing the firing rate of all neurons (the high-dimensional, 'neural state'), $\dot{x}(t)$ is its time derivative, A is a matrix describing the interactions of the neurons, and $u(t)$ is a stochastic external input (i.e., control signal) that is mapped on to the neural state by the matrix B. A lower dimensional projection ($y(t)$) of the high-dimensional neural data ($x(t)$) is

$$y(t) = Cx(t) \tag{1.2}$$

Where the dimensionality (d) of the projected neural data y is substantially less than the number of neurons. The subspace into which the neural data is projected is given by the

matrix C, which is extracted from the neuronal firing rates so that the lower-dimensional projection maximizes some quantity of the original, high-dimensional neural data. We specifically considered the controllability of neural population dynamics under feedforward (**Fig. 1.2**, grey/black) and feedback control schemes (**Fig. 1.2**, pink/red) by deriving quantities that correspond to feedforward and feedback controllability. Importantly, these metrics are intrinsic quantities of the observed neural population data, and do not require knowledge of the inputs and outputs of the neural population (**Fig. 1.2**).

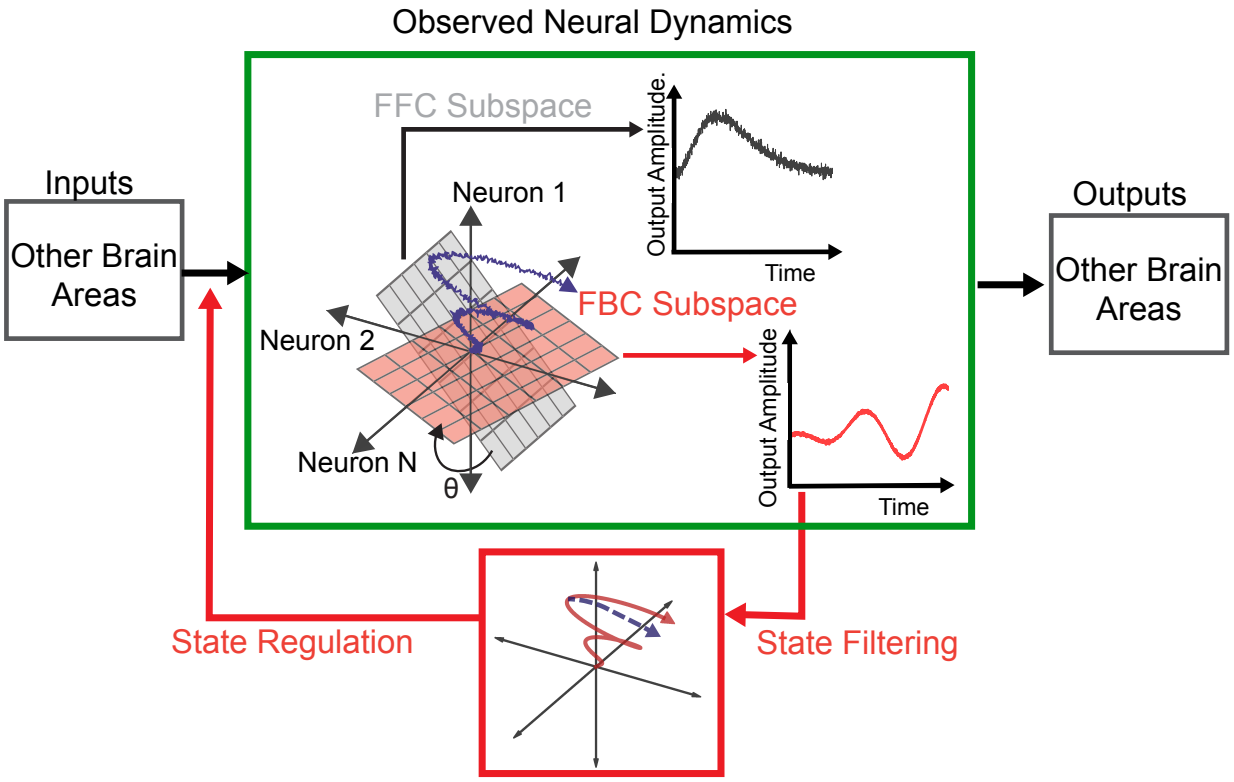


Figure 1.2: Controllable Subspaces of Neural Population Dynamics The firing rates of observed neural dynamics lie in a high dimensional state space (visualized here in 3 dimensions). This high dimensional space can be segregated into subspaces in which dynamics are most feedforward controllable (FFC) and feedback controllable (FBC). Noisy system inputs combine with recurrent dynamics to produce noisy firing rates (blue trace). FFC subspaces (black 1D trace) contain high variance activity that amplifies both signal and noise. FBC subspaces (red 1D trace) contain activity that produces the most accurate, denoised reconstructions of the full state trajectory (red high dimensional trace) through a state filtering step and is most enable to state regulation via feedback (dashed blue trace). Dale's law gives rise to a finite angle (θ) between FFC and FBC subspaces.

In line with prior work in control theory [34], we defined feedforward controllability (FFC) as the "volume of neural state space" that is obtainable by an input control signal.

Put another way, feedforward controllable (FFC) subspaces of neural population dynamics are those within which inputs generate the highest amplitude firing rates (**Fig. 1.2**, FFC subspace). Below, we show that the most feedforward controllable subspace of dimension d for a linear dynamical system coincides exactly with the d dimensional subspace that maximizes variance, and is thus given by PCA. Intuitively, the FFC subspace will coincide with directions of state space that give rise to large amplification of firing rates in response to inputs, as this will maximize the volume of neural state space that the input can explore. However, this amplification will indiscriminately act on both desired signal and undesirable noise (**Fig. 1.2**, black trace in upper right).

In contrast to FFC, the extent to which neural population dynamics can be controlled via feedback depends on two functional stages. First, due to the presence of noise in the neural population activity, the neural state dynamics must be reconstructed from observations provided by the feedback controllable (FBC) subspace (depicted as the denoised red trace in **Fig. 1.2** bottom). Second, an appropriate control signal must be synthesized from these reconstructed dynamics and routed back into the FBC subspace (**Fig. 1.2** bottom, purple dashed trace). We created a novel linear dimensionality reduction method that extracts subspaces of neural population dynamics that simultaneously minimize the cost of state filtering and state regulation (**Fig. 1.2**, red), and thus maximize feedback controllability (Feedback Controllable Components Analysis, FCCA.

Correspondence between Feedforward Controllability and PCA

A categorical definition of controllability for a dynamical system is that for any desired trajectory from initial state to final state, there exists, in principle, a control signal that could be applied to the system to guide it through this trajectory. For a (stable) linear dynamical system, a necessary and sufficient condition for this to hold is that the controllability Gramian, Π, have full rank. Π is obtained from the state space parameters through the solution of the Lyapunov equation:

$$A\Pi + \Pi A^\top = -BB^\top \quad \Pi = \int_0^\infty dt\, e^{At} BB^\top e^{A^\top t} \tag{1.3}$$

The rank condition on Π as a definition of controllability, while canonical [35], is an all or nothing designation; either all directions in state space can be reached by control signals, or they cannot. Furthermore, this definition does not take into account the energy required to achieve the desired transition. While certain directions in state space may in principle be reachable, the energy required to push the system in those directions may be prohibitive.

Thus, given that the system is controllable, we can ask a more refined question: what is the energetic effort required to control different directions of state space? The energy required for control is measured by the norm of the input signal $u(t)$. It can be shown that to reach states that lie along the eigenvectors of Π, the optimal (i.e., minimal) energy is proportional to the inverse of the corresponding eigenvalues of Π. Directions of state space

that have large projections along eigenvectors of Π with small eigenvalues are therefore harder to control. For a unit norm input signal, the volume of reachable state space is proportional to the determinant of Π [34].

We can encode the above intuition into the objective function of a dimensionality reduction problem: for a fixed norm input signal, what choice of C maximizes the reachable volume within the subspace? This volume is measured by the determinant of $C\Pi C^\top$. Identifying subspaces of maximum feedforward controllability is then posed as the following optimization problem-

$$\text{argmax}_C \log\det C\Pi C^\top \tag{1.4}$$

In order to assess this objective in data, we make the further assumption that the dynamics of $x(t)$ are stationary and that the inputs $u(t)$ can be approximated by temporally white noise. In this case, the observed covariance of the data will coincide with the controllability gramian [36]. When Π is the steady state covariance of $x(t)$, the optimization problem 1.4 coincides with the objective function of PCA, as the optimal C of fixed dimensionality d is given by the top d eigenvectors of Π (see Proposition 1 in Section S.1.9).

LQG Singular Values measure feedback controllability

How do we quantify the feedback controllability of a system? The primary distinction between feedforward (i.e., open loop) control (FFC) and feedback (i.e., closed loop) control (FBC) is that FBC utilizes observations of the state to synthesize subsequent control signals. Feedback control therefore involves two functional stages: filtering (i.e., estimation) of the underlying dynamical state ($x(t)$) from the available observations ($y(t)$) and construction of appropriate regulation (i.e., control) signals. For a linear dynamical system, state estimation is optimally accomplished by the Kalman filter, whereas state regulation is canonically achieved via linear quadratic regulation (LQR). It will be crucial in what follows to recall that the Kalman Filter accomplishes nothing more than an efficient, recursive, Gaussian minimum mean square error (MMSE) estimate of $x(t)$ given observations $y(\tau)$ for $\tau \leq t$. These two functional stages optimally solve the following cost functions:

$$\text{Kalman Filter}: \min_{p(x_0|y_{-T:0})} \lim_{T\to\infty} \text{Tr}\left(\mathbb{E}\left[(\mathbb{E}(x_0|y_{-T:0}) - x_0)(\mathbb{E}(x_0|y_{-T:0}) - x_0)^\top\right]\right)$$

$$\text{LQR}: \min_{u\in L^2[0,\infty)} \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T x^\top C^\top C x + u^\top u \, dt\right]$$

where $y_{-T:0}$ denotes observations over the interval $[-T, 0]$. The minima of these cost functions are obtained from the solutions of dual Riccati equations:

$$AQ + QA^\top + BB^\top - QC^\top CQ = 0 \tag{1.5}$$

$$A^\top P + PA + C^\top C - PBB^\top P = 0 \tag{1.6}$$

where

$$Q = \lim_{T\to\infty} \min_{p(x_0|y_{-T:0})} \mathbb{E}\left[(\mathbb{E}(x_0|y_{-T:0}) - x_0)(\mathbb{E}(x_0|y_{-T:0}) - x)^\top\right]$$
$$x_0^\top P x_0 = \min_{u\in L^2[0,\infty)} \left\{ \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T x^\top C^\top C x + u^\top u \, dt\right], \;\; x(0) = x_0, u \in L^2[0,\infty)\right\}$$

Q therefore is the covariance matrix of the estimation error, whereas P encodes the regulation cost incurred for varying initial conditions (x_0). For example, the operator norm of $P : \sup_x ||Px||/||x||$ bounds the initial condition that yields the worst case regulation cost. On the other hand, $\mathrm{Tr}(P)$ is proportional to the average regulation cost over all unit norm initial conditions.

The solutions of the Riccati equations are not invariant under the invertible state transformation $x \to Tx$. The filtering Riccati equation will transform as $Q \to TQT^\top$ whereas P will transform as $(T^{-1})^\top P T^{-1}$. As such, simply by defining new coordinates via T we can shape the difficulty of filtering and regulating various directions of the state space. Therefore Q and P on their own are not suitable cost functions for measuring feedback controllability. However, we notice that the product PQ undergoes a similarity transformation $PQ \to (T^\top)^{-1} Q P T^\top$. Hence, the eigenvalues of PQ are invariant under similarity transformations, and define an *intrinsic* measure of the feedback controllability of a system. Additionally, there exists a particular T that diagonalizes PQ. Following [37], we refer to the corresponding eigenvalues as the LQG (Linear Quadratic Gaussian) singular values. In this basis, the cost of filtering each direction of the state space equals the cost of regulating it. We formalize these statements by restating Theorem 1 from [37]:

Theorem 1 *Let (A, B, C) be a minimal realization of $G(s)$. Then, the eigenvalues of QP are similarity invariant. Further, these eigenvalues are real and strictly positive. If $\mu_1^2 \geq \mu_2^2 \geq \mu_n^2 > 0$ denote the eigenvalues of QP in decreasing order, then there exists a similarity transformation T, $(A, B, C) \to (TAT^{-1}, TB, CT^{-1}) \equiv (\tilde{A}, \tilde{B}, \tilde{C})$ such that:*

$$Q = P = diag(\mu_1, \mu_2, ..., \mu_n)$$

The realization $(\tilde{A}, \tilde{B}, \tilde{C})$ will be called the closed-loop balanced realization.

Proof: Let $Q = LL^\top$ be the Cholesky decomposition of Q and let $L^\top P L$ have Singular Value Decomposition $U\Sigma^2 U^\top$. Then, $T = \Sigma^{1/2} U^\top L^{-1}$ provides the desired transformation:

$$TQT^\top = \Sigma^{1/2} U^\top L^{-1} L L^\top (L^\top)^{-1} U \Sigma^{1/2} = \Sigma$$
$$(T^{-1})^\top P T^{-1} = \Sigma^{-1/2} U^\top \underbrace{L^\top P L}_{U\Sigma^2 U^\top} U \Sigma^{-1/2} = \Sigma$$

□

Hence, as an intrinsic measure of feedback controllability, we take the sum of the LQG singular values μ_i^2, corresponding to the sum of the ensemble cost to filter and regulate each direction of the neural state space:

$$\text{FBC} : \text{Tr}(PQ) \tag{1.7}$$

The Feedback Controllability Components Analysis Method.

Our goal was to develop a dimensionality reduction method, Feedback Controllability Components Analysis (FCCA), that can be readily applied to observed data from typical systems neuroscience experiments. To do so, we constructed estimators of the LQG singular values, and hence $\text{Tr}(PQ)$, directly from the autocorrelations of the observed neural firing rates. The objective function for FCCA arises from the observation that causal Kalman filtering and acausal Kalman filtering are also related via dual Riccati equations. We will first show that through an appropriate variable transformation, we obtain a state variable $x_b(t)$ whose dynamics unfold backwards in time via the same dynamics matrix (A) which evolves $x(t)$ (the neural state) forwards in time. Once established, this fact enables us to use the error covariance matrix of Kalman filtering $x_b(t)$ as a stand-in for the cost of regulating $x(t)$.

In particular, if we have a state space realization of a forward time stochastic linear system (eq. 1.1), then the joint statistics of $(x(t), y(t))$ can be parameterized by a Markovian model that evolves backwards in time [38]:

$$\begin{aligned} -\dot{x}(t) &= A_b x(t) + Bu(t) \\ y &= Cx(t) \end{aligned} \tag{1.8}$$

where $A_b = -A - BB^\top \Pi^{-1} = \Pi A^\top \Pi^{-1}$ and $\Pi = \mathbb{E}[x(t)x(t)^\top]$.

Examination of eq. 1.5 and eq. 1.6 reveals that the filtering and LQR Riccati equations differ primarily in 2 respects. First, the dynamics matrix is transposed ($A \to A^\top$), and second the inputs and outputs have been exchanged ($B \to C^\top$, $C \to B^\top$). To use the error covariance of state filtering as a stand-in for the state regulation cost, we therefore require that the corresponding acausal state dynamics (determined by A_b) respect these differences. To this end, consider the transformed state $x_a(t) = \Pi^{-1}x(t)$. Substituting $x(t) = \Pi x_a(t)$ and $A_b = \Pi A^\top \Pi^{-1}$ into the equations for the backward dynamics result in following dynamics for this *adjoint* state:

$$-\dot{x}_a(t) = A^\top x_a(t) + \Pi^{-1} B u(t)$$

Then, if we construct a readout of this transformed state $y_a(t) = C\Pi x_a(t) = Cx(t)$, the Riccati equation associated with Kalman filtering x_a, whose solution we denote $\tilde{P}$, takes on the form:

$$A^\top \tilde{P} + \tilde{P}A + \Pi^{-1}BB^\top\Pi^{-1} - \tilde{P}\Pi C^\top C\Pi\tilde{P} = 0 \tag{1.9}$$

$$A^\top P + PA + C^\top C - PBB^\top P = 0 \tag{eq 1.6}$$

We see that eq. 1.9 coincides with eq. 1.6 (reproduced for convenience) upon switching the inputs and outputs ($B \to C^\top$, $C \to B^\top$, as with 1.6 and 1.6) and reweighting them by a factor of Π^{-1} and Π, respectively. In fact, eq. 1.9 coincides with the Riccati equation associated with a slightly modified LQR problem:

$$\min_{u\in L^2[0,\infty)} \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T x^\top\Pi^{-1}BB^\top\Pi^{-1}x + u^\top\Pi^2 u\, dt\right] \tag{1.10}$$

This is the regulator problem for the adjoint state $x_a(t) = \Pi^{-1}x(t)$. Therefore, under the assumption that the observed dynamics can be approximated by a linear dynamical system, **we can measure LQG singular values associated with this modified LQR problem directly from measuring the causal minimum mean square error (MMSE) associated with prediction of** $x(t)$**, and the acausal MMSE associated with prediction of** $x_a(t)$.

To explicitly construct an estimator of the quantity $\mathrm{Tr}(\tilde{P}Q) = \mathrm{Tr}(Q\tilde{P})$, we recall the standard formulas for the error covariance of MMSE prediction of a Gaussian distributed variable u given v: $\Sigma_u - \Sigma_{uv}\Sigma_v^{-1}\Sigma_{vu}^\top$ where $\Sigma_u = \mathbb{E}[uu^\top], \Sigma_v = \mathbb{E}[vv^\top]$ and $\Sigma_{uv} = \mathbb{E}[uv^\top]$. The matrix Q is the error covariance of MMSE prediction of the system state $x(t)$ given past observations $y(t)$ over the interval $(t-T, t)$, whereas the matrix $\tilde{P}$ is the error covariance of MMSE prediction of the transformed system state $x_a(t)$ given future observations $y_a(t)$ over the interval $(t, t+T)$. As discussed above, the Kalman Filter is used to efficiently calculate these MMSE estimates given an explicit state space mode of the dynamics. In our case, to keep system dynamics implicit, we instead directly use the formulas for the MMSE error covariance in terms of cross correlations between $x(t), x_a(t)$ and $y(t), y_a(t)$. This gives rise to the FCCA objective function:

$$\text{FCCA}: \quad \mathrm{argmin}_C \mathrm{Tr}\left[\underbrace{\left(\Pi - \Lambda_{1:T}(C)\Sigma_T^{-1}(C)\Lambda_{1:T}^\top(C)\right)}_{\text{causal MMSE covariance } (Q)}\underbrace{\left(\Pi^{-1} - \tilde{\Lambda}_{1:T}^\top(C)\Sigma_T^{-1}(C)\tilde{\Lambda}_{1:T}(C)\right)}_{\text{acausal MMSE covariance } (\tilde{P})}\right] \tag{1.11}$$

where for discretization timescale τ,

$$
\begin{aligned}
\Pi &= \mathbb{E}[x(t)x(t)^\top] \quad \text{(covariance of the neural data)} \\
\Lambda(C)_{1:T} &= \{\Lambda_1 C^\top, \Lambda_2 C^\top, ..., \Lambda_T C^\top\} \\
\Lambda_k &= \mathbb{E}[x(t+k\tau)x(t)^\top] \quad \text{(autocorrelation of the neural data)} \\
\tilde{\Lambda}(C)_{1:T} &= \{\tilde{\Lambda}_1 \Pi C^\top, \tilde{\Lambda}_2 \Pi C^\top, ..., \tilde{\Lambda}_T \Pi C^\top\} \\
\tilde{\Lambda}_k &= \mathbb{E}[x_a(t+k\tau)x_a(t)^\top] \quad \text{(autocorrelations of the adjoint state)} \\
\Sigma_T(C) &= \underbrace{\begin{bmatrix} C\Lambda_0 C^\top & C\Lambda_1 C^\top & C\Lambda_2 C^\top & ... & C\Lambda_T C^\top \\ C\Lambda_1^\top C^\top & C\Lambda_0 C^\top & C\Lambda_1 C^\top & ... & C\Lambda_{T-1} C^\top \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ C\Lambda_T^\top C^\top & C\Lambda_{T-1}^\top C^\top & C\Lambda_{T-2}^\top C^\top & ... & C\Lambda_0 C^\top \end{bmatrix}}_{\text{(space by time covariance of y(t))}}
\end{aligned}
$$

Control-Theoretic Intuition for FCCA

We have shown how the LQG singular values are an intrinsic measure of the cost to filter/regulate a linear dynamical system. We now provide a further intuition for FCCA. In order to control the system state dynamics, the controller itself must implement its own, internal, state dynamics. These dynamics carry out the computations necessary to perform reconstruction of the systems state and synthesis of the required regulator signal. Thus, in addition to the complexity of the system itself, we may inquire about the complexity of the controller. One intuitive measure of this complexity is given by the controller's state dimension (i.e., the McMillan degree), which corresponds to the number of dynamical degrees of freedom it must implement to function. As these controller degrees of freedom must ultimately be implemented via networks of neurons within the brain, it stands to reason that biology may favor performing task-relevant computations via dynamics that require low-dimensional controllers. As we argue below, minimizing the LQG singular values over readout matrices (C) corresponds to a relaxation of the objective of searching for a subspace that enables control via a controller of low dimension. In other words, feedback controllable dynamics can be regulated with controllers of low internal dimensionality.

We recall from Theorem 1 above that there exists a linear transformation that simultaneously diagonalizes both P and Q. Let $(\tilde{A}, \tilde{B}, \tilde{C})$ be the corresponding balanced realization. Let us order the LQG singular values in descending magnitude $\{\mu_1, ..., \mu_N\}$ and divide them into two sets $\{\mu_1, ..., \mu_m\}$ and $\{\mu_{m+1}, ..., \mu_N\}$. Let us assume the system input is of dimensionality p and the output is of dimensionality d (i.e. $\tilde{B} \in \mathbb{R}^{N \times p}$ and $\tilde{C} \in \mathbb{R}^{d \times N}$).Then, one can partition the state matrices $\{\tilde{A}, \tilde{B}, \tilde{C}\}$ accordingly.

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$
$$\tilde{B} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \quad \tilde{C} = \begin{bmatrix} C_1 & C_2 \end{bmatrix}$$

Where $A_{11} \in \mathbb{R}^{m \times m}, A_{22} \in \mathbb{R}^{N-m \times N-m}$, $B_1 \in \mathbb{R}^{m \times p}, B_2 \in \mathbb{R}^{N-m \times p}$, $C_1 \in \mathbb{R}^{d \times m}, C_2 \in \mathbb{R}^{d \times N-m}$. It can be shown that the optimal controller of dimension m is obtained from solving the Riccati equations corresponding to the truncated system (A_{11}, B_1, C_1). If the neglected LQG singular values $\{\mu_{m+1}, ..., \mu_N\}$ are small, then the controller dimension can be reduced with esssentially no loss in regulation performance.

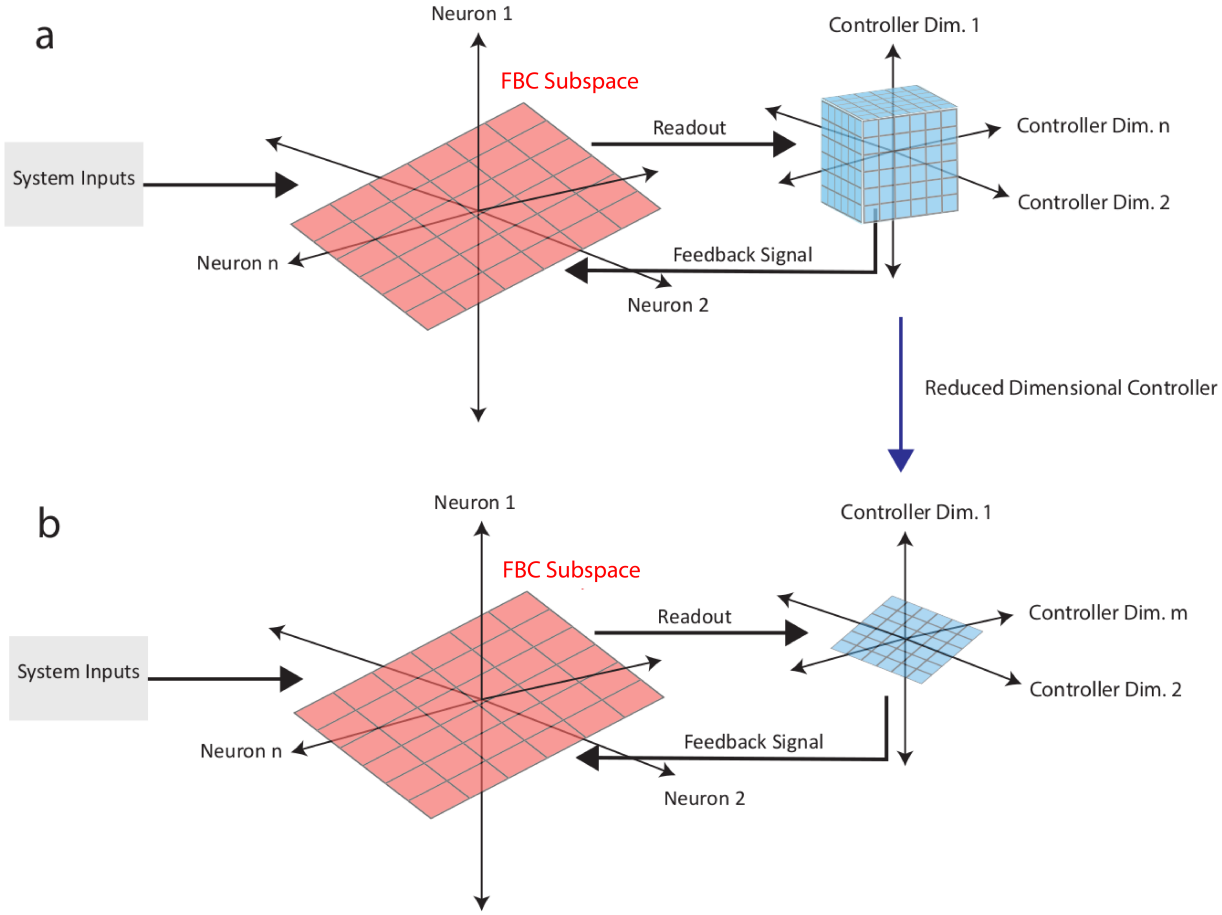


Figure 1.3: In principle, a controller of dimension as large as the neural state space may be required to effectively regulate dynamics within a FBC subspace (a). However, subspaces optimized to minimize either the rank, or more practically, the trace of PQ will require controllers of lower dimensionality to achieve near-optimal performance (b).

In the case when the singular values $\{\mu_{m+1}, ..., \mu_n\}$ are all precisely zero, then the controller dimensionality could be reduced from n to m with no loss in LQG performance. We illustrate this idea schematically in **Figure 1.3**, where the controller state space dimension (blue) is truncated. This suggests that to search for subspaces of neural dynamics that require low dimensional controllers to regulate, we minimize the following objective function:

$$\text{argmin}_C \text{Rank}(\tilde{P}Q)$$

where $\tilde{P}$ and Q are the solutions to the Riccati equations 1.9 and 1.5, respectively. However, rank minimization is an NP hard problem. A convex relaxation of the rank function is the nuclear norm (i.e. the sum of the singular values) [39]. Given that $\tilde{P}Q$ is a positive semi-definite matrix, a tractable objective function that seeks subspaces of dynamics that require low complexity controllers is therefore given by:

$$\text{argmin}_C \text{Tr}(\tilde{P}Q)$$

which is precisely what FCCA minimizes in a data-driven fashion.

Dale's law enables distinguishing feedforward and feedback controllable subspaces.

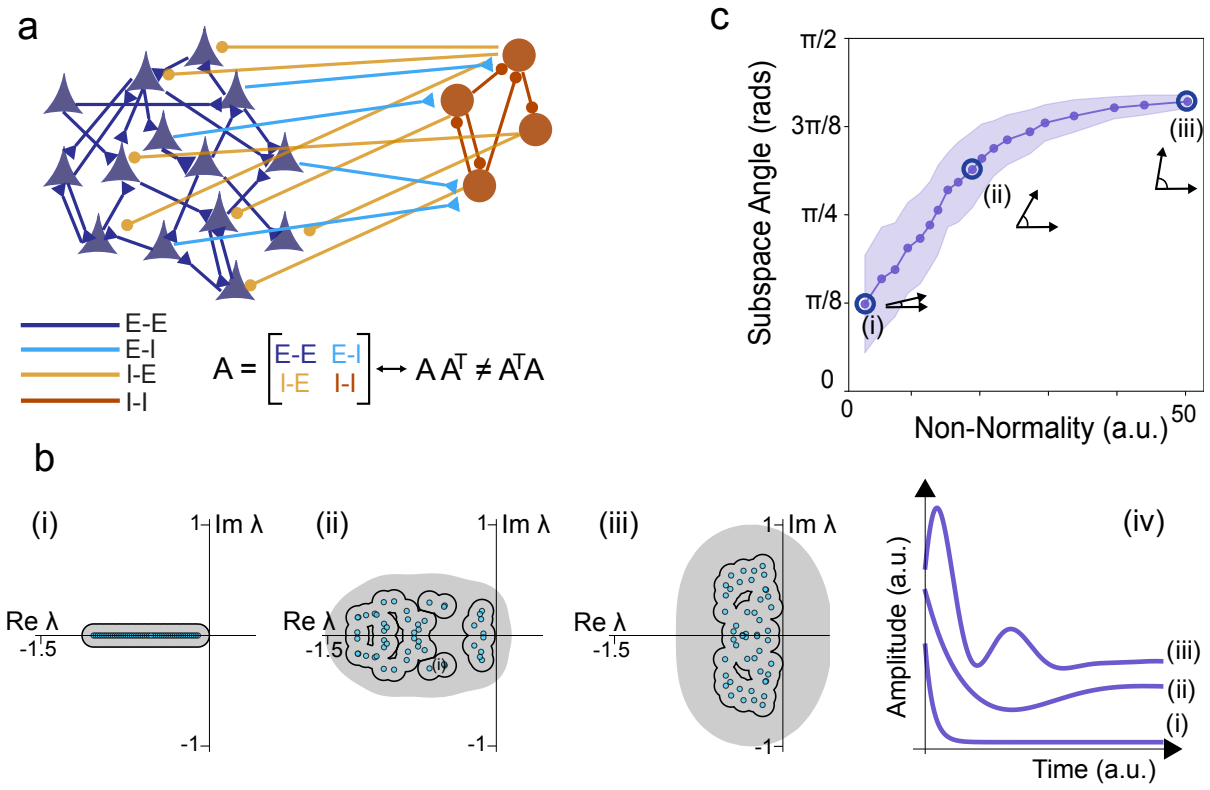


Figure 1.4: (a) Dale's law imposes block constraints on excitatory/excitatory (E-E), excitatory/inhibitory (E-I, I-E) and inhibitory/inhibitory (I-I) connectivity that give rise to a non-normal synaptic connectivity matrix, A. (b, left) Depiction of the eigenvalues (blue scatter points) and pseudospectral contours (boundary of light grey shading) of a synthetic synaptic connectivity matrix A as the degree of non-normality of A is increased from (i) to (iii). Dark contours indicate the pseudospectral contours expected from a normal matrix with matched eigenvalue. (b, right) Time courses of dynamics of the systems depicted in (i)-(ii) projected along the leading PCA directions. (c) Plot of the mean (standard deviation) subspace angle between FFC and FBC subspaces of $d = 6$ vs. non-normality in synthetic linear dynamical systems. Statistics are calculated across 100 repetitions at each level of non-normality.

With data-driven methods for extracting FBC and FFC subspaces, we next sought to understand if, and under what dynamical conditions, these two subspaces differ (i.e., when is the angle θ indicated in **Fig. 1.2** large?). The anatomical structure of neural circuits plays an integral role in shaping neural population dynamics. In cortex, chief amongst these structures is Dale's Law [40], which requires every neuron to exert either excitatory or inhibitory effects on its post-synaptic target, but not both. As such, the synaptic connectivity matrix that

describes the interaction of different neurons (which corresponds to the A matrix in our context) is constrained to have the same sign within each row. Combined with the directed and asymmetric nature of recurrent connectivity, constrains the neural dynamics matrix A to belong to a special class of matrices, so called non-normal matrices (i.e., $AA^\top \neq A^\top A$) (**Fig. 1.4a**) [41, 42]. Systems that evolve according to non-normal matrices exhibit interesting collective dynamics, such as transient amplification [41, 43], extensive memory traces of input signals [44, 45], and efficient information transmission [46]. As we show below, a new control theoretic result of our work is that FBC subspaces and FFC subspaces are different for systems with non-normal dynamics.

The spectrum (i.e., eigenvalues) of the dynamics matrix A determine the long-term collective dynamics produced by a neural population. The real part of eigenvalues describe the strength of growth or decay of the collective dynamics, while the imaginary part describes the strength of oscillations of those dynamics. However, an important property of systems driven by non-normal matrices (e.g. recurrent networks constrained by Dale's Law) is that their short term dynamics can differ from their long-term dynamics. The pseudo-spectrum provides an important mathematical tool to understand the qualities of non-normal systems [42]. The pseudo-spectrum of A identifies regions of the complex plane that, over short time periods, behave like eigenvalues of A. More specifically, the ϵ-pseudospectrum of an n-dimensional square matrix A, λ_ϵ, refers to all values in the complex plane that are eigenvalues of matrices which are ϵ-close to A as measured by any matrix norm. In other words, the set λ_ϵ contains all complex numbers z for which there exists a matrix $E, ||E|| \leq \epsilon$ such that z is an eigenvalue of $A + E$. From this definition, it can be seen that the ordinary eigenvalues coincide with the 0-pseudospectrum. For a normal matrix, the ϵ-pseudospectrum is straightforward to determine: it is given by the union of circles with radius ϵ around the eigenvalues. Deviation from these regular contours gives an indication of the degree to which short term dynamics differ that predicted by the system eigenvalues.

The precise distribution of the pseudospectrum around the eigenvalues can account for many of the nonintuitive behavior of non-normal dynamical systems [42]. For example, a lower bound on the maximum amplitude attained by the time evolution of a state vector x within a linear dynamical system is given by $\max_{\lambda_\epsilon}(\mathrm{Re}(\lambda_\epsilon))/\epsilon$. Thus, when the ϵ-pseudospectrum extends greater than an amount ϵ into the right hand of the complex plane (as in **Fig. 1.4b(iii)**), the corresponding linear dynamical system will exhibit transient amplification.

To build intuition for these concepts, we turn to numerical simulations of systems of the form of eq 1.1 with $B = I$. We sampled 50 dimensional A matrices with i.i.d entries above the diagonal assigned according to $C_{ij}J_{ij}$ where $C_{ij} \sim$ Bernoulli(0.5) and $J_{ij} \sim \mathcal{N}\left(0, \frac{1}{2\sqrt{50}}\right)$. The entries below the main diagonal were given by $A_{ji} = \alpha A_{ij}$. Reducing α from 1 to 0 therefore allows one to systematically tune the asymmetry, and consequentially, the non-normality of A. We set the diagonal elements uniformly to the smallest negative number that ensured system stability. We quantitatively measure non-normality by the Henrici metric: $||AA^\top - A^\top A||_F$ where $||\cdot||_F$ is the Frobenius norm. This metric is zero if and only if A is

normal (i.e. $A^\top A = AA^\top$).

In **Figure 1.4b**, we depict the spectrum (i.e., eigenvalues, small blue circles), $\epsilon = 0.1$ pseudo-spectrum (grey-shaded region), and pseudo-spectral contors expected from an equivalent normal matrix of three dynamical systems with increasing non-normality ((i) $\rightarrow$ (ii) $\rightarrow$ (iii)), as well as the corresponding system dynamics (iv). For the normal A matrix, the eigenvalues are purely real and negative (i.e., the live on the negative x-axis, **Fig. 1.4b**, (i)), giving rise to purely exponential decaying dynamics (curve (i) in **Fig. 1.4b**, (iv)). As the non-normality in A increased, eigenvalues took on imaginary components (ii, iii) that imbue the population dynamics with oscillations (curves (ii) and (iii) in **Fig. 1.4 b**, (iv)). Additionally, the pseudo-spectrum of A (grey shaded region) extends over increasingly larger regions of the complex plane relative to the black contours. In particular, the pseudo-spectrum of (iii) extends significantly into the right hand side of the complex plane, predicting that over short times, the system will behave as if it were unstable. This prediction is borne out by the initial rise and subsequent decay (i.e., transient amplification) of the system dynamics shown in curve (iii) in **Figure 1.4b** (iv).

To test the hypothesis that feedback controllability (FBC) is a normative theory of neural population dynamics, we need to establish the conditions under which FBC subspaces will differ from FFC subspaces. We first provide the following theorem, which shows that, under certain conditions, the FBC and FFC subspaces are identical only for normal dynamical systems:

Theorem 2 *For $B = I_N, A = A^\top, A^{N\times N}$, with all eigenvalues of A distinct and $\max Re(\lambda(A)) < 0$, the projection matrix onto the eigenspace spanned by the d eigenvalues of A with largest real value constitutes a critical point of the FBC objective function and a global maximum of the FFC objective function.*

The proof of the theorem is provided at the end of the chapter. To demonstrate this analytic result, we return to the simulated linear dynamical systems describe above. We extracted FBC (with FCCA) and FFC (with PCA) subspaces ($d = 6$) from the generated data, and measured the difference between those two subspaces as the average angle between them (**Fig. 1.4c**, mean $\pm$ s.d., n=100). Two subspaces partially overlap if at least one subspace angle is zero, and are completely orthogonal if all subspace angles are equal to $\pi/2$. In practice, these angles can be obtained as the cosine of the singular values of the product $C_{\text{FFC}}^\top C_{\text{FBC}} \in \mathbb{R}^{d\times d}$ [47]. For dynamics driven by a completely normal dynamics matrix (i.e., giving rise to purely relaxation dynamics, corresponding to **Fig. 1.4 b** (i)), the average subspace angles were small ($\sim \pi/8$ rads). Increasing non-normality drove these subspace angles apart ($> 3\pi/8$ rads, Fig. **1.4 c** (iii)).

Qualitatively similar results were observed in stability optimized neural circuits, a previously proposed model of networks that resepct Dale's Law in which the degree of non-normality can be systematically tuned [48]. We generated networks with 100 excitatory and 100 inhibitory elements with a uniform connection probability of 0.25 and uniform, sign-constrained weights. Self-decay terms (i.e. diagonal elements of A) were also uniform

across the entire network. The magnitude of non-zero weights determines the spectral radius (i.e. magnitude of the largest eigenvalue) of the dynamics matrix [49]. For sufficiently large weights, the dynamics produced by this network will be unstable. Following [48], we then optimize the inhibitory weights of the network in order to achieve stability. The resulting matrix will have enhanced non-normality, with the degree of resulting non-normality having, empirically, a monotonic relationship with the starting spectral radius. We tuned the initial spectral radius over the same range of values as in [48].

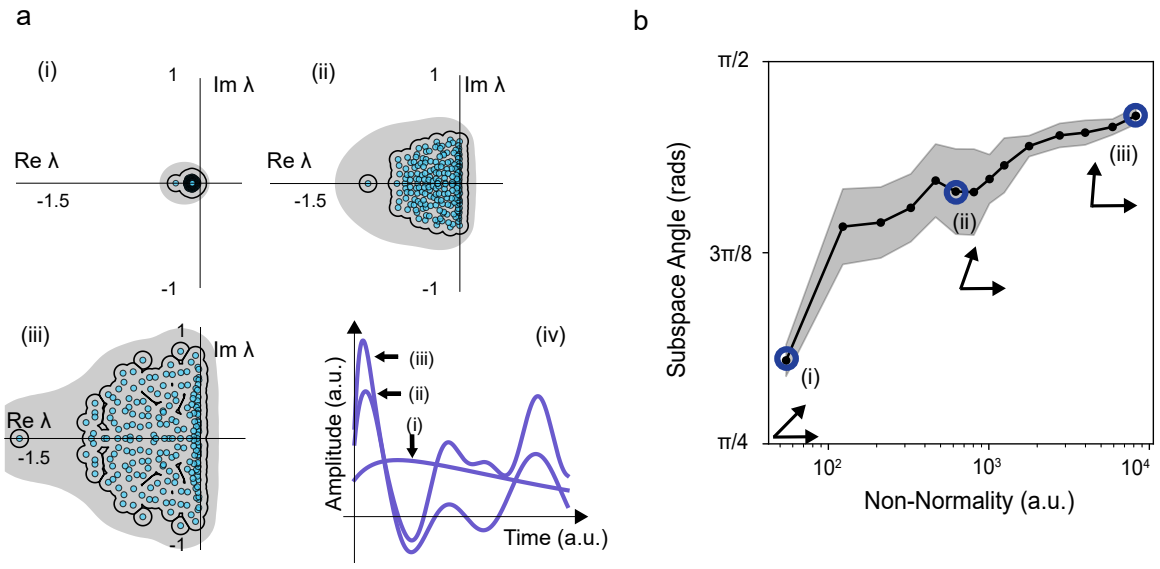


Figure 1.5: FBC/FFC subspaces diverge within stability optimized circuits (a, i-iii) Plot of the eigenvalues and $\epsilon = 0.1$ pseudospectral contours for typical, stabilized weight matrices. Non-normality increases from (i) to (ii) to (iii). (a, iv) Example trajectories from each system projected onto the the leading PCA vector. (b) Plot of the mean angle between FBC and FFC subspaces as the degree of non-normality within synthetic networks is increased. Spread indicates standard deviation over the random generation of 20 synaptic weight matrices and 10 simulations of dynamics for each weight matrix. Example systems from panel (a) are marked along the curve.

Example pseudospectra from these synthetic systems are shown in **Figure 1.5a**, with non-normality increasing from (i) to (ii) to (iii). As non-normality increases, we observe a proliferation of complex eigenvalues, and pseudospectral contours that extend beyond the equivalent normal matrix reference levels (grey vs. black contours). In **Figure 1.5b**, we plot the corresponding average subspace angle between PCA and FCCA subspaces. The indicated spread is the standard deviation across 20 initializations of synaptic weight matrices. The location of the example systems within panel a are labeled along the curve. Paralleling the results in the previously presented synthetic systems, increasing non-normality drives the subspace angles towards $\pi/2$. Together, these novel control-theoretic results establish that FBC subspaces are distinct from FFC subspaces when the underlying dynamics (i.e., A)

are non-normal. Since Dale's Law implies that cortical neural population dynamics should be non-normal, FBC subspaces should be distinguishable from FFC subspaces in neural population data.

Feedback controllable subspaces enable better decoding of behavior than feedforward controllable subspaces.

The key prediction of FBC as a normative theory of neural population dynamics is that task-relevant subspaces are more feedback controllable. Put another way, decoding behavior from FBC subspaces should be more accurate than decoding from FFC subspaces. We tested this prediction in previously recorded single-unit population neural data from monkey primary motor cortex (M1) and somatosensory cortex (S1) during a self-paced reaching task (**Fig. 1.6a**, right). Two macaque monkeys performed reaches on a 6x6 grid of starting and positions. In **Figure 1.6a**, left, we overlay reaches from one recording session aligned to the start location of the reach. These reaches exhibited a range of directions, velocities and lengths [50]. **Figure 1.6b** plots the time-course of single-unit neural activity recorded from primary motor (left) and primary somatosensory cortex (right) during this task. Across the two monkeys, there were 35 distinct recording sessions (35 in M1, 8 in S1), and the number of single units identified within each recording session varied from 97-200 in M1 and 86-187 in S1.

For our hypothesis to be testable, there must be substantial differences between FBC and FFC subspaces extracted from the neural population data. As we have demonstrated, these substantial differences hinge on the underlying dynamics being non-normal. We therefore first assessed the degree of non-normality of the observed neural recordings. We visualized the spectra (**Fig. 1.6c,d**, teal circles) and pseudospectra (grey region) of the neural population dynamics. Across recording sessions In M1, we observed a set of complex eigenvalues with larger imaginary than real part, indicating the presence of robust rotational dynamics [25]. In both M1 and S1, the pseudospectral contours extend beyond what would be expected from equivalent normal matrices (black contours). As with results from simulations (**Fig. 1.4c**), the subspace angles between FBC and FFC subspaces were consistently large across all recording sessions in M1 ($\approx 3\pi/8$ radians, n = 35, **Fig. 1.6b** boxplot shows median $\pm$ IQR) and S1 (n=8, **Fig. 1.6b** median $\pm$ IQR). Thus, in both M1 and S1, FBC subspaces are distinct from FFC subspaces. This distinction allows us to test whether FBC or FFC subspaces provide better bases for decoding arm reaches.

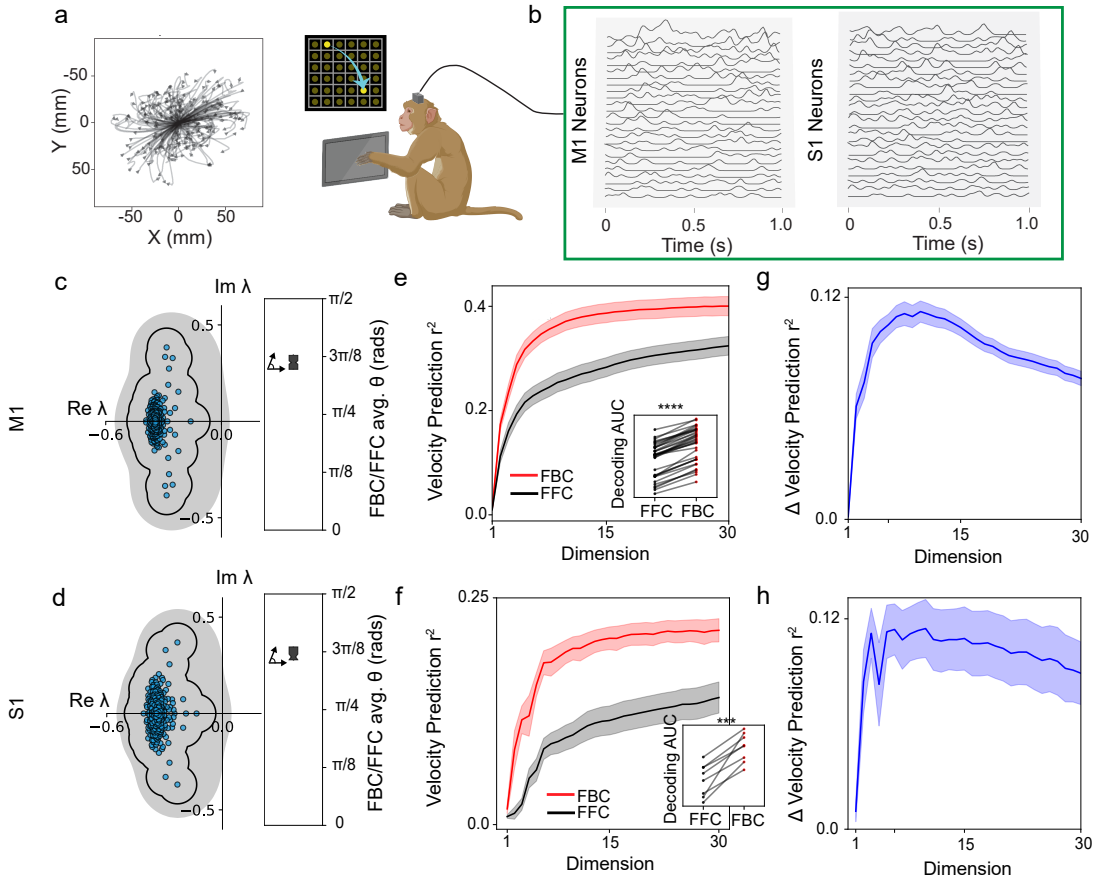


Figure 1.6: Feedback controllable subspaces enable better decoding of behavior than feedforward controllable subspaces. (a) A self-paced reaching task on a 2D grid provides a probe of a complex, naturalistic behavior in monkeys with Utah array recordings. (left) example reaches from one recording session, aligned to the physical start location of the reach. (b) Single-unit neural firing rates from primary motor cortex (M1, left) and somatosensory cortex (S1, right) in macaque recorded via Utah array co-recorded during reaching. (c, d) Left plots: Eigenvalues (blue scatter points) with associated pseudo-spectral contours (grey shaded region) of neural dynamics from one recording session in Macaque M1 and S1, respectively. Black contours indicate pseudo-spectral contours expected from a normal matrix. Right plots: Average subspace angle between FBC/FFC subspaces across recording sessions (median ± IQR) (e, f) Linear prediction of cursor velocity from activity projected into FBC/FFC subspaces within M1 and S1, respectively. Traces indicate mean r^2 of behavioral prediction from projected activity in FBC (red) and FFC (black) subspaces vs. projection dimension averaged across recording sessions (shading indicates standard error). Insets compare the total area under the r^2 vs. dimension curve (AUC) for each recording session between subspace methods (WSRT, $n = 35$ (e), $n = 8$ (f), ***: $p < 10^{-3}$, ****: $p < 10^{-4}$) (g, h) Paired difference in decoding performance between the FBC and FFC subspaces (mean ± s.e.).

To test whether feedforward or feedback controllability provides a better normative account of behaviorally relevant neural population dynamics, we decoded hand velocity from projections of the neural population data into FBC and FFC subspaces of dimensionality 1-30. **Figure 1.6c,d** plots predictive accuracy (cross-validated r^2) as a function of dimension for decoding from M1 (**Fig. 1.6c**) and S1 (**Fig. 1.6d**) using neural population data projected into FBC (red) and FFC (black) subspaces (mean ± s.e., M1, n = 35, S1, n = 8). We found that FBC subspaces were substantially better decoders of behavior than their FFC counterparts across dimensions. This is despite both subspaces being derived from the class of linear dimensionality reduction. We quantified the decoding performance across dimensions as the area under the decoding curve for each recording session separately. The superior decoding of FBC subspaces was consistent across all recording session in M1 (**Fig. 1.6e** inset, one sided paired Wilcoxon signed rank test (WSRT) $p < 10^{-4}$, $n = 35$) and all recording sessions of S1 (inset of **Fig. 1.6f** inset, one sided paired WSRT $p < 10^{-3}, n = 8$). These results demonstrate that behaviorally relevant neural population dynamics are more aligned with FBC than FFC directions.

Additionally, we compared the performance of decoding from unsupervised FBC subspaces to Preferential Subspace Identification, a linear, supervised method that directly identifies the behaviorally relevant neural subspace [51] (PSID). PSID implements subspace identification [52], identifying a latent stochastic, linear dynamical system that simultaneously drive behaviorally relevant dynamics and the behavior itself. PSID is a supervised method, projecting neural activity directly onto observed behavior. As such, it provides a useful upper bound on the decoding performance achievable by linear methods. We asked to what extend could decoders built off of FFC and FBC subspaces, i.e. unsupervised measures of the neural dynamics, identify behaviorally relevant dynamics. In contrast to decoding results presented in **Figure 1.6**, we use a Kalman filter to decode behavior from FFC/FBC projected activity, paralleling the strategy used by PSID to decode behavior from its latent states.

In **Figure 1.7**, we compare the velocity decoding performance of FBC, FFC, and preferential subspace identification in M1 and S1 as a function of dimension. Here, dimension for preferential subspace identification refers to the dimension of the latent, behaviorally relevant state space dynamics. In M1, we found that the decoding performance of FBC subspaces attained 80 % of that of preferential subspace identification by $d = 6$ and 85 % by $d = 10$ (comparing red and maroon traces in **Fig. 1.7a**). In S1 (**Fig. 1.7b**), we found analogously that by $d = 10$, the decoding performance of FBC subspaces attained within 85 % of that of preferential subspace identification performance by $d = 10$. As we emphasize, this is despite the identification of FBC subspaces via FCCA occurring without access to the behavior during dimensionality reduction, in contrast to PSID. FBC subspaces therefore capture the vast majority of behaviorally relevant information available to linear Gaussian methods in this dataset.

Finally, to ascertain the dimension at which feedback controllability of neural population dynamics was most important for behavior, we calculated the paired difference in prediction performance (Δ-Velocity prediction) as a function of dimension (**Fig. 1.6e, f**). The Δ-

velocity prediction reached 90% of its maximum value by dimension 6 in both M1 and S1. The percent improvement in decoding performance at the dimension of peak Δ-velocity prediction was substantial: 43% in M1, and 96 % in S1. This indicates that the FBC subspace most relevant for behavior is simple (i.e., low-dimensional), and we therefore used $d = 6$ as a standardized dimension for subsequent analyses. Together, these results validate the key prediction feedback controllability as a normative theory of neural population dynamics.

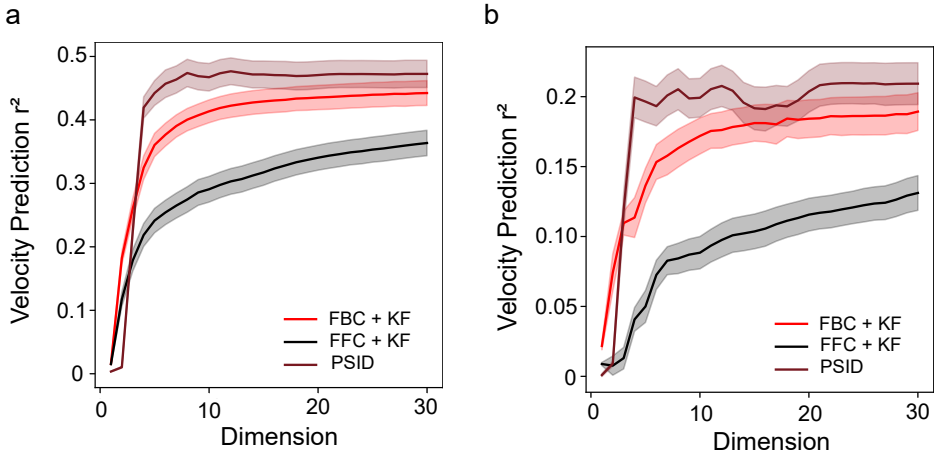


Figure 1.7: Decoding performance from FBC subspaces approximates that of a supervised method. (a) Comparison of velocity prediction r^2 vs. dimension between PSID identified subspaces (purple), FBC subspaces (red) and FFC subspaces (black) in M1. For PSID, dimension refers to the dimension of the latent behaviorally relevant subspace. (b) Analagous curves for behavioral decoding from projected activity in S1.

Time courses of feedback controllability match reach acceleration

Next, to investigate how the mapping between dynamics in the FBC/FFC subspaces and behavior was modulated during the time course of reaches, we segmented the first 1.5 seconds of behavior following reach initiation (defined as the time when the visual target cue switched) into 100 ms windows and trained linear decoders from the projected neural data to predict cursor velocity. In **Figure 1.8a**, we plot the mean $\pm$ s.e of the time resolved velocity prediction across recording sessions within M1 ($n = 35$). The overall decoding performance from activity projected into both FFC (black) and FBC (red) subspaces was strongly correlated with overall reach velocity, with the peak in decoding performance occurring within 50 ms of the peak in cursor velocity at $\sim$ 500 ms after reach start both on average (**Fig. 1.8a**, vertical dashed lines) and individually across recording sessions (not shown). The corresponding results for S1 are shown in **Figure 1.8 b** (mean $\pm$ s.e., $n = 8$). Note that the difference in the average reach velocity time course relative to **Figure 1.8a** is driven by S1 recordings being available for only a subset of recording sessions.

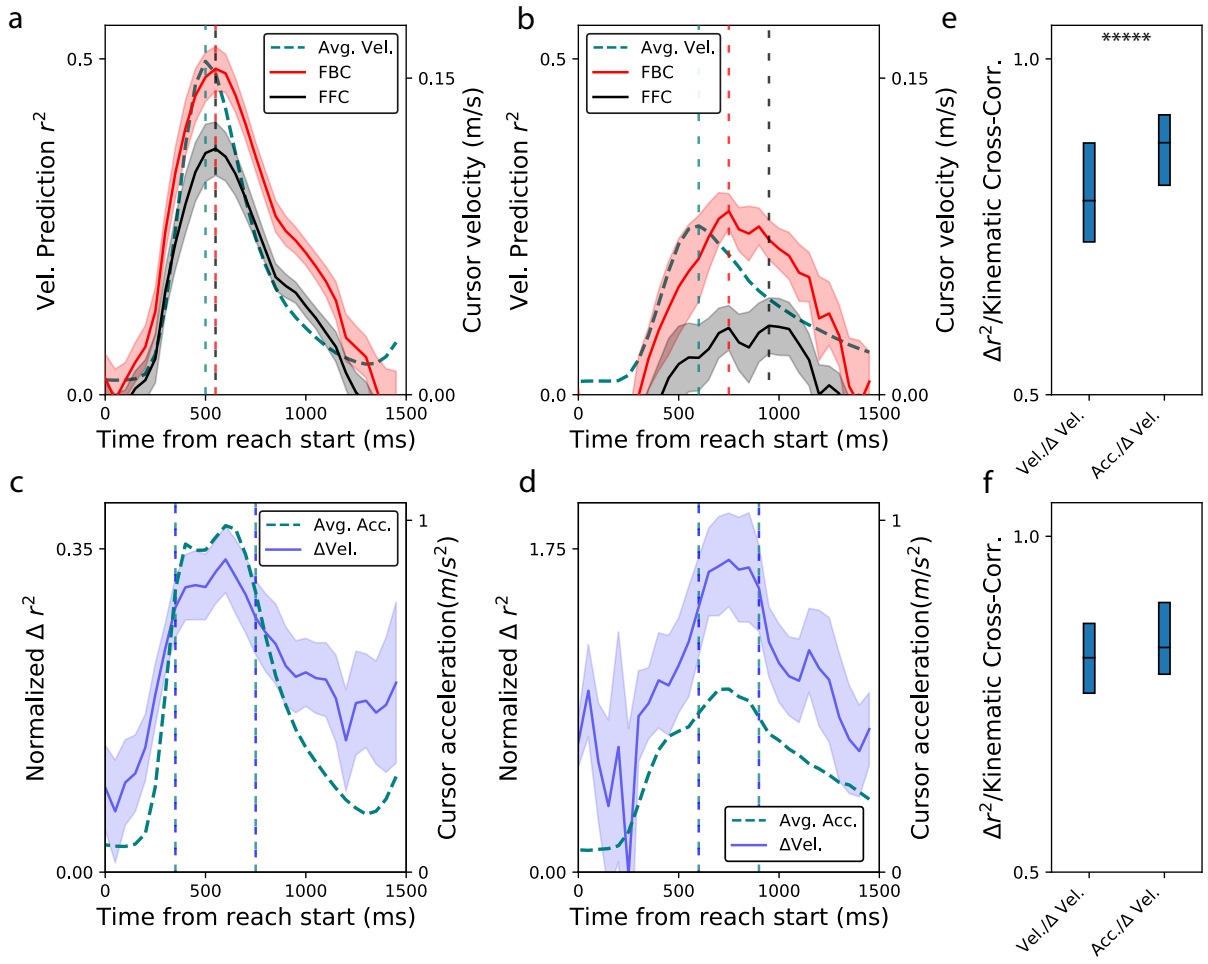


Figure 1.8: Time courses of feedback controllability match reach acceleration (a) Mean and standard error across recording sessions ($n = 35$) of time-resolved prediction r^2 of cursor velocity by M1 FCCA (red) and M1 PCA (black) compared to the average cursor velocity (dashed green line) during reaching. Peaks of all three curves coincide (dashed colored lines). (b) Analogous traces (mean ± s.e., $n = 8$) for prediction of cursor velocity from S1 FCCA (red) and S1 PCA (black). (c) Mean and standard error across recording sessions ($n = 35$) of the paired difference in velocity prediction from M1 activity (blue) co-plotted against the average reach acceleration (green), normalized to the peak FFC derived prediction. Vertical dashed lines indicate when the plateau of both curves begins and ends, defined as 80% relative to maximum. (d) Analogous traces for paired difference in velocity prediction from S1 activity (mean ± s.e. $n = 8$). (e-f) Distribution of cross-correlation coefficients (median ± IQR) between the Δ-velocity prediction curves in M1 and S1 (blue traces in c,d, respectively) and the reach velocity (left boxes) and reach acceleration (right boxes) across recording sessions (one-sided paired WSRT, $p < 10^{-5}, n = 35$).

In contrast to M1, the peak velocity decoding performance from both S1 FFC and S1 FBC subspaces occurred later, between 750 and 1000 ms after reach start (i.e. 250 to 500 ms after

the peak in reach velocity). This timing is measured with respect to the start of reaches. The timing difference between M1 and S1 activity is accentuated by the lag between neural activity and behavior employed within the linear decoders. For M1, we used a window of neural activity centered 100 ms prior to behavior, whereas for S1, we used a window of activity centered 50 ms prior to behavior, in line with when these regions were found to be most predictive of behavior and with results reported in [53] for the same dataset. In sum, we found that S1 was most predictive of behavior approximately 300 ms after M1.

To assess how the relative importance of feedback controllability varied as a function of the reach kinematics, we then calculated the paired difference in decoding performance between FBC and FFC subspaces across time. Analogously to **Figure 1.6g, h**, we useΔ-velocity prediction r^2 (abbr. Δ-r^2) as a measure of the relative importance of feedback controllability vs. feedforward controllability during the time course of a reach. In **Figure 1.8c**, we plot the Δ-r^2 for M1 (mean $\pm$ s.e., $n = 35$) normalized against the peak FFC decoding r^2 from **Figure 1.8a**. We found that the Δ-r^2 exhibited a rapid rise beginning at 200 ms after reach start, saturating at 400 ms, and decaying at 750 ms (**Fig. 1.8 c**, blue trace). These dynamics tracked the magnitude of average reach acceleration (**Fig. 1.8 c** dashed green trace) closely, reproducing the double peak structure visible at 400 and 600 ms in the latter. We calculated the time at which both the Δ-r^2 and magnitude of acceleration rose past 80 % of their maximum value and then declined past this threshold in the terminal period of the reach. These rise and fall times were consistent across recording sessions between the Δ-r^2 and average acceleration (blue and green vertical dashed lines in **Fig. 1.8 c**, respectively). Analogous results held in S1 (**Fig. 1.8d**, mean $\pm$ s.e. of Δ-r^2 shown, $n = 8$), again with an approximately 300 ms time lag.

To quantify the observed similarity between the time courses of reach acceleration and Δ-r^2, we normalized each trace within each recording session to a 0-1 scale and calculated the cross-correlation between them. The 0-1 normalization ensures that this cross-correlation also lies between 0 and 1. In M1, the cross-correlation between the Δ-r^2 time course and reach acceleration consistently peaked at 0 relative lag and exhibited a median of 0.88 across recording sessions (**Fig. 1.8e**, right bar, median $\pm$ IQR indicated, $n = 35$). Similarly, in S1, cross-correlations peaked at zero lag in all recording sessions with the median peak value across recording sessions equaling 0.83 (**Fig. 1.8f**, right bar, median $\pm$ IQR, $n = 8$). Notably, the median cross-correlation between Δ velocity prediction r^2 and the time course of average acceleration was higher across recording sessions than the analogous median zero lag cross-correlation between Δ velocity prediction r^2 curves and the average reach velocity in both M1 (0.88 vs. 0.79) and S1 (0.83 vs. 0.82) (left bars in **Fig. 1.8e,f**, median $\pm$ IQR shown). In M1, the paired difference between these two cross-correlations was found to be statistically significant (one sided paired WSRT $p < 10^{-5}$, $n = 35$). The relatively small effect size is to be expected, as the average velocity magnitude and average acceleration magnitude are themselves highly correlated (normalized correlation $r = 0.95$ in M1 and $r = 0.98$ in S1). In M1, a significant cross-correlation exists between the Δ-r^2 and the average acceleration magnitude after both time series are decorrelated from the average velocity magnitude (one-sided paired WSRT, $p < 0.05$, $n = 35$). Conversely, no significant residual correlation was

detected when first decorrelating the Δ-r^2 and velocity magnitude from the acceleration magnitude. Together, these results that the relative importance of feedback controllable dynamics within both M1 and S1 to behavioral decoding closely tracks the magnitude of reach acceleration. This suggests that acceleration is the kinematic parameter most under feedback control.

Feedback controllability is mediated by a distinct population of neurons

The large angles between FFC and FBC subspaces found in both M1 and S1 (**Fig. 1.6 a, b**) suggested that FFC and FBC dynamics were mediated by distinct populations of neurons. As the FBC and FFC subspaces are composed of additive, weighted combinations of the individual neurons in the recorded population, we are able to assign each neuron an importance score associated with each subspace. These importance scores provide information above and beyond that provided by the subspace angles, as large subspace angles could arise in a number of ways. We therefore directly tested the hypothesis that the populations mediating FBC/FFC dynamics were distinct, evidence for which would allow us to connect functionally defined subspaces of neural dynamics back to the properties of the individual recorded neurons.

To determine the importance score of a neuron i within an FFC or FBC projection matrix $C \in \mathbb{R}^{N \times d}$, we calculate the norm of the i^{th} row of the projection matrix, and normalize across rows, i.e. $||C_{i,:}||^2 / \max_i ||C_{i,:}||^2$ (example projections $[\text{FBC Score}]_j$, $[\text{FFC Score}]_j$ and $[\text{FBC Score}]_k$,$[\text{FFC Score}]_k$ shown in **Fig. 1.9a**). This yields a score for each neuron in the population within FFC and FBC subspaces that we normalize to lie on a 0-1 range (schematically indicated in the partitioned red and black vectors at the bottom of **Fig. 1.9a**). All reported importance scores were obtained from $d = 6$ projections, and averaged across projections fit to 5 folds of the data. We visualize these scores in **Figure 1.9b** and **c** on a log-log scale (M1 in **Fig. 1.9b**, S1 in **Fig. 1.9c**). Each scatter point corresponds to a single unit within a single recording session. We indicate the relative feedback controllability of each neuron, defined as the FBC importance score normalized by the sum of FFC and FBC importance scores, by its color (**Fig 1.9b,c** colorbars). Neurons with high FBC importance scores but low FFC importance scores are shaded red, and those with high FFC but low FBC importance scores are shaded black. We observed that across the entire population of neurons in M1, a neuron's importance score within the FBC subspace was uncorrelated with its importance score in the FFC subspace (spearman rank correlation $\rho = -0.02, p = 0.11, n = 5041$). In S1, we observed a very small correlation magnitude ($\rho = 0.08$) that was statistically significant due to the large sample sizes ($p = 4 * 10^{-3}, n = 1257$). Again, this result is in line with, but not necessarily implied by the large reported FBC/FFC subspace angles in both regions (**Fig. 1.6a,b**). For example, a small number of neurons may have had large importance scores in FFC as opposed to FBC subspaces (and vice-versa), with the remainder of neurons having correlated, but small FFC/FBC importance scores. Our

observation to the contrary further suggested that FBC vs. FFC subspaces were composed of distinct populations of neurons.

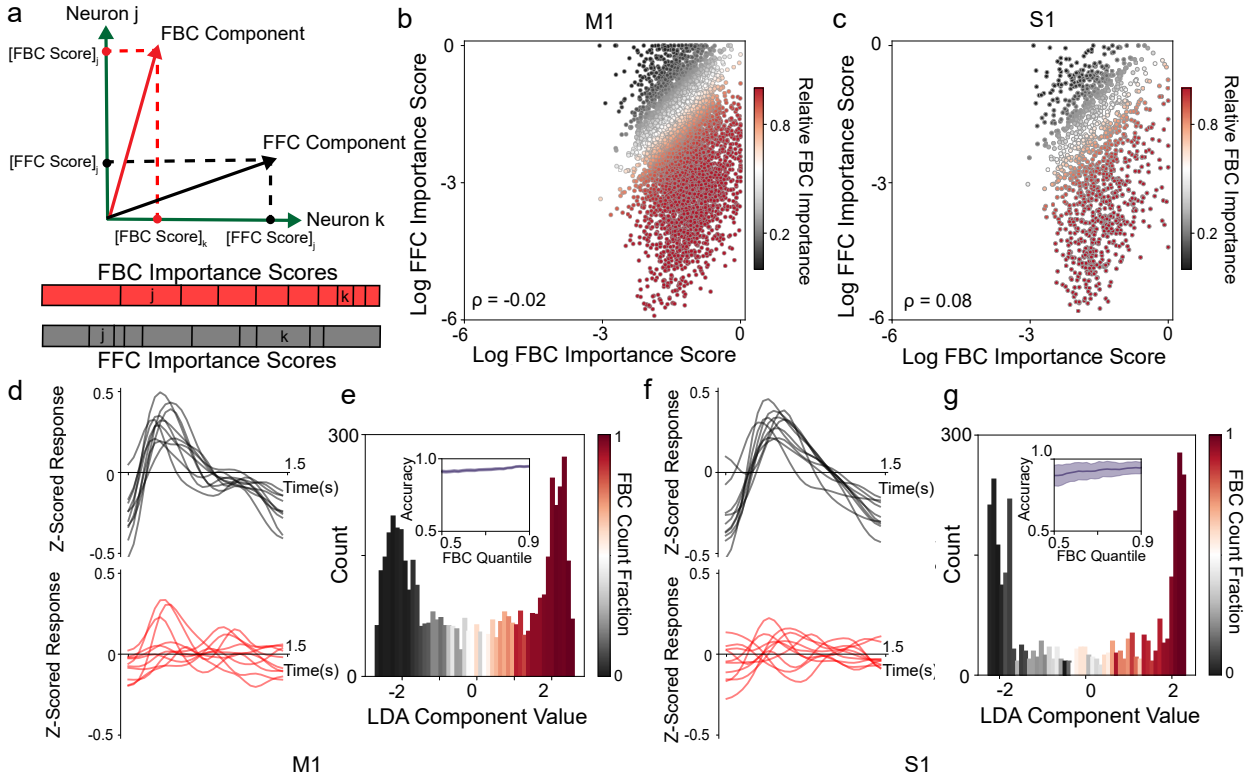


Figure 1.9: Feedback controllability is mediated by a distinct population of neurons. (a) Simplified schematic of how the importance scores of each neuron are derived from FFC/FBC projections. (b) Scatter plot of the importance scores of neurons in FFC vs. FBC subspaces across all M1 recording sessions. Each scatter point corresponds to a single unit from one recording session. Spearman rank correlation ρ between FBC/FFC importance scores indicated. (c) Analogous scatter plot for S1 data. (d) Example trial-averaged, Z-scored firing rates aligned to reach initiation of M1 neurons with the highest relative FFC (black) and FBC (red) importance score. (e) Histogram of transformed firing rates for all M1 neurons across all sessions ($n = 5041$) projected onto the LDA component. Each histogram bin is colored according to the fraction of neurons within it that are designated as either FBC or FFC neurons. (Inset) Cross-validated LDA prediction accuracy of FFC/FBC category (mean $\pm$ s.e. across recording sessions, $n = 35$) as a function of the quantile of relative FBC used to assign neurons to categories. (f) Example trial-averaged, Z-scored, firing rates of S1 neurons with the highest relative FFC (black) and FBC (red) importance scores. (g) Analogous plot to (e) for all S1 neurons across all sessions ($n = 1257$). (Inset) Mean $\pm$ s.e. of cross-validated LDA prediction accuracy across recording sessions ($n = 8$) as a function of relative FBC quantile used for class assignment.

Therefore, we next examined if neurons important for FFC vs. FBC had distinct electro-

physiological activity profiles. We categorized neurons according to whether their relative FBC importance (i.e., colorbar in **Fig. 1.9b,c**) lied above or below the median relative FBC importance within the recording session. We then calculated smoothed, Z-scored, and trial averaged, firing rates for each neuron over the first 1.5 seconds following reach initiation. We plot these firing rate time courses for a set of neurons with high FFC (top, black) and FBC (bottom, red) importance scores in M1 (**Fig. 1.9d**) and S1 (**Fig. 1.9f**). A visual inspection of the trial averaged firing rates indicated that the distinct FFC and FBC populations were characterized by differing dynamics during reaching behavior. FFC neurons exhibited a high degree of similarity amongst themselves, with a robust, large amplitude turn-on effect following reach initiation (peak in black traces in **Fig. 1.9d,f** at approximately 400 ms). By contrast, FBC neurons exhibited low-amplitude, heterogeneous, ongoing dynamics in many cases unassociated with the start of reaches. The dynamic range was significantly higher amongst FFC neurons than FBC neurons in both M1 (WSRT $p < 10^{-5}, n = 35$) and S1 (WSRT, $p < 10^{-2}, n = 8$). To determine whether these differences in activity profiles were sufficient to accurately classify neurons as being important for FFC vs. FBC, we applied Linear Discriminant Analysis (LDA) to features derived from UMAP applied to the firing rates. For a 2 class classification problem, LDA yields a projection of data onto a one dimensional space over which the two classes are most linearly separable. We plot histograms of these projections applied to all neurons across all recording sessions in both M1 (n=5041, **Fig. 1.9e**) and S1 (n=1257, **Fig. 1.9g**). Each histogram bin is colored by the fraction of neurons within that bin that are important for FBC vs. FFC. We observed a clear bimodal structure within the histogram densities in both brain regions, indicating the presence of two clusters of neurons corresponding FFC/FBC neurons that were linearly separable by features derived from their firing rates. The average cross-validated accuracy of classification across recording sessions was 0.91 in M1 and 0.88 in S1. This classification accuracy was significantly higher than chance (0.5, one-sided WSRT $p < 10^{-5}, n = 35$ and $n = 8$ in M1, and S1, respectively), and robust to the choice of FBC quantile used to divide the population into classes (**Fig. 1.9e,g** inset and **Fig. 1.10**). We therefore conclude that FBC and FFC subspaces are comprised of two distinct populations of neurons within both M1 and S1.

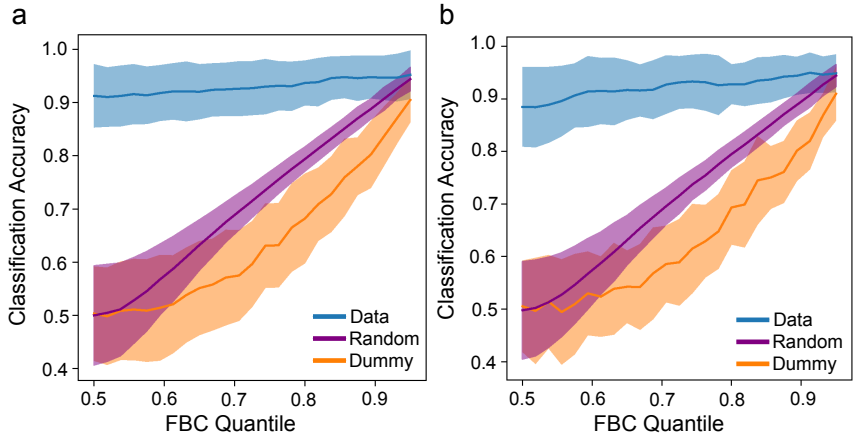


Figure 1.10: (a) Plot of the average classification accuracy of LDA applied to the data and FFC/FBC classification (blue), a dummy classifier (orange), and an LDA classifier trained on random labels (purple) as a function of the relative FBC quantile used to assign neurons to FFC/FBC classes. Spread is the standard error taken across recording sessions in M1 ($n = 35$). (b) Analogous plot across S1 recording sessions (mean $\pm$ s.e. $n = 8$).

Feedback controllability is an emergent, population level property.

The prior analysis revealed that the populations of neurons important for FFC and FBC exhibited disparate firing rate profiles. Visual inspection of the firing rates plotted in **Figure 1.9d, g** suggested further that FFC neurons exhibited a higher degree of pairwise similarity (i.e., cross-correlation) and temporal alignment than FBC neurons. Cross-correlations are a measure of functional interactions between neurons (depicted as blue arrows in **Fig. 1.11a**, top). We sought to determine the extent to which feedback vs. feedforward controllability of neural dynamics relied on these functional interactions, as opposed to being predictable from the functional properties of single neurons taken in isolation (schematically depicted in **Fig. 1.11a**, bottom with dashed arrows). If controllability cannot be reduced to single neuron properties alone, then this would provide strong evidence for it being an emergent, population level phenomena within neural circuits.

We took two complementary approaches towards assessing whether FBC/FFC was an emergent property resulting from population interactions. First, we evaluated whether importance scores within FFC/FBC subspaces could be predicted from a set of single neuron properties frequently assayed in systems neuroscience. For each neuron, we calculated its response variance, weight in a linear decoder of reach velocity trained on the entire population, and r^2 of an encoding model of its firing rate from reach kinematics. We then trained a linear model to predict FBC/FFC importance scores from these features.

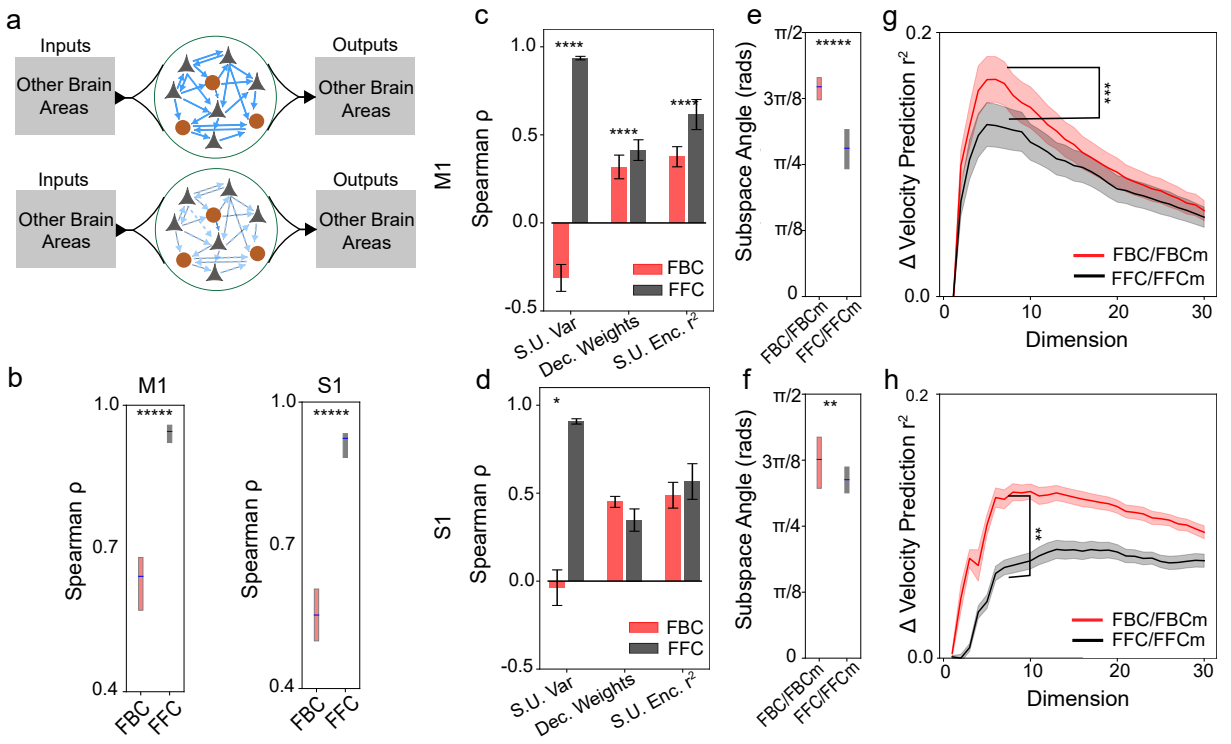


Figure 1.11: Feedback controllability is an emergent, population level property. (a) Schematic of the comparison made between analyses that disregard interactions between neurons (bottom) and those that do not (top). (b) Median ± IQR across recording sessions of the spearman rank correlation (ρ) between actual FBC/FFC importance scores and importance scores predicted from the a linear regression using single unit features for M1 (left) and S1 (right, *****: $p < 10^{-5}$, WSRT $n = 35$ and $n = 8$, respectively). (c) Bar plots of mean ± s.e. across recording sessions of the individual spearman rank correlations between M1 single neuron features utilized to fit models in panel (b) and FBC/FFC importance scores (WSRT, ****: $p < 10^{-4}$, $n = 35$). (d) Analogous plot for S1 (one-sided WSRT, *: $p < 0.05, n = 8$). (e) Median ± IQR across recording sessions of the distribution of average subspace angles between $d = 6$ FBC and FBCm projections (red) and FFC/FFCm projections (black) in M1 (WSRT, $p < 10^{-5}, n = 35$). (f) Analogous distribution of subspace angles across recording sessions in S1 (WSRT, $p < 0.01, n = 8$). (g) Plot of the paired differences (mean ± s.e. across recording sessions, n=35) in cursor velocity prediction r^2 between using activity projected into FBC vs. FBCm (red) and FFC vs. FFCm (black) subspaces as a function of projection dimension. Significance in the difference between peaks in the two curves at $d = 6$ as measured by WSRT indicated ($p < 10^{-3}, n = 35$) (h) Analogous curves for S1 (mean ± s.e. across sessions, $n = 8$). Significance of WSRT similarly indicated ($p < 0.01, n = 8$).

The median spearman rank correlation between the predicted and actual importance scores was 0.95 for FFC subspaces (**Fig. 1.11b** left, median ± IQR across recording sessions in black box) vs. 0.64 for FBC subspaces (median ± IQR across recording sessions in red box) in M1, and 0.92 for FFC vs. 0.55 for FBC in S1 (**Fig. 1.11b** right, median ± IQR

across recording sessions in black and red boxes, respectively). These differences were highly significant in both regions (WSRT, $p < 10^{-5}$. $n = 35$ and $n = 8$, respectively). From this, we conclude that the importance of a neuron to the feedback controllable subspace cannot be well predicted by single neuron properties, in contrast to neurons that mediate feedforward controllability. The individual spearman rank correlations between each property and the FBC/FFC importance scores, which we plot as histograms (mean $\pm$ s.e. across recording sessions) in **Figure 1.11 c, d**, support this conclusion. In both M1 and S1, a high importance in the FFC subspace could be accounted for almost entirely by a neuron's variance (leftmost black bars, $\rho = 0.93$ in M1, $\rho = 0.90$ in S1). By contrast, a high importance in FBC subspaces was weakly, negatively correlated with neuron variance (leftmost right bars, $\rho = -0.28$ in M1, $\rho = -0.07$ in S1). In M1, we additionally found that the decoding weights and r^2 performance of single neuron encoding models were significantly more correlated with the FFC importance scores than FBC importance scores (WSRT $p < 10^{-4}, n = 35$), whereas these differences were not found to be significant amongst S1 neurons.

While the set of single neuron properties considered above were unable to accurately predict the importance scores of neurons within FBC subspaces, it could still be the case that the FBC subspaces could nonetheless be derived from measures of controllability that neglect functional interactions (i.e., cross-correlations) between neurons. To test this, we obtained FFC and FBC subspaces from fits of PCA to FCCA that included only single neuron (i.e., marginal) variance and autocorrelations within their objective functions. This procedure is tantamount to applying PCA and FCCA to surrogate data that has been shuffled to remove cross-unit correlations [54]. We refer to the resulting subspaces as the FFCm and FBCm subspaces, respectively. In **Figure 1.11e** and **f**, we plot the distribution (median $\pm$ IQR) of subspace angles between the FBC/FFC subspaces and their marginal variants in M1 and S1, respectively. The median subspace angle between FBC and FBCm (red boxes) was significantly higher than that between FFC and FFCm (black boxes) in both M1 (one sided Wilcoxon paired difference test, $p < 10^{-5}, n = 35$.) and S1 ($p < 10^{-2}, n = 8$). Finally, similarly to results shown in **Figure 1.6 c, d**, we trained linear decoders of cursor velocity from the marginal subspaces. Compared to the same subspaces methods extracted from the full population statistics, we observed a substantial drop-off in decoding performance. In **Figures 1.11 g, h**, we plot the paired difference in decoding performance (mean $\pm$ s.e.) as a function of dimension between FBC/FBCm (red trace) and FFC/FFCm (black trace) in M1 and S1 respectively. These paired differences peaked in M1 (**Fig. 1.11g**) and saturated in S1 (**Fig. 1.11h**) at $d = 6$. The reduction in decoding performance between FBC and FBCm vs. FFC and FFCm was larger across all dimensions examined, and at $d = 6$ exhibited a statistically significant difference (one sided paired WSRT $p < 10^{-3}, n = 35$ for M1 $p < 0.01, n = 8$ for S1). Taken together, these results demonstrate that accurate assessment of controllability from population dynamics requires incorporating emergent, population level statistics, with this population structure being more important for FBC vs. FFC.

Feedforward and feedback controllable subspaces engage distinct dynamical regimes.

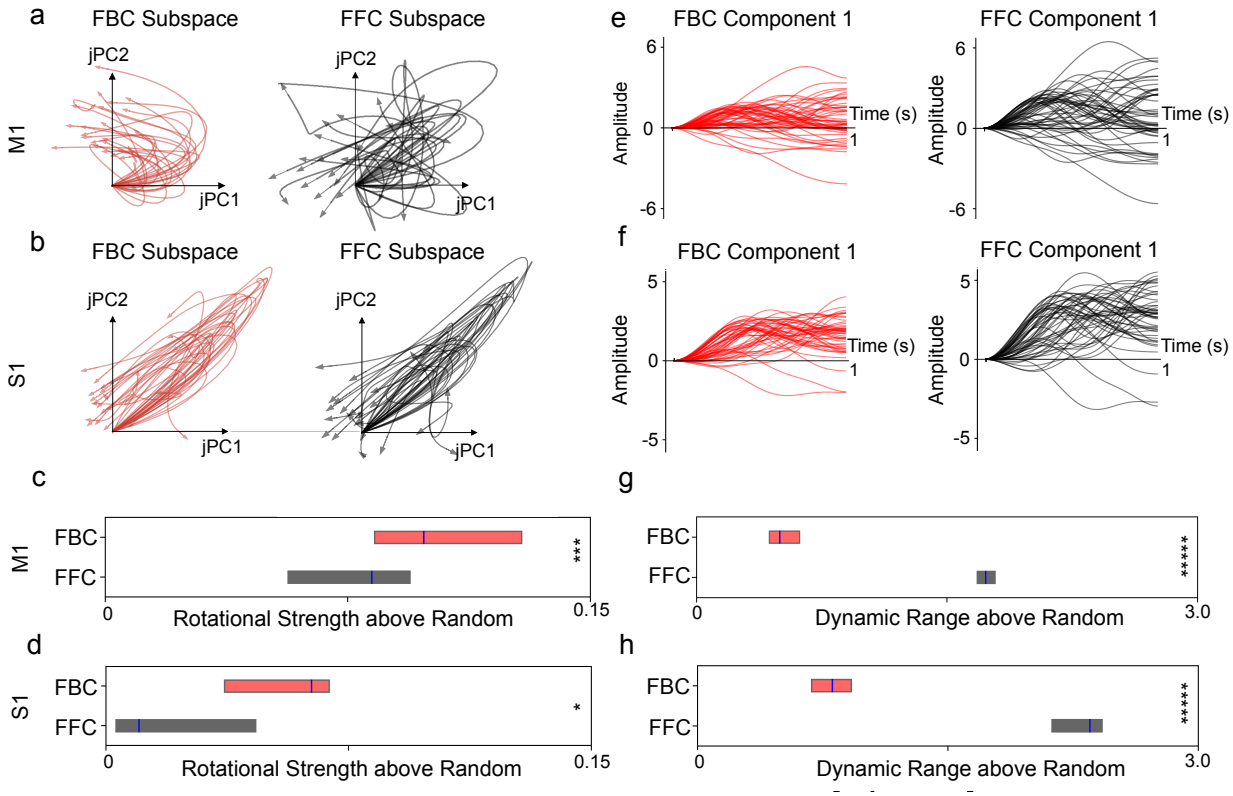


Figure 1.12: Feedforward and Feedback controllable subspaces engage distinct dynamical regimes. (a) Example trajectories in M1 FBC (red) and FFC (black) subspaces projected onto the top jPCs. (b) Analogous example trajectories of S1 FBC/FFC subspaces projected onto the top two jPCs. (c, d) Distribution of rotational strength (median ± IQR of sum of imaginary eigenvalues of jPCA fits across recording sessions) in FFC vs. FBC above average rotational strength in random subspaces in M1 (n=35) and S1 (n=8), respectively (WSRT, ***: $p < 10^{-3}, n = 35$, *: $p < 0.05, n = 8$) (e) Example trajectories in M1 FBC and FFC subspaces projected onto directions of highest amplification. (f) Analogous plots for S1 data. (g, h) Distribution of average dynamic range (median ± IQR across recording sessions) in FFC. vs FBC vs. random subspaces in M1, and S1 respectively (WSRT, *****: $p < 10^{-5}, n = 35$ and $n = 8$, respectively.

Generally speaking, linear models of population dynamics can generate rotations and scalings of the population firing rate vector over time. Rotational dynamics in particular are a robustly observed feature of population dynamics within motor cortex [25], and the presence of large imaginary components within the eigenvalues of functional connectivity matrices within M1 (**Fig. 1.6a**) indicated the presence of rotational dynamics within this dataset. Population level interactions are necessary for both the generation of rotational

dynamics [25] and determination of FBC subspaces (**Fig. 1.11**), while the FFC objective function favors amplifying dynamics and FFC subspaces were associated with high variance neurons (**Fig. 1.11c,d** leftmost black bars). Additionally, the firing rate profiles of FFC vs. FBC neurons were distinct (**Fig. 1.9 d**, and **g**), with the former exhibiting amplification at reach onset, and the latter exhibiting temporally heterogenous, oscillatory dynamics. Given these facts, we hypothesized rotational dynamics would be associated with feedback controllability, while scaling dynamics would be associated with feedforward controllability. Such a correspondence would establish an underlying normative, computational role for these distinct dynamic regimes.

To quantify rotational dynamics, we fit jPCA [25] to projections of the first 1 second of activity following reach initiation into the FFC and FBC subspaces. We assessed the strength of rotational dynamics by taking the sum of imaginary jPCA eigenvalues as compared to jPCA fits within random projections of the data. Examples of smoothed, single trial firing rates projected onto the top two jPCs within FBC and FFC subspaces respectively are shown in **Figure 1.12a** and **b** for M1 and **1.12c** and **d** for S1. Visually, we observed more stereotyped rotational dynamics in M1 than in S1 (**Fig 1.12a** vs. **b**), while within M1, rotations were more cleanly observed within FBC subspaces than FFC subspaces (**Fig 1.12a**, red vs. black traces). Across all recording sessions, the strength of rotational dynamics above that contained within random projections was significantly higher in FBC vs. FFC subspaces in both M1 (median $\pm$ IQR, red vs. black boxes in **Fig. 1.12e**, WSRT, ***: $p < 10^{-3}, n = 35$) and S1 (median $\pm$ IQR, red vs. black boxes in **Fig. 1.12f**, WSRT, *:$p < 0.05, n = 8$). FBC subspaces therefore contained stronger rotational dynamics than FFC subspaces.

On the other hand, to examine scaling, or amplification, of dynamics, we calculated the average dynamic range within projected activity within 1 seconds after reach initiation measured relative to baseline activity prior to reach initiation. Examples of activity from single recording sessions along the top 2 dimensions ordered by dynamic range are shown in **Figure 1.12g-j** (FBC components in red, FFC components in black). We again compared the average dynamic range to that found within random $d = 6$ projections of the data. Across recording sessions, the average dynamic range was significantly higher within FFC subspaces than FBC subspaces in both M1 (median $\pm$ IQR, black vs. red boxes in **Fig. 1.12k**) and S1 (median $\pm$ IQR, black vs. red boxes in **Fig. 1.12l**). Thus, FBC activity within both M1, and S1 contained stronger rotational dynamics than FFC activity, whereas FFC activity contained more amplification in their time courses than FBC activity.

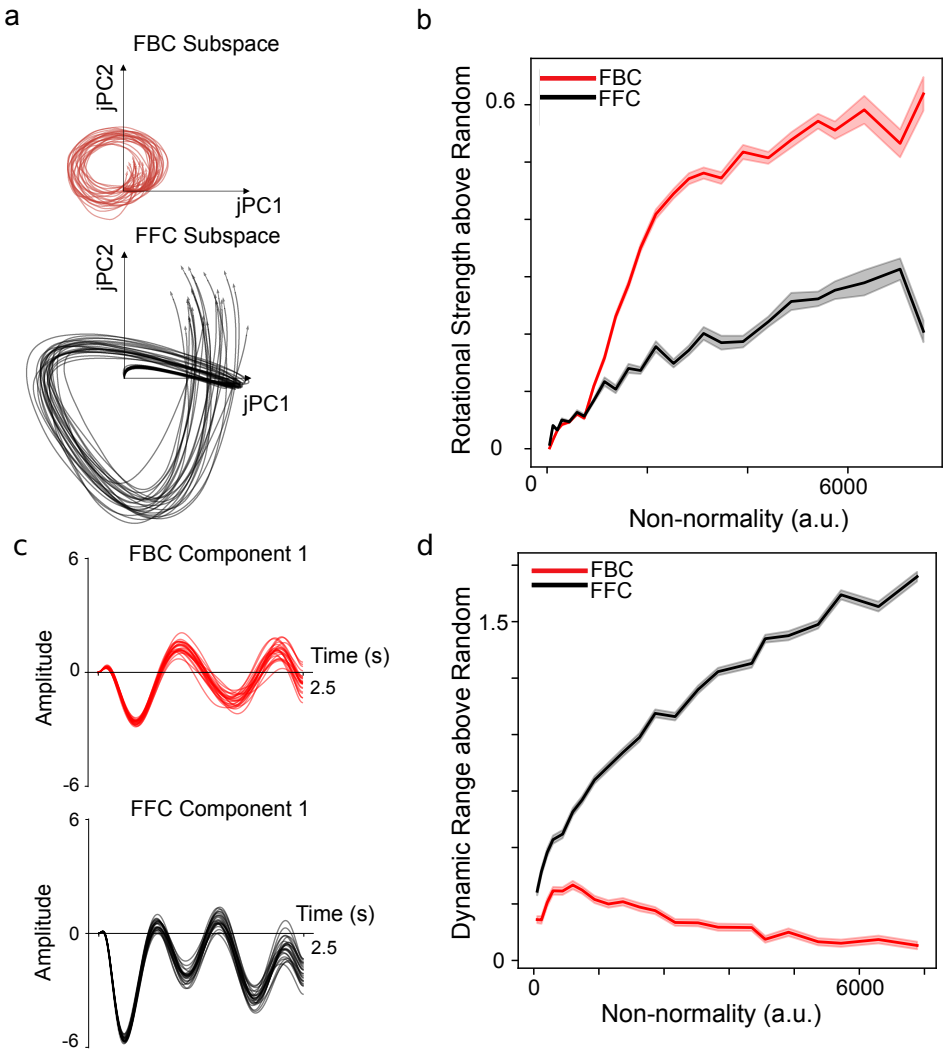


Figure 1.13: Feedback controllable subspaces exhibit stronger rotational dynamics than feedforward controllable subspaces in inhibitory stabilized networks (a) Plot of example trajectories projected to $d = 6$ within FBC (black) and FFC (red) subspaces, and then further projected into the top 2 jPCA dimensions. (b) Plot of the sum of jPCA eigenvalues within $d = 6$ FBC (red), and FFC (black) subspaces relative to random projections as non-normality of the inhibitory stabilized networks is increased. Spread represents the standard deviation taken across 20 initializations of the synaptic weight matrix and 10 trajectories within each stabilized network. (c) Plot of example trajectories projected to $d = 6$ within FBC and FFC subspaces, and then further projected onto the direction of highest amplification. (d) Plot of the mean $\pm$ s.d. dynamic range relative to random projections within FBC/FFC subspaces across the same range of non-normality as panel (b).

We replicate the above findings within synthetic, stability optimized E/I networks. We fit

jPCA onto activity projected to $d = 6$ via PCA and FCCA across the full range of networks explored in **Figure 1.5**. Subspace activity projected onto the top 2 jPCs for an examples system drawn from the most non-normal regime (corresponding to point (iii) in **Fig. 1.5a**) are shown in **Figure 1.13 a** and **b** for FFC and FBC subspaces, respectively. Visually, we observed trajectories within the top 2 jPCs within the FFC subspace exhibited a large degree of both shearing and amplification. By contrast, FCCA trajectories (plotted on the same scale as the PCA trajectories) were found to exhibit comparatively less amplification and were better aligned to pure rotations. We quantified this effect analogously to **Figure 1.12 c, d** by taking the sum of the imaginary eigenvalues associated with the jPCA fit to the FFC/FBC subspace projections relative to 1000 random $d = 6$ projections of the data. In **Figure 1.13b**, we plot this statistic as a function of the underlying non-normality. We find that the degree of underlying non-normality in linear, stability optimized E/I networks increases the strength of rotational dynamics in both FFC and FBC subspaces. Nevertheless, after an initial regime of relatively low non-normality (lower left corner), the strength of rotational dynamics was found to be significantly stronger in FBC subspaces. As in **Figure 1.12**, we then calculated the average dynamic range contained within FFC/FBC subspaces. In **Figure 1.13c**, we plot system trajectories the same system as **Figure 1.13a** along the direction of highest dynamic range within FBC (red, top) and FFC (black, bottom) subspace. FFC trajectory visibly containing stronger amplification in its initial dynamics. In **Figure 1.13d**, we plot the average dynamic range relative to random projections over the same range of non-normality as in panel (b). We observe a monotonic increase in the dynamic range within FFC subspaces relative to random projections, whereas FBC subspaces were found to actually decrease in their dynamic range relative to random projections as non-normality was increased sufficiently. Thus, we are able to recapitulate the key observations that FBC subspaces contain stronger rotational dynamics, whereas FFC subspaces container stronger amplification within a synthetic system that respects Dale's Law.

Rotational dynamics enhance stability and feedback controllability

To better understand why feedback controllability was consistently associated with stronger rotational dynamics than feedforward controllability, we investigated a simplified 2 dimensional linear dynamical system in which the relative contribution of scaling and rotations to the dynamics could be independently varied. We parameterized A as:

$$A = \begin{bmatrix} \epsilon - \delta & \phi \\ -\phi & \epsilon + \delta \end{bmatrix} \quad A_{\text{sym}} = \begin{bmatrix} (\epsilon - \delta) & 0 \\ 0 & (\epsilon + \delta) \end{bmatrix} \quad A_{\text{skew}} = \begin{bmatrix} 0 & \phi \\ -\phi & 0 \end{bmatrix}$$

The corresponding eigenvalues of the symmetric and skew-symmetric components are:

$$\lambda_{\text{sym}} = \{\epsilon + \delta, \epsilon - \delta\} \quad \lambda_{\text{skew}} = \{i\phi, -i\phi\}$$

Thus, as δ is increased, the largest eigenvalue of A_{sym} is also increased, whereas ϕ on the other hand increases the strength of rotational dynamics. The parameter ϵ was set to

-0.05 in order to keep the overall dynamics of A stable for most choices of δ and ϕ. We generated many 2 dimensional dynamics matrices by varying δ and ϕ over the interval $[0, 0.5]$ (corresponding to the x and y axes of **Fig. 1.14a**), spanning a range of dynamics from pure rotations to pure scalings. We set $B = I$ and measured the intrinsic controllability of systems without dimensionality reduction (i.e. $C = I$).

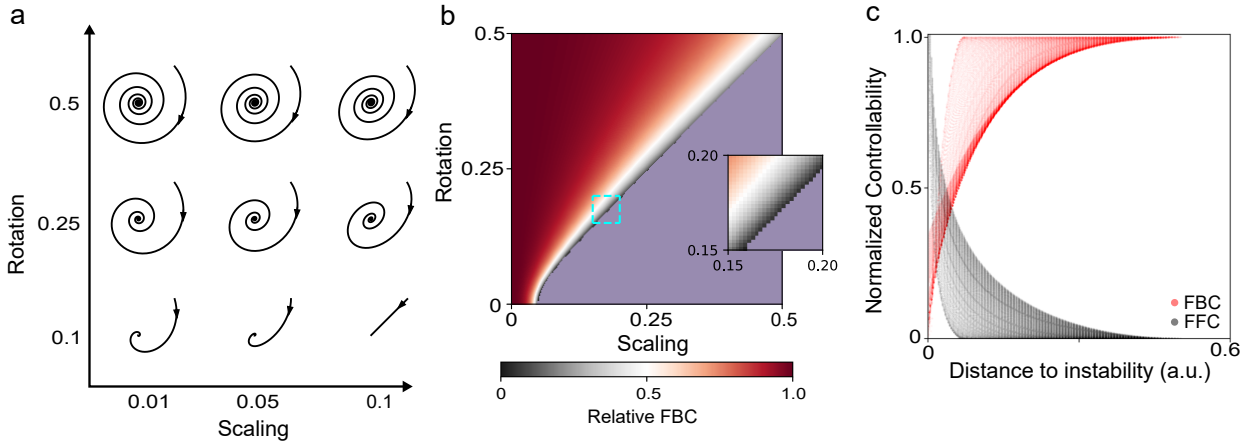


Figure 1.14: Feedback controllability is enhanced within stable dynamical systems. (a) Plot of example trajectories simulated from the 2D dynamical system as a function of scaling and rotational strength. (b) Colormap of ratio of normalized FBC to FFC. Parameters for which systems are more FBC than FFC are shaded red, whereas parameters for which systems are more FFC than FBC are shaded black. Purple region denotes parameters regime for which dynamics are unstable. (b, inset) Zooming into the dashed cyan region close to the instability boundary. (c) Scatter plot of normalized controllability (FBC in red, FFC in black) vs. distance to instability.

We calculated the FFC and FBC of dynamics over this parameter space, normalizing the FBC and FFC values attained to each line on a 0-1 scale. In **Figure 1.14b**, we plot a colormap of the relative FBC, defined as the normalized FBC divided by the sum of normalized FBC and FFC. Regions shaded red correspond to systems with high FBC and (relatively) low FFC, whereas regions shaded grey/black correspond to systems with high FFC and (relatively) low FBC. For fixed strength of rotation and increasing strength of scaling, the FFC increased while the FBC decreased. On the other hand, for fixed strength of scaling and increasing strength of rotations, the FBC increased while the FFC decreased (see also **Fig. ??**). For sufficiently large scaling strength, the system dynamics become unstable (purple region in **Fig. 1.14b**). We found that FFC increases as one approaches this instability boundary, while the FBC decreases (**Fig. 1.14b** inset). This suggested that dynamical systems closer to instability are more FFC, whereas systems that lie far away from this instability boundary are more FBC. Given the orientation of the instability contour in the rotation-scaling plane, we observe that for fixed rotatiationl strength, this distance increases with decreasing scaling strength, while for fixed strength of scaling, the distance increases with increasing rotational

strength. This fact establishes a link between stronger rotational dynamics and greater system stability and feedback controllability for a fixed strength of scaling dynamics.

Accordingly, a larger distance to instability was found to be strongly correlated with higher FBC, while a smaller distance to instability was strongly correlated with higher FFC. In **Figure 1.14c**, we scatter the normalized FBC and FFC as a function of the distance to instability for all points in the parameter space. We observed a sharp increase in FBC as the distance to instability initially departs from zero, followed by an eventual saturation. The FFC followed the opposite trends, diverging close to instability and decreasing rapidly as the distance to instability increases. Overall, the spearman correlation between the distance to instability and the FBC and FFC was found to be 0.90 and -0.84, respectively. These results establish that for a fixed degree of scaling dynamics, rotations enhance dynamical stability, and as a consequence, the controllability of dynamics under feedback. The greater stability, and therefore feedback controllability, afforded by rotations relative to other types of dynamics may therefore provided a normative account of their presence in cortical data.

1.3 Additional Characterization of the FCCA Method

FCCA exhibits low variability across initializations.

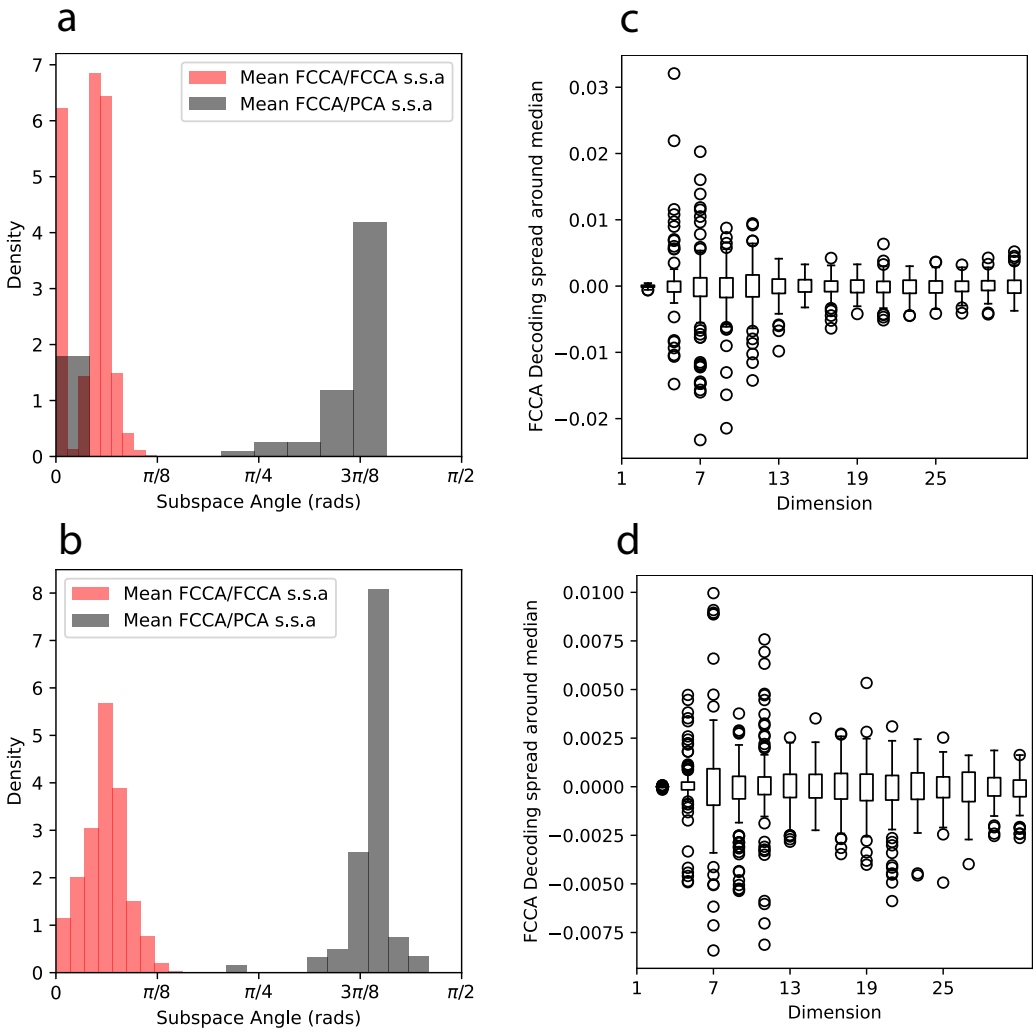


Figure 1.15: FCCA exhibits low variability across initializations. (a,b) Histogram of the average subspace angles between different $d = 6$ FCCA projections (red) and between FCCA and $d = 6$ PCA (black) taken across 20 random initializations of FCCA fit on M1 (a) and S1 (b) data. (c, d) Variation in cursor velocity prediction r^2 from M1 (c) and S1 (d) as a function of projection dimension. Spread indicates the maximum deviation from the median decoding performance over 20 initializations for each recording session.

FCCA is not a convex optimization problem. Throughout the above results, we initialized optimization over 10 different random orthogonal projection matrices and use the solution that returns the best value of the objective function. In order to assess the variability in

behavior of the method across initializations, we individually examined subspace angles and decoding performance across each of 20 random initializations for each projection dimensions in both M1 and S1. In **Figure 1.15 a, b**, we plot histograms of the average pairwise subspace angles between different initializations of FCCA across a subset of recording sessions (M1 in **a**, S1 in **b**). In both M1 and S1, we observe these subspaces angles are tightly clustered and bounded above by $\pi/8$ (red histogram bars). We also measured the average subspace angles between each initialization of FCCA and the corresponding PCA projection. The distribution of these subspace angles in both M1 and S1 was clustered around $3\pi/8$ (black histogram bars), though we did observe a small number of solutions in M1 that aligned very closely with PCA directions (black histogram bars in panel **a**).

We also trained decoders off the basis of each random initialization. As in the above results, we averaged decoding performance across 5 folds of the data. In **Figure 1.15 c, d** we plot the maximum deviation in decoding performance across initializations and recording sessions relative to the median performance at each dimension within each recording session. We observe that in both M1 (**c**) and S1 (**d**), the spread in r^2 about the median is no larger than 0.035 across initializations. We note that this gap is much smaller than the gap between FCCA and PCA (**Fig. 1.6 c, d**). Thus, while FCCA is a non-convex dimensionality reduction method, the key effects of (i) large subspace angle between FCCA and PCA and (ii) the superior decoding performance of FCCA relative to PCA hold consistently across initializations.

FCCA/PCA subspace angles remain large across dimensionality

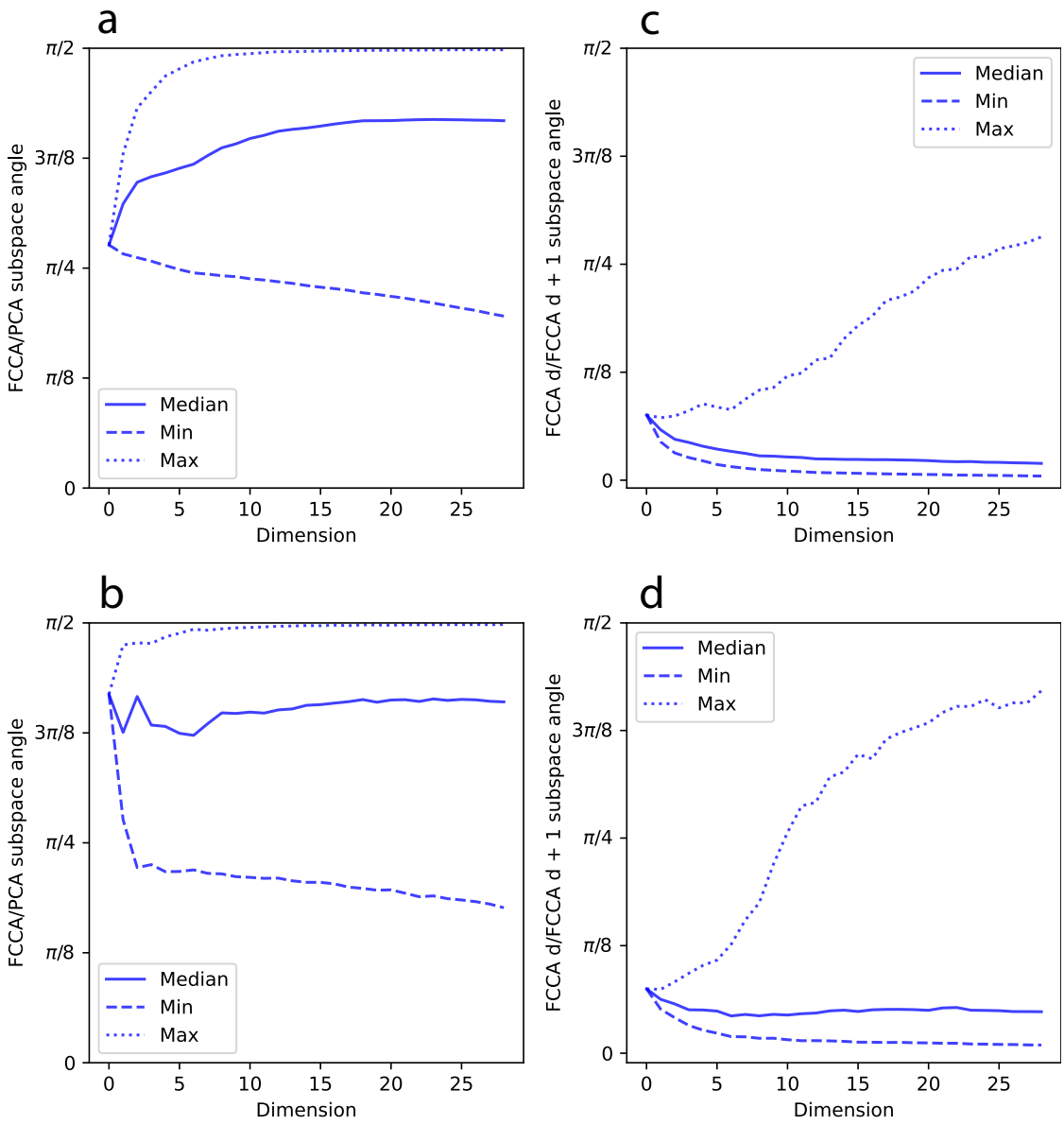


Figure 1.16: FCCA/PCA subspaces subspace angles remain large across dimensionality. (a,b) Comparison of the minimum, median, and maximum subspace angle between PCA and FCCA as a function of projection dimension in M1 (top) and S1 (bottom) (c, d) Comparison of the minimum, median, and maximum subspace angle between FCCA at dimension d vs. dimension $d+1$ within M1 (top) and S1 (bottom). The analogous curves for PCA (or any nested, orthogonal subspace method) would lie at 0 for all 3 statistics across all dimensions.

In **Figure 1.6**, we reported the average subspace angle between FCCA and PCA pro-

jections of dimension $d = 6$. Here, we calculate these angles across a range of projection dimensionalities in both M1 (**Fig. 1.16a**) and S1 (**Fig. 1.16b**). We found that the average subspace angle (solid blue trace) increased as a function of projection dimension, saturating at approximately $2\pi/5$ rads in both M1 and S1. The maximum subspace angle between projections (dotted line) reached $\pi/2$ by dimension 10 in both M1 and S1, whereas the minimum subspace angle (dashed line) decreased as a function of dimensionality. The latter result is to be expected, since as the projection dimension increases to the full dimension of the ambient space, this angle will decrease to zero. Overall then, FCCA and PCA subspaces remain geometrically segregated across projection dimensionality.

The objective function of FCCA is optimized separately for each desired projection dimension d. Furthermore, the optimal projection does not arise from the solution of an eigenvalue problem as in PCA. As a result, it is not necessarily the case that a projection of dimension $d+1$ will contain as a subspace the projection of dimension d (i.e., the subspaces may not be nested). This fact could potentially hamper interpretability, as projections of varying dimensionality may pick out completely disjoint regions of the neural state space. To rule out the presence of this sort of behavior within FCCA, we measured the subspace angles between successive projection dimensionalities. When comparing a projection of dimension d to a projection of dimension $d+1$, it is possible to measure a total of d subspace angles. In **Figure 1.16 c, d**, we plot the min, max, and median of these subspace angles in M1/S1 respectively, as a function of projection dimension. For a pair of nested subspaces, all 3 statistics will always be zero, as the dimension d subspace is entirely contained within the dimension $d+1$ subspace. As in the main text, we considered here the projections with the best FCCA score over 10 initializations. For FCCA, we observe that both the minimum (dotted line) and median (solid line) subspace angle in both M1 and S1 remain negligibly small ($< \pi/10$ radians), approaching zero as the projection dimension is increased. The maximum subspace angle by contrast was found to increase with projection dimensionality. These trends indicates that FCCA subspaces are *partially* nested, with most of the d dimensional subspace lying within the $d+1$ dimensional space.

FCCA returns consistent subspaces across T parameter and with the addition of observational noise.

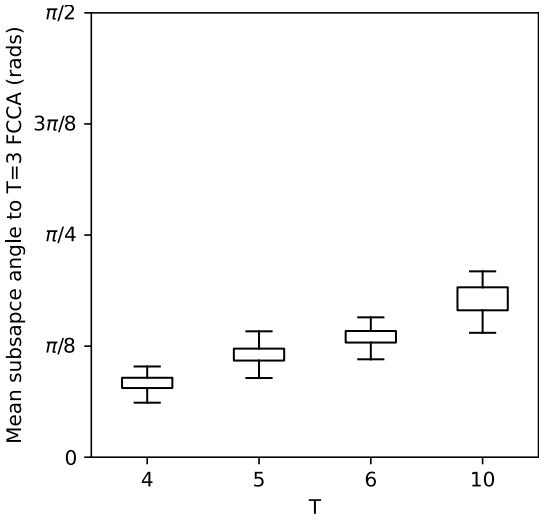


Figure 1.17: FCCA returns consistent subspaces across T parameter. Plot of median ± IQR of the average subspace angle between $d = 6, T = 3$ FCCA projections and FCCA projections that use varying T parameter (increasing along the x-axis). Spread is taken across folds and recording sessions within M1.

The only free hyperparameter within the FCCA method is the T parameter, which controls the numbers of observations within the FBC subspace that are used within the causal and acausal MMSE prediction of the neural state. In **Figure 1.17a**, we plot the average subspace angle between FCCA projections using $T = 3$ (the parameter we use for the main analyses), and various values of T (increasing along the x-axis). The spread is taken across folds and recording sessions within M1. We observe that the subspace angles increase as T increases from $T = 3$, though they remain small relative to the large subspace angles reported between FCCA and PCA in **Figure 1.6a**. In practice, results obtained from large values of T are also likely to be unreliable as the accuracy of autocorrelation matrix estimates at long lags will diminish. Thus, overall, we find FCCA to be relatively insensitive to the choice of T.

1.4 Discussion

The theoretical importance of feedback control for brain function and behavior have been recognized for nearly 80 years [10, 55]. Despite the overwhelming evidence supporting feedback control as a normative theory of behavior, if and how feedback control explains on-going

neural population dynamics has been largely unarticulated (though see: [21, 56]) and completely untested in experimental data. For example, the neural population dynamics of monkey M1 during reaching provides a heavily studied example of a neural circuit executing computations through population dynamics. Recent literature postulates that during reach execution, M1 operates as a feedforward dynamical system: initial conditions are first set within an output null subspace [57] and then movement is supported by the "set and forget" time evolution of neural population dynamics [58]. Neural population dynamics have been proposed to function as a basis set of dynamical motifs that are transformed downstream into muscle commands. An implicit prediction of this view is that in the absence of external behavioral perturbations, the neural population dynamics of M1 generate a reach in a feedforward manner. At the same time, the canonical examples of neural circuits thought to be operating in a feedforward manner are primary sensory cortices (e.g., V1) during simple perception. These circuits are typically conceived as transmitting processed sensory information up the sensory hierarchy (e.g., V1 $\rightarrow$ V2). Indeed, a classic view of sensory perception is that lower-level representations form a basis set for synthesis of higher-level representations [59]. However, from a dynamics perspective, if M1 is operating in a feedforward manner, why then are its neural population dynamics (which have a strong rotational component) so different from, e.g., V1 (which did not have a strong rotational component), during the same task [60]?

In contrast to the predominate view articulated above, we hypothesized that neural population dynamics in a given brain area (e.g., M1 and S1) maybe steered in real-time to maintain trajectories and achieve desired end-states for behavior based on feedback control. A key prediction of this hypothesis is that neural subspaces that are most feedback controllable (FBC) should be more aligned with behavior than neural subspaces that are most feedforward controllable (FFC). Supporting our hypothesis, we found that FBC subspaces of neural population activity within both M1 and S1 were substantially better predictors of reach kinematics than FFC subspaces. Notably, prediction performance within FBC subspaces saturated at a lower dimensionality than FFC subspaces indicating that, relative to FFC, FBC effectively compresses behaviorally relevant information. In this context, the low dimensionality of FBC subspaces implies that the controllers required to steer behaviorally relevant dynamics in M1 and S1 themselves have low state dimension, potentially requiring simpler circuits to implement.

We found that FBC subspaces were nearly orthogonal to FFC controllable subspaces in both M1 and S1. This raises the possibility that these distinct modes of control maybe differentially engaged depending on the area and task demands. The possibility of distinct modes of control is supported by our finding that there are distinct feedforward and feedback controllable neural populations in both M1 and S1. Functionally, the time-courses of the observed neural populations suggest that FFC dynamics may support reach initiation (before sensory feedback has time to re-enter the system), while FBC dynamics support on-going reaching. Neurobiologically, these distinct functions could be implemented by anatomically and/or genetically different populations of neurons.

Harmonizing the neuron doctrine with the theory of computations through

neuronal population dynamics. Neurons are fundamental units of computation in the brain [61, 62]. A central tenet of the neuron doctrine is that neurons are specialized to perform specific functions (Dales law is another tenet). At the same time, it is becoming clear that brain computations result from the dynamics of neural populations [24]. We found that FBC is mediated by a distinct population of single units and is an emergent, population level property of that population. Specifically, we found that the neurons most important for FBC had lower task modulated firing rates and fast time-scales of firing-rate dynamics. Across the population, they had heterogeneous temporal relationships relative to each other. Despite their heterogeneous dynamics, the structure of the FBC subspace was an emergent property depending heavily on inter-neuronal interactions. In contrast, the neurons most important for FFC had higher firing rates and slower time-scales of firing-rate dynamics. Across the population, they homogeneously exhibited a transient burst of activity near the onset of reach initiation, followed by a decay. Despite their homogeneous dynamics, the structure of the FFC subspace was only modestly dependent on inter-neuronal interactions, and could almost entirely be explained by single-unit firing rate variance.

The Utah array recordings we analyzed targeted L5 in both M1 and S1 of macaques. Excitatory pyramidal cells constitute the overwhelming majority of neurons ($> 85\%$ [63]) and there is a known sampling bias of *in vivo* extracellular electrophysiology for excitatory neurons. Therefore, it is very likely that the distinct neuron populations underlying feedforward vs. feedback controllability correspond to different classes of excitatory neurons in L5. Indeed, diverse studies in multiple species and brain areas have found that there are two major distinguishable classes of L5 pyramidal neurons: extratelencephalicephalic (ET) and intratelencephalic (IT) [64, 65]. Anatomically, ET neurons project to the thalamus, mid-brain, and brainstem, with only modest intra-columnar connectivity. Electrophysiologically, these neurons have higher firing rates and a propensity for initial bursts of action potentials followed by sustained firing. Computationally, ET neuron are thought to transmit information and be involved in behavioral initiation. This constellation of properties maps well to the response characteristics and computations of the feedforward controllable neural population. In contrast, anatomically, IT neurons contain both local and long-range connectivity profiles projecting to other cortical areas and the striatum. Electrophysiologicaly these neurons typically have lower firing rates and more complicated and heterogenous firing rate properties. Computationally, IT neurons are thought to be involved in more sophisticated dynamics and information processing such as planning and sampling [64]. This constellation of properties maps well to the response characteristics and computations of the feedback controllable neural population. We note that the FBC population had rapid dynamics and low evoked responses, indicating that Ca2+ imaging studies, with their slow response times and low SNR, maybe insufficient tools to examine these issues. Interestingly, intra- and extratelencephalic sub-populations also have distinct profiles of neurotransmitter receptors, perhaps providing the basis for separable modulation of these populations underlying distinct modes of neural circuit control [64].

Analytic framework to link subspaces to single neurons and networks performing specific computations. Our analytic framework differs from the mathematical

structure of many statistical machine learning methods used to analyze neural population data. In particular, a popular methodological approach aims to reconstruct the neuronal activities and models them as generated by dynamical latent states [66–68]. While mathematically convenient, this approach suffers from several conceptual shortcomings when it comes to neurobiological interpretation and insight. In particular, because the objective function of these methods is typically to reconstruct all the observed neural data, the extracted latent states tend to favor capturing variance in the firing rates. However, capturing variance is not a principle of neural computation *per se*, and *a priori* it is not necessary that the high-variance directions are the ones most important for a specific computation (though, as we found, these subspaces correspond to FFC). In tasks with a handful of known degrees of freedom (e.g., animal location in a maze), the time evolution of the low dimensional latent states can be manually interrogated [69]. However, in general, these latent states in of themselves provide no direct insight into the structure and function of the observed neural dynamics. Additionally, it is the neurons themselves and their networked interactions that generate neural computations that can be summarized as latent states. However, inverting the mapping from a latent state space to the state space spanned by the neurons (i.e., the real physical degrees of freedom of the brain) is often an ill-posed inverse problem. This makes it very challenging to identify which sets of neurons differentially contribute to different computations. As experimental neuroscientists record from ever lager populations of neurons, the observed neural state space is likely to contain within it multiple subpopulations engaged in distinguishable computations. Requiring latent state models to preserve variance across an entire dataset may obfuscate the role played by different populations of neurons, in particular subpopulations with low firing rates.

We took a fundamentally different approach to analyze neural population data that addresses these shortcomings. In particular, we formulated a novel normative computational principle (feedback controllability) and derived an objective function for a dimensionality reduction method (FCCA) encoding this principle. This allowed us to directly identify subspaces generated by the networked interactions of the observed neural activities implementing that computational principle. As we showed, the FBC subspace was nearly orthogonal to the subspace that maximized variance (i.e., PCA). This provides a concrete example that methods that preserve variance across an entire data set (i.e., reconstruct the data) may miss computationally important aspects of neural population dynamics. Additionally, as the FBC subspaces had better decoding performance, this suggest that the FCCA objective maybe a fruitful direction for brain-computer interface methodology. Furthermore, as the neural dynamics matrix (A) encodes the influence of every neuron on the change of every other neuron over time, it can be interpreted as the functional connectivity of the neurons [70]. While not the goal of the current work, analysis of functional connectivity associated with different computations may provide insights into the networks of neurons that generate those computations. This emphasizes the importance of continued methodological development for accurate functional connectivity estimation [71,72]. As the subspaces identified by our methods are obtained from projections of the observed firing rates, rather than through latent variable inference, the mapping between the observed neurons and the subspace is automat-

ically obtained. This allowed us to identify sub-populations of neurons that generated FBC vs. FFC subspaces which in turn had distinct firing rate profiles, functional properties, and strengths of interactions.

Non-normality of neural circuits enables different dynamical regimes for distinct modes of control. We found that feedforward and feedback controllable subspaces are distinguishable in neural population data due to the the non-normality of the underlying dynamics. The work of Hennequin et al [48] suggests that non-normality arising from finely tuned excitatory/inhibitory balance provides a mechanism for rapid amplification of firing rates upon reach initiation and subsequently a rich set of transients for use in the synthesis of movement. Our results significantly generalize this picture: instead of requiring fine-tuning, high dimensional non-normal systems such as the brain generically contain subspaces with distinct controllability properties. That is, as non-normality is a necessary consequence of asymmetric synaptic connectivity implied by Dale's Law, having subspaces with distinct controllability properties may be an unavoidable feature of canonical microcircuits.

We found that FBC subspace dynamics exhibited strong rotations, while FFC subspaces exhibited strong, transient amplifications. We showed that systems with stereotyped, purely rotational dynamics maximize FBC, while feedforward controllability is maximized by FFC. That is, in a high dimensional non-normal system like the brain with a mixture of amplification and rotational dynamics, the latter are more heavily expressed in the feedback controllable subspace. Thus, not only are rotational dynamics consistent with a feedback controller, as shown in [56], they may arise as a necessary consequence of optimization for the stability of feedback control. This distinction between amplification for feedforward control vs. rotations for feedback control provides a normative account of rotational dynamics observed previously [25]. As described above, the population of single-units underlying FBC had modest task-dependent firing rate modulation but complex dynamics, while the population of single-units underlying FFC had large task-dependent firing rate modulation but simple dynamics. Not only do these findings map onto the properties of L5 intra- and extratelencephalic pyramidal neurons, respectively, they intuitively map onto the rotational and amplification dynamics expressed at the population level that we found. Both the non-normality of neural circuits due to Dales law and the distinguishing characteristics of L5 intra- and extratelencephalic neurons are ubiquitous across sensory, cognitive, and motor cortical areas [63–65]. As such, there is every reason to believe that similar principles of control will apply throughout cortex.

We speculate that the degree to which a given brain areas' neural population dynamics are feedforward vs. feedback controllable may depend on task demands. For example, in sensory areas, it could be that when the task is easy (e.g., unambiguous stimuli), the need for top-down feedback control is minimal. In such cases, the area (e.g., V1) may operate in a more feedforward mode of control, in which the stimuli are the primary determinants of neural population dynamics. The transient amplification dynamics characteristic of feedforward control may render sensory features distinguishable, as suggested previously [41]. Conversely, when the task is difficult (e.g., ambiguous or otherwise distorted stimuli), neural population dynamics may operate in a more feedback mode of control, where the neural population

dynamics is driven by both the sensory stimuli as well as top-down signals. In such cases, state feedback must be employed to steer neural population dynamics in real-time in the presence of noisy neural activity and biophysical delays. The rotational dynamics characteristic of FBC may underlie task-relevant processing. Indeed, a theory of sensory perception, predictive coding, can be viewed as a special case of feedback control, in which the dynamics of a lower-level sensory area are controlled to minimize the transmission of redundant information. Understanding if and how the controllability of diverse cortical areas is modulated by top down feedback processes [15, 73] and task demands presents an interesting direction for future work for which our methodology can be deployed.

Furthermore, we have linked two modes of neural data analysis that heretofore have remained disconnected: extraction of subspaces of neural population dynamics and characterization of the functional properties of single neurons. As techniques advance to elucidate the electrophysiological, molecular, and connectomic taxonomy of single neurons co-registered with recordings of their dynamics, analysis frameworks that connect these two levels of description will provide necessary insights into structure-function relationships. In the long-term, this may advance understanding of disruptions of neural population dynamics in the context of the properties of distinct neuronal populations, and thereby enable treatment/control of brain disorders through targeted interventions. Together, these results indicate that feedback control is a unifying theory of brain and behavior, and suggest it maybe a general theory of neural population dynamics across the brain.

1.5 Proof of equivalence of FFC and FBC for stable, normal A

In this section, we prove the equivalence of the solutions of the FFC (eq. 1.4) and FBC objective functions (eq. 1.7) when system dynamics are stable and symmetric. We focus on symmetric matrices as the requirement that dynamics be stable (i.e., all eigenvalues of the dynamics A must have negative real part) essentially reduces the space of normal matrices to that of symmetric matrices. We reproduce these objective functions for convenience:

$$\begin{aligned} C_{\text{FFC}}: &\quad \text{argmax}_C \log\det C\Pi C^\top \\ C_{\text{FBC}}: &\quad \text{argmin}_C \text{Tr}(PQ) \end{aligned}$$

We prove this theorem when the matrix P in the FBC objective function arises from the canonical LQR loss function:

$$\min_u \left\{ \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T x^\top x + u^\top u \, dt\right], \;\; x(0) = x_0, u \in L^2[0,\infty) \right\}$$

and not the variant given in eq. 1.10. When calculating FBC from data within FCCA, we must use the latter LQR loss function as it maps onto acausal filtering, and therefore

may be estimated from data. Recall from the discussion below eq. 1.8 that within the FFC objective function, we assess controllability when the output/observation matrix C is used as the input matrix for the regulator signal (i.e., we make the relabeling $B^\top \to C$. We further work under the assumption that the input matrix B to the open loop system is equal to the identity. The open loop dynamics of $x(t)$ are then given by:

$$\dot{x} = Ax(t) + u(t) \tag{1.12}$$

where $u(t)$ has the same dimensionality as $x(t)$, and is uncorrelated with the past of $x(t)$ (i.e. $u(t) \perp x(\tau), \tau < t$). Formally, $u(t)$ represents the innovations process of $x(t)$. The equations for Q (corresponding to the Kalman Filter, eq. 1.5) and the equation for P (corresponding to the LQR, eq. 1.6) reduce to the following:

$$AQ + QA + I_N - QC^\top CQ = 0 \tag{1.13}$$

$$AP + PA + I_N - PCC^\top P = 0 \tag{1.14}$$

where I_N denotes the $N \times N$ identity matrix.

We observe that under the stated assumptions, the Riccati equations for Q and P actually coincide, and thus the FBC objective function reads $Tr(Q^2)$. We will show that both FFC and FBC objective functions achieve local optima for some fixed projection dimension d when the projection matrix C coincides with a projection onto the eigenspace spanned by the d eigenvalues of A with largest real part, which we denote as V_d. In fact, in the case of the FFC objective function, the eigenspace corresponds to a global optimum. Intuitively, in the case of symmetric, stable, A, perturbations exponentially decay in all directions, and so the maximum response variance is contained in the subspace with slowest decay.

For the FBC objective function, we are able to establish global optimality rigorously for the $2D \to 1D$ dimension reduction. The intuition for the slow eigenspace of A serving as a (locally) optimal projection in this case is then given by the fact that state reconstruction from past observations, the goal of the Kalman filter, will occur optimally using observations that have maximal autocorrelations with future state dynamics. Similarly, for the LQR, for a fixed rank input, the most variance will be suppressed by regulating within the subspace with slowest relaxation dynamics.

We briefly outline the proof strategy. First, we will prove the optimality of V_d for the FFC objective function in section S1.9.1 by showing that (i) V_d is an eigenvector of Π in the case when A is symmetric and (ii) relying on the Ky Fan maximum principle. Then, in section S1.9.2, we will prove that V_d is a critical point of the FBC objective function. The proof relies on an iterative technique to solve the Riccati equation. These iterates form a recursively defined sequence that provide increasingly more accurate approximations to the FBC objective function that converge in the limit. Treating these iterative approximations of the FBC objective function as a function of C, we show that V_d is a critical point of all iterates, and thus in the limit, V_d is a critical point of the FBC objective function.

S1.9.1 FFC Objective Function

Theorem 3 *For $B = I_N, A = A^\top, A \in \mathbb{R}^{N\times N}$, with all eigenvalues of A distinct and $\max Re(\lambda(A)) < 0$, the optimal solution for the feedforward controllability objective function for projection dimension d coincides with the eigenspace spanned by the d eigenvalues with largest real value.*

Proof

Let V_d denote a matrix whose column space coincides with the eigenspace spanned by the d eigenvalues of A with largest real part. We will first show that V_d solves the FFC objective function:

$$\begin{aligned} &\text{argmax}_C \log\det C\Pi C^\top \\ &\Pi = \int_0^\infty dt\; e^{At}BB^\top e^{A^\top t} = \int_0^\infty dt\; e^{2At} \end{aligned} \tag{1.15}$$

Let $A = U\Lambda U^\top$ denote the eigenvalue decomposition of A. Recall that since $A = A^\top$, U is orthogonal. Then we can write:

$$\begin{aligned} \Pi &= U\int_0^\infty dt e^{2\Lambda t}U^\top \\ &= \frac{1}{2}UDU^\top \end{aligned}$$

where D is a diagonal matrix with diagonal entries $\{\frac{1}{-\lambda_1}, \frac{1}{-\lambda_2}, ..., \frac{1}{-\lambda_N}\}$. We conclude that the matrix Π has the same eigenbasis as A. Also, since all λ_j are real and negative, the ordering of the eigenvalues is preserved ($\lambda_i > \lambda_j$ implies $-\frac{1}{\lambda_i} > -\frac{1}{\lambda_j}$). That V_d solves 1.15 follows from the Ky Fan principle [74], which we restate for convenience:

Proposition 1 *Ky Fan Maximum Principle*

Let A be any square matrix, and let $\sigma_1 > \sigma_2 > \sigma_3$ be its singular values. Then:

$$\sup|\det U_1 A U_2| = \prod_{i=1}^{d}\sigma_i$$

where the supremeum is taken with respect to all unitary matrices U_1, U_2 of rank d.

We observe that the choice of $U_1 = U_2 = V_d$ saturates the upper bound. □

S.1.9.2 FBC Objective Function

For the case of the FBC objective function, we show that projection matrices of rank d

that align with the d slowest eigenmodes of A constitute local minima of the objective function. We rely on two simplifying features of the problem. First, the FBC objective function is invariant to the choice of basis in the state space. We therefore work within the eigenbasis of A, as within this basis, the system defined by eq. 1.12 decouples into n non-interacting scalar dynamical systems. Additionally, we rely on the fact that the FBC objective function is also invariant to coordinate transformations within the projected space. In other words, the choice of coordinates in which we express y also makes no difference. Without loss of generality then, we may treat the problem in a basis where A is diagonal with entries given by its eigenvalues and C is an orthonormal projection matrix (i.e. $CC^\top = I_d$). A restatement of the latter condition is that C belongs to the Steifel manifold of $N \times d$ matrices: $\Omega \equiv \{C \in \mathbb{R}^{N\times d} | CC^\top = I_d\}$.

Theorem 4 *For $B = I_N, A = A^\top, A^{N\times N}$, with all eigenvalues of A distinct and $\max Re(\lambda(A)) < 0$, the projection matrix onto the eigenspace spanned by the d eigenvalues of A with largest real value constitutes a critical point of the LQG trace objective function on Ω*

Proof Explicitly calculating the gradient of the solution of the Riccati equation is analytically intractable for $n > 1$, and so we we will rely on the analysis of an iterative procedure to solve the Riccati equation via Newton's method, known as the Newton-Kleinmann (NK) iterations [75]. These iterations are described in the following proposition:

Proposition 2 *Consider the Riccati equation $0 = AQ + QA^\top + BB^\top - QC^\top CQ$. Let $Q_m, m = 1, 2, ...$ be the unique positive definite solution of the Lyapunov equation:*

$$0 = A_k Q_m + Q_m A_k^\top + BB^\top + V_{k-1} C^\top C V_{k-1} \tag{1.16}$$

where $A_k = A - C^\top C V_{k-1}$, and where V_0 is chosen such that A_1 is a stable matrix (i.e. all real parts of its eigenvalues are < 0). For two positive semidefinite matrices M, N, we denote $M \geq N$ if the difference $M - N$ remains positie semidefinite. Then:

1. $Q \leq Q_{m+1} \leq Q_m \leq ..., k = 0, 1$
2. $\lim_{k\to\infty} Q_m = Q$

Thus the Q_m iteratively approach the solution of the Riccati equation from above. Since in our case, the Riccati equations for P and Q coincide, an identical sequence P_k can be constructed using analogous NK iterations that approaches P from above. From this, it follows that $\lim_{k\to\infty} Tr(Q_m P_k) = \lim_{k\to\infty} Tr(Q_m^2) = Tr(Q^2)$. We then use the fact that in addition to the Q_m converging to Q, the sequence $\nabla_C \mathrm{Tr}\,(Q_m^2)$ converges to $\nabla_C \mathrm{Tr}(Q^2)$ as $k \to \infty$, where ∇_C denotes the gradient with respect to C. This is rigorously established in the following lemma, which is the multivariate generalization of Theorem 7.17 from [76]:

Lemma 1 *Suppose $\{f_m\}$ is a sequence of functions differentiable on an interval $h \subset \mathcal{H}$, where $\mathcal{H}$ is some finite-dimensional vector space, such that $\{f_m(x_0)\}$ converges for some point $x_0 \in h$. If $\{\nabla f_m(x_0)\}$ converges uniformly in h, then $\{f_m\}$ converges uniformly on I, to a function f, and*

$$\nabla f(x) = \lim_{m\to\infty} \nabla f_m(x) \quad x \in h$$

Here, the $\{f_m\}$ are the Newton-Kleinmann iterates Q_m, and x_0 corresponds to the C matrix that projects onto the slow eigenspace of A. The NK iterates are known to converge uniformly over an interval of possible C matrices (in fact any such C matrix for which there exists a K such that $A - C^T C K$ is a stable matrix) [75].

We will calculate the gradient $\nabla_C Q_m$ on Ω by explicitly calculating the directional derivatives of Q_m over a basis of the tangent space of Ω at C_{slow}. Any element Ψ belonging to the tangent space at $C \in \Omega$ can be parameterized by the following [77]:

$$\Psi = CM + (I_N - CC^\top)T$$

where M is skew symmetric and t is arbitrary. Let C_{slow} be the projection matrix onto the slow eigenspace of A of dimension d. Since we work in the eigenbasis of A, $C_{\text{slow}} = \begin{bmatrix} I_d & 0 \end{bmatrix}$. At this point, elements of the tangent space take on the particularly simple form

$$\Psi = \begin{bmatrix} M & T \end{bmatrix}$$

where now M is a $d \times d$ skew symmetric matrix and $T \in \mathbb{R}^{d\times(N-d)}$ is arbitrary. A basis for the tangent space is provided by the set of matrices $\{M_{ij}, T_{kl}, i = 2, ...d, j = 1, ..., i-1, k = 1, ..., d, l = 1, ..., N-d\}$ where M_{ij} is a matrix with entry 1 at index (i, j) and -1 at index (j, i) and zero otherwise, and T_{kl} is the matrix with entry 1 at index (k, l) and zero otherwise. Denote by $D_\Psi Q_m$ the directional derivative of Q_m along the direction of Ψ, viewing Q_m as a function of C (denoted $Q_m[C]$):

$$D_\Psi Q_m = \lim_{\alpha\to 0} \frac{Q_m[C_{\text{slow}} + \alpha\Psi] - Q_m[C_{\text{slow}}]}{\alpha} \tag{1.17}$$

Let $\Psi_{ij,kl}$ denote the tangent matrix $\begin{bmatrix} M_{ij} & T_{kl} \end{bmatrix}$. Before calculating $Q_m(C_{\text{slow}} + \alpha\Psi_{ij,kl})$ explicitly, we first observe that as long as the NK iterations are initialized with a diagonal Q_0, then the diagonal nature of $C_{\text{slow}}^\top C_{\text{slow}}$ ensures that all Q_m will subsequently remain diagonal matrices. In fact, it can be shown that $\lim_{k\to\infty} Q_m = Q$ will also be diagonal, in this case. We write A in block form as $\begin{bmatrix} \Lambda_{||} & 0 \\ 0 & \Lambda_\perp \end{bmatrix}$, and similarly $Q_{m-1} = \begin{bmatrix} \mathcal{Q}_{||} & 0 \\ 0 & \mathcal{Q}_\perp \end{bmatrix}$, where $\Lambda_{||}, \mathcal{Q}_{||}$ are $d \times d$ diagonal matrices defined on the image of C_{slow} and $\Lambda_\perp, \mathcal{Q}_\perp$ are diagonal matrices defined on the kernel of C_{slow}. We denote the individual diagonal elements of $\Lambda_{||}, \mathcal{Q}_{||}$ as $\lambda_i, \mathcal{Q}_i, i = 1, ..., d$ and of $\Lambda_\perp, \mathcal{Q}_\perp$ as $\lambda_i, \mathcal{Q}_i, i = d, ..., N-d$. Then, equation 1.16 becomes:

$$\left(\begin{bmatrix}\Lambda_{||} & 0\\ 0 & \Lambda_{\perp}\end{bmatrix} - \begin{bmatrix}(I_d-\alpha^2 M_{ij}^2)\mathcal{Q}_{||} & (\alpha T_{kl}+\alpha^2 M_{ij}^{\top}T_{kl})\mathcal{Q}_{\perp}\\ (\alpha T_{kl}^{\top}+\alpha^2 T_{kl}^{\top}M_{ij})\mathcal{Q}_{||} & \alpha^2 T_{kl}^{\top}T_{kl}V_{\perp}\end{bmatrix}\right) Q_m[C_{\text{slow}}+\Psi_{ij,kl}]$$
$$+ Q_m[C_{\text{slow}}+\Psi_{ij,kl}]\left(\begin{bmatrix}\Lambda_{||} & 0\\ 0 & \Lambda_{\perp}\end{bmatrix} - \begin{bmatrix}\mathcal{Q}_{||}(I_d-\alpha^2 M_{ij}^2) & \mathcal{Q}_{||}(\alpha T_{kl}+\alpha^2 M_{ij}^{\top}T_{kl})\\ V_{\perp}(\alpha T_{kl}^{\top}+\alpha^2 T_{kl}^{\top}M_{ij}) & V_{\perp}\alpha^2 T_{kl}^{\top}T_{kl}\end{bmatrix}\right)$$
$$+ I_N + \begin{bmatrix}\mathcal{Q}_{||}(I_d-\alpha^2 M_{ij}^2)\mathcal{Q}_{||} & \mathcal{Q}_{||}(\alpha T_{kl}+\alpha^2 M_{ij}^{\top}T_{kl})\mathcal{Q}_{\perp}\\ \mathcal{Q}_{\perp}(\alpha T_{kl}^{\top}+\alpha^2 T_{kl}^{\top}M_{ij})\mathcal{Q}_{||} & \alpha^2 V_{\perp}T_{kl}^{\top}T_{kl}V_{\perp}\end{bmatrix} = 0 \tag{1.18}$$

where we have used $M^{\top} = -M$. The equivalent equation for $Q_m(C_{\text{slow}})$ reads:

$$\left(\begin{bmatrix}\Lambda_{||} & 0\\ 0 & \Lambda_{\perp}\end{bmatrix} - \begin{bmatrix}\mathcal{Q}_{||} & 0\\ 0 & 0\end{bmatrix}\right) Q_m[C_{\text{slow}}] + Q_m[C_{\text{slow}}]\left(\begin{bmatrix}\Lambda_{||} & 0\\ 0 & \Lambda_{\perp}\end{bmatrix} - \begin{bmatrix}\mathcal{Q}_{||} & 0\\ 0 & 0\end{bmatrix}\right) + I_N + \tag{1.19}$$
$$\begin{bmatrix}\mathcal{Q}_{||}^2 & 0\\ 0 & 0\end{bmatrix} = 0 \tag{1.20}$$

This latter equation is easily solved to yield:

$$Q_m[C_{\text{slow}}] = \begin{bmatrix}\frac{1}{2}\left(I_d+\mathcal{Q}_{||}^2\right)\left(\mathcal{Q}_{||}-\Lambda_{||}\right)^{-1} & 0\\ 0 & -\frac{1}{2}\Lambda_{\perp}^{-1}\end{bmatrix}$$

To explicitly solve the former equation, we recall that the matrices M_{ij} and T_{kl} have only two and one nonzero terms, respectively. M_{ij}^2 contains two nonzero terms at index (i,i) and (j,j). $T_{kl}^{\top}T_{kl}$ contains one non-zero term at index (l,l). $M_{ij}^{\top}T_{kl}$ contains a single nonzero term at (i,l) or (j,l) only if $k=i$ or $k=j$, respectively. Accordingly, we distinguish between where $k=i$ or $k=j$ (without loss of generality we may assume that $k=j$), and where $k\neq i$ and $k\neq j$.

In what follows, we will denote the (i,j) entry of $Q_m[C_{\text{slow}}+\alpha\Psi_{ij,kl}]$ as q_{ij}.

1. *Case 1:* $k=j$ In this case, careful inspection of eq. 1.18 reveals that it differs from eq. 1.20 only within a 3×3 subsystem:

$$\begin{bmatrix}\mathcal{S}_{11} & \mathcal{S}_{12} & \mathcal{S}_{13}\\ \mathcal{S}_{21} & \mathcal{S}_{22} & \mathcal{S}_{23}\\ \mathcal{S}_{31} & \mathcal{S}_{32} & \mathcal{S}_{33}\end{bmatrix} = 0$$

Note that this matrix is symmetric, yielding 6 equations for 6 unknowns:

$$\begin{aligned}
\mathcal{S}_{11} &= \alpha^2\mathcal{Q}_i^2 + 2\alpha^2\mathcal{Q}_{d+l}q_{i,d+l} + \mathcal{Q}_i^2 + 2q_{ii}\left(-\alpha^2\mathcal{Q}_i + \lambda_i - \mathcal{Q}_i\right) + 1 \\
\mathcal{S}_{12} &= \alpha^2\mathcal{Q}_{d+l}q_{j,d+l} - \alpha\mathcal{Q}_{d+l}q_{i,d+l} + q_{ij}\left(-\alpha^2\mathcal{Q}_i + \lambda_i - \mathcal{Q}_i\right) + q_{ij}\left(-\alpha^2\mathcal{Q}_j + \lambda_j - \mathcal{Q}_j\right) \\
\mathcal{S}_{13} &= -\alpha^2\mathcal{Q}_i\mathcal{Q}_{d+l} + \alpha^2\mathcal{Q}_i q_{ii} + \alpha^2\mathcal{Q}_{d+l}q_{d+l} - \alpha\mathcal{Q}_j q_{ij} + \\
&\quad q_{i,d+l}\left(-\alpha^2\mathcal{Q}_{d+l} + \lambda_{d+l}\right) + q_{i,d+l}\left(-\alpha^2\mathcal{Q}_i + \lambda_i - \mathcal{Q}_i\right) \\
\mathcal{S}_{22} &= \alpha^2\mathcal{Q}_j^2 - 2\alpha\mathcal{Q}_{d+l}q_{j,d+l} + \mathcal{Q}_j^2 + 2q_{jj}\left(-\alpha^2\mathcal{Q}_j + \lambda_j - \mathcal{Q}_j\right) + 1 \\
\mathcal{S}_{23} &= \alpha^2\mathcal{Q}_i q_{ij} + \alpha\mathcal{Q}_j\mathcal{Q}_{d+l} - \alpha\mathcal{Q}_j q_{jj} - \alpha\mathcal{Q}_{d+l}q_{d+l,d+l} + q_{j,d+l}\left(-\alpha^2\mathcal{Q}_{d+l} + \lambda_{d+l}\right) + \\
&\quad q_{j,d+l}\left(-\alpha^2\mathcal{Q}_j + \lambda_j - \mathcal{Q}_j\right) \\
\mathcal{S}_{33} &= 2\alpha^2\mathcal{Q}_i q_{i,d+l} + \alpha^2\mathcal{Q}_{d+l}^2 - 2\alpha\mathcal{Q}_j q_{j,d+l} + 2q_{d+l,d+l}\left(-\alpha^2\mathcal{Q}_{d+l} + \lambda_{d+l}\right) + 1
\end{aligned}$$

Direct solution is still infeasible, but noting our interest is in the behavior of solutions as $\alpha \to 0$, and only terms of $O(\alpha)$ will survive in the limit in eq. 1.17, we consider solving these equations perturbatively. That is, we express each q_{ij} in a power series in α: $q_{ij} = q_{ij}^{(0)} + q_{ij}^{(1)}\alpha + O(\alpha^2)$. One obtains each coefficient in the expansion by plugging this form into the above matrix and setting all terms of the corresponding order in α to 0. The lowest order term, $q_{ij}^{(0)}$, coincides with the solution of the unperturbed system, eq. 1.20. Plugging in the expansion into the 3×3 subsystem above, as well as the solution of the unperturbed system, and collecting all coefficients proportional to α yields the following system of equations:

$$\begin{bmatrix} \mathcal{S}_{11}^{(1)} & \mathcal{S}_{12}^{(1)} & \mathcal{S}_{13}^{(1)} \\ \mathcal{S}_{21}^{(1)} & \mathcal{S}_{22}^{(1)} & \mathcal{S}_{23}^{(1)} \\ \mathcal{S}_{31}^{(1)} & \mathcal{S}_{32}^{(1)} & \mathcal{S}_{33}^{(1)} \end{bmatrix} = 0$$

$$\begin{aligned}
\mathcal{S}_{11}^{(1)} &= 2\lambda_i q_{ii}^{(1)} - 2\mathcal{Q}_i q_{ii}^{(1)} \\
\mathcal{S}_{12}^{(1)} &= \lambda_i q_{ij}^{(1)} + \lambda_j q_{ij}^{(1)} - \mathcal{Q}_i q_{ij}^{(1)} - \mathcal{Q}_j q_{ij}^{(1)} \\
\mathcal{S}_{13}^{(1)} &= \lambda_i q_{i,d+l}^{(1)} + \lambda_{d+l} q_{i,d+l}^{(1)} - \mathcal{Q}_i q_{i,d+l}^{(1)} \\
\mathcal{S}_{22}^{(1)} &= 2\lambda_j q_{jj}^{(1)} - 2\mathcal{Q}_j q_{jj}^{(1)} \\
\mathcal{S}_{23}^{(1)} &= \lambda_j q_{j,d+l}^{(1)} + \lambda_{d+l} q_{j,d+l}^{(1)} + \mathcal{Q}_j\mathcal{Q}_{d+l} - \mathcal{Q}_j q_{j,d+l}^{(1)} - \frac{\mathcal{Q}_j\left(\mathcal{Q}_j^2 + 1\right)}{-2\lambda_j + 2\mathcal{Q}_j} + \frac{\mathcal{Q}_{d+l}}{2\lambda_{d+l}} \\
\mathcal{S}_{33}^{(1)} &= 2\lambda_{d+l} q_{d+l}^{(1)}
\end{aligned}$$

Solving this system yields the following solutions for the $q_{ij}^{(1)}$:

$$\begin{aligned}
q_{ii}^{(1)} &= 0 \\
q_{jj}^{(1)} &= 0 \\
q_{d+l,d+l}^{(1)} &= 0 \\
q_{ij}^{(1)} &= 0 \\
q_{i,d+l}^{(1)} &= 0 \\
q_{j,d+l}^{(1)} &= \frac{-2\lambda_j\lambda_{d+l}\mathcal{Q}_j\mathcal{Q}_{d+l} - \lambda_j\mathcal{Q}_{d+l} - \lambda_{d+l}\mathcal{Q}_j^3 + 2\lambda_{d+l}\mathcal{Q}_j^2\mathcal{Q}_{d+l} - \lambda_{d+l}\mathcal{Q}_j + \mathcal{Q}_j\mathcal{Q}_{d+l}}{2\lambda_j^2\lambda_{d+l} + 2\lambda_j\lambda_{d+l}^2 - 4\lambda_j\lambda_{d+l}\mathcal{Q}_j - 2\lambda_{d+l}^2\mathcal{Q}_j + 2\lambda_{d+l}\mathcal{Q}_j^2}
\end{aligned}$$

2. *Case 2*: $k \neq i, k \neq j$. In this case, we must again consider the 3×3 subsystem indexed by $i, j, d+l$, but since $M_{ij}T_{kl}$ is a matrix of all zeros, the expression simplifies considerably:

$$\begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_{12} & \mathcal{S}_{13} \\ \mathcal{S}_{21} & \mathcal{S}_{22} & \mathcal{S}_{23} \\ \mathcal{S}_{31} & \mathcal{S}_{32} & \mathcal{S}_{33} \end{bmatrix} = 0$$

$$\begin{aligned}
\mathcal{S}_{11} &= \alpha^2\mathcal{Q}_i^2 + \mathcal{Q}_i^2 + 2q_i\left(-\alpha^2\mathcal{Q}_i + \lambda_i - \mathcal{Q}_i\right) + 1 \\
\mathcal{S}_{12} &= q_{ij}\left(-\alpha^2\mathcal{Q}_i + \lambda_i - \mathcal{Q}_i\right) + q_{ij}\left(-\alpha^2\mathcal{Q}_j + \lambda_j - \mathcal{Q}_j\right) \\
\mathcal{S}_{13} &= \lambda_{d+l}q_{i,d+l} + q_{i,d+l}\left(-\alpha^2\mathcal{Q}_i + \lambda_i - \mathcal{Q}_i\right) \\
\mathcal{S}_{22}\alpha^2\mathcal{Q}_j^2 + \mathcal{Q}_j^2 + 2q_j &\left(-\alpha^2\mathcal{Q}_j + \lambda_j - \mathcal{Q}_j\right) + 1 \\
\mathcal{S}_{23} &= \lambda_{d+l}q_{j,d+l} + q_{j,d+l}\left(-\alpha^2\mathcal{Q}_j + \lambda_j - \mathcal{Q}_j\right) \\
\mathcal{S}_{33} &= 2\lambda_{d+l}q_{d+l} + 1
\end{aligned}$$

Plugging in the power series expansion $q_{ij} = q_{ij}^{(0)} + q_{ij}^{(1)}\alpha + O(\alpha^2)$, one finds the lowest order terms in α within this system of equations occurs at $O(\alpha^2)$, and thus to $O(\alpha)$, the solution of $Q_m[C_{\text{slow}} + \alpha\Psi_{ij,kl}]$ coincides with $Q_m[C_{\text{slow}}]$.

To complete the proof of Theorem 3, we must calculate the following quantity:

$$D_{\Psi_{ij,kl}}\text{Tr}\left(Q_m^2\right) = \lim_{\alpha\to 0}\frac{\text{Tr}(Q_m[C_{\text{slow}+\alpha\Psi_{ij,kl}}]^2) - \text{Tr}(Q_m[C_{\text{slow}}]^2)}{\alpha}$$

From the case-wise analysis above, we see that the only matrix element of Q_m that differs between $Q_m[C_{\text{slow}+\alpha\Psi_{ij,kl}}]$ and $Q_m[C_{\text{slow}}]$ to $O(\alpha)$ is an off-diagonal term ($q_{j,d+l}^{(1)}$). However, this term does not contribute to the trace of Q_m^2 at $O(\alpha)$. Thus, we conclude that along a complete basis for the tangent space of Ω at C_{slow}, $D_{\Psi_{ij,kl}}\text{Tr}\left(Q_m^2\right) = 0$. From this, we

conclude that $\nabla_C \mathrm{Tr}(Q_m[C_{\mathrm{slow}}]^2) = 0$ on Ω. The proof of Theorem 3 follows from application of Lemma 1. $\square$

We again note that the FCCA objective function differs from the LQG trace by the factor of Π and Π^{-1} in the regulator Riccati equation. The presence of these re-weighting factors lead to a small, but non-zero subspace angle between PCA and FCCA even when A is symmetric (as we report in **Figure 1.4c**).

Chapter 2

Maximum Entropy Random Graph Models for Large Scale Connectomics

2.1 Introduction

A grand challenge in neuroscience is link the structure of the brain, as specified by the connectivity between neurons, to its functionality. As modern connectomics yields increasingly complete descriptions of the microscale connectivity between networks of neurons, a key open avenue for research remains the development of computational tools that can leverage this data to uncover the underlying specific wiring principles that shape the dynamics and functionality of these networks.

Obtaining insight into why observed neural circuits are wired in the way they are requires considering their relation to the broader space of possible patterns of connectivity between neurons. This broader "network morphospace" [8] is bounded by physical and biological constraints that can be conceptualized as operating in a "top-down", or global, and "bottom-up", or local fashion. Examples of the former include the fact that brains are constrained by the physical volume they inhabit (i.e., they are spatially embedded networks [78]), and that the construction of long range axonal connections incurs significant energetic cost. An example of the latter type of constraint is the reproducible observation of cell-type dependent connection probabilities between genetically defined cell types [79,80]. Within the bounds of these constraints, neural circuits must further be wired together in a manner which enables them to carry out their respective computational functions. This requirement constitutes an additional "top-down" constraint on the space of possible networks.

Given these set of global and local constraints on neural connectivity, the space of possible network configuration is still vast. Locating observed connectomes within this space provides direct insight into the underlying wiring principles of neural circuits. Indeed, a major thrust of prior work in connectomics has been to assess the degree to which observed patterns of connectivity at various spatial scales are emergent and potentially selected for, as opposed to being expected consequences of the local constraints on the morphospace [81–83]. With re-

spect to top down influences, the characteristics of empirically observed networks relative to the potential alternatives contained within the morphospace reveals the extent to which biology has balanced competing requirements. To this end, connectomes across diverse species have been found to tradeoff between wiring cost with communication efficiency [84]. Finally, we argue that the specific connectivity patterns exhibited by real connectomes relative to a particular equivalence class of networks *reveals mechanism*, i.e. how biology has balanced the particular competing bottom up and top down constraints under consideration.

The idea of a constrained network morphospace is closely related to well known issues surrounding parameter degeneracy in systems biology models [85]. The most well studied example of this in phenomena is the crab STG system, in which multiple circuit configurations are a priori consistent with the sequence of neural activity that gives rise to the pyloric rhythm [86]. Parameter degeneracy is also a well known phenomenon in artificial recurrent neural networks [87], which are increasingly used as mechanistic models of neural computation. Developing methods to explore the set of functionally equivalent network architectures is therefore a problem with broad importance in computational and systems neuroscience.

Connectomes can be modelled as weighted, directed graphs in which edges represent the presence and strength of synaptic connections between neurons. The constrained network morphospace, which we will subsequently refer to as a null model, can be modelled using tools borrowed from statistical physics [88]. Thermodynamic systems are subject to a set of macroscopic constraints on system properties such as energy and particle number. Subject to these constraints, a particular set of microscopic configurations are possible, and their occurrence at equilibrium is governed by a probability distribution that satisfies the macroscopic constraints but is otherwise maximally random (or maximum entropy). These constrained maximum entropy distributions have found wide application across the sciences, and provide a means of parameterizing and sampling from null models. However, working with these distributions is not without difficulty, as their normalization constants often cannot be explicitly calculated. This challenge has impeded connectomic analysis, as thus far studies have only considered the consequences of highly local constraints such as the network degree distribution and pairwise distances between neurons. However, as articulated above, neural circuits are shaped under the influence of both local constraints and global, functional constraints. In [89], the authors address the challenge of fitting maximum entropy probability distributions including non-trivial, global constraints to computational neuroscience models by maximizing entropy within a restricted, tractable set of distributions. However, the particular class of distributions employed (normalizing flows), are only applicable to continuous parameter spaces, and thus cannot be applied as models of distributions over graphs. **A key motivation of this work is to therefore develop computational methods to enable inference and sampling from maximum entropy models with global constraints on network function**.

The particular functions carried out by neural circuits diverse and difficult to directly quantify, and thus some simplifying assumptions are clearly required to make progress. As in the previous chapter, we will assume linear dynamics for our system. In this regime, it is possible to assess the controllbility measures developed in the previous chapter, this time on

the ground truth connectivity between neurons. We also consider the related, but distinct, notion of signal propagation between nodes of the connectome via diffusion. This model has been widely used to interrogate the structure of connectivity between neurons [84]. We apply our methods to the recently released connectome of the *Drosophila* hemibrain connectome [1] (**Fig. 2.1**). This dataset contains the complete connectivity between 21,737 neurons, comprising $>$ 25 million synapses and 61 brain regions (regions of interest, ROIs). Our preliminary results, which compare the hemibrain connectome to structurally constrained null models, reveal a great deal of heterogeneity across ROIs in the ability of pairwise connectivity rules between neurons to account for controllability and diffusivity in those ROIs.

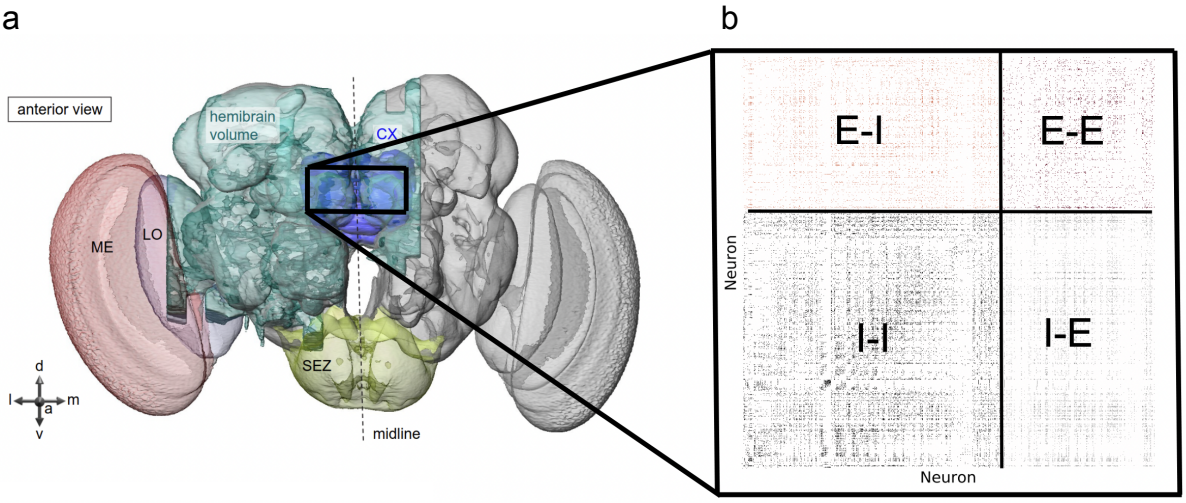


Figure 2.1: (a) Overview of the *Drosophila* brain and the region mapped within the hemibrain connectome. Reproduced from [1]. (b) Weighted adjacency matrix of the Fan Shaped Body ROI ordered by excitatory (E) and inhibitory (I) neurons.

2.2 Sampling and Inference within Structurally Constrained Null Models.

In this section, we detail our approach to fitting from null models for networks that encode bottom up structural constraints. These constraints are operationalized as terms in an "energy function", $H_\theta(G)$ that defines a maximum entropy probability distribution over random graphs G [88]:

$$P(G) = \frac{1}{Z}\exp(H_\theta(G))$$

We consider the most general models of exponential random graphs within which inference remains tractable (i.e., the partition function Z can be exactly calculated), including

constraints on the empirical degree distribution, average connectivity between biologically defined cell types, and average weighted connectivity between cell types. In contrast to previous use of exponential random graphs in connectomics (schneidermann ref, null models for network neuroscience), we include constraints on both the binary adjacency matrix of the connectome as well as the weights along edges. In the *Drosophila* connectome, the strength of connectivity between neurons is determined primarily by the number of synapses formed between two neurons [1]. We relax this ordinal edge weight to a continuous valued edge weight for ease of inference. The energy function of our structurally constrained model reads:

$$
\begin{aligned}
H_\theta(G) = H_\theta(A,W) &= H_{\text{configuration}}(A) + H_{\text{rSBM}}(A) + H_{\text{wSBM}}(W) \\
H_{configuration}(A) &= \sum_{i<j} (\alpha_i + \beta_j) a_{ij} + (\alpha_j + \beta_i) a_{ji} \\
H_{rSBM}(A) &= \sum_{i<j} \omega_{g_i g_j} a_{ij}(1 - a_{ji}) + \omega_{g_j g_i} a_{ji}(1 - a_{ij}) + \omega^{\leftrightarrow}_{g_i g_j} a_{ij} a_{ji} \\
H_{wSBM}(W,\mu) &= \sum_{i \neq j} \mu_{g_i g_j} w_{ij}
\end{aligned}
$$

where the graph G is represented by its binary and weighted adjacency matrices (A and W). Each entry of these matrices is denoted a_{ij} and w_{ij}, respectively. The parameters θ of the model are $(\alpha_i, \beta_i, \omega_{g_i,g_j}, \omega^{\leftrightarrow}_{g_i,g_j}, \mu_{g_i,g_j})$ for $i, j = 1...N$, where N is the number of nodes in the network. These parameters constrain, respectively, the in degree, our degree, average undirectional connectivity between types, average bidirectional connectivity between types, and average weighted connectivivity between types.

The probability distribution over (A, W) induced by the above energy function factorizes over dyads (i.e. the pair of edges pointing from node i to node j and node j to node i). This fact renders the partition function analytically tractable:

$$
\begin{aligned}
Z &= \prod_{(i,j)} Z_{(i,j)} \\
Z_{(i,j)} &= 1 + \int dw_{ij} \exp\left((\alpha_i + \beta_j) + \omega_{g_i g_j} - \mu_{g_i g_j} w_{ij}\right) + \\
&\quad \int dw_{ji} \exp\left((\alpha_j + \beta_i) + \omega_{g_j g_i} - \mu_{g_j g_i} w_{ji}\right) + \\
&\quad \int dw_{ij} dw_{ji} \exp\left((\alpha_i + \beta_j) + (\alpha_j + \beta_i) + \omega^{\leftrightarrow}_{g_i g_j} - \mu_{g_i g_j} w_{ij} - \mu_{g_j g_i} w_{ji}\right) \\
&\quad = 1 + \frac{1}{\mu_{g_i g_j}} \exp\left((\alpha_i + \beta_j) + \omega_{g_i g_j}\right) + \frac{1}{\mu_{g_j g_i}} \exp\left((\alpha_j + \beta_i) + \omega_{g_j g_i}\right) + \\
&\quad \frac{1}{\mu_{g_i g_j} \mu_{g_j g_i}} \exp\left((\alpha_i + \beta_j) + (\alpha_j + \beta_i) + \omega^{\leftrightarrow}_{g_i g_j}\right)
\end{aligned}
$$

where (i, j) refers to the dyad of edges between node i and node j. The energy function of our model is linear in the sufficient statistics (i,e., the constraints): $H_\theta(G) = \sum_i \theta_i c_i(G)$. For such a model, the gradient of the log likelihood of takes on a particularly intuitive form:

$$\frac{\partial \log \mathcal{L}}{\partial \theta_i} = \langle t_i(G) \rangle_{\text{data}} - \langle t_i(G) \rangle_\theta$$

In other words, we update model parameters until the expected value of the sufficient statistics under the model ($\langle t_i(G) \rangle_\theta$) match the observed values of the sufficient statistics under the data ($\langle t_i(G) \rangle_{\text{data}}$). Analytic calculation of the expectation values of each $\langle t_i(G) \rangle$ requires the values of each $\langle a_{ij} \rangle, \langle a_{ij} a_{ji} \rangle$, and $\langle w_{ij} \rangle$. These values can again be derived exactly:

$$\langle a_{ij} \rangle = \frac{1}{Z_{(i,j)}} \left(\frac{1}{\mu_{g_i g_j} \mu_{g_j g_i}} \exp\left((\alpha_i + \beta_j) + (\alpha_j + \beta_i) + \omega^{\leftrightarrow}_{g_i g_j} \right) + \frac{1}{\mu_{g_i g_j}} \exp\left((\alpha_i + \beta_j) + \omega_{g_i g_j} \right) \right)$$

$$\langle a_{ij} a_{ji} \rangle = \frac{1}{Z_{(i,j)}} \frac{1}{\mu_{g_i g_j} \mu_{g_j g_i}} \exp\left((\alpha_i + \beta_j) + (\alpha_j + \beta_i) + \omega^{\leftrightarrow}_{g_i g_j} \right)$$

$$\langle w_{ij} \rangle = \frac{1}{Z_{(i,j)}} \left(\frac{1}{\mu^2_{g_i g_j} \mu_{g_j g_i}} \exp\left((\alpha_i + \beta_j) + (\alpha_j + \beta_i) + \omega^{\leftrightarrow}_{g_i g_j} \right) + \frac{1}{\mu^2_{g_i g_j}} \exp\left(\alpha_i + \beta_j + \omega_{g_i g_j} \right) \right)$$

With these quantities in hand, we can fit the structurally constrained model to the connectome using exact maximum likelihood. Due to the linearity of the energy function in its sufficient statistics, the optimization is guaranteed to be convex [90]. To enable rapid and reliable convergence, we use accelerated gradient descent with a backtracking line search.

Sampling

To sample from fitted models, we draw samples first from the marginal distribution of the binary adjacency matrix, and then the conditional distribution of (w_{ji}, w_{ij}) given (a_{ij}, a_{ji}).

First, we derive the marginal distribution $p(a_{ij}, a_{ji})$. This is a vector of four probabilities, one for each possible outcome: $(0, 0), (1, 0), (0, 1), (1,)$, and can essentially be read off from the partition function above:

$$
\begin{aligned}
p(a_{ij}, a_{ji}) &= \int dw_{ij} dw_{ji} P_{ij}(a_{ij}, a_{ji}, w_{ij}, w_{ji}) \\
&= \frac{1}{Z_{(i,j)}} \Bigg((1 - \delta_{a_{ij}})(1 - \delta_{a_{ji}}) + \frac{1}{\mu_{g_i g_j}} \exp\Big((\alpha_i + \beta_j) + \omega_{g_i g_j} \Big) \delta_{a_{ij}} (1 - \delta_{a_{ji}}) + \\
&\frac{1}{\mu_{g_j g_i}} \exp\big((\alpha_j + \beta_i) + \omega_{g_j g_i} \big) + \\
&\frac{1}{\mu_{g_i g_j} \mu_{g_j g_i}} \exp\Big((\alpha_i + \beta_j) + (\alpha_j + \beta_i) + \frac{1}{2} \omega^{\leftrightarrow}_{g_i g_j} \Big) \delta_{a_{ij}} \delta_{a_{ji}} \Bigg)
\end{aligned}
$$

The conditional distribution is then the ratio of the joint and marginal distributions for each binary dyad outcome. We notice that these essentially reduce to the corresponding exponential distributions:

$$
\begin{aligned}
p(w_{ij}, w_{ji} | a_{ij} = 1, a_{ji} = 0) &= \delta(w_{ij}) \delta(w_{ji}) \\
p(w_{ij}, w_{ji} | a_{ij} = 1, a_{ji} = 0) &= \mu_{g_i, g_j} \exp(-\mu_{g_i g_j} w_{ij}) \delta(w_{ji}) \\
p(w_{ij}, w_{ji} | a_{ij} = 0, a_{ji} = 1) &= \mu_{g_j, g_i} \exp(-\mu_{g_j g_i} w_{ji}) \delta(w_{ij}) \\
p(w_{ij}, w_{ji} | a_{ij} = 1, a_{ji} = 1) &= \mu_{g_j, g_i} \mu_{g_i, g_j} \exp(-\mu_{g_i g_j} w_{ij} - \mu_{g_j, g_i} w_{ji})
\end{aligned}
$$

Preliminary Results

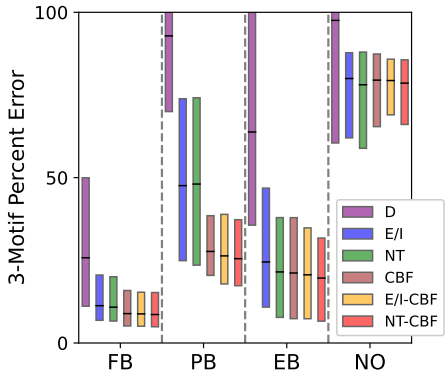


Figure 2.2: Box plots (median ± IQR) of the aggregated percent error in prediction of the frequency of all directed 3 node motifs across model resolution and ROI.

At the time of writing this thesis, I had not yet fully implemented the machinery described above. Thus, the results presented below are derived from the ergm R package [91]. Inference was performed using the pseudolikelihood approach, which is inexact for energy functions that contain the bidirectional constrain imposed by $\omega^{\leftrightarrow}$.

We considered models that constrained the average (unidirectional and bidirectional) connectivity between cell types, exploring a hierarchy of resolutions in cell typing categorization. At the coarses resolution, we constrained just the overall edge density (i.e. no type, denoted D), followed by the typing according to excitatory vs. inhibitory cell type (E/I), then the specific neurotransmitter expression

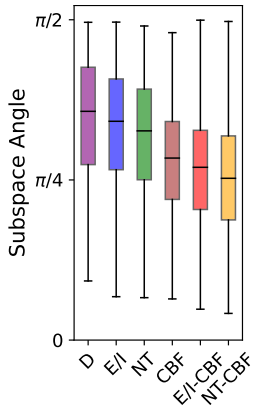


Figure 2.3: Box Plots (median ± 95th CI) of the subspace angle between the leading eigenvector of the controllability Gramian of the empirical network and the model derived networks. Angles are aggregated across ROIs for each model.

(NT), the cell body fiber cluster a neuron's soma resides in (CBF), and finally finer resolutions that considered combinations of these features (EI-CBF, NT-CBF). These models

resulted in distributions over directed, weighted graphs, with edge signs determined by neurotransmitter expression. As our interest was in determining how well these models which add constraints at various levels of resolution could account for the controllability of the network, we considered two measures of function - linear controllability and information diffusion. We consider the controllability of network dynamics given particular choices of input region of interest (ROI) and controlled ROI. Within this framework, we consider simple linear dynamics for neuronal firing rates: $\dot{x} = -Ax + Bu$, where A is the adjacency matrix of the controlled ROI, B is a rank 1 matrix encoding the connectivity from the input ROI to the controlled ROI. The single neurons that are most energetically easy to control are determined by the entries of the eigenvectors corresponding to the largest eigenvalues of the controllability Gramian. Information diffusion, which relates to controllability as it measures the efficacy by which input signals can propagate across the network [92], is quantified by the spectral properties of the (weighted, directed) graph Laplacian, $L = D - A$, where D is a diagonal matrix with weighted node degrees along the diagonal. In particular, the eigenvalues of L with smallest non-zero real part probe the slowest time scales of diffusion, and thereby the properties of large scale connectivity within the network.

We first determined the degree to which models with pairwise constraints could explain the frequency of all observed directed three node motifs (**Fig. 2.2**). We measured the percent error between the count of empirically observed motifs and the average count of motifs across 1000 graph samples from each distribution.

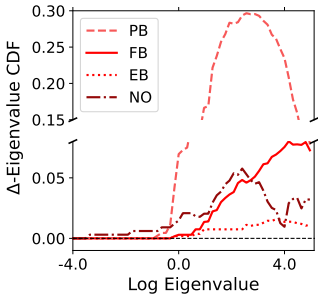


Figure 2.4: Difference between the empirical and NT-CBF model derived CDF of graph Laplacian eigenvalues across ROIs.

Across 4 ROIs from the Drosophila central complex (Fan-Shaped Body (FB), Proto-Cerebral Bridge (PB), Ellipsoid Body (EB), and the Noduli (NO)), we observed that models that incorporated cell type specific connectivity significantly outperformed the baseline density constrained model (purple bars vs. others, Mann-Whitney U test, $p < 10^{-5}, n = 1000$). The error in 3 node motif reconstruction saturated at the cell body fiber level of resolution (brown bars). Furthermore, there were ROI dependent effects, with NO and PB exhibiting singificantly higher error across all models considered, indicating the presence of higher-order statistical dependencies between edges in these regions. Next, we measured how well pairwise constrained models could reproduce network controllability. We compared the angles

between the eigenvectors corresponding to the dominant controllability Gramian eigenvalues between the empirical network and the ensemble average from each model. There was a significant mismatch between empirical and model predicted eigenvectors ($> \pi/4$ median angle across all models and all combinations of input/controlled ROIs), with the baseline density model performing worst (leftmost bar). Additionally, there was substantial heterogeneity in error across combinations of controlled and input ROIs, as evidenced by the large spread in all boxplots in **Figure 2.3**. The capacity of pairwise models to explain third order network structure, which depends purely on mesoscale connectivity statistics, but not the most controllable directions, and therefore populations of neurons, suggested that these models failed to capture the macroscale organization of connectivity and edge weights. To test this hypothesis, we finally compared the ability of pairwise models to recapitulate the diffusion Laplacian spectrum. In **Figure 2.4**, we plot the CDF of the difference (Δ-CDF) between the empirical spectral density and the NT-CBF model spectral density across ROIs. In line with our hypothesis, we found pairwise models underestimated the density in the lower tail of the distribution (log eigenvalues < 4), reflected in the Δ-CDF being > 0 across ROIs. This effect was particularly pronounced within FB and PB. In sum, our results indicate that connectivity is structured with respect to genetically specified cell types in Drosophila, but that global network function relies on emergent structure beyond these pairwise interactions.

2.3 Sampling and Inference within Functionally Constrained Null Models

We now describe our strategy for incorporating top-down, functional constraints on our random graph ensembles. As a specific example, we discuss additions to the energy function dervied from the linear controllability Gramian, but the approach described below does not rely on this specific choice. We envisage the incoporation of top-down functional contraints to be a general tool to assay the impact of hypothesized computations on the possible network morphospace.

The energy function for our problem now reads:

$$H_\theta(G) = H_\theta(A, W) = H_{\text{configuration}}(A) + H_{\text{rSBM}}(A) + H_{\text{wSBM}}(W) + \gamma \log\det \Pi_C$$

where Π_C is the controllablity Gramian, defined with particular choices of input ROIs serving as the "controller" for a particular output ROI. Any global functional measure necessarily couples all degrees of freedom in the network to each other. In this context, calculation of normalization constants is rendered intractable, and we must rely on MCMC techniques. It is vitally important that the associated MCMC chains mix. The failure of the ergm package to reliably fit models with 3 node motif terms suggests that relying on naive (random walk) Metropolis Hastings is insufficient. Recent work has shown that it is possible to dramatically improve upon the efficiency of Metropolis Hastings in discrete spaces by addressing the following shortcomings of Metropolis-Hastings:

1. Naive MH is slow because most proposal steps are likely to be rejected.

2. Naive MH is slow because all proposals are local perturbations to the current state (e.g. they differ by a single edge swap or a single spin flip).

In continuous state spaces, both issues are addressed by the use of Hamiltonian Monte Carlo, Langevin sampling, or some combination thereof [93]. Our configuration space has support over both a continuous valued random variable (edge weights) and a discrete valued random variable (binary adjacency). In discrete state spaces, analogues of Langevin sampling have recently been developed for the case when the energy function of the model is differentiable [94, 95]. These algorithms form a key component of our approach.

The maximum entropy distributions described above enforce "soft constraints", referring to the fact that the prescribed values of the sufficient statistics $\{T_i(G)\}$ hold only on average. Alternatively, "hard constrained" ensembles may be considered, though these are infeasible to sample from directlty in high dimensions as one cannot use standard MCMC approaches. Nonetheless, soft-constrained ensembles are known to have certain degeneracy issues [96]. Since constraints are forced to hold only on average, it may be that the support of the distribution lies on nearly disconnected clusters of the configuration space. While the correct value of sufficient statistics is obtained on average, samples from the model will almost surely belong to one of these clusters, whose properties may differ strongly from the mean. An example of this phenomena is provided the triangular model, which constrains the frequency of directed 3-cycles within the enesemble. With no further constraints, typical samples from this ensemble will contain either almost no directed cycles or a pathologically large number of directed cycles ([97]).

Interestingly, solutions to this degeneracy problem, which also can also impede MCMC mixing as they give rise to disconnected high probability regions in configuration space, rely on restricting the support of the probability distribution [96, 98, 99] . One role played by the local, tractable constraints is then to achieve this narrowing of support by requiring graphs to adhere closely to the observed statistics of connectivity between cell types.

To this end, we consider the use of the Gaussian ensemble (not to be confused with a Gaussian distribution), a thermodynamic ensemble that interpolates cleanly between the microcanonical (hard-constrained) and canonical (soft-constrained) ensembles [100]. In a traditional statistical mechanical setting, the microcanonical ensemble describes a closed system with fixed energy (hence the sufficient statistics take on exactly fixed values). The canonical ensemble describes a system coupled to an infinitely large heat bath. While the bath and the ensemble as a whole have fixed energy, the energy of the system is allowed to fluctuate by interaction with the bath to a degree set by the temperature (or more generally, by a set of thermodynamic parameters analogous to the θ_i). Conceptually, interpolating between the canonical and microcanonical ensembles therefore involves varying the size of the bath, with the latter ensemble corresponding to vanishing thermal bath. Operationally, the probability of a system state in the Gaussian ensemble is distributed according to the $P(G) \propto \exp(-a(H(G) - E_t)^2)$ with E_t a fixed constant and the value of a controlling

the interpolation between canonical and microcanonical ensembles ($a \to 0$ limits to the canonical ensemble, while $a \to \infty$ limits to the microcanonical ensemble). Alternatively, one may consider separate constraints on both the mean and variance of the desired sufficient statistic, which leads to a slightly different ensemble [101].

Our motivation for considering these generalized canonical ensembles is to provide a flexible "reference" measure as a basis for the additional, global functional constraints. There are numerous works suggesting that restricting the support of exponential family distributions significantly alleviates barriers to inference [96,98,99]. Our use of low-level, biologically interpretable constraints restricts the support of distribution on graphs to a space that biologically could plausibly explore given our observations of real connnectomes. The ability to tune between canonical and microcanonical ensembles allows one to tune the scale of fluctations away from the observed connectivity statistics, while providing a means of incorporating approximate sampling from microcanonical ensembles into standard MCMC pipelines.

Gradient-Based acceleration of Monte Carlo sampling in discrete spaces

Recalling the gradient of the log likelihood for an exponential random graph model:

$$\frac{\partial \log \mathcal{L}}{\partial \theta_i} = \langle t_i(G) \rangle_{\text{data}} - \langle t_i(G) \rangle_\theta$$

Inference requires accurate estimates of $\langle t_i(G) \rangle_\theta$, which in the present case must be obtained via MCMC. In this section, we illustrate why incorporating information about the gradient of the energy function can speedup Monte Carlo simulation in the simplified context of binary random variable models. First, fixing some notation, let G again denote the random variable of interest and $\mathcal{G}$ its configuration space. Let G_i denote the i^{th} coordinate of G, and $p(i|\neg i)$ the conditional distribution of G_i given $G_j, j \neq i$.

Consider Gibbs sampling, a canonical MCMC algorithm that proposes sequentially updating the i^{th} degree of freedom in the configuration space according the conditional probability distributions $p(i|\neg i)$. Often, for most i, most probability mass in the distribution $p(i|\neg i)$ will be concentrated around the current state of i. Thus, most proposed changes of state will be likely to be rejected, wasting computational effort. Rather than sequentially flipping (or randomly in the case of the random walk Metropolis Hastings), we can consider a particular proposal distribution $q(G'|G)$ over configurations that has high probability mass on states that are likely to change during the Monte Carlo step. It can be shown that an optimal tradeoff between these two goals that only uses local information is given by:

$$q(G'|G) = \exp\left(\frac{1}{2}(H(G') - H(G))\right) 1_{|G-G'|=1}$$

where $1_{()}$ is the indicator function. For many problems of interest, the difference $H(G') - H(G)$ can be approximated via $\nabla H(G)$, which is much more efficient to calculate. This proposal distribution underlies the Gibbs with Gradient (GWG) algorithm [94].

In continuous state spaces, using gradient information to accelerate MCMC chains by biasing moves in the directions of large changes in the energy is the motivation for Langevin Monte Carlo (LMC). Letting $p(X), X \in \mathbb{R}^d$ be a probability distribution from which we wish to draw samples, the LMC algorithm runs the dynamics:

$$\dot{X} = \nabla_X \log p + dW_t \tag{2.1}$$

where dW_t is a Wiener process on $\mathbb{R}^d$. The stationary solution to this stochastic differential equation then yields samples from $p(X)$. LMC may be derived by noting that the gradient of the Kullback-Liebler divergence between a distribution q and the target distribution p with respect to q is given by:

$$\partial_q D_{\mathrm{KL}}(q; p) = \nabla_X \cdot [\log p(x) q(x)] - \Delta q(x)$$

This can be used to define a gradient flow over the space of probability distributions:

$$\frac{\partial q}{\partial t} = \nabla_X \cdot [\log p(x) q(x)] - \Delta q(x) \tag{2.2}$$

Langevin dynamics then result by observing that eq. 2.2 is a Fokker-Plank eqution for which the particle level trajectories are defined by eq. 2.1.

The key obstacle to implementing Langevin dynamics over discrete configuration spaces is the lack of a well defined gradient ∇_X. Nevertheless, it is possible to derive an analogous expression for $\partial_q D_{\mathrm{KL}}(q; p)$ over discrete state spaces. Using this to define a gradient flow over probability distributions over discrete spaces, and subsequently deriving a particle level realization yields the Discrete Langevin Monte Carlo (DLMC) algorithm [95]. The GWG algorithm can be seen to be a special case of DLMC, the key difference being that in DLMC, changes may be proposed to states at a larger Hamming distance from the current state [95].

In the plot below, we show the effective sample size (over 50000 MC steps) associated with sampling from a stochastic block model via different sampling techniques:

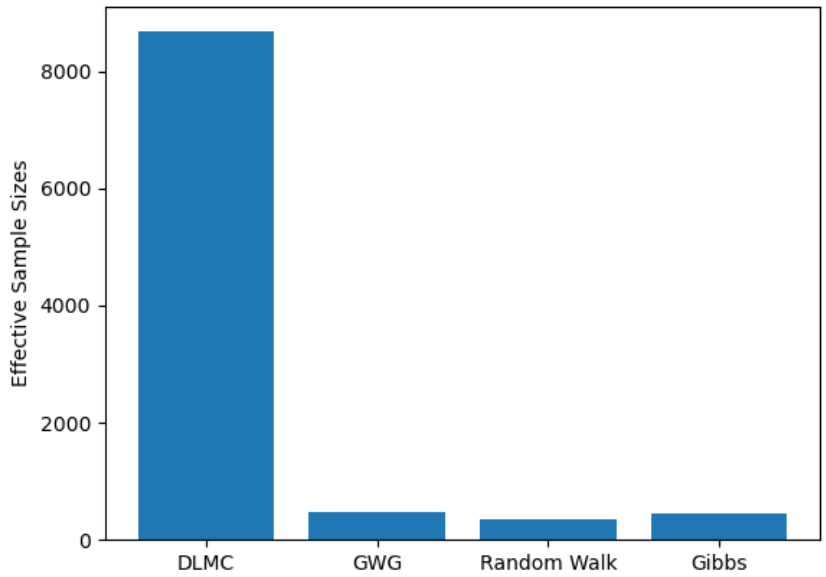


Figure 2.5: Effective Sample Size over 50000 MC steps associated with sampling from a 100 node stochastic block model via different sampling algorithms.

We observe that DLMC is able to provide a dramatically higher effective sample size as compared to traditional sampling methods (Gibbs, Random Walk Metropolis Hastings). The DLMC algorithm therefore is a promising approach to performing inference within functionally constrained null models.

Testing for Convergence of MCMC and Importance Weighting via Stein Divergences

While MCMC is asymptotically unbiased, mixing may take a prohibitively long time in high dimensions even when using gradient based proposal distributions. Furthermore, actually assessing its convergence to the desired distribution is challenging. This raises a challenge for MCMC based MLE - how can we determine whether our Markov chain has run long enough to yield a good estimate of $\langle H_\theta \rangle_p$, and therefore the gradient of the likelihood?

A solution to this issue is provided by the Stein discrepancy [102]. The Stein discrepancy is a type of integral probability metric (IPM). These metrics between probability distribution take the following form:

$$\text{IPM}(p, q) = \sup_{f \in \mathcal{F}} [E_q f - E_p f] \tag{2.3}$$

where $\mathcal{F}$ is a suitably rich class of test functions. The Stein discrepancy is an IPM that may be computed when expectations under p are unavailable due to, for example, intractability of its normalization constant. The function class $\mathcal{F}$ is chosen to be one that satisfies Stein's identity $\forall f \in \mathcal{F}$:

$$\mathbb{E}_p \mathcal{A} f = 0$$

where $\mathcal{A}$ is known as a Stein operator. There exist many techniques for constructing Stein operators [102]. One choice involves the use of the score function of the distribution, $\nabla_x \log p$:

$$\mathcal{A} f = \nabla_x \log p(x) f(x) + \nabla_x f(x) \tag{2.4}$$

Plugging eq. 2.4 into eq. 2.3, one obtains the Stein discrepancy:

$$\mathbb{D}(q, p) = \sup_{f \in \mathcal{F}} \mathbb{E}_q \left[\nabla_x \log p(x) + \nabla_x f(x) \right]$$

As the score function does not depend on the normalization constant, the Stein discrepancy may be calculated given just access to the energy function of p and samples from q. The supremum over $\mathcal{F}$ may be obtained in two ways. If we take $\mathcal{F}$ to be a unit norm Reproducing Kernel Hilbert Space (RKHS) with associated kernel k, then the Stein discrepancy actually takes on a closed form:

$$\mathbb{D}(q, p) = \mathbb{E}_{x, x' \sim q} \Big[s_p(x)^\top k(x, x') s_p(x') + s_p(x)^\top \nabla_x k(x, x') + \nabla_{x'} k(x, x')^\top s_p(x') + \mathrm{Tr} \nabla_x \nabla_{x'} k(x, x') \Big]$$

where $s_p(x) = \nabla_x \log p$ is the score function of p. We refer to this form of the Stein discrepancy as the Kernel Stein Discrepancy (KSD).

The above discrepancy is only applicable to continuous valued random variables due to the use of gradients with respect to the random variables. However, we recover computable Stein discrepancies for binary random variables if we replace the gradient ∇_{x_i} with the inversion operator: $\Delta_{x_i} f(x) = f(x_1, x_2, ...\neg x_i, ..., x_n)$ [103]. In **Figure 2.6a**, we verify the ability of the Stein discrepancy to detect convergence when samples have been drawn from a degree corrected stochastic block model ($H(G) = H_{\text{configuration}} + H_{\text{rSBM}}$) given access only to the energy functions of models. Specifically, we draw 1000 exact samples from a target model over 10 nodes with parameters θ_0 and evaluate the Stein discrepancy between the empirical distribution defined by the samples and a set of distributions with parameters θ_i linearly interpolated between θ_0 and a different, randomly initialized parameter vector θ_1. We observe that the log KSD is sharply sensitive to deviations of the energy function from

θ_0, spanning over 30 orders of magnitude as the energy functions are interpolated between θ_0 and θ_1. This suggests that the Stein discrepancy is a promising tool for assessing convergence of samples generated by MCMC to a target distribution of interest.

In addition to measuring sample quality without a need to calculate the normalization constant, the Stein discrepancy can also be used as a black box importance sampler [104]. Consider a set of samples $\{x_i\}$ generated from *any* mechanism (e.g., an MCMC chain run for a finite amount of time). In general, an estimate of $\langle H_\theta \rangle$ from these samples will be biased. This bias will be reflected in the Stein discrepancy.

One can then use the Stein discrepancy to obtain a set of importance weights to improve the accuracy of the ensemble average by solving a quadratic program:

$$w_i = \text{argmin}\left(w^\top \mathcal{D}(q,p)w, \quad \sum w_i = 1, w_i \geq 0\right)$$

In **Figure 2.6b**, we plot the improvement in subspace angle between a DLMC estimated gradient of the degree corrected sochastic block model over 10 nodes and that obtained from importance weighting the MCMC samples via minimizing the KSD. The plot is taken over 10 repetitions of MCMC. We observe that the KSD importance weighting is able to consistently improve the estimate of the gradient direction, across all reptitions.

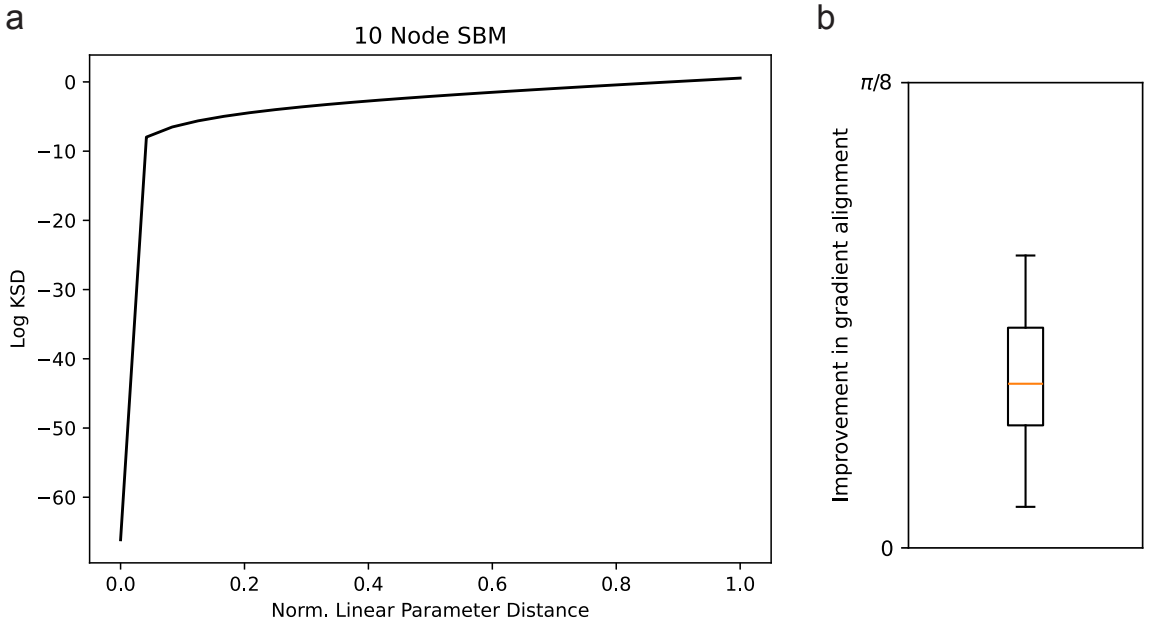


Figure 2.6: (a) Plot of the log KSD between samples initialized from a target model as a function of the normalized, linear distance in parameter space away from the target model. (b) Boxplot (median ± 95 CI) of the difference in alignment with the ground truth gradient between a DLMC estimated gradient and a gradient estimated from KSD derived importance weights.

Overall Proposed Algorithm

Given functional measure $f(G)$, local sufficient statistics t_i, their empirically observed values f^*, t_i^*, construct the loss function:

$$\mathcal{L} = \left\langle (f(G) - f^*) + \sum_i (t_i(G) - t_i^*) \right\rangle_\theta \tag{2.5}$$

Until convergence,

1. Initialize exponential random graph parameters θ randomly.
2. Draw samples using the following Block-Gibbs sampling algorithm:
   a) Conditional on the weights, update the binary adjacency adjacency matrix using DLMC.
   b) Conditional on the binary adjacency pattern, update weights using Hamiltonian Monte Carlo.

   The MCMC is run until the KSD achieves a pre-defined threshold.
3. Calculate importance weights for MCMC samples via KSD minimization.
4. Update θ via gradient descent

Implementation of this algorithm and application to the *Drosophila* connectome is the focus of ongoing work at the time of writing this thesis.

Chapter 3

Numerical Characterization of Support Recovery in Sparse Regression with Correlated Design

3.1 Introduction

While connectomics provides an exciting opportunity to probe the exact, anatomical connectivity between neurons, construction of these datasets remains an expensive and laborious process that is limited to small model organisms or localized spatial scales [105]. A long-standing alternative approach to understanding how the brain is wired together has been to extract functional or effective connectivity from distributed neural activity. The functional connectivity between neurons or populations of neurons encodes the statistical dependencies between firing activity. The intepretability of this connectivity depends on the accuracy of the underlying statistical inference process used to derive adjacency matrices from data. In particular, as functional connectomic studies frequently analyze graph-theoretic properties [106], understanding if and when the sparsity pattern (i.e., which edges are and are not present in the adjacency matrix) can be reliably estimated is of paramount scientific importance.

Abstractly, the inference problem can be formulated as that of reconstructing a k-sparse vector from noisy observations. In its simplest form, one is concerned with inference within the following model:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon \tag{3.1}$$

with $\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^p$ is a k-sparse vector. The noise is i.i.d, $\epsilon \in \mathbb{R}^n, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and the observational model is Gaussian, $y_i \sim \mathcal{N}(\boldsymbol{X}_i\beta, \epsilon_i)$. The sparse linear model is employed in diverse scientific fields [107–111]. In real world applications, it is also commonly the case that the design or covariate matrix X is correlated, so that the columns of

$\boldsymbol{X}$ can not be taken to be i.i.d. In this setting, the correct identification of non-zero elements of β, which is crucial for scientific interpretability, is especially challenging. Yet, a systematic exploration of the effect of correlations between the covariates on the recoverability of β is lacking.

Statistically optimal sparse estimates of β within (3.1) are returned by the solution to the following constrained optimization problem:

$$\begin{aligned} &\min ||y - \boldsymbol{X}\beta||_2^2 \\ &||\beta||_0 \leq \lambda \end{aligned} \tag{3.2}$$

Finding the global minima of problem (3.2) is NP-hard, though recent progress has been made in computationally tractable approaches [112, 113]. The most common approach is to relax the l_0 regularization. In this work, we focus on the Lasso, Elastic Net, SCAD, MCP [114–117], and $\text{UoI}_{\text{Lasso}}$, an inference framework we introduced in [118] that combines stability selection and bagging approaches to produce low variance and nearly unbiased estimates. To select the regularization strength or otherwise compare between candidate models returned between these estimators, one must employ a model selection criteria such as cross-validation or BIC. While the literature on sparsity inducing estimators and model selection criteria is vast, studies that consider the interaction of particular choices of estimator and model selection criteria are lacking. In particular, no systematic exploration of the impact of choice of estimator *and* model selection criteria on the selection accuracy of the resulting procedure when the predictive features exhibit correlations has been carried out. In this work, we address this gap by performing systematic numerical investigations of the selection accuracy performance of several estimators and model selection criteria across a broad range of regression designs, including diverse correlated design matrices.

3.2 Review of Prior Work

The statistical theory of the sparse estimators considered in this chapter is vast and we do not attempt to review it all here. Our particular focus is on characterizing finite sample selection accuracy, especially in the context of correlated design. The asymptotic oracular selection performance of the SCAD and MCP are well known [116, 117] and require only mild conditions on the design matrix. For the Lasso, one must impose an irrepresentible condition to guarantee asymptotic selection consistency [119]. Specifically, if we let $S \in \mathcal{I}_k$ index the true model support and let $\bar{S} := \{1, ..., p\} \setminus S$ index the complement of the model support, we can partition the feature covariance matrix as follows:

$$\Sigma = \begin{bmatrix} \Sigma_{SS} & \Sigma_{S,\bar{S}} \\ \Sigma_{\bar{S},S} & \Sigma_{\bar{S}\bar{S}} \end{bmatrix}$$

Estimator	Regularization
Lasso	$\lambda\|\beta\|_1$
Elastic Net	$\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$
SCAD	$\int_0^{\|\beta\|} dx \left(\lambda \mathbb{I}(\|\beta\| \leq \lambda) + \frac{(\gamma\lambda - x)_+}{(\gamma-1)\lambda}\mathbb{I}(\|\beta\| > \lambda)\right)$
MCP	$\int_0^{\|\beta\|} dx \left(1 - \frac{x}{\gamma\lambda}\right)_+$
$\text{UoI}_{\text{Lasso}}$	$\lambda\|\beta\|_1$ across bootstraps, see [118]

Model Selection Criteria	
Cross-Validation	R^2 averaged over 5 folds
BIC	$2\log\|y - X\hat{\beta}\|_2^2 - \log(n)\|\hat{\beta}\|_0$
AIC	$2\log\|y - X\hat{\beta}\|_2^2 - 2\|\hat{\beta}\|_0$
gMDL [129]	$\begin{cases} \frac{\hat{k}}{2}\log\left(\frac{n-\hat{k}}{\hat{k}}\frac{y^\top y - \|y-\hat{y}\|_2^2}{\|y-\hat{y}\|_2^2}\right) + \log n & \text{if } R^2 > \frac{\hat{k}}{n} \\ \frac{n}{2}\log\left(\frac{y^\top y}{n}\right) + \frac{1}{2}\log(n) & \text{otherwise} \end{cases}$
Empirical Bayes [130]	$2\log\|y - X\hat{\beta}\|_2^2 - \begin{cases} \hat{k} + \hat{k}\log(\hat{y}^\top\hat{y}) - \hat{k} - 2((p-\hat{k})\log(p-\hat{k}) + \hat{k}\log\hat{k}) & \text{if } \hat{y}^\top\hat{y}/\hat{k} > 1 \\ \hat{y}^\top\hat{y} - 2((p-\hat{k})\log(p-\hat{k}) + \hat{k}\log\hat{k}) & \text{otherwise} \end{cases}$

Table 3.1: (Top) Sparsity inducing regularized estimators. λ and γ denote regularization parameters. In this study, we keep γ for SCAD and MCP fixed to 3. (Bottom) Model selection criteria. Here and throughout, $\hat{k}$ refers to the estimated support size, $\hat{y}$ the model predictions of y, and p is the total number of features.

Letting β_S denote the vector of non-zero coefficients. The irrepresentable constant (section 3.2 in [119]) is then given by $\eta = 1 - |\Sigma_{\bar{S},S}\Sigma_{SS}^{-1}\,\text{sign}(\beta_S)|_\infty$. For $\eta < 0$, the Lasso is not asymptotically selection consistent.

The finite sample implications of these differing requirements have not been explored. A series of works have addressed the correlated design problem by devising regularizations that tend to assign correlated covariates similar model coefficients [120–125]. In fact, the Elastic Net was the first estimator introduced to exhibit this type of "grouping" effect [115]. However, this type of behavior can be undesirable in many real data applications where covariates may be correlated, yet still contribute heterogenously to a response variable of interest.

When the true model generating the data is contained amongst the candidate model supports, the BIC and gMDL have asymptotic guarantees of selection consistency [119]. Extensions of these results to the high dimensional case are available [126], but fall outside the scope of this work. Implicit in these theoretical results is that one can evaluate the penalized likelihoods on all 2^p candidate model supports [127]. Practically, one first assembles a much smaller set of candidate model supports using a regularized estimators. To this end, the use of the BIC with SCAD has been shown to be selection consistent [128].

A more recent body of work has focused on non-asymptotic analyses of model (3.1) in the framework of compressed sensing rather than regression. Here, the sparsity level of β is a priori known, and the sensing matrix X is typically drawn from a random ensemble. In

this setting, it is possible to establish sharp transitions in the mean square error distortion of the signal vector as a function of measurement density (i.e., asymptotic n/p ratio) [131]. Necessary and sufficient conditions on the number of samples needed for high probability recovery of the support of β by the Lasso was treated in [132]. Subsequently, a series of works examined the information theoretic limits on sparse support recovery by forgoing analysis of computationally tractable estimators in favor of establishing the sample complexity of exhaustive evaluation of all $\binom{p}{k}$ possible supports via maximum likelihood decoding [133–140]. This approach provides information theoretic bounds on the selection performance of any inference algorithm, and a measure of the suboptimality of existing algorithms.

Of particular relevance to this work are [133] and [138], whose analyses permit correlated sensing (i.e., design) matrices. Let $\beta_{\min}$ be the minimum non-zero coefficient of β, σ^2 be the additive noise variance, and Σ be the covariance matrix of the distribution from which columns of $\boldsymbol{X}$ are drawn. Denote the set of all subsets of $\{1, 2, ..., p\}$ of size k as $\mathcal{I}_k$. $\mathcal{I}_k$ indexes possible model supports. Given $S, T \in \mathcal{I}_k$ we define the matrix $\Gamma(S, T)$ to be the Schur complement of $\Sigma_{S\cup T, S\cup T}$ with respect to Σ_{TT}, $\Gamma(T, S) = \Sigma_{S\setminus T, S\setminus T} - \Sigma_{S\setminus T, T}(\Sigma_{TT})^{-1}\Sigma_{T, S\setminus T}$. Let $\rho(\Sigma, k)$ be the smallest eigenvalue this matrix can have for any T: $\rho(\Sigma, k) = \min_{T\in\mathcal{I}_k\setminus S} \lambda_{\min}(\Gamma(T, S))$. From these quantities, we define α:

$$\alpha = \frac{\beta_{\min}^2 \rho(\Sigma, k)}{\sigma^2} \tag{3.3}$$

In Theorem 1 of [133], sufficient conditions on the sample size required for an exhaustive search maximum likelihood decoder to recover the true model support with high probability are given in terms of p, k, and α:

Theorem 5 *Theorem 1 of [133]. Define the function* $g(c_1, p, k, \alpha)$*:*

$$g(c_1, p, k, \alpha) := (c_1 + 2048) \max \left\{ \log \binom{p-k}{k}, \log(p-k)/\alpha \right\}$$

If the sample size n *satisfies* $n > g(c_1, p, k, \alpha)$ *for some* $c_1 > 0$*, then the probability of correct model support recovery exceeds* $1 - \exp(-c_1(n-k))$.

If $\alpha^{-1} > p\log(p-2k) + 2k/p$, then g, and therefore the sample complexity of support recovery, will be modulated by α for p large enough. Many of the design matrices considered in our numerical study (see Section 3) satisfy this condition.

In contrast to compressed sensing, the sparsity level of β (i.e., k) is typically unknown in applications of regression. Furthermore, sufficient conditions on high probability theory such as Theorem 1 above rely on concentration inequalities, which may formally hold in the non-asymptotic setting, but are rarely tight. As a result, the applicability of these results for practitioners evaluating the robustness of support recovery in finite sample regression is unclear. The main contribution of this chapter is to address this gap through extensive numerical simulations. We find α to be a useful measure of the difficulty of a particular

regression problem, and find selection accuracy performance to be modulated by α even when it does not satisfy the condition stated above.

Previous empirical works have evaluated the effects of collinearity on domain specific regression problems [141, 142] and evaluate the efficacy of various information critera for model selection [143–145]. Finally, the performance scaling of a series of sparse estimators with sample size is evaluated in [146].

In contrast, we specifically consider the differing effects on selection accuracy of *joint* choices of estimators and model selection criteria. We demonstrate that the choice of model selection criteria significantly modulates the selection performance of estimators, and that there are empirically identifiable transition points in the value of α beyond which the selection performance of all inference procedures degrades.

3.3 Description of Simulation Study

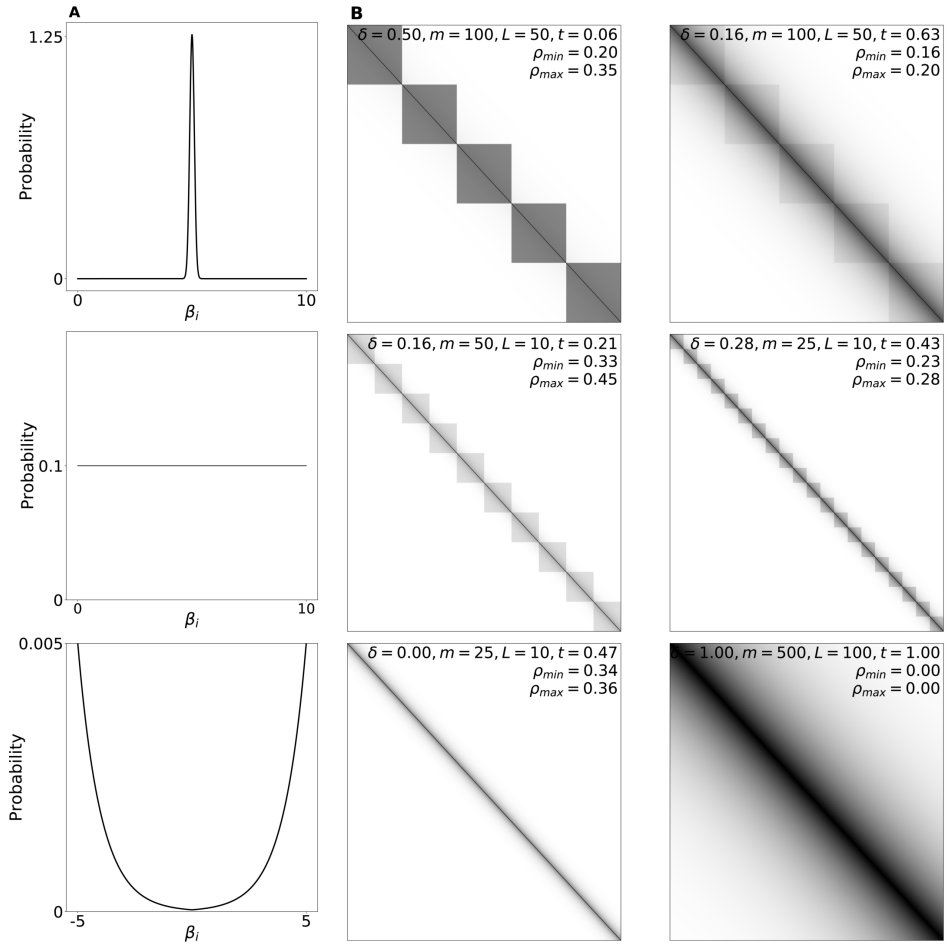


Figure 3.1: Design of Simulation Study. (a) (Right column) Coefficients β are drawn from a narrowly peaked Gaussian, uniform, and inverse exponential distribution. (b) (Left column) Design matrices are parameterized as $\Sigma = t \oplus_i \delta I_{m \times m} + (1-t)\Lambda(L)$ where $\Lambda(L)_{ij} = \exp(-|i-j|/L)$ and $I_{m \times m}$ is the m-dimensional identity matrix. Parameters δ, m, t and L are shown for each example design matrix. Also shown are bounds for the minimum and maximum $\rho(\Sigma, k)$ across k.

We consider regression problems with 500 features with 15 different model densities (i.e., $|\beta|_0$) logarithmically distributed from 0.025 to 1. Additionally, we vary over the following design parameters:

1. 80 covariance matrices Σ of exponentially banded, block diagonal, or a structure that interpolates between the two (see **Figure 3.1**).

2. Three different β distributions: a sharply peaked Gaussian, a uniform, and an inverse exponential distribution (see **Figure 3.1**)

3. Signal to noise (SNR) ratios of 1, 2, 5, 10. We define signal to noise as $|X\beta|_2^2/\sigma^2$.

4. Sample to feature (n/p) ratios of 2, 4, 8, and 16.

To simplify the presentation, we often restrict the analysis to the following three combinations of SNR and n/p ratio that represent ideal signal and sample, SNR starved, and sample starved scenarios, respectively:

1. Case 1: SNR 10 and n/p ratio 16

2. Case 2: SNR 1, and n/p ratio 4

3. Case 3: SNR 5 and n/p ratio 2

A distinct model design is comprised of a particular model density, predictor covariance matrix, a coefficient distribution drawn from one of the three β-distributions, an SNR, an n/p ratio. Each distinct model is fit over 20 repetitions with each repetition being comprised of a new draw of $X \sim \mathcal{N}(0, \Sigma)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, with σ^2 set by the desired SNR. We use the term estimator to refer to a particular regularized solution to problem 3.1 (e.g. Lasso) and model selection criteria to refer to the method used to select regularization strengths (e.g. BIC). The estimators and model selection criteria we consider are listed in Table 3.1. We use the term inference algorithm to refer to particular choices of estimator and model selection criteria.

Let $S = \{i|\beta_i \neq 0\}$ in eq. 3.1, and $\hat{S} = \{i|\hat{\beta}_i \neq 0\}$, i.e. the true and estimated model supports. Then, we evaluate regression on the basis of selection accuracy $(1 - \frac{|(S\setminus\hat{S})\cup(\hat{S}\setminus S)|_0}{|S|_0+|\hat{S}|_0})$, false negative rate $(\frac{|S\setminus\hat{S}|_0}{|S|_0})$ and false positive rate $(\frac{|\hat{S}\setminus\hat{S}|_0}{p-|S|_0})$. We use α to associate a single scalar to measure the difficulty of a regression problem. Smaller α correspond to harder regression problems.

In practice, we do not calculate $\rho(\Sigma, k)$ explicitly, but rather lower bound it. Let S be the true model support and T an alternative model support. First, observing that $\Gamma(S, T)$ is just the inverse of the subblock of the precision matrix $\Sigma^{-1}_{S\setminus T, S\setminus T}$, we seek to bound the largest eigenvalue of this subblock:

$$(\rho(\Sigma, k))^{-1} \leq \max_{T\in\mathcal{I}_k\setminus S} \lambda_{\max}(\Sigma^{-1}_{S\setminus T, S\setminus T})$$

We do this via the use of Brauer-Cassini sets [147]. For an arbitrary $n \times n$ complex matrix A with entries a_{ij}, let $R_i = \sum_{j\neq i} |a_{ij}|$. Then, define the Brauer sets $K_{ij} : K_{ij} = \{z \in \mathbb{C} : |z - a_{ii}||z - a_{jj}| \leq R_i R_j, i \neq j\}$. The eigenvalues of A lie within $\bigcup_{i,j} K_{ij}$

To bound specifically the largest eigenvalue of $\Sigma^{-1}_{S\setminus T, S\setminus T}$, we use the following proposition:

Proposition 3 *Let $A \in \mathbb{R}^{n\times n}$ be a positive semidefinite matrix and let $\tilde{A}$ be the the matrix that results from sorting the rows of $|A|_{ij} = |a_{ij}|$ in descending order. Define the truncated*

row sums $\tilde{R}_i = \sum_{j=1}^{m} |\tilde{a}_{ij}|$ *where* $\tilde{a}_{ij}$ *are the entries of* $\tilde{A}$*. Let* $B \in \mathbb{R}^{m \times m}$ *be a principal submatrix of* A*. The largest eigenvalue of* B *is bounded from above by:*

$$\max_{i,j:i \neq j} \left[\sqrt{\tilde{R}_i \tilde{R}_j + \frac{1}{4}(|\tilde{a}_{i0}| - |\tilde{a}_{j0}|)^2} + \frac{1}{2}(|\tilde{a}_{i0}| + |\tilde{a}_{j0}|) \right]$$

Proof: Since A is positive semidefinite, by Proposition 1, it follows that the largest eigenvalue of A can be no larger than the rightmost boundary of the rightmost Brauer set on the real axis. As a principal submatrix of a positive semidefinite matrix is also positive semidefinite, this holds analogously for the matrix B and the Brauer sets $\hat{K}_{ij} = \{z \in \mathbb{C} : |z - b_{ii}||z - b_{jj}| \leq \hat{R}_i \hat{R}_j, i \neq j\}$ where $\hat{R}_i = \sum_{j=1, j \neq i}^{m} |b_{ij}|$. In Cartesian coordinates, the Brauer set is defined on the real axis by $(x - b_{ii})(x - b_{jj}) = \hat{R}_i \hat{R}_j$. The rightmost root of this equation is given by $\frac{1}{2}(b_{ii} + b_{jj}) + \sqrt{\hat{R}_i \hat{R}_j + \frac{1}{4}(b_{ii} - b_{jj})^2}$

By sorting A to obtain $\tilde{A}$, we necessarily have

$$\max_{i,j \in \{1,\ldots,n\}, i \neq j} \frac{1}{2}(|\tilde{a}_{i0}| + |\tilde{a}_{j0}|) + \sqrt{\tilde{R}_i \tilde{R}_j + \frac{1}{4}(|\tilde{a}_{i0}| - |\tilde{a}_{j0}|)^2} \geq$$

$$\max_{i,j \in \{1,\ldots,m\}, i \neq j} \frac{1}{2}(b_{ii} + b_{jj}) + \sqrt{\hat{R}_i \hat{R}_j + \frac{1}{4}(b_{ii} - b_{jj})^2} \quad \blacksquare$$

Proposition 2 enables us to bound the largest eigenvalue of a subblock of a matrix of a given size. Depending on the overlap between sets T and S, the dimension of the matrix $\Sigma^{-1}_{S \setminus T, S \setminus T}$ will vary. However, by the Cauchy interlacing theorem, the largest eigenvalue of a proper submatrix of dimension k' is bounded by the largest eigenvalue of subbmatrices of dimension $k > k'$. Therefore, we use the results of Proposition 2 to bound the largest eigenvalue of subblocks of Σ^{-1} of dimension k, corresponding to searching over T that are completely disjoint from S. Inverting this bound then gives a lower bound on $\rho(\Sigma, k)$.

3.4 Results of Simulation Study

False Positive/False Negative Characteristics

We first visualized support selection performance across estimators by scattering the false negative rate vs. false positive rate of each fit for several representative model densities (**Fig. 3.2** for BIC and AIC selection, (other criteria are visualized below in **Figure 3.3**). Each scatter point represents the selection characteristics of fits to a distinct model design averaged over its 20 instantiations. The boundaries of the grayscale partitions of the false positive false negative rate plane correspond to contours of equal selection accuracy. The rotation of these contours with the true underlying model density reflects the relative importance of false negative and false positive control in modulating selection accuracy. Specifically,

rotation towards the horizontal implies larger sensitivity to false positives, while conversely rotation towards the vertical implies greater sensitivity towards false negatives.

The accuracy of estimators exhibited clear structure that depends on the characteristics of the model design described above. We observe in panel A of **Figure 3.2** that estimators that more aggressively promote sparsity (SCAD, MCP, UoI in red, green, and dark blue, respectively) featured better selection accuracy at low model densities (i.e. scatter points for these estimators lie in the white to light gray shaded regions), whereas those that control false negatives less aggressively, namely the Elastic Net (orange) and to a lesser extent the Lasso (cyan), fared better in denser true models (panel C). The scatter points for each estimator formed bands that span the false negative rate. This banding effect was most pronounced for SCAD/MCP/UoI.

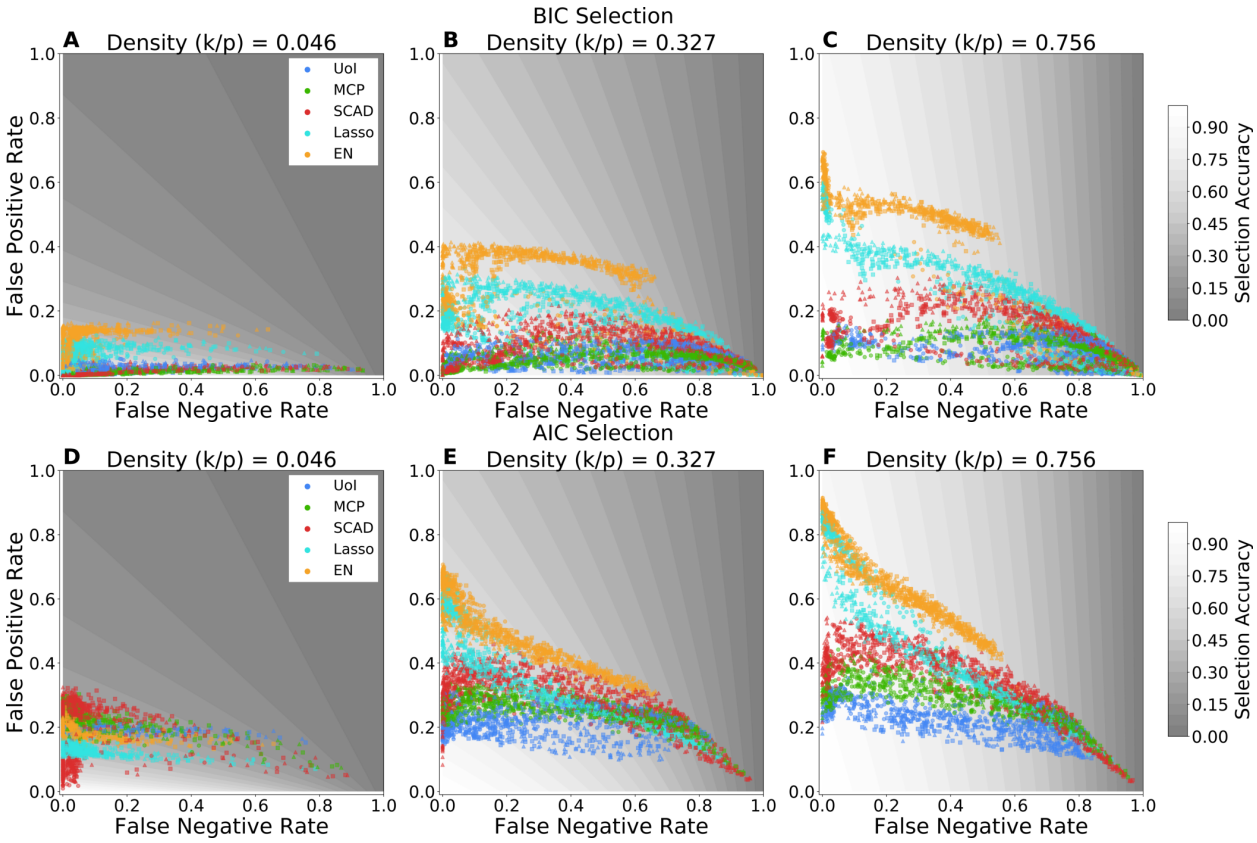


Figure 3.2: Scatter plots of the false negative rate vs. false positive rate for BIC selection (A-C) and AIC selection (D-F) across 3 different model densities (n/p ratio = 4, all signal to noise parameters included). Each scatter point represents a single fit. β distributions are encoded in marker shapes (square: uniform distribution, triangular: inverse exponential distribution, circular: Gaussian distribution). Shaded regions represent regions of equal selection accuracy. The orientation of these regions for different model densities illustrates the differing contributions of false negatives vs. false positives, with false positive control being far more important for sparser models, and conversely false negatives being more important for denser models.

Comparing the BIC selection (**Fig. 3.2 A-C**) to AIC (**Fig. 3.2 D-F**), these scatter plots also revealed that varying model selection methods also systematically shifted false negative & false positive characteristics of estimators. Selection methods with lower complexity penalties (i.e., AIC, CV) lifted the bands up along the false positive direction. Comparing the location of the blue/red/green scatter points between panels B and E, for example, we note that this effect was most dramatic for the set of estimators that most aggressively control false positives (SCAD/MCP/UoI). Consequently, similar tradeoffs as described before arose, with empirically better selection accuracy when models are dense obtained for AIC/CV, and vice versa for larger complexity penalties (BIC).

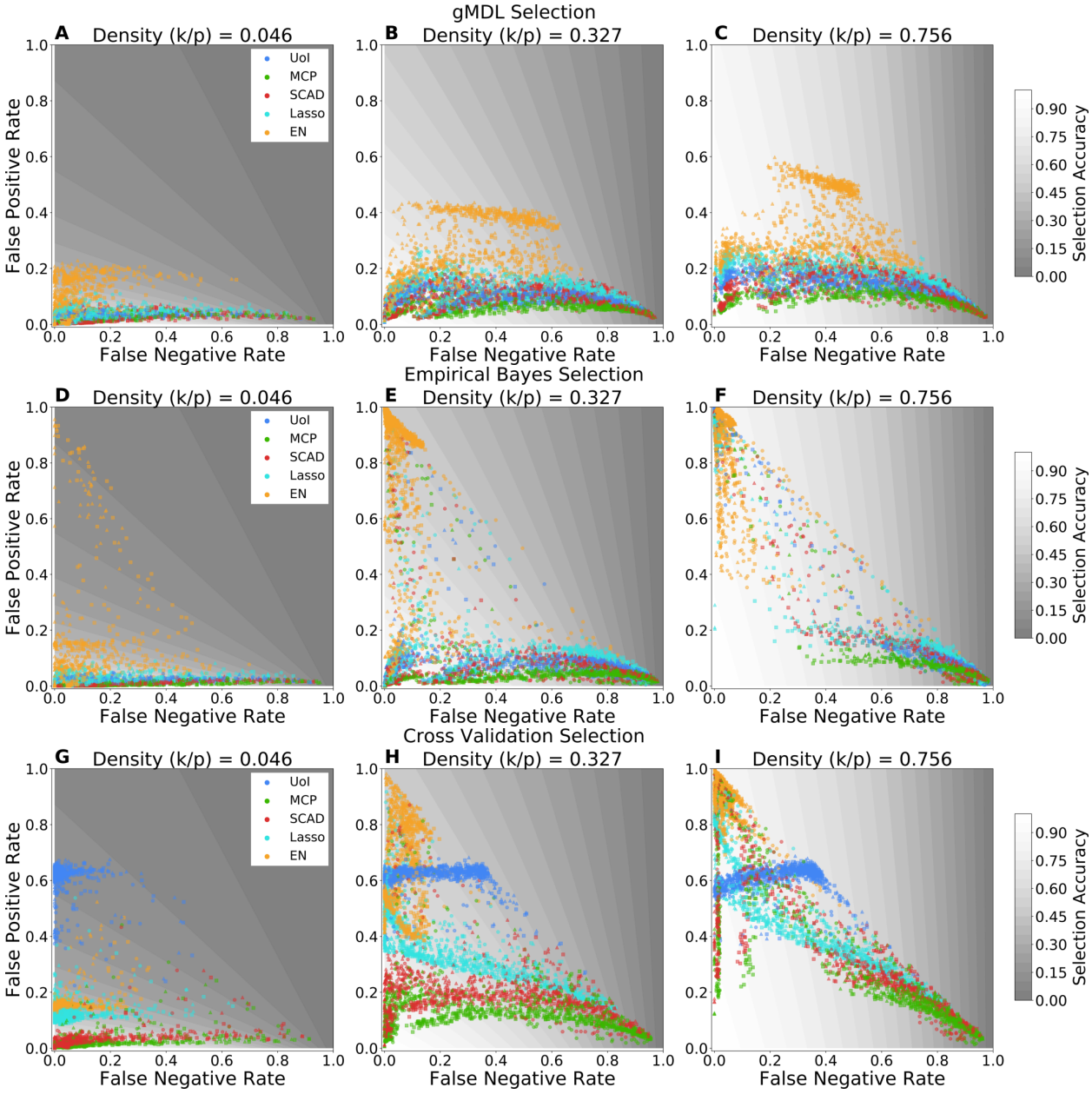


Figure 3.3: Scatter plots of the false negative rate vs. false negative rate for gMDL model selection (panels A-C), empirical Bayes model selection (panels D-F), and cross-validation selection (panels G-I) for 3 different model densities (0.03, 0.33, 0.76). The n/p ratio displayed is 4, all signal to noise parameters are included.

We note the qualitative similarity of the profile of scatter points for gMDL selection panels A-C to that of BIC (**Figure 3.2**, panels A-C). The gMDL selection method, while nominally sensitive to the underlying model sparsity, gave rise to tight false positive control

for all estimators, save for the Elastic Net (orange scatters). In contrast to the BIC at dense model density (panel C, both figures), the gMDL selection criteria provided tighter false positive control for the Lasso (cyan scatter points), at the expense of increased false negatives.

In panels A, B, and D, E of **Figure 3.3**, we observe that the gMDL and empirical Bayes selection method led to similar selection profiles for UoI, SCAD, MCP, and Lasso, with nearly all scatter points staying at false positive rates < 0.25. However, we also observe that supports selected by using the Elastic Net, in particular (orange), and other estimators for particular sets of parameters, became very dense (false positive rate $\to 1$) at model density 0.33 and especially model density 0.76 (panel F). This led to overall better selection accuracy (white regions) in denser models. We conclude that the choice of estimator *and* model selection criteria are both important in determining the false positive/false negative rate behavior of inference strategies.

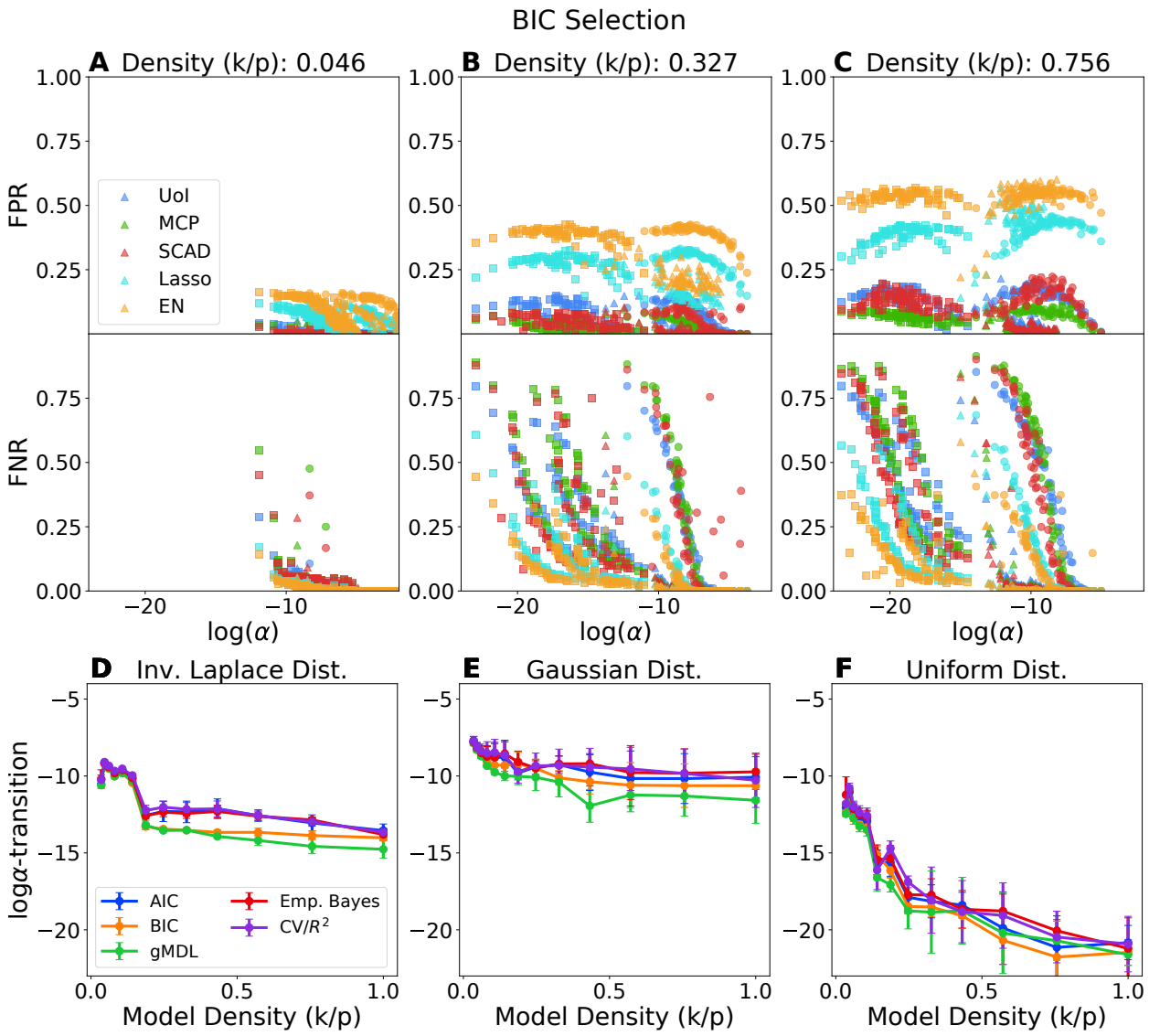


Figure 3.4: (A-C) Scatter plot of the false positive rate and the false negative rate vs. α for each estimator using BIC as a selection criteria for three different model densities. β distributions are encoded in marker shapes (square: uniform distribution, triangular: inverse exponential distribution, circular: Gaussian distribution). (D-F) Plot of the α-transition point associated with an inference algorithm's false negative rate as a function of model density, separated by β distribution and selection method. Errorbars are standard deviations taken across repetitions and estimator. The different numerical regimes of the α-transition (highest in panel E, intermediate in panel D, and lowest in panel F) are attributable to the different characteristic value of β_{min} for the different β distributions.

Recalling that the parameter α tunes the difficulty of the selection problem, we scattered the false positive and false negative rate vs. α for each inference algorithm across different model densities. A representative set of such plots for BIC selection is shown in **Figure 3.4A-C**; other selection methods are shown in Figure 3.4. Cross-Validation is not included due to space considerations but behaves similarly to the AIC. There was broadly large variation in performance modulated by the selection method employed. Furthermore, β-distributions are separately resolvable due to their different typical values of β_{min}. For example, in the bottom axes of **Figure 3.4C**, for each estimator, the uniform distribution scatter points (squares) lie to the left of the inverse exponential distribution (triangular), which in turn lies to the left of the Gaussian distribution (circular).

In line with **Figure 3.2**, the false positive rate was not modulated by α (**Fig. 3.4 A-C**, top axes). In fact, for some estimators, the highest false positive rate was achieved for intermediate α, followed by a decline in false positive rate for smaller α (e.g. Lasso in **Figure 3.4C**). The false positive rate is instead a characteristic of each estimator. The SCAD/MCP/UoI class of estimators achieved lower false positives than Lasso, which in turn featured lower false positives than the Elastic Net.

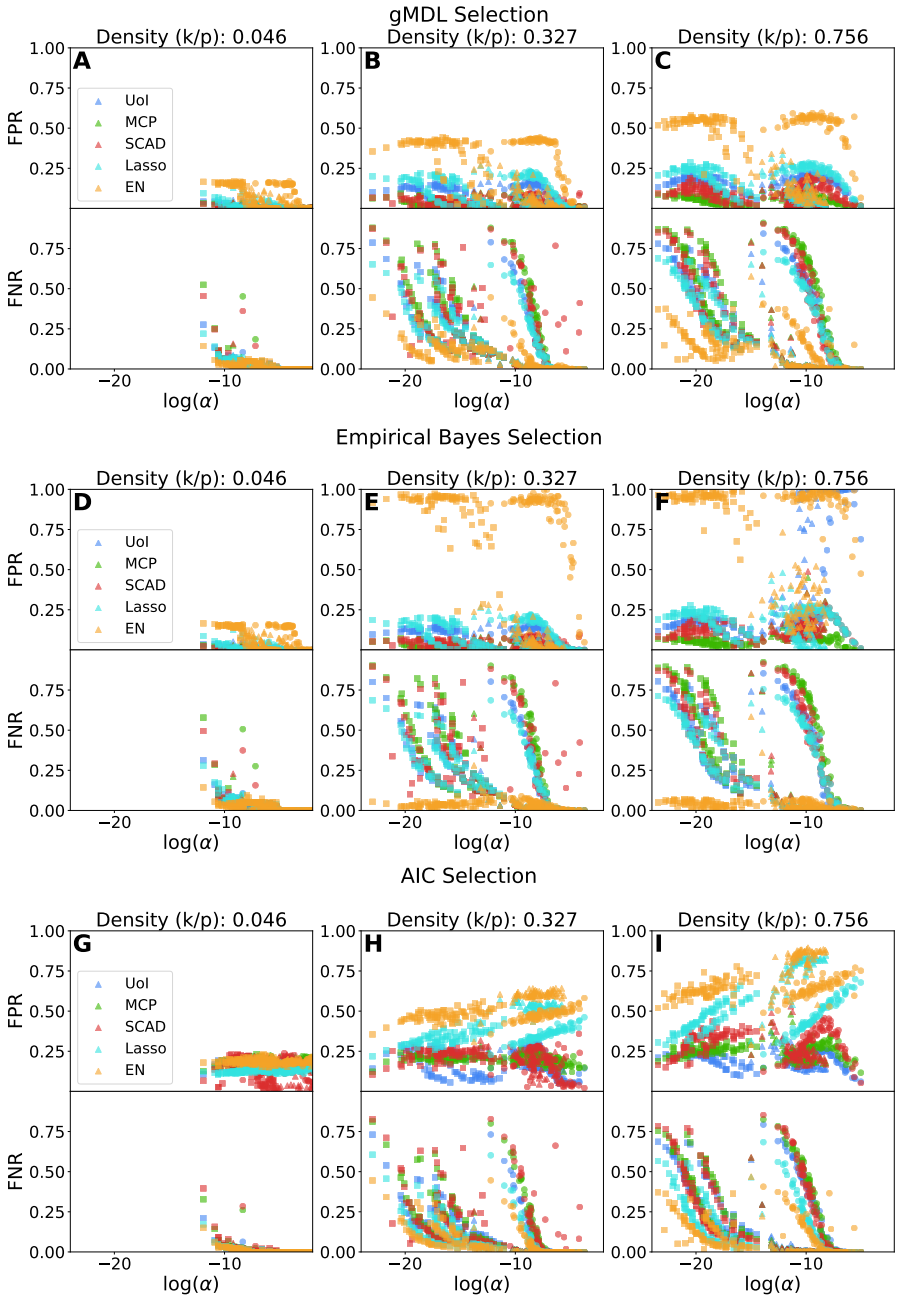


Figure 3.5: Plot of the false positive rate (FPR) and false negative rate (FNR) vs. $\log \alpha$ for signal case 1 across several model densities. gMDL selection was used in panels A-C, empirical Bayes in panels D-F, and AIC in panels H-I. The cross-validation selection method is not shown, but exhibited similar characteristics to AIC.

Model selection criteria can also be classified into a set that led to low false positive rates (gMDL, empirical Bayes, and BIC) vs. those that lead to high false positive rates (AIC,

CV), although the Elastic Net with empirical Bayes selection featured the highest false positive rate of any inference algorithm (**Fig. 3.3**, panels D-F). We note that the inverse exponential distribution (triangular points) induced very false negatives by any inference algorithms, likely due to its coefficient magnitudes being concentrated towards larger values.

On the other hand, the false negative rate scatter points, when separated by β-distribution, featured consistent behavior across inference algorithms. Focusing on BIC selection, all estimators achieved low false negative rates at the low model densities (**Fig. 3.4A**). At intermediate model densities (**Fig. 3.4B**), the false negative rate remained low until $\log \alpha$ became sufficiently small, at which point it rapidly increases. This value of $\log \alpha$ varied by β-distribution due to the differing characteristic values of $\beta_{\min}$, occuring around $\log \alpha \approx -7.5$ for the Gaussian distribution at model density 0.327, $\approx \log \alpha = -10$ for the inverse exponential distribution, and $\approx \log \alpha = -15$ for the uniform distribution. Otherwise, this transition point is fairly universal across inference algorithms.

To produce summary statistics of false negative rates across model densities, selection methods, and n/p ratio/SNR cases, we fit sigmoidal curves to data for each inference algorithm and for each β distribution. The sigmoid curve is described by 4 parameters:

$$S(\alpha) = c + \frac{a}{1 + \exp(-b(\alpha - \alpha_0))}$$

In particular, we use the fitted value for the sigmoid midpoint α_0, which we refer to as the α-transition point, to quantify the value of α at which false negative rate has begun to increase appreciably. We found a large degree of universality in this transition point across estimators and selection methods. In **Figure 3.4D-F** we have averaged curves across estimators and plotted the mean and standard deviation of the resulting α transition points. Colors now represent each selection method. The curves for each selection method were strikingly similar within a β distribution, with small standard deviations within each selection method indicating universality across estimators. The decrease of the α-transition point with increasing model density can be explained by the overall shift of α towards smaller values due to the increase of $\rho(\Sigma, k)$ with k.

α-dependence of False Positive/False Negative Coefficient Magnitude

In the preceding analysis we treated false positives and false negatives as hard thresholded quantities. On the other hand, one can ask whether false negatives primarily arise from setting support elements with small signal strength to zero, and conversely whether false positives are associated with small coefficient estimates. Thus, while exact model support recovery in most cases is unattainable, one would hope that support inconsistencies produce low distortion of the desired coefficient vector. To evaluate this supposition, we calculated the average magnitude of false negatives and false positives, and normalized these quantities by the average magnitude of ground truth β. In the case of false negative magnitudes,

we focused on the uniform β distribution, as this provides the most "edge" cases of small coefficient magnitudes. Raw scatter plots of these quantities (not shown) revealed that at low correlations, the hoped for low distortion effect largely holds true, but that there is an α transition point for both false negative and false positives after which significantly larger ground truth β_i are selected out, and erroneously selected β_i are assigned much larger values relative to the true signal mean.

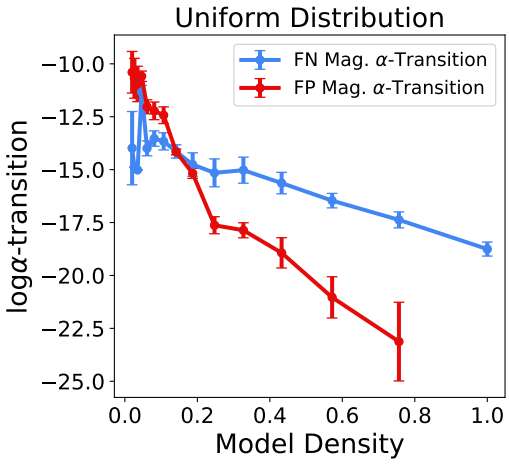


Figure 3.6: Plot of the average α transition point for estimation distortion across all inference algorithms and selection methods vs. model density for signal case 1. Errorbars represent standard deviation. After a model density of > 0.15, the transition generally occurs at lower correlations (smaller α) for the false negative magnitude. Furthermore, the variance across inference algorithms is consistently smaller for false negatives as opposed to false positives.

We again fit sigmoidal functions to the raw scatter points of normalized false negative & false positive magnitude vs. $\log \alpha$ and extracted the α-transition points as in **Figure 3D-F**. In **Figure 3.6**, we plot the transition point as a function of model density averaged across all estimators, selection criteria, and fit repetitions. For model densities > 0.15, the transition point occurs at much smaller correlation strengths for false negative distortions than for false positive distortions. The variance in the location of this transition point for false negative distortions is noticeably smaller than for false positive distortions. Nevertheless, similarly to the behavior exhibited by the α-transition points associated with the false negative rate, the α-transition points for false positive/false negative coefficient magnitudes is saliently uniform across inference strategies. Overall, these results highlight the usefulness in the parameter α, which emerges out of tail bounds on the performance of the exhaustive maximum likelihood decoder, as a quantifier of the difficulty of a sparse regression problem.

Overall Selection Accuracy

An inference algorithm deployed in practice must employ both an inference estimator and model selection criteria. We have therefore determined what the best performing combina-

tion is as a function of underlying model density and α. To set an overall scale for these comparisons, one can use an oracle selection criteria that simply chooses the support along a regularization path of maximum selection accuracy. For each value of α and model density, the maximum of this oracular selection across all estimators gives a proxy for the best achievable selection accuracy in principle at finite sample size and SNR.

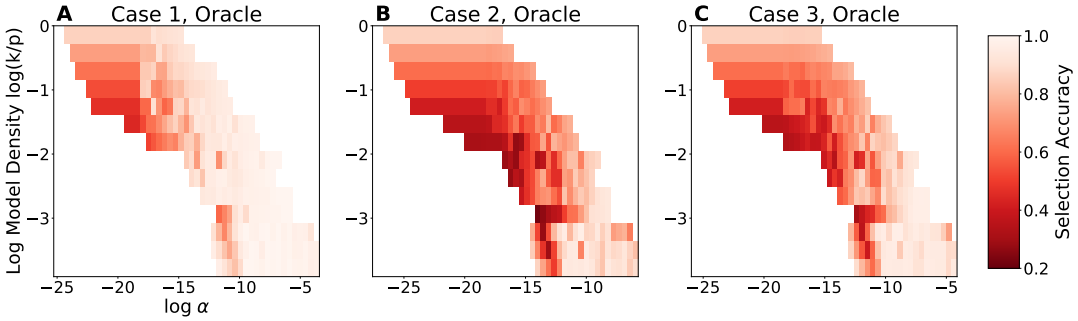


Figure 3.7: Oracle selection accuracy as a function of the log model density and α for each of the 3 signal cases described in Section 3. Each pixel in the colormap is the maximum oracle performance across all estimators for the particular combination of density and α. For ideal signal characteristics in Case 1 (panel A), near perfect support recovery is in principle possible for a broad range of correlation strengths for log model densities < -1.9. The similar oracle selection accuracies between cases 2 and 3 (panels B and C) suggest that the sample starved and signal starved regression problems behave similarly. As compared to Case 1, worst case performance for intermediate model densities ($log(k/p) > -2.3$ and < -0.69) is lower, especially for large correlations. For the densest models ($log(k/p) > 0.5$), oracle performance is relatively insensitive to correlation strength, reflecting the insensitivity of the FPR to α. Near-perfect support recovery is empirically still possible for the sparsest models ($log(k/p) < -3$).

In **Figure 3.7** we plot the oracle selector for each signal case. In the ideal signal and sample size case (case 1), the oracle selector was able to achieve near perfect selection accuracy in the fully dense models (top row, panel C) and those models with with density < 0.14 (log model densities < -2) even in model designs with very small α. The oracle selector suffered moderate loss of selection accuracy in intermediate model densities for model designs with small α (darker orange regions of panel C). A similar structure is present in the adequate sample but high noise and low sample size but adequate SNR cases (cases 2 and 3 in panels B and C, respectively), but the magnitude of selection accuracy performance loss and regions of α and model densities for which the loss occurred expanded. In particular, only in the very sparsest models (density < 0.05, log model density < -3) with larger α was perfect selection possible in principle.

For each estimator and selection criteria combination, we take the average deviation of its selection accuracy from the oracular performance shown in **Figure 3.7** as a measure of sub-optimality. We divide the analysis into an overall measure of sub-optimality, averaging over all model densities and α, as well as restricting the averaging to only sparse generative

	Case 1, All Densities				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp. Bayes	gMDL
EN	27.000	19.228	25.214	14.483	11.964
Lasso	23.867	14.634	27.151	5.840	5.982
MCP	23.408	**4.325**	6.948	4.717	5.220
SCAD	16.947	**3.233**	8.051	**3.534**	**4.039**
UoI Lasso	22.163	5.659	33.795	5.020	5.134

	Case 1, Sparse Models Only				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp. Bayes	gMDL
EN	35.139	25.146	33.098	17.533	15.371
Lasso	30.622	18.402	35.220	4.601	5.046
MCP	30.319	**1.121**	7.511	**1.033**	**2.184**
SCAD	21.267	**0.815**	9.361	**0.728**	**1.756**
UoI Lasso	29.558	3.290	44.522	3.077	3.396

	Case 2, All Densities				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp. Bayes	gMDL
EN	22.615	22.473	18.382	13.092	13.581
Lasso	22.495	19.004	19.524	16.708	14.185
MCP	26.411	13.521	**11.869**	14.382	14.671
SCAD	26.453	**11.789**	12.003	12.628	12.266
UoI Lasso	23.366	17.505	27.575	15.959	13.351

	Case 2, Sparse Models Only				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp. Bayes	gMDL
EN	29.940	18.539	25.556	16.691	13.503
Lasso	26.464	14.120	22.749	9.593	8.965
MCP	30.789	**3.879**	5.971	4.659	8.013
SCAD	31.579	**3.485**	8.154	4.213	6.434
UoI Lasso	27.151	7.287	36.588	9.150	7.024

	Case 3, All Densities				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp. Bayes	gMDL
EN	18.290	26.087	15.339	**10.420**	12.985
Lasso	19.424	19.653	17.955	17.632	15.227
MCP	23.590	16.526	14.600	17.211	18.156
SCAD	21.125	15.241	14.119	15.007	15.873
UoI Lasso	22.030	19.866	24.080	17.420	15.268

	Case 3, Sparse Models Only				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp. Bayes	gMDL
EN	22.596	15.357	21.305	13.346	11.668
Lasso	20.213	10.948	18.151	8.921	8.636
MCP	24.455	**5.059**	6.754	6.695	11.546
SCAD	22.070	**5.020**	9.135	**5.884**	9.905
UoI Lasso	21.729	7.765	31.733	9.615	7.832

Table 3.2: Table of summed deviation in selection accuracy from oracular performance. (Top) Case 1 signal conditons (SNR 10, n/p ratio 16). (Middle) Case 2 Signal Conditions (SNR 1, n/p ratio 4). (Bottom) Case 3 Signal Conditions (SNR 5 and n/p ratio 2). (Left column) All model densities. (Right column) Sparse models only. Best performers are highlighted in bold.

models (model densities < 0.3). The results are summarized in Table 2. The best performing inference algorithms are bolded. When taken across all model densities, in signal case 1 (Table 2 top left), the SCAD with BIC selection and SCAD with empirical Bayesian selection emerged as the best inference algorithms with respect to feature selection. When restricted to low SNR or low sample sizes (cases 2 and 3, Tables 2 middle and bottom, left), these strategies remained amongst the best performing, with cross-validated SCAD/MCP exhibiting robust selection in case 2, the Elastic Net with empirical Bayesian selection performing the best in case 3. When restricting to sparse models only, false positive control becomes paramount, and the Elastic Net was no longer competitive. Instead, the SCAD with BIC or empirical Bayes is near optimal in case 1 (Table 2, top right), and still the best performing in cases 2 and 3 (Table 2, top and middle, right). MCP exhibited similar performance, with UoI Lasso trailing slightly behind. Thus, in general, the SCAD estimator with BIC or empirical Bayesian model selection led to the most robust algorithm for feature selection.

3.5 Comparison of Bias/Variance of UoI vs. SCAD/MCP

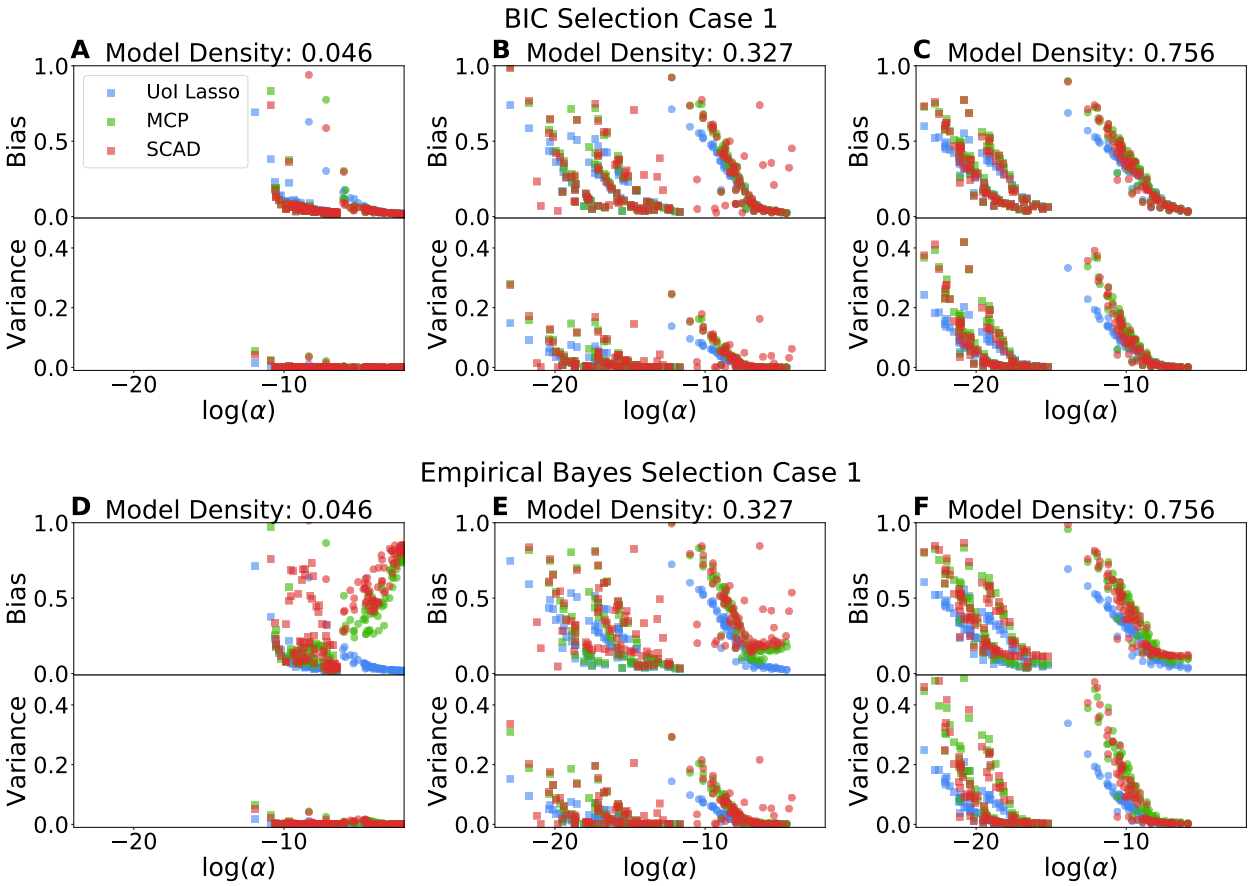


Figure 3.8: Plot of estimator Bias and Variance normalized by the number of non-zero true model coefficients vs. $\log \alpha$ for the BIC model selection (A-C) and the empirical Bayes model selection (D-F). UoI exhibits lower bias and variance than MCP and SCAD as α becomes smaller

The UoI_{Lasso}, SCAD, and MCP estimators, especially when combined with BIC or empirical Bayes model selection achieve state of the art model selection performance in the presence of correlated variability (Tables 2-7). The UoI algorithm separates estimation and selection by fitting OLS models to non-zero support coefficients, and uses bootstrapped aggregation to average together several model estimates. In **Figure 3.8** we compare the bias and variance between UoI/MCP/SCAD for the BIC and empirical Bayes model selection criteria. The bias ($\mathbb{E}(\hat{\beta}) - \beta$), where $\hat{\beta}$ are the estimated coefficients) and variance ($\mathbb{E}(\hat{\beta} - \mathbb{E}(\beta))^2$), was estimated by averaging over 20 fit repetitions, and further normalized by number of true non-zero model coefficients. When using BIC and empirical Bayes selection, (panels A-C of 3.8), UoI was able to reduce per coefficient bias/variance over SCAD and MCP at smaller α. Curiously, with empirical Bayesian selection, SCAD and MCP featured very high bias even at large $\log \alpha$ (panel D) in sparse models. These results highlight the ability of model averaging and re-estimation procedures to reduce estimation bias and variance.

3.6 Comparison with the Irrepresentable Constant

In [119], the importance of the irrepresentable constant, η, in ensuring the (asymptotic) selection consistency of the Lasso was established. To determine how η tunes the finite sample selection accuracy of the Lasso and the other estimators considered, we calculated η for the design matrices considered in this study and plot the selection accuracy vs. η for BIC selection and the Gaussian coefficient distribution (top axes of each panel in **Figure 3.9**). Simultaneously, we scatter η vs. $\rho(\Sigma, k)$ for the regression problems consdiered (bottom axes of each panel of **Figure 3.9**). For low model densities (panels A, B), the relationship between η and selection accuracy was as expected - selection accuracy of algorithms montonically decays as $\eta \to 0$. This decline was more gradual for the Lasso and Elastic Net (cyan and orange scatters), while it is quite dramatic for UoI/SCAD/MCP. Concommitantly, the relationship between η and $\rho(\Sigma, k)$ is monotonic, with smaller η corresponding to smaller $\rho(\Sigma, k)$. As the model density increases to 0.25 and above, the model selection performance declined as $\eta \to 0$ from the right, but rebounds for $\eta < 0$. At model density 0.25 (panel C), this effect was especially pronounced for the Lasso and Elastic Net. As the model density increases, a higher proportion of the feature covariance matrices considered in this study corresponded to $\eta < 0$, while the selection accuracy was no longer monotonically related to η. In panels D-F, one observes that the selection accuracy declined as $\eta \to 0$ from both the left and right, but that the selection accuracy was only slightly reduced from its maximum for the most negative values of η. This observation holds for all estimators. To explain this effect, we observe that beginning in **Figure 3.9C** and continuing in panels D-F, design matrices with $\eta < 0$ actually yielded relatively large $\rho(\Sigma, k)$, with small $\rho(\Sigma, k)$ corresponding to matrices with the smallest $|\eta|$. As this pattern mirrors that of the selection accuracy observed in **Figure 3.9**, we conclude that η tracks the *finite sample* selection accuracy performance of the Lasso (and to a lesser extent other estimators) only insofar as it is monotonically related to $\rho(\Sigma, k)$. In other words, $\rho(\Sigma, k)$ is a more reliable measure of

how feature covariance matrices modulate selection accuracy. Note that in [119], empirical evaluation was done on the probability that the entire Lasso solution path would contain the true support, *not* on the selection accuracy after employing a model selection criteria.

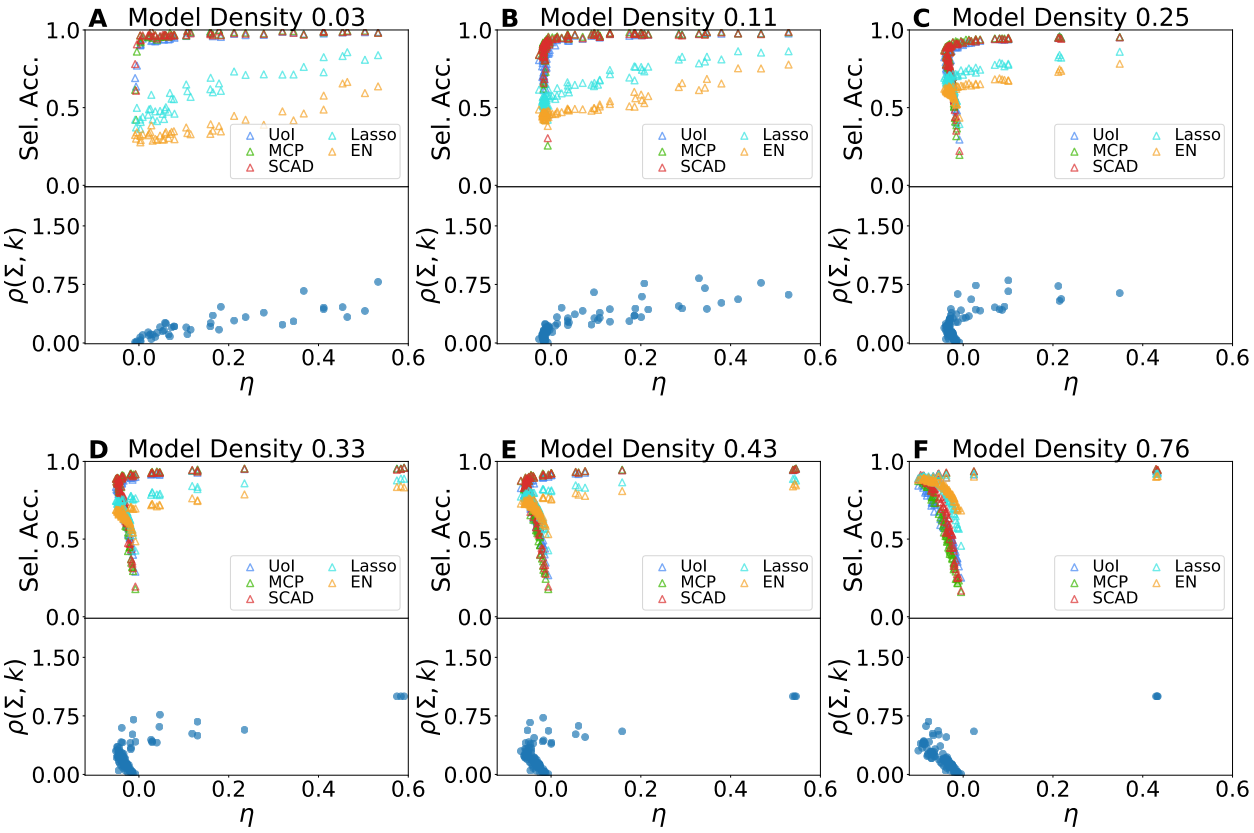


Figure 3.9: Plot of selection accuracy vs. η (top axes in each panel) and $\rho(\Sigma, k)$ vs η (bottom axes in each panel) for BIC selection criteria and Gaussian β distribution for different model densities. Note that $k = [\text{Model Density} \times 500]$. At low model densities (panels A, B), the decay in selection performance is monotonic as $\eta \to 0$, whereas for higher model densities (panels C-F), the selection accuracy decays rapidly with $|\eta|$, but selection accuracies for regression problems arising from design matrices correspond to $\eta < 0$ are high. In parallel, the relationship between η and $\rho(\Sigma, k)$ is monotonic at low model densities, but reverses back on itself at higher model densities, such that there are many design matrices for which $\rho(\Sigma, k)$ is large (corresponding to easier regression problems) but $\eta < 0$.

3.7 Discussion

Connections to Prior Work

Our numerical work corroborates and extends several results from the statistical literature in a non-asymptotic setting. We found the frequently employed cross-validated Lasso to be amongst the worst performing selection strategies. It has been shown that using predictive performance as a criteria for regularization strength selection with the Lasso leads to inconsistent support recovery [148]. A necessary and sufficient condition for asymptotically consistent model selection by the Lasso is for the irrepresentable condition to hold [119]. In the non-asymptotic setting of this study, we find that the parameter α is a more useful modulator of selection accuracy, and that the irrepresentable constant of [119] tracks the selection accuracy of Lasso only insofar as it tracks α (**section S5**). We find that the SCAD/MCP and UoI Lasso select model supports more robustly in the presence of correlated design. It is known that the SCAD/MCP do not require any strong conditions on the design matrix for oracular properties to hold [149], and neither does the BoLasso [150], upon which the selection logic of UoI is partially based on. Our work demonstrates that the choice of model selection criteria is as important as the choice of estimator to achieve good selection accuracy. The model selection criteria we have considered can all be categorized as penalized likelihood methods. Cross-validation is known to behave asymptotically like the AIC ([127]. The magnitude of this complexity penalty can be interpreted as a prior on the model size. We correspondingly find that the BIC performs best in sparse models, whereas the AIC and CV perform best in dense models. The tension between the BIC and AIC has been noted in the literature [151]. The asymptotic selection consistency of using BIC to select SCAD regularization strength has been noted in [128]. Our numerical investigations reveal that this remains one of the best extant selection strategies in non asymptotic settings with mild correlated variability as well.

The empirical Bayesian and gMDL procedures were devised with complexity penalties nominally adaptive to the underlying model density. We find that these methods lead to good model selection performance across model densities, but only in ideal signal conditions (i.e. case 1) and low design matrix correlations. There is therefore possible room for methodological development of adaptive complexity penalties. We leave this for future work.

Best Practices in Real Data

Proper model selection is essential for interpretability of parametric models. While sufficient conditions for model selection are available in the literature, they do not provide actionable results for the practitioner in real data. Our extensive numerical simulations reveal best practices. Non-convex optimization estimators such as the SCAD and MCP generically perform better at selection than the Lasso and Elastic Net when the underlying model is sparse. This in line with both prior numerical work and the understanding that asymptotically, these estimators are oracular selectors [117], [116]. Our work reveals that this performance

gap remains even as design matrices become increasingly correlated. While the SCAD and MCP are nonconvex problems, recent work has shown that the statistical performance of all stationary points is nearly equivalent [152]. Furthermore, development of the optimization algorithms for these estimators has matured to the point where regularization paths for the SCAD and MCP can be computed in the same order of magnitude of time as the Lasso/Elastic Net (see for e.g. [153]). Our work provides further motivation for the adoption of these algorithms. The $\text{UoI}_{\text{Lasso}}$ algorithm has selection performance competitive with MCP and SCAD in many cases. Furthermore, as we show in **section S4**, the OLS-bagging procedure used in coefficient estimates in UoI leads to lower bias/variance estimates than SCAD/MCP.

There is a tradeoff between false positive and false negative control achieved by model selection strategies. False positive control is largely insensitive to the degree of design correlation. Practitioners seeking tight control of false negatives in model selection may be inclined to use the Elastic Net estimator. The presence of a number of fairly generic α transition points after which selection accuracy degrades, and false negative & positive magnitude inflates suggests a heuristic criteria that could be estimated from the sample covariance. Specifically, combining empirical estimates of the precision matrix with empirical estimates of β_{min} and σ^2 allows one to estimate α, and therefore have a rough sense of whether selection and estimation performance is likely to have degraded due to correlated covariates or low signal strength.

3.8 Conclusions

Our empirical results reveal that the joint choice of sparse estimator and model selection criteria significantly modulates selection performance. Nevertheless, with the exception of the previously mentioned [128], theoretical results that capture non-asymptotic behavior of regularization strength selection via specific model selection criteria are lacking.

We found no inference algorithm to be dominant across underlying model density in the presence of correlated covariates, including the nominally adaptive empirical Bayes and gMDL selection criteria. Whether these reflect information theoretic constraints or methodological gaps is a potentially avenue of future work. We also believe our observation of a universal α-transition point across false negatives and coefficient distortion to be novel. This phenomena is reminiscent of the well known reconstructability transition in compressed sensing as a function of noise level and sampling density [131]. An average case analysis of coefficient support distortion as a function of α or other spectral parameters of the design matrix will be the topic of future work.

Chapter 4

Conclusion and Future Directions

Understanding the principles by which the brain functions requires studying phenomena across levels of description and spatiotemporal scales. In this thesis, I addressed this challenge through the development and characterization of novel theory and statistical analysis methods. In chapter 1, I presented a normative characterization of behaviorally relevant neural population dynamics through the lens of feedback control theory. I established fundamental links between the anatomy of neural circuits, as constrained by Dale's Law, their dynamics, as manifest through non-normal dynamics, to the optimality of various subspaces with respect to feedback and feedforward control. In chapter 2, I examined the relationship between local connectivity statistics between neurons and global measures of network function in the Drosophila hemibrain connectome. I also proposed an algorithmic approach to directly impose global, or top-down functional measures on networks as constraints within maximum entropy network models. Chapter 3 empirically investigates the fundamental limits of sparse recovery in linear statistical models with correlated design matrices. The results provide prescriptive approaches to choosing sparsity-inducing estimators and model selection criteria with applications to the estimation of functional connectivtiy from neural recordings.

In closing, I describe several avenues along which the work of this thesis could be extended. While the work described in chapter 1 examined measures of controllability and confined its analysis to single brain areas, analysis of the co-recorded activity from multiple brain areas allows for the analysis of how population activity in one area communicates with or is controlled by activity in another area. There are currently several proposed mechanisms by which brain areas communicate [154], including the idea of communication through coherence [155] and the use of communication subspaces [156]. While rigorously establishing the mechanism of causal effects by one brain area on another requires perturbation experiments wherein one area is activated or silenced, there is in parallel a need to develop statistical methods that can identify the coupling between high dimensional population recordings.

The transfer of information between dynamical systems is a necessary component of the control of one dynamical system by another. Paralleling the development of dimensionality reduction methods to identify feedforward and feedback controllable subspaces in chapter 1, a promising research direction is to develop dimensionality reduction methods that can

distinguish between feedforward and feedback communication across multiple co-recorded brain areas. To do so, I propose using the directed information (DI) [157] as the objective function for dimensionality reduction. Directed information, closely related to measures such as transfer entropy and Granger causality [158], is a measure of (directed) information transfer between dynamical systems and plays a fundamental role in the theory of communication over channels with feedback. In the case of linear stochastic systems, the existence of DI in one or both directions between two brain areas is equivalent to testing for the absence or existence of state feedback [159]. Using this fact, a dimensionality reduction algorithm immediately suggets itself. Intuitively, the joint activity within two co-recorded brain regions (denote as $\Psi_1(t)$ and $\Psi_2(t)$) is composed of purely local dynamics, feed-forward interaction subspaces (FFIS) directed from one area to another, and a shared feedback interaction subspace (FBIS). An interaction subspace is comprised of subspaces in both areas $[U\Psi_1(t), V\Psi_2(t)]$. An algorithm to identify the FFIS and FBIS takes on the following form:

(1) Find projections U_1, V_1 applied to $\Psi_1(t)$ and $\Psi_2(t)$ respectively, such that the dynamics in the projected subspace maximizes the total DI from $\Psi_1(t)$ to $\Psi_2(t)$: DI($U_1\Psi_1(t) \rightarrow V_1\Psi_2(t)$).

(2) Find a second pair of projections U_2, V_2 applied to $\Psi_1(t)$ and $\Psi_2(t)$ respectively, such that the dynamics in the projected subspace maximizes the total DI from $\Psi_2(t)$ to $\Psi_1(t)$: DI($V_2\Psi_2(t) \rightarrow U_2\Psi_1(t)$).

(3) The FBIS between $\Psi_1(t)$ and $\Psi_2(t)$ is defined by the intersections $[U_1 \cap U_2, V_1 \cap V_2]$. These are the subspaces with bi-directional information above and beyond self-predictive information within an area. Correspondingly, the FFISs from $\Psi_1(t)$ to $\Psi_2(t)$ and $\Psi_2(t)$ to $\Psi_1(t)$ contain the complementary DI that is unidirectional, and are obtained from the intersections $[U_1 \cap U_2^c, V_1 \cap V_2^c]$ and $[U_2 \cap U_1^c, V_2 \cap V_1^c]$, respectively ($S^c$ denotes the complement set to S). These FFISs contain DI that is strictly from one area to another, excluding the FBIS. These definitions formalize the intuitions described above.

The decomposition of co-recorded brain areas into FBIS and FFIS would significantly enrich the idea of communication subpaces. Analogously to FCCA, the use of linear projection matrices permits the identification of the single neurons important for mediating the dynamics within the respective subspaces. An interesting analysis would therefore be to compare the subspaces and single neurons involved in inter-area communication across areas with those involved with feedforward and feedback controllable dynamics within areas.

There are also numerous further directions to pursue regarding the relationship between anatomical structures of neural circuits arising from Dale's Law and the ability of these circuits to be controlled. In particular, a widespread observed feature of cortical circuits is that excitatory subnetworks on their own are unstable; the rapid action of inhibition is required to maintain overall stability in response to input perturbations [160]. Inhibitory circuits therefore serve as a local feedback controller for excitatory subnetworks, balancing these latter circuits' sensitivity to inputs with homeostatic regulation. The fact that many circuits in cortex are inhibitory stabilized may have consequences for their ability to be controlled under feedback. In control theory, the stabilization of an open loop unstable

system gives rise to an overall system that possesses non-minimum phase characteristics [161]. In turn, there are well known fundamental limitations in the feedback controllability of non-minimum phase systems [162]. This raises the possibility of tradeoffs that cortical circuits must make between feedforward sensitivity to inputs and controllability via top-down feedback. These tradeoffs could be investigated in models informed by the increasingly well resolved patterns of local connectivity between excitatory and inhbitory cell types [105, 163].

One widely believed computational role played by feedback within cortex is to enable the predictive coding of sensory inputs. In this normative model of hierarhical sensory processing, a generative model of the external world emerges as higher processing layers learn to predict the activity of lower processing layers in response to stimuli [73]. In the parlance of control theory, predictive coding is a model of hierarchical output regulation or stabilization around quiescence. It is a well known fact in control theory that output regulation, or what might alternatively be called distubrance rejection, requires that the controller contain an internal model that can effectively simulate the dynamics of the external disturbance system [7]. In fact, the presence of an internal model is both necessary and sufficient for robust regulation across both linear and nonlinear systems [164–166]. This connection suggests a novel mechanism for predictive coding through the feedback control of population dynamics. Internal models of the external world may be learned by cortex serving as a homeostatic regulator of the collective dynamics of lower levels of processing.

This framework could dispense with a key challenge faced by traditional predictive coding models, which posit that prediction and prediction error computations are performed by specialized single neurons [167]. To date, experimental evidence for these neurons has been scant. Predictive coding through population level feedback control would not necessarily require such neurons, as internal models could be encoded in a distributed fashion. Prediction errors would be subsumed by residual output firing, which would also be distributed across the population of output neurons within a processing layer. Such a reframing of the computation to occur through population dynamics mirrors the transition made in motor control, where models in which individual neurons tuned to particular kinematic variables have given way to computation through population dynamics [24]. Population level predictive coding also suggests a normative role for feedback controllable dynamics within sensory cortex, as one interpretation of these dynamics is that they require controllers of low complexity to regulate. In a hierarchical setting, structuring dynamics to be feedback controllable could help control the implementational cost of top down regulation in deeper layers. A simple computational study in which these ideas could be instantiated would involve training an RNN to serve as a regulator for the dynamics of another RNN perturbed by an external stimulus. During training, the emergence of an internal model could be tracked by periodically decoupling the controller RNN, stimulating its open loop dynamics and applying information bottleneck approaches [168] to identify subspaces predictive of the open loop dynamics of the controlled RNN.

Predictive coding is just one of many proposed normative theories of sensory processing. Others include, but are not limited to, the infomax principle [169], sparse coding [3], and predictive information coding [4]. While theories abound, a key challenge remains adjudi-

cating which of these normative theories provides a better characterization of experimentally recorded neural responses. Ideally, one could construct a hypothesis test that would determine, at a particular level of statistical significance, whether observed responses were optimal according to a particular normative theory, and further design stimuli that would maximally discriminate between competing hypotheses. The framework proposed in chapter 2 to place functional constraints on maximum entropy models provides a means of constructing null distributions for these hypothesis tests. Similarly to [170], to each normative theory, one could formulate an energy based model for observed neural responses that in addition to modelling low firing rate statistics, contained a term penalizing suboptimality according to each normative theory. For example, a model corresponding to the infomax theory would contain a term proportional to the mutual information between the stimulus distribution and the neural responses. Hypothesis testing could then be conducted via likelihood ratio comparison of models induced by competing normative theories. Such a framework could put head to head competing first principles explanations for neural activity.

It is the hope of the author that these future directions, among others, can be pursued in the ensuing years.

Bibliography

[1] Louis K Scheffer, C Shan Xu, Michal Januszewski, Zhiyuan Lu, Shin-ya Takemura, Kenneth J Hayworth, Gary B Huang, Kazunori Shinomiya, Jeremy Maitlin-Shepard, Stuart Berg, and others. A connectome and analysis of the adult Drosophila central brain. *Elife*, 9:e57443, 2020. Publisher: eLife Sciences Publications, Ltd.

[2] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

[3] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[4] Stephanie E. Palmer, Olivier Marre, Michael J. Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908, June 2015.

[5] Paul B Cook and John S McReynolds. Lateral inhibition in the inner retina is important for spatial tuning of ganglion cells. *Nature neuroscience*, 1(8):714–719, 1998.

[6] Norbert Wiener. Cybernetics. *Scientific American*, 179(5):14–19, 1948.

[7] Bruce A Francis and Walter Murray Wonham. The internal model principle of control theory. *Automatica*, 12(5):457–465, 1976.

[8] Andrea Avena-Koenigsberger, Joaquin Goni, Ricard Solé, and Olaf Sporns. Network morphospace. *Journal of the Royal Society Interface*, 12(103):20140881, 2015. Publisher: The Royal Society.

[9] Daniel M Wolpert and Zoubin Ghahramani. Computational principles of movement neuroscience. *Nature neuroscience*, 3(11):1212–1217, 2000.

[10] Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 2019.

[11] Emanuel Todorov and Michael I. Jordan. Optimal feedback control as a theory of motor coordination. *Nature neuroscience*, 5(11):1226–1235, November 2002.

[12] Karl Friston, Spyridon Samothrakis, and Read Montague. Active inference and agency: optimal control without cost functions. *Biological cybernetics*, 106:523–541, 2012.

[13] John Houde and Srikantan Nagarajan. Speech Production as State Feedback Control. *Frontiers in Human Neuroscience*, 5, 2011.

[14] Emanuel Todorov. Optimality principles in sensorimotor control. *Nature Neuroscience*, 7(9):907–915, September 2004.

[15] Tiago Marques, Julia Nguyen, Gabriela Fioreze, and Leopoldo Petreanu. The functional organization of cortical feedback inputs to primary visual cortex. *Nature Neuroscience*, 21(5):757–764, May 2018.

[16] Giovanni Pezzulo and Paul Cisek. Navigating the Affordance Landscape: Feedback Control as a Process Model of Behavior and Cognition. *Trends in Cognitive Sciences*, 20(6):414–424, 2016.

[17] Jon T. Sakata and Michael S. Brainard. Online Contributions of Auditory Feedback to Neural Activity in Avian Song Control Circuitry. *Journal of Neuroscience*, 28(44):11378–11390, 2008.

[18] Leopoldo Petreanu, Diego A. Gutnisky, Daniel Huber, Ning-long Xu, Dan H. O'Connor, Lin Tian, Loren Looger, and Karel Svoboda. Activity in motor–sensory projections reveals distributed coding in somatosensation. *Nature*, 489(7415):299–303, September 2012.

[19] Julie A. Harris, Stefan Mihalas, Karla E. Hirokawa, Jennifer D. Whitesell, Hannah Choi, Amy Bernard, Phillip Bohn, Shiella Caldejon, Linzy Casal, Andrew Cho, Aaron Feiner, David Feng, Nathalie Gaudreault, Charles R. Gerfen, Nile Graddis, Peter A. Groblewski, Alex M. Henry, Anh Ho, Robert Howard, Joseph E. Knox, Leonard Kuan, Xiuli Kuang, Jerome Lecoq, Phil Lesnar, Yaoyao Li, Jennifer Luviano, Stephen McConoughey, Marty T. Mortrud, Maitham Naeemi, Lydia Ng, Seung Wook Oh, Benjamin Ouellette, Elise Shen, Staci A. Sorensen, Wayne Wakeman, Quanxin Wang, Yun Wang, Ali Williford, John W. Phillips, Allan R. Jones, Christof Koch, and Hongkui Zeng. Hierarchical organization of cortical and thalamic connectivity. *Nature*, 575(7781):195–202, November 2019.

[20] J. Andrew Pruszynski and Stephen H. Scott. Optimal feedback control and the long-latency stretch response. *Experimental Brain Research*, 218(3):341–359, May 2012.

[21] Ta-Chu Kao, Mahdieh S. Sadabadi, and Guillaume Hennequin. Optimal anticipatory control as a theory of motor preparation: A thalamo-cortical circuit model. *Neuron*, 109(9):1567–1581.e12, 2021.

[22] Gordon M Shepherd. *The synaptic organization of the brain*. Oxford university press, 2003.

[23] Kenneth D Harris and Gordon M G Shepherd. The neocortical circuit: themes and variations. *Nature Neuroscience*, 18(2):170–181, February 2015.

[24] Saurabh Vyas, Matthew D. Golub, David Sussillo, and Krishna V. Shenoy. Computation Through Neural Population Dynamics. *Annual Review of Neuroscience*, 43(1):249–275, July 2020.

[25] Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.

[26] Samuel J Sober and Philip N Sabes. Flexible strategies for sensory integration during motor planning. *Nature neuroscience*, 8(4):490–497, 2005.

[27] Janis K Hesse and Doris Y Tsao. The macaque face patch system: a turtle's underbelly for the brain. *Nature Reviews Neuroscience*, 21(12):695–716, 2020.

[28] Gideon Rothschild, Elad Eban, and Loren M Frank. A cortical–hippocampal–cortical loop of information processing during memory consolidation. *Nature neuroscience*, 20(2):251–259, 2017.

[29] Fabio Pasqualetti, Sandro Zampieri, and Francesco Bullo. Controllability Metrics, Limitations and Algorithms for Complex Networks. *arXiv:1308.1201*, 2013.

[30] Jason Z Kim, Jonathan M Soffer, Ari E Kahn, Jean M Vettel, Fabio Pasqualetti, and Danielle S Bassett. Role of graph architecture in controlling dynamical networks with applications to neural systems. *Nature physics*, 14(1):91–98, 2018.

[31] Kenji Kashima. Noise Response Data Reveal Novel Controllability Gramian for Nonlinear Network Dynamics. *Scientific Reports*, 6(1):27300, June 2016.

[32] Stephen H. Scott. A Functional Taxonomy of Bottom-Up Sensory Feedback Processing for Motor Actions. *Trends in neurosciences*, 39(8):512–526, August 2016.

[33] John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

[34] Tyler H. Summers, Fabrizio L. Cortesi, and John Lygeros. On Submodularity and Controllability in Complex Dynamical Networks. *IEEE Transactions on Control of Network Systems*, 3(1):91–101, 2016.

[35] Thomas Kailath. *Linear systems*, volume 156. Prentice-Hall Englewood Cliffs, NJ, 1980.

[36] D Mitra. Wmatrix and the geometry of model equivalence and reduction. In *Proceedings of the Institution of Electrical Engineers*, volume 116, pages 1101–1106. IET, 1969. Issue: 6.

[37] E. Jonckheere and L. Silverman. A new set of invariants for linear systems–Application to reduced order compensator design. *IEEE Transactions on Automatic Control*, 28(10):953–964, 1983.

[38] L. Ljung and T. Kailath. Backwards Markovian models for second-order stochastic processes (Corresp.). *IEEE Transactions on Information Theory*, 22(4):488–491, July 1976.

[39] M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *Proceedings of the 2004 American Control Conference*, volume 4, pages 3273–3278 vol.4, 2004.

[40] Piergiorgio Strata, Robin Harvey, and others. Dale's principle. *Brain research bulletin*, 50(5):349–350, 1999.

[41] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.

[42] Lloyd N Trefethen and Mark Embree. *Spectra and pseudospectra*. Princeton university press, 2020.

[43] Madhura R. Joglekar, Jorge F. Mejias, Guangyu Robert Yang, and Xiao-Jing Wang. Inter-areal Balanced Amplification Enhances Signal Propagation in a Large-Scale Circuit Model of the Primate Cortex. *Neuron*, 98(1):222–234.e8, April 2018.

[44] Surya Ganguli, Dongsung Huh, and Haim Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105(48):18970–18975, 2008.

[45] Giulio Bondanelli and Srdjan Ostojic. Coding with transient trajectories in recurrent neural networks. *PLoS computational biology*, 16(2):e1007655, 2020. Publisher: Public Library of Science San Francisco, CA USA.

[46] Giacomo Baggio and Sandro Zampieri. Non-normality improves information transmission performance of network systems. *IEEE Transactions on Control of Network Systems*, 8(4):1846–1858, 2021. Publisher: IEEE.

[47] Andrew V Knyazev and Merico E Argentati. Principal angles between subspaces in an A-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040, 2002. Publisher: SIAM.

[48] Guillaume Hennequin, Tim P Vogels, and Wulfram Gerstner. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–1406, 2014.

[49] Kanaka Rajan and Larry F Abbott. Eigenvalue spectra of random matrices for neural networks. *Physical review letters*, 97(18):188104, 2006.

[50] Joseph E. O'Doherty, Mariana M. B. Cardoso, Joseph G. Makin, and Philip N. Sabes. Nonhuman primate reaching with multichannel sensorimotor cortex electrophysiology, Nov 2018.

[51] Omid G. Sani, Hamidreza Abbaspourazad, Yan T. Wong, Bijan Pesaran, and Maryam M. Shanechi. Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification. *Nature Neuroscience*, 24(1):140–149, January 2021.

[52] Wallace E Larimore. Canonical variate analysis in identification, filtering, and adaptive control. In *29th IEEE Conference on Decision and control*, pages 596–604. IEEE, 1990.

[53] Kristopher Jensen, Ta-Chu Kao, Jasmine Stone, and Guillaume Hennequin. Scalable Bayesian GPFA with automatic relevance determination and discrete noise models. *Advances in Neural Information Processing Systems*, 34:10613–10626, 2021.

[54] Gamaleldin F. Elsayed and John P. Cunningham. Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310–1318, September 2017.

[55] Ta-Chu Kao and Guillaume Hennequin. Neuroscience out of control: control-theoretic perspectives on neural circuit dynamics. *Current opinion in neurobiology*, 58:122–129, 2019. Publisher: Elsevier.

[56] Hari Teja Kalidindi, Kevin P. Cross, Timothy P. Lillicrap, Mohsen Omrani, Egidio Falotico, Philip N. Sabes, and Stephen H. Scott. Rotational dynamics in motor cortex are consistent with a feedback controller. *eLife*, 10, November 2021.

[57] Matthew T Kaufman, Mark M Churchland, Stephen I Ryu, and Krishna V Shenoy. Cortical activity in the null space: permitting preparation without movement. *Nature neuroscience*, 17(3):440–448, 2014.

[58] Mark M. Churchland, John P. Cunningham, Matthew T. Kaufman, Stephen I. Ryu, and Krishna V. Shenoy. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron*, 68(3):387–400, November 2010. Place: United States.

[59] Tomaso Poggio. A theory of how the brain might work. In *Cold Spring Harbor symposia on quantitative biology*, volume 55, pages 899–910. Cold Spring Harbor Laboratory Press, 1990.

[60] JS Seely, MT Kaufman, A Kohn, MA Smith, JA Movshon, NJ Priebe, SG Lisberger, SI Ryu, KV Shenoy, JP Cunningham, et al. Comparing visual and motor cortex: representational coding versus dynamical systems. In *Front. Comput. Neurosci. Conference Abstract: Bernstein Conference*, 2012.

[61] Gordon M Shepherd. *Foundations of the neuron doctrine*. Oxford University Press, 2015.

[62] Theodore H. Bullock, Michael V. L. Bennett, Daniel Johnston, Robert Josephson, Eve Marder, and R. Douglas Fields. The Neuron Doctrine, Redux. *Science*, 310(5749):791–793, 2005.

[63] Henry Markram, Eilif Muller, Srikanth Ramaswamy, Michael W. Reimann, Marwan Abdellah, Carlos Aguado Sanchez, Anastasia Ailamaki, Lidia Alonso-Nanclares, Nicolas Antille, Selim Arsever, Guy Antoine Atenekeng Kahou, Thomas K. Berger, Ahmet Bilgili, Nenad Buncic, Athanassia Chalimourda, Giuseppe Chindemi, Jean-Denis Courcol, Fabien Delalondre, Vincent Delattre, Shaul Druckmann, Raphael Dumusc, James Dynes, Stefan Eilemann, Eyal Gal, Michael Emiel Gevaert, Jean-Pierre Ghobril, Albert Gidon, Joe W. Graham, Anirudh Gupta, Valentin Haenel, Etay Hay, Thomas Heinis, Juan B. Hernando, Michael Hines, Lida Kanari, Daniel Keller, John Kenyon, Georges Khazen, Yihwa Kim, James G. King, Zoltan Kisvarday, Pramod Kumbhar, Sébastien Lasserre, Jean-Vincent Le Bé, Bruno R.C. Magalhães, Angel Merchán-Pérez, Julie Meystre, Benjamin Roy Morrice, Jeffrey Muller, Alberto Muñoz-Céspedes, Shruti Muralidhar, Keerthan Muthurasa, Daniel Nachbaur, Taylor H. Newton, Max Nolte, Aleksandr Ovcharenko, Juan Palacios, Luis Pastor, Rodrigo Perin, Rajnish Ranjan, Imad Riachi, José-Rodrigo Rodríguez, Juan Luis Riquelme, Christian Rössert, Konstantinos Sfyrakis, Ying Shi, Julian C. Shillcock, Gilad Silberberg, Ricardo Silva, Farhan Tauheed, Martin Telefont, Maria Toledo-Rodriguez, Thomas Tränkler, Werner Van Geit, Jafet Villafranca Díaz, Richard Walker, Yun Wang, Stefano M. Zaninetta, Javier DeFelipe, Sean L. Hill, Idan Segev, and Felix Schürmann. Reconstruction and Simulation of Neocortical Microcircuitry. *Cell*, 163(2):456–492, October 2015. Publisher: Elsevier.

[64] Sara Moberg and Naoya Takahashi. Neocortical layer 5 subclasses: From cellular properties to roles in behavior. *Frontiers in Synaptic Neuroscience*, 14, 2022.

[65] Manfred Oswald, Malinda Tantirigama, Ivo Sonntag, Stephanie Hughes, and Ruth Empson. Diversity of layer 5 projection neurons in the mouse motor cortex. *Frontiers in Cellular Neuroscience*, 7, 2013.

[66] Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Maneesh Sahani. Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *Journal of Neurophysiology*, 102(1):614–635, July 2009.

[67] Yuan Zhao and Il Memming Park. Variational latent gaussian process for recovering single-trial dynamics from population spike trains. *Neural computation*, 29(5):1293–1316, 2017.

[68] Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 15(10):805–815, 2018.

[69] David Sussillo and Omri Barak. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.

[70] Danielle S Bassett and Olaf Sporns. Network neuroscience. *Nature neuroscience*, 20(3):353–364, 2017. Publisher: Nature Publishing Group US New York.

[71] Trevor Ruiz, Sharmodeep Bhattacharyya, Mahesh Balasubramanian, and Kristofer Bouchard. Sparse and Low-bias Estimation of High Dimensional Vector Autoregressive Models. In *Learning for Dynamics and Control*, pages 55–64. PMLR, 2020.

[72] Pratik S Sachdeva, Jesse A Livezey, Maximilian E Dougherty, Bon-Mi Gu, Joshua D Berke, and Kristofer E Bouchard. Improved inference in coupling, encoding, and decoding models and its consequence for neuroscientific interpretation. *Journal of Neuroscience Methods*, 358:109195, 2021.

[73] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, January 1999.

[74] K. Fan. Maximum Properties and Inequalities for the Eigenvalues of Completely Continuous Operators. *Proceedings of the National Academy of Sciences of the United States of America*, 37(11):760–766, November 1951.

[75] D. Kleinman. On an iterative technique for Riccati equation computations. *IEEE Transactions on Automatic Control*, 13(1):114–115, 1968.

[76] Walter Rudin and others. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

[77] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[78] L. Barnett, E. Di Paolo, and S. Bullock. Spatially embedded random networks. *Phys. Rev. E*, 76:056115, Nov 2007.

[79] Brad K Hulse, Hannah Haberkern, Romain Franconville, Daniel Turner-Evans, Shinya Takemura, Tanya Wolff, Marcella Noorman, Marisa Dreher, Chuntao Dan, Ruchi Parekh, Ann M Hermundstad, Gerald M Rubin, and Vivek Jayaraman. A connectome of the *Drosophila* central complex reveals network motifs suitable for flexible navigation and context-dependent action selection. *eLife*, 10:e66039, October 2021.

[80] Xiaolong Jiang, Shan Shen, Cathryn R Cadwell, Philipp Berens, Fabian Sinz, Alexander S Ecker, Saumil Patel, and Andreas S Tolias. Principles of connectivity among morphologically defined cell types in adult neocortex. *Science*, 350(6264):aac9462, 2015.

[81] František Váša and Bratislav Mišić. Null models in network neuroscience. *Nature Reviews Neuroscience*, 23(8):493–504, 2022. Publisher: Nature Publishing Group UK London.

[82] Adam Haber, Adrian Wanner, Rainer W Friedrich, and Elad Schneidman. The structure and function of neural connectomes are shaped by a small number of design principles. *bioRxiv*, pages 2023–03, 2023.

[83] Mikail Rubinov. Constraints and spandrels of interareal connectomes. *Nature communications*, 7(1):13812, 2016.

[84] Ed Bullmore and Olaf Sporns. The economy of brain network organization. *Nature Reviews Neuroscience*, 13:336, April 2012.

[85] Bryan C Daniels, Yan-Jiun Chen, James P Sethna, Ryan N Gutenkunst, and Christopher R Myers. Sloppiness, robustness, and evolvability in systems biology. *Current opinion in biotechnology*, 19(4):389–395, 2008.

[86] Judith S Eisen and Eve Marder. Mechanisms underlying pattern generation in lobster stomatogastric ganglion as determined by selective inactivation of identified neurons. iii. synaptic connections of electrically coupled pyloric neurons. *Journal of neurophysiology*, 48(6):1392–1415, 1982.

[87] Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. *Advances in neural information processing systems*, 32, 2019.

[88] Giulio Cimini, Tiziano Squartini, Fabio Saracco, Diego Garlaschelli, Andrea Gabrielli, and Guido Caldarelli. The statistical physics of real-world networks. *Nature Reviews Physics*, 1(1):58–71, 2019. Publisher: Nature Publishing Group UK London.

[89] Sean R Bittner, Agostina Palmigiano, Alex T Piet, Chunyu A Duan, Carlos D Brody, Kenneth D Miller, and John Cunningham. Interrogating theoretical models of neural computation with emergent property inference. *Elife*, 10:e56265, 2021.

[90] Bradley Efron. *Exponential families in theory and practice*. Cambridge University Press, 2022.

[91] David R Hunter, Mark S Handcock, Carter T Butts, Steven M Goodreau, and Martina Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software*, 24(3):nihpa54860, 2008.

[92] Isaac Klickstein and Francesco Sorrentino. Control distance and energy scaling of complex networks. *IEEE Transactions on Network Science and Engineering*, 7(2):726–736, 2018. Publisher: IEEE.

[93] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

[94] Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris Maddison. Oops i took a gradient: Scalable sampling for discrete distributions. In *International Conference on Machine Learning*, pages 3831–3841. PMLR, 2021.

[95] Haoran Sun, Hanjun Dai, Bo Dai, Haomin Zhou, and Dale Schuurmans. Discrete Langevin Samplers via Wasserstein Gradient Flow. In *International Conference on Artificial Intelligence and Statistics*, pages 6290–6313. PMLR, 2023.

[96] Michael Schweinberger, Pavel N. Krivitsky, Carter T. Butts, and Jonathan R. Stewart. Exponential-Family Models of Random Graphs: Inference in Finite, Super and Infinite Population Scenarios. *Statistical Science*, 35(4):627–662, November 2020.

[97] Sourav Chatterjee and Persi Diaconis. Estimating and understanding exponential random graph models. 2013.

[98] Vahid Rostami, PierGianLuca Porta Mana, Sonja Grün, and Moritz Helias. Bistability, non-ergodicity, and inhibition in pairwise maximum-entropy models. *PLoS computational biology*, 13(10):e1005762, 2017. Publisher: Public Library of Science San Francisco, CA USA.

[99] Vishesh Karwa, Sonja Petrović, and Denis Bajić. DERGMs: Degeneracy-restricted exponential family random graph models. *Network Science*, 10(1):82–110, 2022. Publisher: Cambridge University Press.

[100] Murty SS Challa and JH Hetherington. Gaussian ensemble as an interpolating ensemble. *Physical review letters*, 60(2):77, 1988. Publisher: APS.

[101] Arsham Ghavasieh and Manlio De Domenico. Generalized network density matrices for analysis of multiscale functional diversity, 2022.

[102] Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, et al. Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139, 2023.

[103] Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570. PMLR, 2018.

[104] Qiang Liu and Jason Lee. Black-box importance sampling. In *Artificial Intelligence and Statistics*, pages 952–961. PMLR, 2017.

[105] MICrONS Consortium, J Alexander Bae, Mahaly Baptiste, Caitlyn A Bishop, Agnes L Bodor, Derrick Brittain, JoAnn Buchanan, Daniel J Bumbarger, Manuel A Castro, Brendan Celii, et al. Functional connectomics spanning multiple areas of mouse visual cortex. *BioRxiv*, pages 2021–07, 2021.

[106] Xi-Nian Zuo, Ross Ehmke, Maarten Mennes, Davide Imperati, F Xavier Castellanos, Olaf Sporns, and Michael P Milham. Network centrality in the human functional connectome. *Cerebral cortex*, 22(8):1862–1875, 2012.

[107] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, May 2015.

[108] Patrik Waldmann, Gábor Mészáros, Birgit Gredler, Christian Fürst, and Johann Sölkner. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*, 4:270, 2013.

[109] ROBERT Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, February 1997. Publisher: John Wiley & Sons, Ltd.

[110] Ewout W. Steyerberg and Yvonne Vergouwe. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal*, 35(29):1925–1931, August 2014.

[111] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse Representation for Computer Vision and Pattern Recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[112] Junxian Zhu, Canhong Wen, Jin Zhu, Heping Zhang, and Xueqin Wang. A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences*, 117(52):33117, December 2020.

[113] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *Ann. Statist.*, 44(2):813–852, April 2016. Publisher: The Institute of Mathematical Statistics.

[114] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[115] Hui Zou and Trevor Hastie. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[116] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, April 2010. Publisher: The Institute of Mathematical Statistics.

[117] Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, December 2001. Publisher: Taylor & Francis.

[118] Kristofer Bouchard, Alejandro Bujan, Fred Roosta, Shashanka Ubaru, Mr Prabhat, Antoine Snijders, Jian-Hua Mao, Edward Chang, Michael W Mahoney, and Sharmodeep Bhattacharya. Union of intersections (uoi) for interpretable data driven discovery and prediction. *Advances in Neural Information Processing Systems*, 30, 2017.

[119] Peng Zhao and Bin Yu. On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006.

[120] Yuan Li, Benjamin Mark, Garvesh Raskutti, and Rebecca Willett. Graph-based regularization for regression problems with highly-correlated designs. *arXiv:1803.07658*, 2018.

[121] Mário A. T. Figueiredo and Robert D. Nowak. Ordered Weighted L1 Regularized Regression with Strongly Correlated Covariates: Theoretical Aspects. In *AISTATS*, 2016.

[122] Peter Bühlmann, Philipp Rütimann, Sara van de Geer, and Cun-Hui Zhang. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858, November 2013.

[123] Malgorzata Bogdan, Ewout van den Berg, Weijie Su, and Emmanuel Candes. Statistical estimation and testing via the sorted L1 norm. *arXiv:1310.1969*, 2013.

[124] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108, 2005.

[125] Daniela M Witten, Ali Shojaie, and Fan Zhang. The Cluster Elastic Net for High-Dimensional Regression With Unknown Variable Grouping. *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences*, 56(1):112–122, February 2014.

[126] Yongdai Kim, Sunghoon Kwon, and Hosik Choi. Consistent Model Selection Criteria on High Dimensions. *Journal of Machine Learning Research*, 13(36):1037–1057, 2012.

[127] Jun Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7(2):221–242, 1997. Publisher: Institute of Statistical Science, Academia Sinica.

[128] Hansheng Wang, Runze Li, and Chih-Ling Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, August 2007.

[129] Mark H Hansen and Bin Yu. Model Selection and the Principle of Minimum Description Length. *Journal of the American Statistical Association*, 96(454):746–774, June 2001.

[130] EdwardI. George and Dean P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, December 2000.

[131] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914, November 2009.

[132] M. J. Wainwright. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using l_1 -Constrained Quadratic Programming (Lasso). *IEEE Transactions on Infor mation Theory*, 55(5):2183–2202, 2009.

[133] M. J. Wainwright. Information-Theoretic Limits on Sparsity Recovery in the High-Dimensional and Noisy Setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, December 2009.

[134] George Kamal Atia and Venkatesh Saligrama. Boolean Compressed Sensing and Noisy Group Testing. *arXiv:0907.1061*, 2009.

[135] S. Aeron, V. Saligrama, and M. Zhao. Information Theoretic Bounds for Compressed Sensing. *IEEE Transactions on Information Theory*, 56(10):5111–5130, October 2010.

[136] Cem Aksoylar and Venkatesh Saligrama. Information-Theoretic Characterization of Sparse Recovery. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 38–46. PMLR, April 2014.

[137] J. Scarlett, J. S. Evans, and S. Dey. Compressed Sensing With Prior Information: Information-Theoretic Limits and Practical Decoders. *IEEE Transactions on Signal Processing*, 61(2):427–439, January 2013.

[138] J. Scarlett and V. Cevher. Limits on Support Recovery With Probabilistic Models: An Information-Theoretic Framework. *IEEE Transactions on Information Theory*, 63(1):593–620, January 2017.

[139] C. Aksoylar, G. K. Atia, and V. Saligrama. Sparse Signal Processing With Linear and Nonlinear Observations: A Unified Shannon-Theoretic Approach. *IEEE Transactions on Information Theory*, 63(2):749–776, February 2017.

[140] K. Rahnama Rad. Nearly Sharp Sufficient Conditions on Exact Sparsity Pattern Recovery. *IEEE Transactions on Information Theory*, 57(7):4672–4679, July 2011.

[141] Carsten F. Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell, and Sven Lautenbach. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, January 2013. Publisher: John Wiley & Sons, Ltd.

[142] Kristina P Vatcheva, MinJae Lee, Joseph B McCormick, and Mohammad H Rahbar. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale, Calif.)*, 6(2):227, April 2016. Edition: 2016/03/07.

[143] Anneli Schöniger, Thomas Wöhling, Luis Samaniego, and Wolfgang Nowak. Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water Resources Research*, 50(12):9484–9513, December 2014. Publisher: John Wiley & Sons, Ltd.

[144] Mark J. Brewer, Adam Butler, and Susan L. Cooksley. The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, 7(6):679–692, June 2016. Publisher: John Wiley & Sons, Ltd.

[145] John J Dziak, Donna L Coffman, Stephanie T Lanza, Runze Li, and Lars S Jermiin. Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, 21(2):553–565, March 2020.

[146] Dimitris Bertsimas, Jean Pauphilet, and Bart Van Parys. Sparse Regression: Scalable Algorithms and Empirical Performance. *Statist. Sci.*, 35(4):555–578, November 2020. Publisher: The Institute of Mathematical Statistics.

[147] Richard S. Varga. Geršgorin-Type Eigenvalue Inclusion Theorems. In Richard S. Varga, editor, *Geršgorin and His Circles*, pages 35–72. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[148] Chenlei Leng, Yi Lin, and Grace Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284, 2006. Publisher: Institute of Statistical Science, Academia Sinica.

[149] Po-Ling Loh and Martin J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *Ann. Statist.*, 45(6):2455–2482, December 2017.

[150] Francis R. Bach. Bolasso: Model Consistent Lasso Estimation through the Bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 33–40, New York, NY, USA, 2008. Association for Computing Machinery. event-place: Helsinki, Finland.

[151] Yuhong Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, December 2005.

[152] Po-Ling Loh and Martin J. Wainwright. Regularized M-Estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima. *J. Mach. Learn. Res.*, 16(1):559–616, January 2015. Publisher: JMLR.org.

[153] Tuo Zhao, Han Liu, and Tong Zhang. Pathwise coordinate optimization for sparse learning: Algorithm and theory. *Ann. Statist.*, 46(1):180–218, February 2018. Publisher: The Institute of Mathematical Statistics.

[154] Adam Kohn, Anna I Jasper, João D Semedo, Evren Gokcen, Christian K Machens, and M Yu Byron. Principles of corticocortical communication: proposed schemes and design considerations. *Trends in Neurosciences*, 43(9):725–737, 2020.

[155] Andre M Bastos, Julien Vezoli, and Pascal Fries. Communication through coherence with inter-areal delays. *Current opinion in neurobiology*, 31:173–180, 2015.

[156] João D Semedo, Amin Zandvakili, Christian K Machens, M Yu Byron, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.

[157] Haim H Permuter, Young-Han Kim, and Tsachy Weissman. Interpretations of directed information in portfolio theory, data compression, and hypothesis testing. *IEEE Transactions on Information Theory*, 57(6):3248–3259, 2011.

[158] Pierre-Olivier Amblard and Olivier JJ Michel. The relation between granger causality and directed information theory: A review. *Entropy*, 15(1):113–143, 2012.

[159] P Caines and C Chan. Feedback between stationary stochastic processes. *IEEE Transactions on Automatic Control*, 20(4):498–508, 1975.

[160] Sadra Sadeh and Claudia Clopath. Inhibitory stabilization and cortical computation. *Nature Reviews Neuroscience*, 22(1):21–37, 2021.

[161] John C Doyle, Bruce A Francis, and Allen R Tannenbaum. *Feedback control theory*. Courier Corporation, 2013.

[162] Jesse B Hoagg and Dennis S Bernstein. Nonminimum-phase zeros-much to do about nothing-classical control-revisited part ii. *IEEE Control Systems Magazine*, 27(3):45–57, 2007.

[163] Elodie Fino, Adam M Packer, and Rafael Yuste. The logic of inhibitory connectivity in the neocortex. *The Neuroscientist*, 19(3):228–237, 2013.

[164] Bruce A Francis and William M Wonham. The internal model principle for linear multivariable regulators. *Applied mathematics and optimization*, 2(2):170–194, 1975.

[165] Walter Murray Wonham. Towards an abstract internal model principle. *IEEE Transactions on Systems, Man, and Cybernetics*, (11):735–740, 1976.

[166] Christopher I Byrnes, Francesco Delli Priscoli, Alberto Isidori, Christopher I Byrnes, Francesco Delli Priscoli, and Alberto Isidori. *Output regulation of nonlinear systems*. Springer, 1997.

[167] Fabian A Mikulasch, Lucas Rudelt, Michael Wibral, and Viola Priesemann. Where is the error? hierarchical predictive coding through dendritic error computation. *Trends in Neurosciences*, 46(1):45–59, 2023.

[168] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[169] Jim W Kay and WA Phillips. Coherent infomax as a computational goal for neural systems. *Bulletin of mathematical biology*, 73:344–372, 2011.

[170] Wiktor Młynarski, Michal Hledík, Thomas R Sokolowski, and Gašper Tkačik. Statistical analysis and optimality of neural systems. *Neuron*, 109(7):1227–1241, 2021.