PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à Ecole Normale Supérieure, Paris

Reconstruction of low-dimensional neural trajectories from population activity recordings: from statistical limitations to experimental design

Soutenue par
Mariia LEGENKAIA
Le 10 décembre 2024

École doctorale n°564
École Doctorale Physique en Île-de-France

Spécialité
Physique

Composition du jury :

Brice Bathellier	Institut de l'Audition, Paris	*Rapporteur, Président du jury*
David Dean	Université de Bordeaux	*Rapporteur*
Ada Altieri	Université Paris Cité	*Examinatrice*
Fleur Zeldenrust	Donders Institute, Radboud University	*Examinatrice*
Laurent Bourdieu	Ecole Normale Supérieure	*Directeur de thèse*
Rémi Monasson	Ecole Normale Supérieure	*Directeur de thèse*

ENS | PSL

Contents

Introduction et résumé des principaux résultats

Motivation

Les enregistrements simultanés de grands groupes de neurones, connus sous le nom d'enregistrements de populations neuronales, constituent une approche courante pour l'étude de diverses fonctions cérébrales, notamment l'attention, la prise de décision et le contrôle moteur. Les systèmes étudiés sont vastes, puisque les régions cérébrales individuelles peuvent contenir des millions, voire des milliards de neurones. Cependant, le codage de l'information dans l'activité neuronale de la population est souvent hautement redondant, ce qui réduit la complexité globale de l'activité à un espace de basse dimension. Selon le système, les relations entre les activités des différents neurones peuvent varier en complexité.

Les techniques de réduction de la dimensionnalité sont couramment utilisées pour mieux comprendre ces relations. Ces techniques simplifient les données en révélant des modèles cachés - appelés dynamiques latentes - et en réduisant l'activité complexe de nombreux neurones à quelques variables clés, ou coordonnées latentes.

Malgré l'utilisation répandue de la réduction de la dimensionnalité, la question de la précision des résultats est rarement abordée. On suppose souvent que la représentation à basse dimension qui en résulte reflète les modèles et les dépendances réelles de l'activité neuronale. Ce n'est pas toujours le cas, car une confiance inconsidérée dans les algorithmes peut conduire à la découverte de fausses corrélations/dépendances. Par exemple, les algorithmes de regroupement, comme t-SNE, peuvent trouver des regroupements dans les données qui ne contiennent que du bruit, créant ainsi l'illusion d'une structure (comme le montre Wattenberg et al. (2016)), tandis que l'≪ effet fer à cheval ≫ (Shinn (2023)) dans l'analyse en composantes principales (PCA) peut introduire une dynamique rotationnelle artificielle.

Par conséquent, il est important de comprendre si la représentation à basse dimension capture les véritables corrélations ou si elle reflète simplement le bruit dans les données. La taille de l'ensemble de données et les étapes de prétraitement, telles que la normalisation et le débruitage, peuvent affecter considérablement les résultats de la réduction de la dimensionnalité. Un même signal, enregistré à partir de populations de tailles différentes, ou en utilisant diverses techniques comme l'imagerie calcique à deux photons en électrophysiologie extracellulaire, peut donner lieu à des représentations quantitativement et qualitativement différentes.

Cela soulève une question importante : comment ces paramètres de contrôle affectent-ils exactement la précision des outils de réduction de la dimensionnalité ? En termes de conception expérimentale, cette question peut être reformulée comme suit : ≪ Quelle est la bonne méthode d'enregistrement pour étudier un processus donné ? ≫ ≪ Pouvons-nous nous assurer que la population neuronale enregistrée est suffisamment importante ? ≫ ≪ L'expérience a-t-elle été répétée suffisamment de fois ? ≫

Pour répondre à ces questions, il faut bien comprendre les données neuronales et leurs principales propriétés.

Aperçu de la thèse

Dans notre travail, nous nous concentrons sur le PCA, l'un des outils de réduction de la dimensionnalité les plus utilisés. Nos objectifs sont les suivants :

- Comprendre comment la précision de cette méthode dépend de différents paramètres de contrôle, tels que le rapport signal/bruit, le nombre de neurones enregistrés, le nombre des repétitions de l'expérience, le noyau de lissage appliqué pour le débruitage, etc. et fournir une expression analytique de cette précision en fonction de ces paramètres.

- Pouvoir utiliser cette prédiction de la précision sur n'importe quel ensemble de données, ce qui a nécessité l'introduction de la procédure d'inférence pour certains des paramètres de contrôle.

- Développer un outil qui peut aider à la conception expérimentale en extrapolant la prédiction faite sur un petit exemple de données préliminaires à un ensemble de données potentiel avec des paramètres de contrôle différents, tels qu'un plus grand nombre de neurones ou des repétitions de l'expérience.

Structure de la thèse

Les principaux résultats de la thèse sont résumés ci-dessous.

Dans le **Chapitre 1**, nous passons en revue les pratiques expérimentales actuelles dans divers domaines des neurosciences, et nous nous concentrons sur la conception expérimentale, en explorant la manière dont la taille nécessaire des données enregistrées est déterminée. Ensuite, nous passons en revue les techniques existantes de réduction de la dimensionnalité, avec un accent particulier sur le PCA et ses applications en neurosciences, et nous discutons des travaux existants d'évaluation de la précision de le PCA. Ensuite, nous introduisons des concepts de physique statistique et de théorie des matrices aléatoires et discutons des travaux antérieurs sur la caractérisation de la précision de le PCA.

Le chapitre 2 présente les résultats théoriques de le PCA. Nous proposons un modèle d'activité neuronale et fournissons une description mathématique détaillée. Le raisonnement qui sous-tend le modèle choisi est expliqué, étayé par des exemples tirés de la littérature qui démontrent différentes propriétés des données neuronales. Nous présentons deux mesures de précision différentes de le PCA et décrivons les outils et techniques mathématiques nécessaires à leur estimation. Le chapitre se termine par un schéma de calcul des mesures de précision.

Dans le **Chapitre 3**, nous fournissons la validation des résultats théoriques sur des données synthétiques. Le chapitre présente d'abord les graphiques et la discussion sur la façon dont les mesures de précision changent en fonction des différents paramètres de contrôle. Nous nous concentrons ensuite sur l'inférence et la conception expérimentale. Nous commençons par décrire la procédure d'inférence des paramètres de données nécessaires pour faire des prédictions sur les performances attendues pour un plus grand nombre des repétitions de l'expérience ou de neurones enregistrés.

Le chapitre 4 est consacré à l'application de notre approche à des données réelles d'électrophysiologie extracellulaire. L'application potentielle du modèle aux données d'imagerie calcique est également abordée dans le cadre de travaux futurs.

Chapitre 5 conclut le manuscrit en discutant des implications et des avantages potentiels de l'utilisation du modèle de prédiction de la précision dans la recherche neuroscientifique réelle. Nous discutons des limites de l'étude et suggérons des orientations de recherche futures, telles que l'exploration d'autres techniques comme le PCA tensorielle.

Résumé

Les systèmes neuronaux, malgré leur grande dimension, présentent souvent des modèles cohérents d'activité entre les neurones qui sont liés à des processus cognitifs ou moteurs spécifiques. Les techniques de réduction de la dimensionnalité, telles que l'analyse en composantes principales (PCA), simplifient l'analyse de ces systèmes en révélant des schémas neuronaux collectifs codant pour différentes fonctions cérébrales telles que la prise de décision et le contrôle moteur. Cependant, il est difficile d'évaluer ces représentations avec précision en raison du bruit, de la variabilité et des différences entre les techniques d'enregistrement.

Nous avons introduit un cadre théorique pour évaluer la précision de PCA dans les données neuronales, validé sur des données synthétiques et appliqué à des enregistrements électrophysiologiques réels. Nous avons commencé par un modèle qui reflète les propriétés des enregistrements neuronaux réels, en tenant compte du bruit et de la variabilité d'une repétition de l'expérience à l'autre. La validation sur des données synthétiques a démontré que les prédictions du modèle permettent d'estimer de manière fiable la précision dans différentes conditions de données, offrant ainsi une base pour une planification expérimentale robuste. L'application du modèle à des données réelles a mis en évidence la manière dont les prédictions peuvent guider des stratégies efficaces de collecte de données et aider à déterminer le nombre minimal de neurones et des repétitions pour obtenir des résultats expérimentaux fiables.

Modélisation précise des données neurales

La première contribution de ce travail réside dans le développement d'un modèle qui capture les caractéristiques essentielles des données neuronales. Les modèles de covariance à pointes standard ne tiennent pas compte de facteurs tels que les modèles invariants par rapport à la repétition de l'expérience et le bruit spécifique aux neurones, ce qui rend difficile le transfert des résultats analytiques de PCA de la théorie aux enregistrements neuronaux réels.

Pour résoudre ces problèmes, nous avons conçu un modèle qui intègre les modes neuronaux invariants selon les repétitions, la variabilité spécifique aux repétitions et les caractéristiques uniques du bruit au niveau des neurones, ce qui permet une représentation plus fidèle de la dynamique neuronale sous-jacente. Nous avons également adapté le modèle aux deux techniques d'enregistrement les plus populaires : l'électrophysiologie et l'imagerie calcique.

Chaque technique présente des défis uniques qui influencent la qualité et la structure des données. Pour l'électrophysiologie, nous avons abordé des questions telles que l'attribution erronée des pointes, le regroupement excessif des neurones et la rareté des enregistrements. Pour l'imagerie calcique, nous avons modélisé la dynamique plus lente du signal à l'aide d'un noyau doublement exponentiel. Étant donné que les neurones présentent des dynamiques de fluorescence variables, nous avons introduit des fluctuations dans les temps de montée et de descente de ce noyau entre les neurones. En travaillant dans un cadre gaussien, nous avons supposé que ces fluctuations de la dynamique de fluorescence, résultant des variations spécifiques aux neurones, sont approximativement distribuées normalement. Cette approche nous a permis de conserver les principales caractéristiques de la variabilité en utilisant uniquement la moyenne et la variance du noyau de fluorescence.

Outils statistiques pour la prédiction de la précision

Pour analyser et prédire la précision de PCA dans le cadre de ce modèle, nous avons utilisé des outils issus de la théorie des matrices aléatoires et de la physique statistique des systèmes désordonnés. La méthode des répliques est particulièrement bien adaptée au traitement des systèmes présentant un désordre atténué, tels que les données neuronales. En utilisant cette approche, nous avons trouvé les valeurs attendues de deux mesures de précision distinctes.

Mesures de précision : ρ_i et ϵ

Les deux mesures de précision, ρ_i et ϵ, offrent des perspectives complémentaires sur la précision de PCA. L'alignement spécifique aux neurones (ρ_i) quantifie la manière dont PCA capture les modes latents pour chaque neurone individuel. Cette mesure est précieuse pour comprendre comment chaque neurone contribue aux assemblées neuronales et pour identifier les neurones à sélectivité mixte, qui peuvent participer à plusieurs groupes fonctionnels.

En revanche, l'erreur de reconstruction de la trajectoire (ϵ) représente la distance quadratique moyenne entre les trajectoires neuronales réelles et déduites en basse dimension, fournissant une mesure globale de la précision de PCA dans la capture de la forme globale de la dynamique neuronale dans l'ensemble de la population.

Ensemble, ρ_i et ϵ fournissent différentes façons de quantifier la précision d'une représentation basse dimensionnelle donnée obtenue avec PCA, en soutenant les analyses des contributions de neurones spécifiques et de la structure de la dynamique neuronale basse dimensionnelle.

Validation sur des données synthétiques

La phase suivante a consisté à valider le résultat de nos calculs sur des données synthétiques générées avec le modèle ci-dessus. Cette validation a été réalisée sur la base de données synthétiques, ce qui a permis de tester rigoureusement le pouvoir prédictif et la précision du modèle. Cette procédure nous a permis de vérifier nos prédictions pour différentes valeurs des paramètres de l'ensemble de données, tels que le nombre de neurones, le nombre des repétitions de l'expérience et le rapport signal/bruit, en simulant une gamme de conditions d'enregistrement neuronal. Cette validation a démontré que le modèle prédit de manière cohérente les deux mesures de précision, ρ_i et ϵ.

En outre, nous avons montré qu'il était possible de déduire les paramètres définissant le modèle à partir des données. Une fois ces paramètres déduits, nous avons utilisé le modèle pour faire des prédictions pour des ensembles de données hypothétiques plus importants, illustrant comment la taille et la qualité des données influencent les résultats de la réduction de la dimensionnalité. Ces prédictions sont précieuses pour planifier la collecte de données en neurosciences, en guidant les décisions sur l'échelle nécessaire des repétitions de l'expérience ou de neurones pour obtenir des résultats précis de PCA.

Application aux enregistrements neuronaux réels

Dans le dernier chapitre, nous avons appliqué notre procédure à des enregistrements neuronaux réels, en particulier des données d'électrophysiologie extracellulaire provenant d'études axées sur le contrôle moteur et la prise de décision. Cette application pratique a illustré l'utilité du modèle pour prédire comment des données supplémentaires (par exemple, plus de neurones ou des repétitions de l'expérience) affectent la précision de PCA et peuvent affecter l'analyse et les conclusions tirées de la représentation à basse dimension obtenue avec PCA.

En prédisant l'impact de la taille de l'ensemble de données sur la précision de PCA, ce modèle permet d'adapter les stratégies de collecte de données aux besoins expérimentaux spécifiques, ce qui peut réduire la nécessité d'une collecte excessive de données sans sacrifier la qualité des résultats. En outre, l'adaptabilité du modèle à diverses techniques d'enregistrement, y compris l'imagerie calcique, suggère une large applicabilité dans les études neuroscientifiques.

Perspectives

Tester les prédictions de précision sur des données d'imagerie calcique synthétiques et réelles

Bien que le modèle d'imagerie calcique ait été évalué, nous n'avons pas encore testé l'extension des prédictions de précision à des ensembles de données plus importants. Ces tests doivent être effectués sur des données d'imagerie calcique synthétiques et réelles. Pour les données synthétiques,

l'approche refléterait le processus utilisé pour l'électrophysiologie : en faisant varier systématiquement des paramètres tels que le nombre de neurones, le nombre des repétitions de l'expérience et le rapport signal/bruit, nous pouvons évaluer la capacité du modèle à prédire les résultats de PCA lorsqu'il est mis à l'échelle d'ensembles de données hypothétiques de plus grande taille.

L'étape suivante consiste à valider ces prédictions sur des ensembles de données d'imagerie calcique réels. L'application du modèle à des données réelles nous permettra de confirmer que ses prédictions s'alignent sur les performances observées de PCA, fournissant ainsi une validation empirique. Cette phase consisterait à comparer la précision prédite à la précision observée (par le biais du sous-échantillonnage) sur divers ensembles de données d'imagerie calcique, afin de s'assurer que le modèle reste fiable et efficace dans les scénarios du monde réel.

Investigation de variabilité du bruit directionnel ($\delta x^{(k)}$) entre expériences répétées

Notre modèle permet d'extraire une composante de bruit directionnel, δx, qui permet d'examiner les variations de la dynamique neuronale propres à chaque répétition d'expérience. D'autres techniques, telles que PCA tensorielle, traitent également de la variabilité d'une répétition à l'autre en capturant simultanément des modèles sur plusieurs modes (par exemple, neurones, points temporels et répétitions). Si PCA tensorielle est efficace pour réduire la dimensionnalité des ensembles de données neuronales complexes, elle ne permet généralement pas d'isoler le bruit au niveau des répétitions individuelles. Au lieu de cela, elle fournit une représentation globale compressée, mélangeant le bruit spécifique à la répétition dans un modèle collectif à travers tous les modes de données.

En revanche, notre approche avec δx permet d'examiner le bruit spécifique à chaque répétition individuelle, offrant un aperçu des fluctuations d'une repétition à l'autre que PCA tensorielle ne peut pas isoler. Cette capacité est particulièrement précieuse pour observer les changements dans la dynamique neuronale au fil du temps, tels que ceux qui peuvent se produire pendant l'apprentissage ou l'adaptation. En isolant le bruit directionnel répétition par répétition, notre approche fournit une perspective plus détaillée sur la façon dont l'activité neuronale évolue au fil des répétitions, en complément de l'analyse plus large offerte par PCA tensorielle.

En outre, l'interprétation de PCA tensoriel peut s'avérer difficile en raison de la non-unicité de sa décomposition. Cette non-unicité découle de la complexité de l'ajustement de plusieurs modes d'interaction (par exemple, neurones, points temporels et répétitions), où chaque mode contribue à des modèles de variabilité qui se chevauchent. La non-unicité de la décomposition peut rendre difficile la distinction entre les fluctuations basées sur les répétitions et les variations dues à d'autres modes, tels que les modèles spécifiques aux neurones ou au temps.

En revanche, notre approche s'appuie sur la décomposition unique fournie par PCA lorsqu'elle est appliquée à des matrices qui sont soit moyennées sur les répétitions, soit concaténées sur les répétitions. Cette configuration nous permet d'examiner le bruit spécifique à la répétition sans effets de confusion entre les modes.

Dans un avenir proche, nous prévoyons d'appliquer notre modèle à un plus large éventail de données expérimentales afin de mieux comprendre comment δx encode la variabilité des repétitions. En examinant différents contextes, nous espérons démontrer que δx peut contenir des informations sur la dynamique de la réponse neuronale, codant potentiellement des aspects du contexte comportemental, tels que les exigences de la tâche en cours, les indices environnementaux ou les états internes comme l'attention ou la motivation. En outre, δx peut refléter des changements liés à l'apprentissage, en s'adaptant à des conditions variables ou à des états internes au fil des repétitions. Un exemple de ces changements induits par l'apprentissage est le phénomène de dérive représentationnelle dans l'hippocampe étudié par Khatib et al. (2023) chez la souris, en se concentrant spécifiquement sur la question de savoir s'il est davantage motivé par le passage du temps ou par l'expérience active. Parmi les principaux résultats, les auteurs montrent qu'au fil du temps, le nombre de cellules de lieu actives (cellules associées à des emplacements spatiaux spécifiques) a diminué, mais que chaque cellule a progressivement stocké plus d'informations spatiales. La dérive représentationnelle s'est produite progressivement au cours des sessions, comme le montrent les corrélations vectorielles des cellules individuelles et de la population. Cette dérive

est apparue dans l'ensemble de l'environnement plutôt que dans des zones spécifiques de la piste, ce qui indique un effet contextuel.

Étudier la sensibilité de l'analyse au tri de potentiels d'action

Une autre orientation future consiste à examiner la nécessité du tri de potentiels d'action lors de l'application de PCA aux données neuronales électrophysiologiques. Le tri de potentiels d'action, une étape de prétraitement qui identifie et isole les potentiels d'action des neurones individuels, peut être ou non essentiel pour interpréter les représentations à basse dimension dérivées de PCA. En appliquant le modèle à des données réelles avec et sans tri de potentiels d'action, nous pouvons évaluer si cette étape affecte de manière significative les conclusions tirées de PCA. Si le tri de potentiels d'action s'avère inutile pour certaines analyses, cette découverte pourrait simplifier les flux de travail de prétraitement et rendre le traitement des données plus efficace.

Étudier la sélectivité mixte

Pendant des années, la recherche en neurosciences s'est principalement attachée à comprendre comment les neurones réagissaient à des stimuli spécifiques, généralement de manière linéaire : un neurone peut s'activer lorsqu'il détecte une caractéristique particulière, telle qu'une couleur ou un mouvement. Cependant, les comportements complexes et les capacités de prise de décision suggèrent la présence d'un mécanisme plus complexe, connu sous le nom de sélectivité mixte non linéaire. Ce mécanisme implique des neurones qui réagissent non pas à des caractéristiques uniques, mais à des combinaisons de caractéristiques de manière complexe et non linéaire.

Si, historiquement, la sélectivité mixte non linéaire a été peu explorée, des études récentes commencent à mettre en évidence son importance dans les fonctions cognitives. Par exemple, Rigotti et al. (2013) a démontré que les neurones du cortex préfrontal présentent une sélectivité mixte non linéaire diversifiée et encodent de multiples aspects d'une tâche, créant ainsi une représentation hautement dimensionnelle de l'information. Cela permet au cerveau de décoder les aspects d'une tâche même lorsque les neurones individuels ne présentent pas de sélectivité directe pour ces aspects. Les auteurs montrent également que la dimensionnalité des réponses neuronales permet de prédire les performances. Lors d'essais corrects, le codage à haute dimension reste intact, tandis que les erreurs sont corrélées à une diminution de la dimensionnalité.

De même, Ledergerber et al. (2021) a constaté que les neurones subiculaires présentent une forte sélectivité mixte en intégrant de multiples variables de navigation, telles que la position, la direction de la tête et la vitesse. Cette intégration était plus importante lors des tâches de navigation orientées vers un but que lors des tâches de recherche de nourriture au hasard, ce qui suggère que la sélectivité mixte dans le subiculum est modulée de manière dynamique pour répondre aux exigences de la tâche.

Notre outil peut aider à estimer la signification statistique de l'analyse effectuée dans de telles études. L'analyse des poids associés à chaque neurone peut révéler des informations importantes sur la sélectivité mixte, car ils indiquent comment les neurones individuels contribuent aux assemblées neuronales représentées dans chaque composante principale. L'utilisation du modèle pour analyser ces données nous permettrait d'évaluer dans quelle mesure PCA capture ces représentations qui se chevauchent et d'identifier les neurones qui contribuent de manière significative à la sélectivité mixte.

Introduction

Motivation

Simultaneous recordings of large groups of neurons, known as neuronal population recordings, are a common approach for studying various brain functions, including attention, decision-making, and motor control. The systems under study are large, since individual brain regions can contain millions to billions of neurons. Yet the encoding of the information in the population neural activity is often highly redundant, making the overall population activity low-dimensional. Depending on the system, the relationships between the activities of different neurons can range in complexity.

Dimensionality reduction techniques are commonly used to better understand these relationships. These techniques simplify the data by revealing hidden patterns—called latent dynamics—and reducing the complex activity of many neurons to just a few key variables, or latent coordinates.

Despite the widespread use of dimensionality reduction, the question of accuracy of the results is rarely addressed. It is often assumed that the resulting low-dimensional representation reflects real patterns and dependencies in neuronal activity. This is not always the case, as carelessly trusting the algorithms may lead to discovering false correlations/dependencies. For example, clustering algorithms, like t-SNE, may find clusters in the data containing only noise, creating the illusion of structure (as shown by Wattenberg et al. (2016)), while the "horseshoe effect" (Shinn (2023)) in Principal Component Analysis (PCA) can introduce artificial rotational dynamics.

Therefore, it is important to understand whether the low-dimensional representation captures true correlations or merely reflects noise in the data. The size of the dataset and preprocessing steps, such as normalization and denoising, can greatly affect the results of dimensionality reduction. Same signal, recorded from populations of different sizes, or using various techniques like two-photon calcium imaging extracellular electrophysiology, can yield quantitatively and qualitatively different representations.

This raises an important question: how exactly do these control parameters affect the precision of dimensionality reduction tools? In terms of experimental design, that question can be reformulated as "What is the correct recording method to study a given process?" "Can we make sure that the recorded neural population is large enough?" "Was the experiment repeated enough times, producing sufficient number of trials?"

Answering these questions requires a good understanding of the neural data, and what are its important properties.

Thesis overview

In our work we focus on PCA, one of the most used dimensionality reduction tools. We aimed at

- Understanding how the accuracy of this method depends on different control parameters, such as signal-to-noise ratio, number of neurons recorded, number of trials, the smoothing kernel applied for denoising etc., and providing an analytical expression for this accuracy as a function of these parameters.

- Being able to use this prediction of the accuracy on any given dataset, which required introducing the inference procedure for some of the control parameters.
- Developing a tool that can help in experimental design by extrapolating the prediction done on a small example of preliminary data to a potential dataset with different control parameters, such as larger number of neurons or trials.

Thesis structure

The main results of the Thesis are summarized below.

In **Chapter 1** we provide a review of current experimental practices in various fields of neuroscience, and focus on experimental design, exploring how the necessary size of the recorded data is determined. Next, we give a review of existing dimensionality reduction techniques, with a particular focus on PCA and its applications in neuroscience, and discuss existing attempts to evaluate the accuracy of PCA. Afterwards, we introduce concepts from statistical physics and random matrix theory and discuss the previous works on characterizing the accuracy of PCA.

Chapter 2 presents the theoretical results on PCA. We propose a model of neural activity and provide its detailed mathematical description. The rationale behind the chosen model is explained, supported by examples from the literature that demonstrate different properties of neural data. We introduce two different accuracy measures of PCA, and describe the mathematical tools and techniques necessary for their estimation. We finish the chapter by providing a scheme of the calculation of the accuracy measures.

In **Chapter 3**, we provide the validation of theoretical results on synthetic data. The chapter first presents the plots and discussion of how accuracy measures change with different control parameters. We then focus on inference and experimental design. We start by describing the procedure of inference of data parameters necessary for making predictions about the expected performance for larger number of trials or of recorded neurons.

Chapter 4 is devoted to the application of our approach to real extracellular electrophysiology data. The potential application of the model to calcium imaging data is also discussed as future work.

Chapter 5 concludes the manuscript by discussing the potential implications and benefits of using the accuracy prediction model in real-world neuroscience research. We discuss the limitations of the study, and suggest future research directions, such as including the exploration of other techniques such as Tensor PCA.

CHAPTER 1

Literature review

1.1 Experiments

The brain is an incredibly complex organ capable of performing a wide range of functions, thanks to its ability to process vast amounts of information using millions of interconnected neurons. These neurons communicate through electrical and chemical signals, enabling everything from basic reflexes to higher-order cognitive processes. To understand how various functions are encoded within the brain, scientists often conduct experiments that involve simultaneously recording an animal's behavior and the neural activity within specific brain regions. Such recordings and analysis of the covariation of populations of neurons has been done for many different brain functions, such as sensory coding and processing (Panzeri et al. (2017), Kayser et al. (2009), Saha et al. (2013), Pillow et al. (2008)) working memory (Wasmuht et al. (2018), Panichello and Buschman (2021), Machens et al. (2010)), decision-making (Raposo et al. (2014) Jacobs et al. (2020) Briggman et al. (2005)), evidence accumulation (Morcos and Harvey (2016)) and motor control (Churchland et al. (2012),Sadtler et al. (2014),Elsayed et al. (2016)).

Recording neural populations, rather than focusing solely on individual neurons, has become a common approach in neuroscience due to the assumption that brain function arises not from isolated neuronal activity, but from the complex interactions among large networks of neurons. This shift in focus has its roots in the evolution of our understanding of how the brain processes information and generates behavior. Early electrophysiological experiments, such as those by Hodgkin and Huxley in the 1950s, focused on the biophysical mechanisms underlying the generation and propagation of action potentials in neurons (Hodgkin and Huxley (1952)). In the 1950s, Horace Barlow's studies on the frog's retina suggested that individual neurons could serve as feature detectors, encoding specific aspects of visual stimuli (Barlow (1953)). Later, Hubel and Wiesel's experiments in the 1960s revealed that neurons in the visual cortex of cats respond selectively to specific orientations of visual stimuli, further solidifying the idea that individual neurons encode distinct features of sensory inputs (Hubel and Wiesel (1962)). These studies established the foundational concept that individual neurons could use a firing-rate code to convey information about the external world.

However, the idea that neural circuits are built for an emergent function appeared as early as the 1930s (Yuste (2015)). Rafael Lorente de Nó, a student of Santiago Ramón y Cajal, proposed that the structural design of many parts of the nervous system was characterized by recurrent connectivity. He suggested that this recurrent wiring might serve to generate functional reverberations—patterns of neuronal activity that persist even after the initial stimulus has ceased (De Nó (1938)). This idea was further developed by Donald Hebb, who proposed that neural circuits operate by sequentially activating groups of neurons, which he called "cell assemblies". According to Hebb, such patterns of neuronal activation, firing in closed loops, were crucial for generating functional states of the brain, such as memories or specific behaviors (Hebb (1949)).

In the following decades, as neuroscience advanced and new experimental techniques emerged, the limitations of single-neuron approaches became more apparent. It was shown that in primary sensory areas neurons do not always respond in the same way to identical sensory stimuli (Gallant et al. (1998)), suggesting that brain relies on distributed coding, where patterns of activity across neural populations convey information. This understanding was further solidified by studies like those of Schwartz et al. (1988), which demonstrated that motor cortex neurons encode movement

direction through population vectors — a collective representation generated by the activity of multiple neurons working together.

To fully capture the complexity of neural population dynamics, it is important to have a recording method that can simultaneously track the activity of numerous neurons with high temporal and spatial resolution. The diverse nature of neural activity and the specific demands of various research questions have led to the development of a range of recording techniques. Each method offers unique advantages and is suited to different aspects of neuroscience research. Below we will explore some of the primary techniques used in modern neuroscience, discussing their strengths, limitations, and the contexts in which they are most effective.

- **Intracellular recording.** Intracellular recording is one of the early techniques for measuring neuronal activity, with foundational work dating back to the early 20th century (Kelly and Woodbury (2003)). This technique involves inserting a fine electrode into a neuron to measure the voltage or current across the neuron's cell membrane, with a reference electrode placed outside the cell. These electrodes are connected to an amplifier to record the membrane potential. Two primary recording modes are commonly employed:

  1. **Current clamp**: In this mode of recording a constant or time-varying current is injected into the cell, and the resulting changes in membrane potential (voltage) are recorded. Often the goal of such recordings is to understand how neurons respond to stimuli, integrate inputs, and generate action potentials under physiological conditions.
  2. **Voltage Clamp**: This mode involves holding the membrane potential at a constant level while measuring the ionic currents flowing across the membrane. Voltage clamp is particularly useful for dissecting ionic mechanisms, studying the behavior of ion channels, and analyzing specific conductances in detail. This technique was used by Hodgkin and Huxley for their famous experiment on recording giant squid axons (Hodgkin and Huxley (1952)).

  Both of these recording modes allow for the indirect measurement of ionic and synaptic conductances across the cell membrane. While intracellular recordings are often used for measuring the electrical activity of individual neurons, this type of recording is impractical for studying large neuronal populations. Inserting an electrode into a neuron without causing damage is a delicate and challenging process. The invasiveness of the technique results in cell damage over time, making it impractical for long-term studies (Brette and Destexhe (2012)).

- **Extracellular electrophysiology** is one of the main approaches to measure neural activity *in vivo*. In single-unit recordings, sharp electrodes are placed near the soma of a neuron to monitor its firing rate by detecting the standardized extracellular signatures of action potentials, commonly referred to as spikes (Fig. 1.1). This method of recording has excellent temporal resolution, allowing data acquisition at the rate of 30 kHz (Steinmetz et al. (2021)). Given the typical spike duration of the order of few milliseconds (Henze et al. (2000)), this results in a very precise reporting of spike times.

  However, attributing a spike to the correct neuron becomes challenging when multiple neurons contribute to the recorded extracellular potential, particularly if two neurons of the same type are equidistant from the electrode. To overcome this challenge, tetrodes were developed as a more sophisticated tool for extracellular recordings. A tetrode is composed of four fine wires—often made of materials like platinum-iridium or tungsten—twisted together or arranged in a small bundle. Each wire functions as an individual electrode, allowing for the recording of neural activity from slightly different spatial locations within the brain. The four electrodes of a tetrode are typically arranged in a diamond or square pattern, with inter-electrode distances on the order of 20-40 micrometers. This small spacing ensures that the electrodes are close enough to detect signals from the same group of neurons, but far enough apart to pick up differences in signal amplitude and waveform due to the spatial positioning of each neuron relative to the individual wires. (Pettersen et al. (2012)). Tetrodes are often implanted in specific brain regions using a microdrive system that allows precise control over their depth, enabling researchers to target particular layers of the brain or neural circuits.

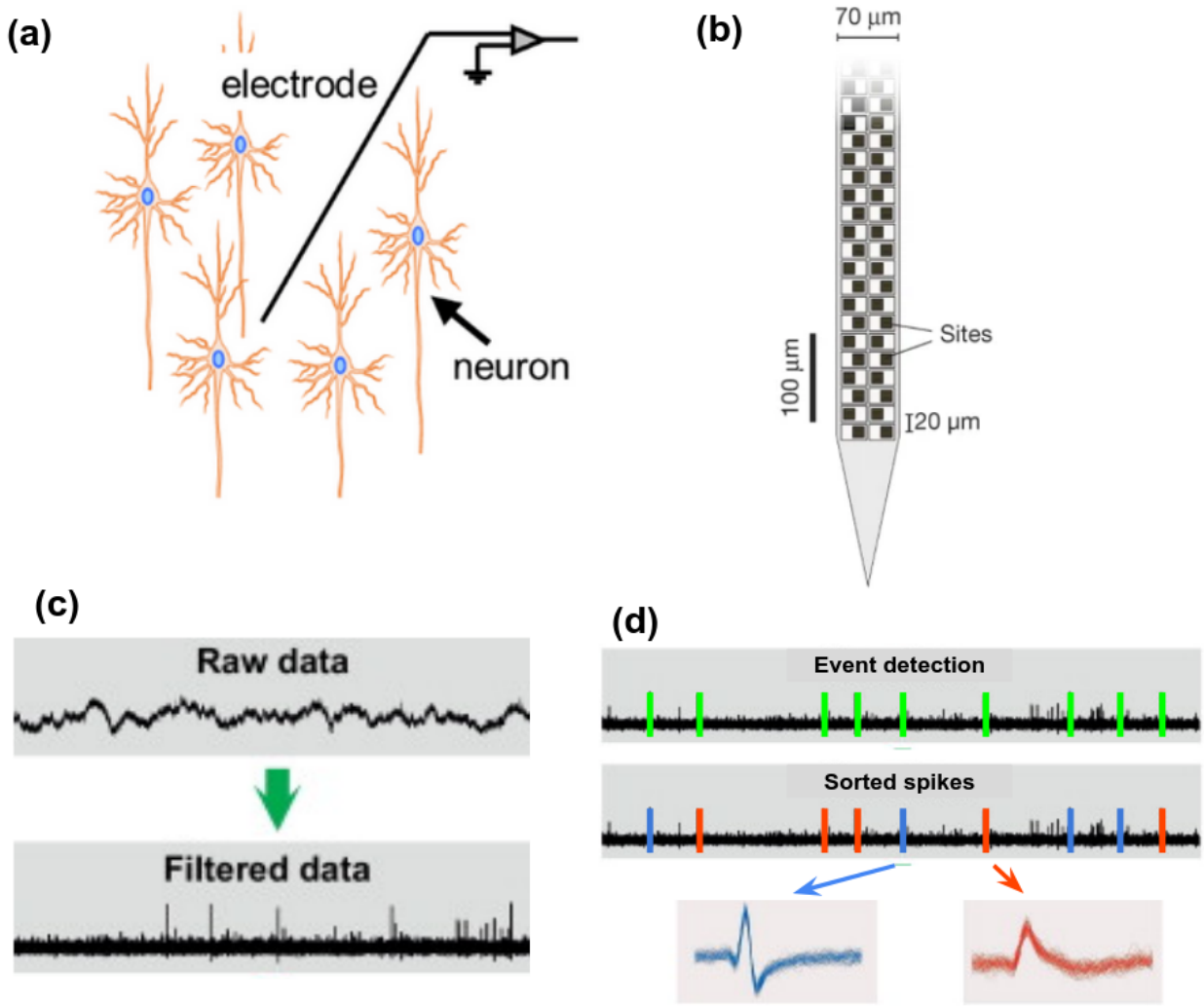

Figure 1.1: Extracellular eletrophysiology. **(a)**: Scheme of the extracellular electrophysiology recording (illustration from Noguchi et al. (2021)). **(b)**: Recording site placement on a shank of a neuropixel probe. Image from Jun et al. (2017). **(c)**: An example of a voltage profile from one of the recording sites. Top: raw profile. Bottom: after high-pass filtering. Adapted from Rey et al. (2015). **(d)**: Main idea of the spike sorting. First, the spikes are detected, usually through setting a threshold. Then, the detected events are sorted into groups based on their shape similarity.

A significant leap in the technology of neuronal recording was achieved with the development of Neuropixels probes (Jun et al. (2017), Steinmetz et al. (2021)). These state-of-the-art devices feature approximately 1,000 recording sites on a single, slender shank 70×20 μm cross-section. The probes can simultaneously record hundreds of neurons across different brain regions, including both cortical and subcortical areas.

However, despite these advancements, Neuropixels probes still face certain limitations. One significant drawback is that they do not provide information about the genetic identity of the recorded cells, limiting the ability to correlate specific genetic markers with functional activity. Additionally, while Neuropixels can record from a large number of neurons, it remains challenging to perform long-term recordings of the same individual cells over extended periods. This limitation arises due to potential shifts in tissue positioning relative to the probe, making it difficult to consistently track the same neurons across days or weeks.

- Over the past several decades, **optical methods** have become increasingly powerful tools for recording neural population activity in neuroscience. These methods are particularly effective due to the use of genetically encoded indicators, such as calcium or voltage indicators, which fluoresce in response to changes in neural activity. Introduced into neurons using

viral vectors or transgenic techniques, these indicators offer several significant advantages, making optical methods a popular choice for researchers.

One of the primary benefits of optical methods is their ability to facilitate chronic recordings. Genetically encoded indicators can be stably expressed in neurons over extended periods, remaining functional for weeks or even months. This long-term stability allows researchers to perform repeated measurements from the same population of neurons, providing valuable longitudinal data. Advanced imaging techniques, like two-photon microscopy, enhance this capability by offering a stable and minimally invasive setup for long-term observations.

Optical methods also excel in targeting and identifying specific cell types within complex neural networks. By selectively expressing genetically encoded indicators in particular cell populations based on genetic markers, researchers can study the distinct roles that different cell types play in brain function. This specificity is crucial for understanding the complex dynamics of neural circuits.

Another significant advantage of optical methods is their capacity to record activity from entire populations of neurons, rather than just a sparse subset. Wide-field or multi-photon imaging techniques enable the simultaneous capture of neural activity across large areas of brain tissue. Since genetically encoded indicators label all neurons within the region of interest, the activity of every cell is recorded, including signals from neurons that may otherwise remain undetected due to their low activity levels.

Calcium imaging relies on fluorescent indicators that change their conformation based on the level of calcium in the cell. This concentration depends on many processes in the cell, but notably changes a lot during action potential generation, when ions enter the soma. Recording the temporal changes in fluorescence results in calcium trace, which can be used to infer the underlying action potential firing.

Traditional imaging techniques, such as confocal microscopy, use single-photon excitation to capture fluorescence from these indicators. However, confocal microscopy requires the use of a pinhole to block out-of-focus light, which limits its depth of penetration in biological tissues and reduces its effectiveness in imaging deep brain structures.

Instead of using a single photon to excite the indicator, it is possible to use multiple photons, such as in the case of two-photon excitation. Two-photon excitation involves the nearly simultaneous absorption of two low-energy photons. Two-photon imaging uses longer wavelengths (typically in the near-infrared range) compared to the visible light used in one-photon imaging. A key advantage of two-photon imaging over confocal microscopy is that excitation occurs only within a small volume around the focal point of the laser. This precision allows all emitted photons, both scattered and ballistic, to be collected and used to reconstruct the image without the need for a pinhole. Consequently, two-photon imaging provides clearer images at greater depths and is more resilient to scattering. (Fig.1.2)

Despite its advantages, two-photon imaging still encounters challenges related to depth penetration. While scattered photons are effectively collected to reconstruct the image, the incident ballistic light decreases with depth due to scattering. Although increasing the incident power can compensate for this loss, there is a limit to how much depth can be achieved before surface signal generation becomes a significant issue, ultimately limiting the effective imaging depth. In addition, it elevates the risk of tissue damage due to two main factors: photobleaching, when fluorescent molecules lose their ability to emit light after being exposed to intense laser light, and phototoxicity, where the energy absorbed from the laser light causes damage to cellular structures (Svoboda and Yasuda (2006)). Overall, two-photon calcium imaging is used to record populations up to $500\mu m$ from the surface of the brain. (Kondo et al. (2017), Tischbirek et al. (2015), Dana et al. (2016)) The field of view (FOV) of a two-photon microscope is a key factor in determining the number of neurons that can be simultaneously recorded. Conventional two-photon microscopy typically offers an FOV of around $500 \times 500\ \mu m^2$ (Ohki et al. (2005), Peron et al. (2015)). Recent advancements have enabled the development of microscopes with much larger FOVs, extending several millime-

ters in diameter (Bumstead (2018)). Overall, the number of neurons recorded simultaneously can reach thousands. Stringer et al. (2019) achieved simultaneous recording of ten thousands neurons from mouse visual cortex.

While larger FOVs are advantageous, they also present challenges related to both speed and signal-to-noise ratio (SNR). Two-photon microscopy requires sequential scanning of individual resolution elements within the FOV. Consequently, larger FOVs can result in slower acquisition speeds, as the laser must scan across a broader area. Even if the same sampling rate is maintained, the SNR decreases with the expansion of the FOV. This is because the time spent per neuron (T) scales inversely with the size of the FOV, leading to reduced SNR. Therefore, there is a trade-off between capturing a larger number of neurons in a single frame and maintaining both high temporal resolution and SNR.

In a standard raster scanning approach, the entire field of view is scanned line by line. However, much of this scanned area often contains elements like blood vessels and neuropil, which do not provide valuable data relevant to the experiment. A prominent method for increasing the image acquisition speed consists of scanning only specific regions of interest instead of the entire field of view. This approach requires making an initial raster scan of the entire field of view. This allows to identify specific areas of interest that may correspond to active neural regions or specific cellular structures. Once these regions are identified, subsequent scans focus exclusively on these selected areas. Acousto-optic deflectors are among the most common tools used for random access scanning in two-photon imaging (Salomé et al. (2006), Kremer et al. (2008)). AODs use sound waves to diffract and steer the laser beam very quickly. AODs operate by using sound waves to diffract and rapidly steer the laser beam. By precisely controlling the frequency of the sound waves, the angle of diffraction can be finely tuned, enabling the laser to be directed to specific points in the sample almost instantaneously. This allows for high-speed recording from selected points of interest (Katona et al. (2012), Otsu et al. (2014), Nadella et al. (2016)).

Optical methods are generally optimized for imaging neurons within a single focal plane, which poses a significant limitation when attempting to capture three-dimensional structures. Extending imaging to 3D environments introduces increased complexity, particularly in terms of maintaining high speed and signal-to-noise ratio (SNR). Akemann et al. (2022) developed the three-dimensional custom-access serial holography (3D-CASH), which enables sampling at 40 kHz from selected neurons of interest.

Calcium imaging serves as a complementary method to electrophysiological recordings, offering several distinct advantages. It allows for the simultaneous recording of larger populations of neurons and does not require spike sorting. Additionally, calcium imaging enables chronic recordings over extended periods, the ability to target genetically identified cells, and the exhaustive sampling of all neurons within the field of view. Unlike electrophysiological recordings, calcium imaging does not exhibit a bias towards neurons with high firing rates, allowing even nearly silent cells to be recorded. However, the temporal resolution of calcium imaging is lower compared to electrophysiological recordings, as action potentials are detected indirectly through changes in fluorescence, which has slower dynamics. Additionally, the recording depth is limited by fluorescence scattering within the tissue.

- Another optical recording technique that has been actively developing in the recent years is **voltage imaging**. This technique uses fluorescent indicators that that are directly sensitive to changes in voltage, resulting in better temporal resolution, allowing to see the changes of neural activity at the scale of milliseconds (Knöpfel and Song (2019)). Despite this advantage, voltage imaging has not yet become widely adopted due to several technical challenges. The rapid changes in fluorescence associated with voltage-based indicators result in very brief signals, requiring image acquisition at very high rates to capture these transient events effectively. To address these challenges, two main approaches have been developed:
  - Widefield Approaches (Gong et al. (2015)): These involve sparse labeling of neurons and are typically limited to imaging superficial layers of the brain. Fast cameras are

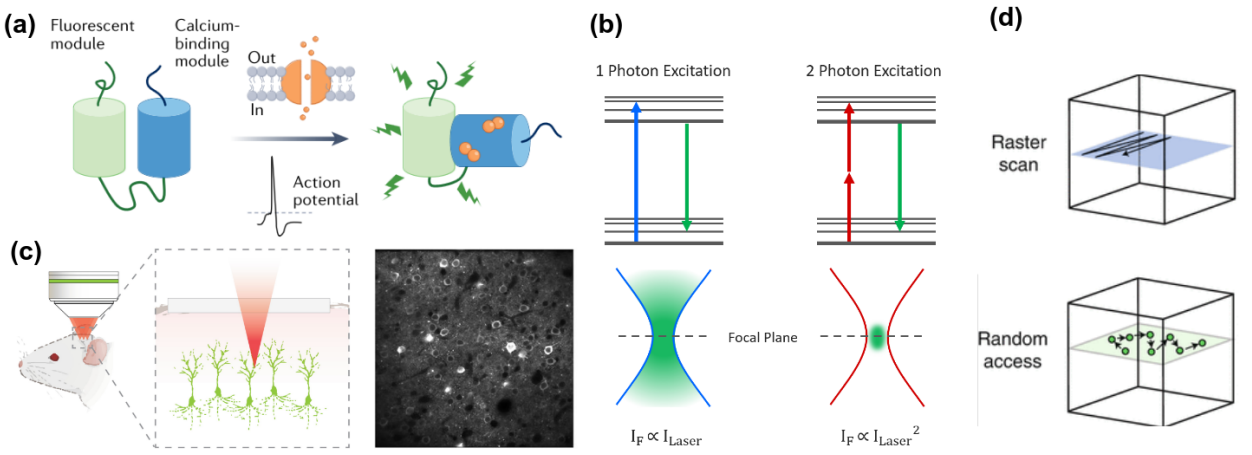

Figure 1.2: Two-photon calcium imaging. **(a)**: Mechanism of work of calcium indicators. The binding of Ca^{2+} ions to the calcium-sensing module of the indicator induces a conformational change in the fluorescent protein, resulting in an alteration of its brightness. Illustation from Grienberger et al. (2022). **(b)**: A comparison between one- and two-photon excitation schemes. In the two-photon case, the fluorescence is localized to a small volume in the focal plane. **(c)**: Left: Scheme of the imaging. The excitation by laser and recording of the fluorescence are done through a cranial window. Illustrartion from Nechyporuk-Zloy (2022). Right: An example of the image obtained with the two-photon calcium imaging. **(d)**: Different scannning techniques. Top: raster scanning. The entire field of view is scanned sequentially. Bottom: random access scanning. Only regions of interest are being recorded. Adapted from Grewe et al. (2013).

employed to capture the rapid fluorescence changes, but this method is constrained by its inability to penetrate deeper tissue layers.

– Deep Cellular Recordings with Two-Photon Microscopy (Villette et al. (2019)): This approach enables imaging of neurons located deeper within the brain but is generally limited to recording from 5-10 neurons within a single plane. Achieving the necessary temporal resolution for voltage imaging in these deeper layers is made possible through the use of acousto-optic deflectors, as in the case with two-photon calcium imaging described before.

Overall, while the current limitations of voltage imaging result in the ability to record from significantly fewer cells compared to recordings using calcium indicators, this technique is still in its early stages of development. With the potential advancement of acquisition methods, the number of neurons that can be simultaneously recorded with voltage imaging may increase in the following years.

There are other, less invasive methods for recording neural activity, such as Electroencephalography (EEG) and Magnetoencephalography (MEG), which are done at the surface of the scalp. We will not focus on them here as they do not provide single-neuron resolution in the recording.

1.2 Experimental design

Even when the overall scheme of the experiment (the type of animal behavior to be studied, the area of interest in the brain from which the neural activity must be recorded) is clear, many choices need to be made about the technical realization of the experiment. Here we list the most common choices to be made about the experimental design:

- **The type of recording technique.** It is important to make sure that the selected recording technique is compatible with the studied system. For example, if the study requires broad coverage of neuron populations near the surface of the brain, with the ability to track activity over time and target specific cell types, calcium imaging may be the preferred method.

However, if the experiment demands high temporal precision or involves deep brain regions, electrophysiological recordings might be more appropriate.

- **Number of animals.** Even among genetically similar animals, such as inbred strains of mice, there exists a natural variability in gene expression, brain anatomy, and neural connectivity. This genetic diversity can lead to differences in neural activity patterns. Using multiple animals ensures that the findings are not specific to an individual animal. While the use of multiple animals is vital for the integrity of scientific research, it must be balanced with ethical considerations. Researchers are obligated to adhere to strict ethical guidelines that aim to minimize harm and distress to animal subjects. This includes the implementation of the 3Rs principle — Replacement, Reduction, and Refinement (Törnqvist et al. (2014)):

  1. Replacement: Whenever possible, researchers should seek alternatives to animal models, such as computational simulations or in vitro studies, to reduce the reliance on live animals.
  2. Reduction: Scientists must carefully justify the number of animals used in their experiments, ensuring that it is sufficient to achieve statistical significance while minimizing unnecessary use.
  3. Refinement: Researchers should employ the most humane techniques available, reducing pain and stress for the animals. This can include improvements in surgical procedures, housing conditions, and care practices.

  Following these principles not only ensures the ethical treatment of animals but also improves the quality of the obtained data, as stressed or unhealthy animals may not provide reliable or representative results (Garner (2005)).

  In addition to ethical considerations, practical aspects play a significant role in the use of animals in research. Maintaining a colony of laboratory animals requires substantial financial investment and logistical effort. Costs associated with animal research include housing, feeding, veterinary care, and facility maintenance. These expenses can quickly accumulate, making it essential for researchers to carefully plan and justify the scale of their studies. Moreover, each animal in an experiment has to go through complicated procedures such as surgeries or specific training protocols. These tasks demand significant time and human resources, from skilled personnel to conduct the procedures to technicians and researchers who oversee daily care and training. The complexity of these tasks further emphasizes the need for efficient study designs that optimize the use of animals while ensuring meaningful outcomes.

- **Number of neurons being recorded in every animal.** Neural information encoding often involves a relatively small proportion of the total neuronal population being active at any given moment ((Hölscher and Munk 2008, Chapter 4), Field (1987), Vinje and Gallant (2000), Yu et al. (2013)). Due to the selective nature of sparse coding, if only a small number of neurons are recorded, there is a risk of missing the critical neurons involved in representing the information of interest. Therefore, increasing the number of recorded neurons can help in capturing the relevant neural activity patterns. Additionally, a larger sample size improves the signal-to-noise ratio, allowing researchers to discern meaningful signals from background noise. The capacity to record a large number of neurons simultaneously depends on the selected recording method, which is another important choice that we mentioned before. Another restriction comes from the fact that recording large numbers of neurons generates vast amounts of data, requiring efficient data management and analysis pipelines. The challenge of handling and interpreting these complex datasets can be a bottleneck in utilizing the full potential of extensive recordings.

- **Number of trials to be recorded for every condition.** Neural activity often exhibits variability across trials due to several factors, including noise, context-dependent changes, and the

inherent stochastic nature of brain processes. Understanding and accounting for this variability is crucial for accurately interpreting how neural circuits encode and process information. There are two primary sources of trial-to-trial variability:

  - Stochastic variability: This type of variability arises from the inherent randomness in neural processes, such as the timing of action potentials (spike timing). Even under identical conditions, the exact timing of spikes can vary from trial to trial due to the probabilistic nature of neurotransmitter release, the dynamic state of ion channels, and other cellular mechanisms. This stochasticity introduces variability in the neural response, making it important to record multiple trials to average out these fluctuations and obtain a clearer signal.

  - Contextual or latent process-linked variability: This form of variability occurs when there are unaccounted factors or latent processes influencing the neural response. For example, different trials might belong to different classes or states that are not explicitly controlled for in the experiment, such as varying levels of attention, internal states (e.g., motivation or arousal), or even subtle differences in sensory input or task execution. These latent processes can lead to systematic differences in neural responses across trials, complicating the interpretation of the data.

  While trial-averaging of neural responses is an effective strategy for mitigating stochastic variability—such as random fluctuations in spike timing—it may not be as effective in addressing variability related to unaccounted latent processes. These latent processes can introduce systematic differences between trials that are not captured by simple averaging, potentially obscuring important aspects of neural function.

- This brings us to another critical decision in experimental design: **The way of treating and analyze the data from different trials.** The two most common approaches are to perform analysis on trial-averaged data or on trial-concatenated data, each with its own strengths and limitations.

  - Trial-Averaged Data: Averaging neural responses across trials is particularly useful when the goal is to identify consistent, robust patterns of activity that are shared across all trials. This method effectively reduces random noise and stochastic variability, allowing researchers to focus on the core neural signal. It is especially beneficial when the experiment is designed to study the average response to a stimulus or task, where variability due to external factors or internal states is considered noise that should be minimized. By reducing trial-to-trial variability, trial-averaging can reveal clear, repeatable patterns of activity that are essential for understanding the basic principles of neural coding. However, trial-averaging may mask important information if different trials are influenced by varying latent processes or if there are subtle, but systematic, differences between them. In such cases, averaging can lead to the loss of meaningful variability, potentially overlooking important neural dynamics that are context-dependent.
  - Trial-Concatenated Data: Analyzing trial-concatenated data involves treating each trial as a separate instance, preserving the full extent of trial-to-trial variability. This approach is particularly powerful when the goal is to explore the influence of latent processes or to identify patterns of activity that vary systematically across different conditions or states. By maintaining the individuality of each trial, researchers can analyze the data to uncover correlations, state-dependent neural dynamics, or even identify subgroups of trials that represent different neural states or responses. This method is important in experiments where understanding the variability itself is key to uncovering the underlying mechanisms of neural processing. The main challenge with trial-concatenated data is that it retains all sources of variability, including both stochastic and systematic differences, which can make it more difficult to identify clear patterns or draw robust conclusions. Without careful analysis and appropriate statistical tools,

this approach can lead to noisy results that are harder to interpret, particularly when the primary interest lies in common features across trials.

A significant portion of neuroscience research is empirical, relying on experimental and observational methods to gather data about the brain and its functions. This data is frequently analyzed through the lens of hypothesis testing — a statistical method used to infer the validity of a specific hypothesis derived from theory. In practical terms, many neuroscience studies present their results as responses to specific hypotheses. Each of these scenarios involves formulating a null hypothesis (typically, that there is no effect or difference) and an alternative hypothesis (that there is an effect or difference), with outcomes reported in terms of statistical significance (Fig. 1.3, a). The size of this sample plays a crucial role in determining the reliability and validity of the study's conclusions.

The limiting factors of cost, time and effort mentioned above, as well as complications introduced by the selected recording technique often result in studies that have very small sample size: it is common to see experiments with very limited number of animals, sometimes as little as two (Fries and Maris (2022)). Similarly, the number of neurons recorded in a given animal in a given session may be small due to technical issues, such as unreliable spike sorting or electrode drift, for electrophysiology, or limited number of neurons due to motion artifacts in calcium imaging.

It is a well-documented concern in neuroscience and other scientific fields that many studies may lack sufficient statistical power due to small sample sizes. This issue has been highlighted in various critiques of scientific research, including the influential paper by Ioannidis (2005) titled "Why Most Published Research Findings Are False".

The core issue is that although the results of the research might show statistical significance, often marked by a low p-value (for example, a p-value below 0.05, indicating that there's less than a 5% chance the observed results are due to random variation), this significance can be misleading if the sample size is too small. A small sample size increases the risk that the observed effect is not truly representative of the population, but rather a random artifact of the limited data.

To enhance the reliability of statistical significance in research, two critical concepts have been introduced: effect size and statistical power (Ellis (2010)).

- **Effect size** is a measure that describes the strength or magnitude of a relationship between variables or the size of a difference between groups in a study. Unlike p-values, which only indicate whether an effect is likely to be real or due to chance, effect size quantifies how large that effect is. Various measures of effect size exist, each suited to different types of data and research designs. Importantly, many of these measures can be converted between one another because they often estimate the degree of separation between two distributions, making them mathematically related.

  One of the most commonly used measures of effect size is Cohen's d (Fig. 1.3, c). Cohen's d is specifically used to measure the difference between two means (let's call them m_1 and m_2) , expressed in standard deviation units. It is calculated as the difference between the means of two groups divided by the pooled standard deviation s of those groups:

$$d = \frac{m_1 - m_2}{s} \tag{1.1}$$

- **Power**, on the other hand, is the probability that a study will correctly reject a false null hypothesis. It is a function of several factors, including the sample size, the magnitude of the effect size, and the chosen significance level (commonly set at 0.05). A study with high statistical power is more likely to avoid Type II errors, or false negatives, where a true effect is missed. (Fig. 1.3, b)

To ensure that a study is adequately designed to detect meaningful effects, researchers often conduct a power analysis before collecting data (Bramlett et al. (1997), Titus et al. (2016), Button et al. (2013)). Power analysis is a statistical technique used to determine the sample size required to achieve a desired level of power, typically around 0.8. This means that the study would have

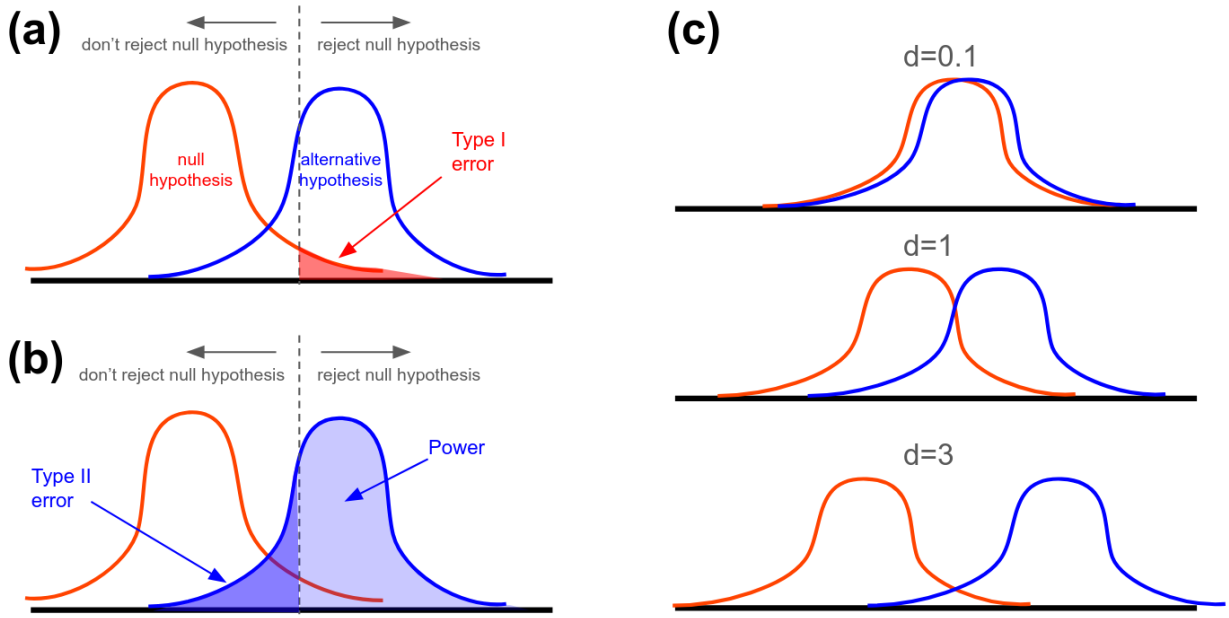

Figure 1.3: Hypothesis testing, effect size and power. **(a)**: Distributions for two hypotheses. The boundary defines whether the null hypothesis will be selected or rejected. **(b)**: A power is defined as the probability to correctly reject a false null hypothesis. **(c)**: Example of effect size calculated for three different cases.

an 80% chance of detecting an effect if one truly exists. The power analysis takes into account the expected effect size, the significance level, and the required power, thereby helping to minimize the risks of both Type I errors (false positives) and Type II errors (false negatives).

However, the effect size is not always known in advance. This can be a significant challenge, as inaccurate estimates can lead to under- or overpowered studies. If similar studies have been conducted before, reviewing existing literature to estimate the expected effect size can be helpful.

Another suggested approach is the "exploration-estimation" two-step experiment scheme, which emphasizes the importance of exploratory studies to identify potential effects and the effect size, followed by confirmatory studies designed to rigorously test these effects with adequate statistical power (Gelman & Carlin, 2014). This approach helps ensure that findings are robust and reproducible.

1.3 Dimensionality reduction techniques

As it was shown in the previous section, modern recording techniques allow recording the activity of many neurons simultaneously. The resulting dataset consists of many time frames of firing dynamics for each recorded neuron. Thus, the dataset is high-dimensional, making it challenging to work with and interpret. To address this challenge, dimensionality reduction methods are often employed. The main goals of dimensionality reduction include:

- **Feature Extraction and interpretation.** Dimensionality reduction can be used to extract meaningful features from neural data by taking different combinations of firing dynamics of the recorded neurons. Assuming that the true neural population dynamics is low-dimensional, only a few of such combinations will be required to provide an accurate approximation of the initial high-dimensional dataset (Cunningham and Yu (2014)). The resulting features are often easier to correlate with the behavior of the animal and the external stimuli than the firing activity of individual neurons (Gallego et al. (2017)).

  The extracted features can be also helpful for understanding how one group of neurons can carry out several different types of computations. For example, to understand how the motor cortex (M1) can prepare movements without executing them, Kaufman et al. (2014) used

Principal Component Analysis (PCA) to identify a six-dimensional neural manifold. They connected the obtained principal components with the activity of groups of muscles that work together, also identified using PCA by using a linear model. The study revealed that the neural manifold could be divided into two distinct spaces: a ”potent” space, which directly influences muscle activity, and a ”null” space, which does not affect muscle activity. Their findings show that preparatory neural activity occurs in a subspace different from the one where the activity responsible for the movement execution happens, allowing the neural population dynamics to evolve without causing any muscle activation.

- **Denoising.** Assuming that the real neural activity is low-dimensional, the overall noise in the recording can be reduced by discarding many dimensions that our method of dimensionality reduction presents as irrelevant, and thus, containing only noise. PCA is commonly used for such application of dimensionality reduction (Altan et al. (2021), Gallego et al. (2018)). In this case it is assumed that the hidden latent dynamics of the population explains most of the variance in the data. Then, assuming a certain threshold of the variance explained by the latent dynamics only leaves several first principal components.

  In their work, Pellegrino et al. (2024) introduce sliceTCA (slice tensor component analysis). In one application, they used sliceTCA on a dataset consisting of fluorescence traces from granule cells in the cerebellum and pyramidal neurons in the premotor cortex of mice. These neurons were recorded simultaneously while the mice performed a motor task that required to execute a sequence of two perpendicular motions on a virtual track: a forward movement followed by a lateral shift to either the left or right. The authors projected the representation of data obtained with sliceTCA onto the axis that best separated left from right trials. This approach revealed more interpretable and denoised representations compared to those derived from raw data, grouping behaviorally similar trajectories in an unsupervised way and enhancing the separation between distinct trials (Fig. 1.4, b).

- **Data Visualization.** The obtained simplified representation of the activity has only a few components. Combining them together in 2D and 3D plots often gives a better and more intuitive understanding of the dynamics of the neural population as a whole.

  For example, Russo et al. (2020) studied the differences in neural activity between the supplementary motor area (SMA) and the primary motor cortex (M1) in monkeys. It is believed that the SMA contributes to higher-order aspects of motor control, a function not attributed to M1. The authors focused on a specific higher-order function: tracking progress during an action. They recorded neural activity in both regions while the monkeys performed a task that involved navigating a virtual environment by moving a pedal with their hands to advance through the space. Using dimensionality reduction (PCA), they obtained neural trajectories for both regions. The results showed that in M1, the trajectory formed a circle, winding on itself with each cycle of pedaling (Fig. 1.4, c). In contrast, in the SMA, the trajectory formed a helix, with the progression along the helix encoding the advancement in the task. The observed difference in trajectories is consistent with the hypothesis that the SMA indeed encodes higher-order aspects of motor control, while M1 does not.

- **Compressing the data and improving computational efficiency.** High-dimensional datasets can be computationally expensive to process and analyze. Sometimes it is possible to first compress the data to a description using only a few components, and then apply any further analysis. This application of dimensionality reduction is often found in brain-computer interface (BCI) systems (Nicolas-Alonso and Gomez-Gil (2012)). The goal of such systems is to establish a direct communication pathway between the brain and an external device, typically a computer or robotic system. The transformation of the neural activity to the movement of the controlled system is typically done through a decoder. To ensure that the prosthetics can be controlled in real time, the delay between the signal sent by the brain and the response of the controlled system, such as a movement of a cursor on a screen or bending of an artificial limb should be minimal. A good way to minimize the response time is to train a decoder not

directly on a high-dimensional neural firing signal, but on a simplified representation of that signal obtained by a dimensionality reduction technique.

Additionally, as shown by Gallego et al. (2020), dimensionality reduction can be used to improve the stability of decoding over time. In this study, they recorded neural activity from the premotor, primary motor, and somatosensory cortices of monkeys while animals performed a center-out reaching task. In this task, the monkeys moved a cursor on a screen to one of eight possible targets by controlling a planar manipulandum with their hands. The neural activity was recorded using multielectrode arrays and was later used to infer the hand velocity by training a decoder. The recordings were repeated over many days, spanning a period of two years. Due to electrode shift in the tissue the neural activity recorded on different days involved different neurons, meaning that a decoder trained on data from the first day would not be effective on subsequent days, as the recorded neurons changed over time. However, the authors hypothesized that because neural activity is inherently low-dimensional, it might be possible to align the low-dimensional representations of data from different days. They developed an alignment method based on canonical correlation analysis and demonstrated that using this approach, decoding accuracy could be maintained over a very long timescale. (Fig. 1.4)

One of the most common ways to classify dimensionality reduction techniques is by the nature of the mathematical relationship between input and output spaces. Linear methods assume that the data lies on a lower-dimensional hyperplane within the original high-dimensional space. These methods linearly transform the data to reduce its dimensions. Typical examples of linear methods include:

1. Principal Component Analysis (PCA) (Joliffe (2002)): Projects data onto orthogonal axes that maximize variance.

2. Factor Analysis (FA) and Gaussian Process Factor Analysis (GPFA) (Yu et al. (2009)): Models observed variables as linear combinations of latent variables plus noise. GPFA extends FA by incorporating temporal correlations, allowing it to capture more complex dynamics in sequential data.

3. Linear Discriminant Analysis (LDA) ((Izenman 2013, Chapter 8)): Finds linear combinations of features that best separate classes.

4. Independent Component Analysis (ICA) (Hyvärinen et al. (2001)): Separates a multivariate signal into additive independent components.

5. Latent Linear Dynamics (LDS) (Cheng and Sabes (2006)): This method is based on linear state-space models that describe the system's evolution over time. LDS assumes that latent states evolve according to linear equations, characterized by a state transition matrix. This provides a discrete-time model that captures step-by-step dynamics.

Nonlinear methods are used when the data structure is assumed to be more complex than a linear manifold. They aim to uncover the intrinsic nonlinear structure of the data. Examples of nonlinear methods include:

1. Latent nonlinear dynamical systems (NLDS) (Wang et al. (2022)): This method can be viewed as an extension of LDS. In NLDS latent states evolve according to nonlinear functions, providing the ability to capture complex temporal patterns and dependencies.

2. Isomap: A method that learns the global geometry of a dataset (Tenenbaum et al. (2000)). For each data point, Isomap identifies a set of nearest neighbors, and then represents the entire dataset as a graph, where each data point is a vertex connected to its nearest neighbors. The graph is then used to estimate the geodesic distances between all pairs of data points. The result is then embedded in a low-dimensional space.

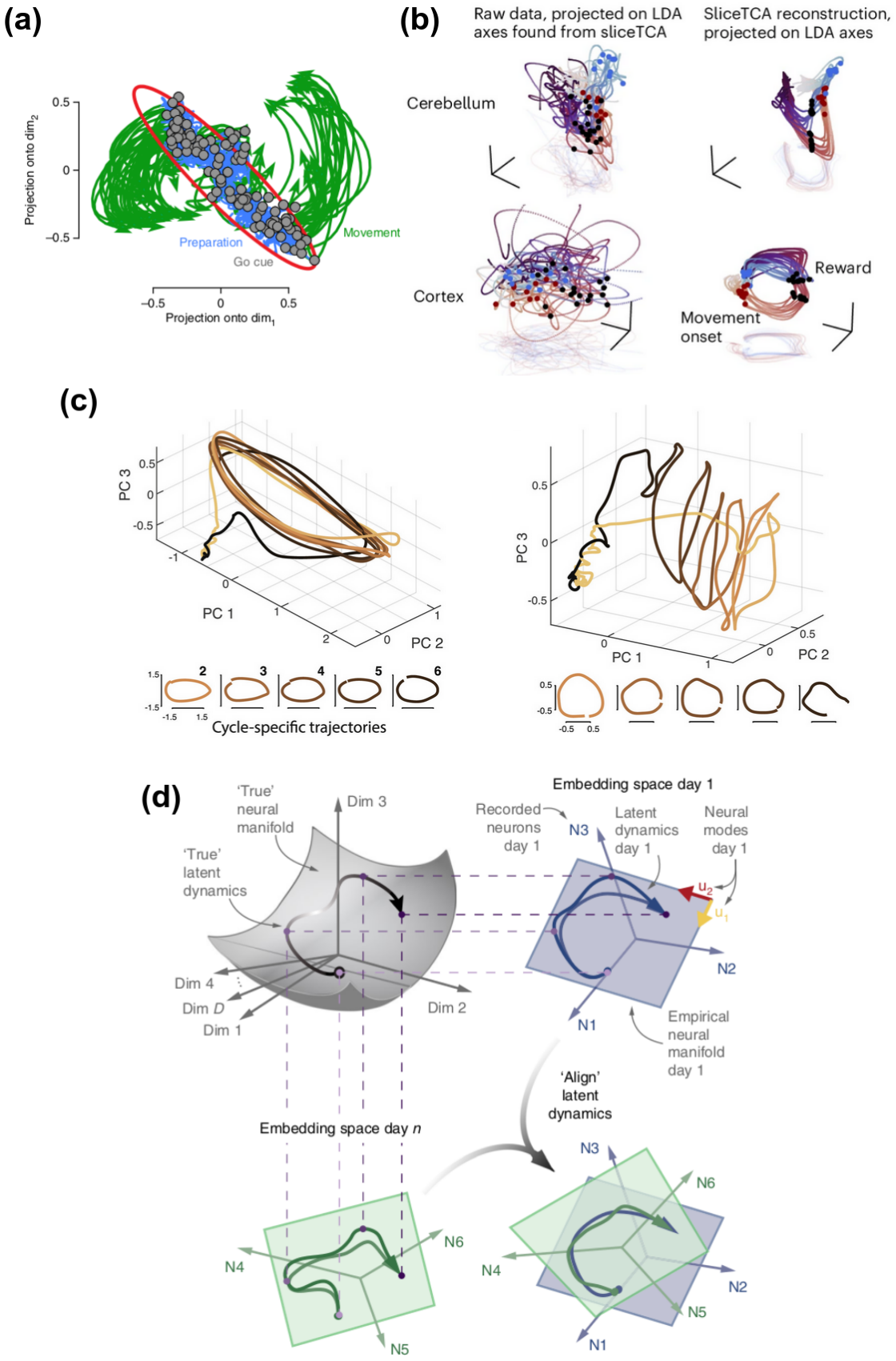

Figure 1.4: Different application of dimensionality reduction methods. **(a)**: Output-null and output-potent spaces of neural activity found in motor cortex of monkeys (from Kaufman et al. (2014)). **(b)**: Comparison of neural trajectories extracted from raw data and SliceTCA reconstruction of the data (from Pellegrino et al. (2024)). **(c)**: Different shapes of neural trajectories in M1 and SMA (from Russo et al. (2020)). **(d)**: Alignment of neural trajectories obtained on different days (from Gallego et al. (2020)).

3. Locally Linear Embedding (LLE) (Roweis and Saul (2000)): Preserves local relationships by reconstructing each data point using its neighbors.

4. Kernel PCA (Schölkopf et al. (1998)): Extends PCA by applying kernel tricks to capture nonlinear structures.

5. Autoencoders (Hinton and Salakhutdinov (2006)): Neural networks designed to learn efficient data representations.

6. Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. (2018)): Preserves both local and global data structure with high efficiency.

While more complex nonlinear methods allow capturing intricate neural dynamics, the combination of simplicity, robustness, and interpretability, along with its utility as a preprocessing step, makes PCA a particularly appealing choice for many neuroscience studies. For these reasons, our work focuses on evaluating the accuracy and effectiveness of PCA in analyzing neural data.

1.4 Accuracy of PCA in neuroscience

It is important to understand how much the low-dimensional representation obtained by PCA reflects the true latent low-dimensional activity of the neural population. . While the reliability and accuracy of PCA as a dimensionality reduction technique for identifying neural activity patterns have been discussed in the literature, these discussions are relatively sparse. However, there are several key aspects that have been explored:

- **Number of dimensions to use:** The effectiveness of PCA in capturing the essential variance in neural activity is highly dependent on the number of principal components chosen. This choice is influenced by the size of the recorded neural population and the complexity of their interactions. Larger and more complex neural systems may require more components to adequately represent the relevant variance. Although PCA offers a systematic approach for reducing dimensionality based on variance, it does not inherently provide a method to determine the optimal number of dimensions. Several methods have been employed in the literature to address this issue:

    1. Arbitrary variance threshold. As described by Altan et al. (2021), some researchers choose to retain enough components to explain a fixed percentage of the total variance (e.g., 90% or 95%). While straightforward, this method is somewhat arbitrary and may not always capture the most relevant aspects of the data.
    2. Scree Plot (Elbow Method) (Cattell (1966)). This visual method involves plotting the eigenvalues (or the variance explained by each component) and looking for an "elbow" where the explained variance starts to level off. The number of components before this point is typically retained. However, identifying the elbow can be subjective, especially in cases where there is no clear inflection point.
    3. Parallel Analysis (Horn (1965)). A more statistically grounded method, parallel analysis involves generating random data with the same structure as the original dataset and comparing the eigenvalues from the original data with those from the random data. The random data can be generated either by shuffling the initial dataset (retaining the mean values and variances for each neuron while erasing correlations between neurons) or by sampling from a given distribution, typically Gaussian, with the appropriate mean and covariance for each neuron. Components are retained only if their eigenvalues exceed those from the random data, providing a more robust criterion for dimensionality selection.
    4. Cross-validation leave-neuron-out. This method, inspired by Yu et al. (2009), involves splitting the dataset into training and test sets. PCA is performed on the training set, and the resulting directions are retained. In the test set, one neuron is left out at a time,

and the fitted model is used to predict the activity of that neuron based on the activity of the remaining neurons. For PCA, this prediction is a simple geometric projection. The accuracy of these predictions can be compared across different choices of retained dimensions, providing a basis for selecting the most appropriate dimensionality.

- **Preservation of the latent activity structure.** Reducing dimensionality of the data may result in losing some of the important aspects or structure of low-dimensional activity. One of the indirect ways to test whether the latent activity structure was preserved well is by performing another type of analysis on both initial, high-dimensional data and its low-dimensional representation, and then to compare the results. A practical example of this approach comes from brain-computer interface (BCI) research. For instance, one might train a decoder to predict limb movements (such as position and velocity) from the neural activity. By comparing the performance of the decoder when trained on the original data versus the PCA-reduced data, one can assess how well the low-dimensional representation preserves the relevant features of the neural activity.

Although discussions on the reliability and accuracy of PCA in neuroscience are relatively limited, there is a substantial body of mathematical and theoretical work that addresses the accuracy estimation of PCA. However, many of these theoretical insights have not yet been widely applied in the analysis of neural data. In the following section, we provide an overview of the key theoretical results related to the accuracy of PCA, highlighting why these results are challenging to apply directly to population neural activity data. This overview also sheds light on how these challenges can be addressed by extending the theoretical frameworks, thereby bridging the gap between theory and practical application in neuroscience.

1.5 Statistical Physics and Random Matrix Theory: theoretical results on PCA

1.5.1 Random Matrix Theory

Random Matrix Theory (RMT) was introduced in the 1950s as a mathematical framework to study large and complex systems, particularly within the field of nuclear physics. The theory was developed by Eugene Wigner, who was motivated by the need to understand the intricate energy levels of heavy atomic nuclei. During this period, nuclear physicists faced significant challenges in analyzing the energy spectra of atomic nuclei, especially those with many protons and neutrons. The interactions within these nuclei are highly complex, making it difficult to predict the exact energy levels using conventional methods. Wigner proposed a new approach to address this complexity by employing random matrices as a statistical model for the energy levels. Wigner hypothesized that the spacings between the energy levels in a heavy atomic nucleus could be modeled by the spacings between the eigenvalues of a random matrix. He suggested that these spacings would depend primarily on the symmetry class of the system's underlying dynamics, rather than the specific details of the nuclear forces.

Consider a matrix $\mathbf{J}$ that represents the system under study. In the approach of Random Matrix Theory, instead of treating $\mathbf{J}$ as a fixed and deterministic entity, it is viewed as a random sample drawn from a specific probability distribution. This means that if one is interested in a particular property of $\mathbf{J}$, such as the distribution of its eigenvalues or the spacing between these eigenvalues, this property is not analyzed for a single matrix. Instead, it is understood as an average over all possible realizations of the random matrix drawn from the specified probability distribution.

The reasoning behind the effectiveness of ensemble averaging in large systems is deeply rooted in the concept of self-averageness. When a system is self-averaging, the fluctuations in macroscopic quantities due to different realizations of the random parameters $\mathbf{J}$ diminish as the system size increases. This means that, for sufficiently large systems, the behavior of almost any single realization of $\mathbf{J}$ will closely match the behavior predicted by averaging over all possible realizations of $\mathbf{J}$.

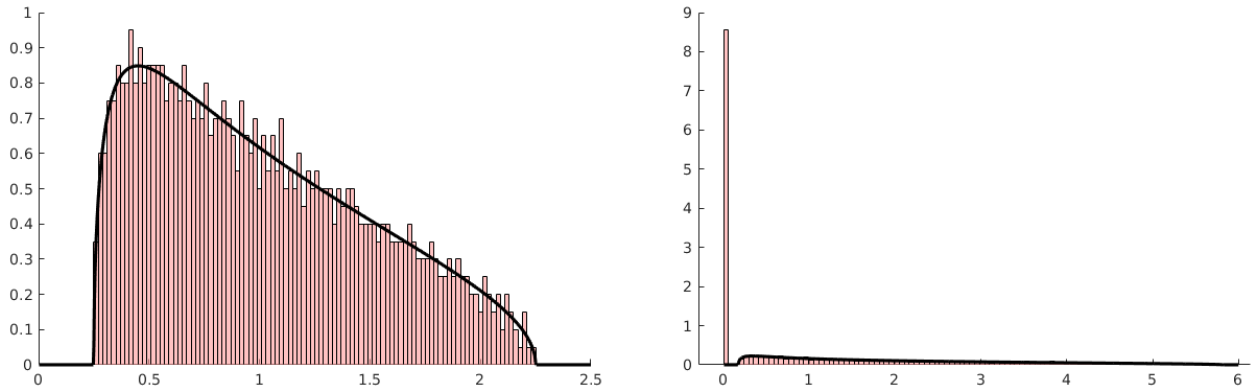

Figure 1.5: Marchenko-Pastur distribution. Left: Distribution for $\sigma = 1$, $\alpha = T/N = 4$ compared to the eigenvalues of a sample random matrix with $N = 1000$, $T = \alpha N = 4000$. Right: same for $\alpha = 0.5$, $N = 1000$, $T = 500$. Note the presence of zero eigenvalues.

1.5.2 Application of random matrix theory to the case of symmetric data

Early analytical results on the accuracy of PCA focused on datasets consisting of multivariate Gaussian noise. Suppose we take such a dataset consisting of T observations $\mathbf{s}_t$, each observation being an N-dimensional random Gaussian vector with the population covariance matrix Σ. In scenarios where N is fixed while T approaches infinity, the sample covariance matrix

$$C_{i,j} = \frac{1}{T}\sum_{t=1}^{T} s_{i,t}s_{j,t} - \left(\frac{1}{T}\sum_{t=1}^{T} s_{i,t}\right)\left(\frac{1}{T}\sum_{t=1}^{T} s_{j,t}\right) \tag{1.2}$$

converges to the population covariance matrix Σ. This convergence implies that PCA effectively identifies the true directions of variance since both the eigenvalues and eigenvectors of the sample covariance matrix closely approximate those of the population covariance matrix. The underlying mathematical framework for this convergence is rooted in the Law of Large Numbers, which asserts that as T increases, sample estimates become more accurate representations of population parameters.

The situation changes significantly when both N and T grow large simultaneously. Marchenko and Pastur (1967) discovered that in the case of the Gaussian data with variance σ^2 for every dimension ($\Sigma = \sigma^2 \mathrm{Id}_N$), when the ratio $\alpha = T/N$ is held constant as both T and N approach infinity, the eigenvalue distribution of the sample covariance matrix converges to a distinct probability distribution, now known as the Marchenko-Pastur distribution. This distribution is characterized by the following probability density function:

$$p(\lambda) = \frac{\alpha}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{1.3}$$

where λ_+ and λ_- define the bounds of the eigenvalue spectrum:

$$\lambda_+ = \sigma^2\left(1 + \sqrt{\frac{1}{\alpha}}\right)^2, \quad \lambda_- = \sigma^2\left(1 - \sqrt{\frac{1}{\alpha}}\right)^2 \tag{1.4}$$

with additional delta-peak at zero for the poor sampling situation ($T < N$) (Fig. 1.5).

The Marchenko-Pastur distribution highlights a critical aspect: as T and N become comparable, the eigenvalue spectrum of the sample covariance matrix diverges from the population covariance matrix, indicating that PCA may no longer reliably identify the true directions of variance in the data.

Marchenko and Pastur derived this distribution without using the formalism of statistical mechanics; however, it can also be approached through the statistical mechanics framework (see the review of Advani et al. (2013)).

Beyond the general description of the eigenvalue spectrum of data coming from purely symmetric distribution done by Marchenko and Pastur (1967), there are results describing the distribution of individual largest eigenvalues in that case. In his work, Johnstone (2001) explored these dynamics and demonstrated that in the limit where both T and N grow large, with $T/N \geq 1$, the top eigenvalue, after centering and scaling approaches the Tracy-Widom distribution. This finding implies that the largest eigenvalue of the sample covariance matrix, in high-dimensional settings, follows a specific probabilistic behavior, distinct from classical Gaussian distributions.

1.5.3 Results on PCA on data with non-symmetrical distribution

Independently of the work within statistical mechanics, the phase transition of the top eigenvalue in PCA was also explored by researchers outside this domain. Johnstone (2001) introduced the spiked covariance model of high-dimensional data, aiming at understanding how out-of-bulk eigenvalues affect the distribution of top eigenvalues. In this model, most eigenvalues of the population covariance matrix are set to one, representing noise, while a small number of eigenvalues (called spikes) are significantly larger, representing signal dimensions. This configuration mirrors real-world datasets, where only a subset of features contributes meaningfully to the data's variance. Mathematically, the spiked covariance model can be expressed as:

$$\Sigma = \text{diag}(l_1, \ldots, l_r, 1, \ldots, 1) \tag{1.5}$$

The spiked covariance model became widely used for analyzing the accuracy of PCA. Baik et al. (2004) showed that in this model as both T,N tend to infinity, the largest eigenvalue λ_1 of the empirical covariance matrix undergoes a phase transition depending on the difference between the spiked and non-spiked population eigenvalues. Informally speaking, if the non-unit eigenvalues are too close to the critical value λ_+, then by looking at the empirical covariance matrix, it would not be possible to know that the population covariance matrix is spiked, since the top eigenvalues of the sample covariance matrix will behave in roughly the same way as if the true covariance were the identity. However, when the true spiked eigenvalues are well above λ_+, the sample eigenvalues have a different asymptotic property.

More precisely, let us define $r = N/T = 1/\alpha$. Then,

- If $r > r_i = (l_i - 1)^2$, the sampling depth is not sufficient to unveil the presence of the non-trivial signal eigenvalue l_i. For instance, if $r > r_1$, the largest eigenvalue of the correlation matrix coincides with λ_+.
- If $r < r_i = (l_i - 1)^2$, the sample eigenvalue λ_i separates significantly from the bulk eigenvalues, as indicated by the centering around $\left(l_i + \frac{l_i\, r}{l_i - 1}\right)$, a factor depending on the spiked value l_i. This reflects a detectable signal that stands out against the noise.

While in their paper Baik, Ben Arous, and Péché focused on eigenvalues, the paper of Paul (2007) provides a comprehensive asymptotic analysis of eigenvectors in the spiked covariance model. Paul derives results regarding the convergence of sample eigenvectors to population eigenvectors. Specifically, he examines how well the sample eigenvectors estimate the true directions of variance in the presence of noise and spikes. This includes analysis of the angles between sample and population eigenvectors and the rate of convergence. Paul's analysis reveals a phase transition phenomenon similar to eigenvalues, but applicable to eigenvectors. He shows how the angle between sample and population eigenvectors changes dramatically depending on whether the corresponding eigenvalue is above or below the threshold.

1.5.4 Statistical mechanics framework

In the previous section, we presented the key analytical results concerning the accuracy of PCA, supported by rigorous mathematical proofs from the referenced works. However, it is important to note that some of these results were initially derived through non-rigorous methods from statistical

mechanics. Given that these methods are also used in the current work, we will now introduce them, and provide an outline of the calculation.

In statistical physics, it is common to encounter probability distributions or optimization problems that are influenced by external, fixed parameters. Let's assume that we have a system with the energy $E_{\mathbf{J}}$ of the states s that depends on some fixed parameters $\mathbf{J}$:

$$p(\mathbf{s}) = \frac{1}{Z(\beta, \mathbf{J})} e^{-\beta E_{\mathbf{J}}(\mathbf{s})} \tag{1.6}$$

where, $Z(\beta, \mathbf{J})$, known as the partition function, serves as a normalization constant. One might be interested in calculating the average value of some observable $f(\mathbf{s})$ over this probability distribution:

$$\langle f(\mathbf{s}) \rangle_{\mathbf{s}} = \int f(\mathbf{s}) p(\mathbf{s}) d\mathbf{s} = \frac{1}{Z(\beta, \mathbf{J})} \int f(\mathbf{s}) e^{-\beta E_{\mathbf{J}}(\mathbf{s})} d\mathbf{s} \tag{1.7}$$

To calculate such averages, one can introduce a generating function for the observable $f(\mathbf{s})$. By modifying the energy function $E_{\mathbf{J}}$ to include a term $\epsilon f(\mathbf{s})$ where ϵ is a small parameter, the partition function becomes:

$$Z(\beta, \mathbf{J}, \epsilon) = \int e^{-\beta E_{\mathbf{J}}(\mathbf{s}) + \epsilon f(\mathbf{s})} d\mathbf{s} \tag{1.8}$$

The free energy corresponding to this modified partition function is:

$$I(\beta, \mathbf{J}, \epsilon) = -\frac{1}{\beta} \ln Z(\beta, \mathbf{J}, \epsilon) \tag{1.9}$$

The average value $\langle f(\mathbf{s}) \rangle_{\mathbf{s}}$ can then be extracted by differentiating the free energy with respect to ϵ and evaluating at $\epsilon = 0$:

$$\langle f(s) \rangle_s = -\frac{\partial I(\beta, \mathbf{J}, \epsilon)}{\partial \epsilon}\bigg|_{\epsilon=0} \tag{1.10}$$

However, as the number of parameters $\mathbf{J}$ increases, or as the energy function $E_J(s)$ becomes more complex, this task becomes increasingly difficult. The high-dimensional nature of the parameter space and the intricate structure of the energy landscape can make direct calculations of such averages intractable.

In the approach of Random Matrix Theory described before, instead of calculating an average value $\langle f(s) \rangle_s$ for a fixed realisation of J, we will replace it with an average value over all possible realisations of $\mathbf{J}$. For the free energy, this will translate into calculating the averaged free energy over the ensemble of all possible realizations of $\mathbf{J}$:

$$\langle I(\beta, \mathbf{J}, \epsilon) \rangle_{\mathbf{J}} = -\frac{1}{\beta} \langle \ln Z(\beta, \mathbf{J}, \epsilon) \rangle_{\mathbf{J}} \tag{1.11}$$

Here, the average $\langle \cdot \rangle_{\mathbf{J}}$ represents the expectation over the distribution of $\mathbf{J}$.

1.5.5 Replica method and its application to PCA

Directly averaging the logarithm of the partition function (1.11) is challenging. The replica method, proposed by Edwards and Anderson (1975), has proved useful in this task. The method involves "creating" multiple copies of the system. Such idea allows us to find the expressions of the type $\langle Z^n \rangle_{\mathbf{J}}$, where n is the number of replicas. Then, the average of the logarithm can be expressed as the following limit:

$$\langle \ln Z(\beta, \mathbf{J}) \rangle_{\mathbf{J}} = \lim_{n \to 0} \frac{\langle Z^n(\beta, \mathbf{J}) \rangle_{\mathbf{J}} - 1}{n} \tag{1.12}$$

The calculation of this is carried as following:

1. Write Z^n as the product of (1.8) for the same data realization $\mathbf{J}$.

2. Evaluate the average over $\mathbf{J}$.

3. Now, previously non-interacting replicas are coupled. Introduce the overlaps that describe the resulting interactions between replicas.

4. Apply replica-symmetric ansatz. The replica-symmetric ansatz assumes that the overlaps between all pairs of replicas are the same.

5. Evaluate the rest of the integrals under this ansatz.

6. Test whether the results obtained with this ansatz match experimental or numerical data. If not, one must consider the more complex scenario of replica symmetry breaking.

The replica method has been particularly powerful in the study of spin glasses—a class of disordered magnetic systems characterized by randomly varying local couplings, both in sign and magnitude. Sherrington and Kirkpatrick (1975) applied this method to their mean-field model of spin glasses, but they found that the solution derived under the assumption of replica symmetry was incorrect at low temperatures. Specifically, the replica symmetric approximation led to unphysical results, such as negative entropy, indicating that it failed to capture the true nature of the spin glass phase. To address this issue, Giorgio Parisi introduced the concept of replica symmetry breaking (RSB), which provided a more accurate and nuanced understanding of the complex, hierarchical organization of states in these systems. Parisi's approach resolved the inconsistencies of the earlier model and has since become a widely used tool in the study of disordered systems.

Replica method has been applied to study the problems of learning, such as the storage problem of perceptron (Gardner and Derrida (1988)), where the goal is to find the weights that reproduce the correct classification of given patterns, supervised learning (Theumann and Köberle (1990)), where a model learns to classify input patterns by being trained on labeled examples provided by a teacher (or dataset), and unsupervised learning (Biehl and Mietzner (1994), Watkin and Nadal (1994)), where the model tries to identify patterns, structures, or relationships within the data on its own, without explicit guidance on what the correct outputs should be.

In their work, Reimann and Van den Broeck (1996) provide a replica calculation for unsupervised learning from T independent data points s^μ, $(\mu = 1 \dots T)$, each of them being a sample from an N-dimensional distribution with a single symmetry-breaking orientation B. The model learns the direction of the symmetry breaking. In this case, the learning relies on minimizing the quadratic energy function (see eq. (1.6)), and is defined as

$$E(r^\mu) = \frac{c}{2}(r^\mu)^2 - dr^\mu \tag{1.13}$$

where r^μ is defined as an overlap (cosine of the angle) between a data point s^μ and the direction B. The parameters c and d can be set to different values, allowing to study different learning rule cases, such as the adaline ($c > 0$) (Widrow (1962)) and the Hebb rule ($c = 0$) (Hebb (1949)). The case of $c < 0, d = 0$ corresponds to the maximal variance principle, making the results derived for the general model applicable to studying the analytical properties of PCA.

The authors show that the cosine overlap $R = \frac{J \cdot B}{N}$ between the learned direction J and the true direction B of symmetry breaking goes though a phase transition as the ratio T/N of available data points over the data dimension N changes. This phenomenon, referred to as retarded learning (Watkin and Nadal (1994)), shows that in the limit of large dimensionality of the system, the right direction can not be inferred unless the sufficient amount of data is accumulated (Fig. 1.6). This phenomenon corresponds to the phase transition for the top eigenvalue described by Baik et al. (2004) and mentioned in the previous section.

Overall, there are many analytical results on PCA, including those related to the convergence properties of eigenvalues and eigenvectors. These results provide a theoretical understanding of how PCA behaves under various conditions of sample size and dimensionality. However, the direct application of these results to neuroscience, especially in terms of experimental design and population neural activity recordings, has not really been explored so far, and for several reasons:

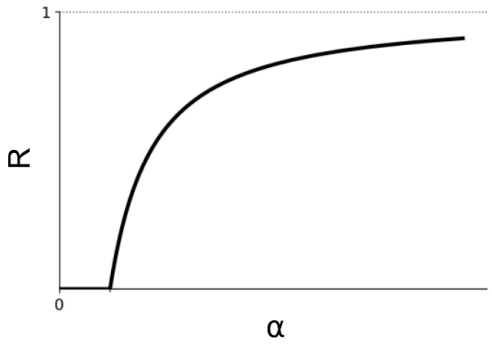

Figure 1.6: Phenomenon of the retarded learning. Below the critical value of $\alpha = T/N$ the alignment of learned direction to the true symmetry breaking direction does not occur.

- Many theoretical results for PCA are derived under simplifying assumptions. Analytical models often assume a basic structure where data is divided into a signal component (e.g., principal components) and a noise component. The signal is assumed to be the same across samples or neurons, and noise is typically modeled as independent and identically distributed (i.i.d.) Gaussian noise. In practice, neural data can exhibit much more complex structures. For instance, neural signals may include temporally correlated noise, non-stationarity, and spatial correlations among neurons. Additionally, preprocessing steps like temporal smoothing can alter the noise characteristics and the data structure, making the direct application of theoretical results challenging.
- Analytical estimation of accuracy of PCA make use of some of the parameters of the data, that in a real setup are not directly accessible, such as signal-to-noise ratio. To apply theoretical predictions to real data, these parameters must be estimated or inferred from the data. Techniques such as empirical estimation, cross-validation, or Bayesian inference might be employed. However, these methods can introduce their own uncertainties and potential biases.

This highlights the motivation for our work. We aim to bridge the gap between analytical results on PCA and their application to neuroscience by addressing the aforementioned challenges. Our approach involves developing a more sophisticated model of neural data that reflects real-world observations and takes the selected recording technique into consideration. We provide an analytical calculation of PCA for this model, expressing the accuracy on both eigenvalues and eigenvectors through the parameters of the model. To ensure applicability to real neural data, we include a procedure for inferring unknown model parameters from any given dataset. We give the interpretation of the accuracy measures in terms of relevant concepts used in neuroscience. The analytical nature of the result allows us to expand it for any other hypothesized dataset size, making it a useful tool for exparimental design.

CHAPTER 2

Theoretical results on Principal Component Analysis

2.1 Towards a realistic spike-covariance model

As stated before, one of the main reasons why the theoretical results on PCA are hard to apply to the data in neuroscience is the simplistic nature of the modelling of the data. One of the most frequently used theoretical settings is the so-called spiked covariance model, presented in the previous chapter. Let us recall how the data matrix is generated with this model: it consists of T data points in dimension N and is given by

$$s_{i,t} = \sum_{k=1}^{K} \sqrt{\lambda_k}\, x_t^{(k)}\, e_i^{(k)} + z_{i,t}, \quad Var(z_{i,t}) = \sigma^2, \;\; Var(x_t^{(k)}) = 1, \;\; |e^{(k)}|^2 = N \tag{2.1}$$

This formula can be interpreted as follows. We consider the dimensions $i = 1 \dots N$ as neurons, observations $t = 1 \dots T$ as time bins, and the directions k with larger variance as the dimensions of latent dynamics in the data. Then, $s_{i,t}$ naturally represents the activity of the neuron i at time t. The variable $z_{i,t}$ represents the (Gaussian) noise on the activity, while $x_t^{(k)}$ is the signal at time t along the direction k. For simplicity, this signal is generally assumed to be centered (zero mean) and variance unity, such that the amplitude of the k^{th} component of the signal is expressed through the factor $\sqrt{\lambda_k}$.

While the classical spiked-covariance model presented above introduces special directions with large variance, it may not fully capture the complexity and peculiarity of real neural activity and of the artifacts introduced by the measurement techniques.

2.1.1 Trial-invariant low-dimensional activity

In real neural systems, the observed latent dynamics often exhibit structured patterns that can be described in terms of a small number of smoothly varying modes, which remain stable across different trials. We denote the number of these modes as K.

To represent this, we introduce latent modes $x_t^{(k)}$, where $k = 1 \dots K$ indicates the specific mode. These latent modes capture the deterministic part of the signal that contributes to the observed neural activity. Thus, the updated model reads (we absorb the unknown scales of the signal components in the signal x itself):

$$s_{i,t} = \sum_{k=1}^{K} x_t^{(k)} e_i^{(k)} + z_{i,t}, \quad Var(z_{i,t}) = \sigma^2 \tag{2.2}$$

The latent modes $x_t^{(k)}$ are not directly observable from the data; their inference is a key objective of dimensionality reduction techniques. Our prediction of accuracy of PCA will rely on the model of the data where we assume that $x_t^{(k)}$ are known. To address this, in the later parts of this work we describe an inference procedure that estimates these modes and their variance from the data.

2.1.2 Trial-to-trial variability in low-dimensional activity

Before we have stated that in neural recordings the observed latent dynamics can be described in terms of a small number of smoothly varying modes, which remain stable across different trials. Despite the stability of these latent modes, neural recordings often show considerable variability between trials. This trial-to-trial variability can stem from several factors:

- **Physiological States:** The animal's physiological state encompasses a range of factors, including attention levels, fatigue, hunger, stress, hormonal fluctuations, and circadian rhythms. These states can fluctuate from trial to trial, leading to variability in neural activity. For example, an animal that is more attentive or less fatigued during one trial might exhibit more consistent and precise neural responses compared to a trial where it is less engaged or more tired. Similarly, hunger or satiety, stress levels, and other physiological conditions can influence how the animal processes stimuli and performs tasks, contributing to variability in the recorded neural data.

- **Learning and Adaptation:** As an animal repeatedly performs a task, its neural responses may change due to learning and adaptation. Early in training, neural activity might be more variable as the animal learns the task. Over time, as the animal becomes more proficient, the variability might decrease, or new patterns might emerge as the animal optimizes its strategy. This can introduce variability across trials, especially in experiments where learning is a significant factor.

- **Sensory Noise:** In tasks involving sensory stimuli, slight variations in the presentation of stimuli can lead to trial-to-trial variability. For example, minor differences in visual or auditory stimulus properties, such as brightness or sound intensity, can affect neural responses. Even in carefully controlled experiments, sensory systems might process the same stimulus slightly differently due to intrinsic noise in sensory pathways.

- **Internal Neural States:** The brain is a dynamic system with ongoing internal processes that may not be directly related to the task at hand. These include spontaneous neural activity, arousal fluctuations, and other cognitive processes such as thoughts or memories, which can vary between trials and contribute to variability in the observed neural activity.

To account for this trial-to-trial variability within the existing framework, we introduce an additional noise term that acts within the same low-dimensional subspace as the neural dynamics. Specifically, for each latent mode k, we introduce a noise term $\delta x_t^{(k)}$. This term represents the variability in the neural dynamics that cannot be attributed to the stable latent modes alone. We assume that this noise term $\delta x_t^{(k)}$ has a variance $(\xi^{(k)})^2$ and can be temporally correlated, and so characterized by a time correlation matrix $\Delta_{t,t'}$. For simplicity we assume that this matrix is the same for all modes k:

$$\left\langle \delta x_t^{(k)} \right\rangle = 0, \quad \left\langle \delta x_t^{(k)} x_{t'}^{(k')} \right\rangle = (\xi^{(k)})^2 \delta_{k,k'} \Delta_{t,t'} \tag{2.3}$$

2.1.3 Different noise amplitude for different neurons

When modeling neural activity across a population of neurons, it is important to recognize that not all neurons will have the same level of noise. This variability arises from the distinct intrinsic properties of each neuron, such as membrane resistance, capacitance, and the distribution of ion channels. These properties influence how each neuron responds to external inputs and, as a result, contribute to differences in the noise characteristics observed in the recorded activity.

To account for this variability, we can model the noise for each neuron i with a neuron-specific variance σ_i^2:

$$\langle z_{i,t} \rangle = 0, Var(z_{i,t}) = \sigma_i^2 \tag{2.4}$$

2.1.4 Summary of the base model

With all of the effects described above, the base model for the neural activity reads

$$s_{i,t} = \sum_{k=1}^{K} (x_t^{(k)} + \delta x_t^{(k)}) e_i^{(k)} + z_{i,t}, \quad Cov(z_{i,t_1}, z_{j,t_2}) = \sigma_i^2 \delta_{ij} \delta_{t_1 t_2} \tag{2.5}$$

Parameter	Description	Additional information
N	Number of neurons	
T	Number of time bins in the recording	
K	Dimensionality of latent activity	
i	Neuron index	in range $1 \dots N$
t	Time bin index	in range $1 \dots T$
k	index of a latent mode	in range $1 \dots K$
$e^{(k)}$	k-th mode of activity	Normalized: $\lvert e^{(k)} \rvert^2 = N$
$x^{(k)}$	temporal dynamics of k-th mode of activity	Trial-invariant
$\delta x^{(k)}$	fluctuations of $x^{(k)}$	$\text{Cov}\left(\delta x_t^{(k)}, x_{t'}^{(k')}\right) = (\xi^{(k)})^2 \delta_{k,k'} \Delta_{t,t'}$
z	noise unrelated to latent processes	$\text{Cov}\left(z_{i,t}, z_{j,t'}\right) = \sigma_i^2 \delta_{i,j} \delta_{t,t'}$

2.1.5 Accuracy measures

The accuracy of PCA can be evaluated in various ways, depending on the specific aspect of the low-dimensional representation that is of interest. For instance, one might focus on the variance explained by each principal component or on how precisely the relevant latent dimensions are captured. Below, we define two measures that may be particularly relevant for researchers studying neural trajectories.

Let's define the notation for the output of PCA: let $v^{(k)}$ represent k-th principal component (normalized to $|v^{(k)}|^2 = N$, and $y_t^{(k)}$ the score, or the projection of our data on the k-th principal component:

$$y_t^{(k)} = \frac{1}{N} \sum_i s_{i,t} v_i^{(k)} \tag{2.6}$$

Error of the estimation of the direction of the principal component

When analyzing population neural activity, researchers are often interested in understanding the collective behavior of neurons—specifically, identifying groups of neurons that tend to fire together, indicating coordinated activity. This coordinated activity can reflect underlying functional relationships, such as shared involvement in processing specific stimuli, participating in the same neural circuit, or contributing to a particular cognitive or motor function.

PCA determines such groups by finding the dimensions that correspond to directions in the data where the variability of neural activity is greatest. The weights (or loadings) for each neuron in a principal component indicate how much the activity of this neuron contributes to the component. If several neurons have high loadings on the same principal component, it suggests that these neurons tend to co-activate or fire together, contributing to the same underlying pattern of population activity.

In terms of the notation introduced above, it means that we are interested in finding the components $e_i^{(k)}$. To measure how well it can be done with PCA, we can compare the modes $v^{(k)}$ obtained with PCA (principal components) with the true latent modes $e^{(k)}$. A standard way to do it is by introducing cosine similarity, which measured how much a mode obtained by PCA is orthogonal to the other true latent modes :

$$R^{(k_1,k_2)} = \frac{v^{(k_1)} \cdot e^{(k_2)}}{|v^{(k_1)}||e^{(k_2)}|} \tag{2.7}$$

However, in cases where neurons respond to combinations of stimuli or tasks (i.e., mixed selectivity), it may be useful to examine how well PCA captures these complex response patterns on an individual neuron basis. This can reveal whether the principal components are accurately reflecting the multidimensional nature of activity of individual neurons. To do this, we can calculate the error in estimation for each neuron separately using the squared difference:

$$\rho_i^{(k_1,k_2)} = \frac{1}{2}\left(v_i^{(k_1)} - e_i^{(k_1)}\right)\left(v_i^{(k_2)} - e_i^{(k_2)}\right) \tag{2.8}$$

We see that it vanishes when $v_i^{(k_1)}$ and $e_i^{(k_2)}$ are equal. We can see the connection of this measure with the cosine product if we average this error over the neurons:

$$\left\langle \rho^{(k_1,k_2)} \right\rangle_i = \frac{1}{N}\sum_{i=1}^{N} \rho_i^{(k_1,k_2)} = \frac{1}{2N}\sum_{i=1}^{N}\left(v_i^{(k_1)}v_i^{(k_2)} + e_i^{(k_1)}e_i^{(k_2)} - v_i^{(k_1)}e_i^{(k_2)} - v_i^{(k_2)}e_i^{(k_1)}\right) \tag{2.9}$$

Since by definition all $v^{(k)}$ and $e^{(k)}$ are normalized to N, this can be simplified into

$$\left\langle \rho^{(k_1,k_2)} \right\rangle_i = \delta_{k_1,k_2} - \frac{R^{(k_1,k_2)} + R^{(k_2,k_1)}}{2} \tag{2.10}$$

This means that on average $\rho_i^{(k_1,k_2)}$ will be smallest when $v^{(k_1)}$ is aligned with $e^{(k_1)}$, and $v^{(k_2)}$ is aligned with $e^{(k_2)}$. Thus, we have introduced a neuron-level error measure for estimating latent modes, which naturally relates to the more widely used cosine similarity.

Error of the estimation of the shape of the neural trajectory

A significant portion of research on neural population activity involves extracting and interpreting neural trajectories. Researchers often compare the shapes of neural trajectories across different conditions, such as varying sensory inputs, behavioral tasks, or stages of a decision-making process. For instance, in decision-making tasks, a key focus might be on identifying the point at which neural trajectories diverge under different choices or stimuli, which can shed light on when and how the brain commits to a particular decision. In other contexts, researchers may search for attractor points—stable states toward which the neural system evolves—indicating a potential resting or default mode of brain activity.

The reliability of these analyses, however, is heavily dependent on the accuracy of the reconstructed neural trajectories. If the reconstructed trajectory shapes are dominated by noise rather than reflecting true underlying neural dynamics, any conclusions drawn from these analyses will be misleading. For example, noise-driven fluctuations could be mistakenly interpreted as meaningful divergences in decision-making processes, or as spurious attractor points that do not correspond to real stable states in the brain.

Therefore, it is important to understand how closely the reconstructed neural trajectories match the true latent dynamics of the neural system. In terms of the model notation that we have introduced before, we would like to know how much $y_t^{(k_1)}$ diverges from $x_t^{(k_2)}$. An easy way to introduce an error measure is to take the average squared distance:

$$\epsilon^{(k_1,k_2)} = \frac{1}{T}\sum_{t=1}^{T}(y_t^{(k_1)} - x_t^{(k_1)})(y_t^{(k_2)} - x_t^{(k_2)}) \tag{2.11}$$

2.2 Electrophysiology model

Before we introduced modifications to the model that account for the structure and characteristics common for all types of neural population recordings, regardless of the method used to obtain the

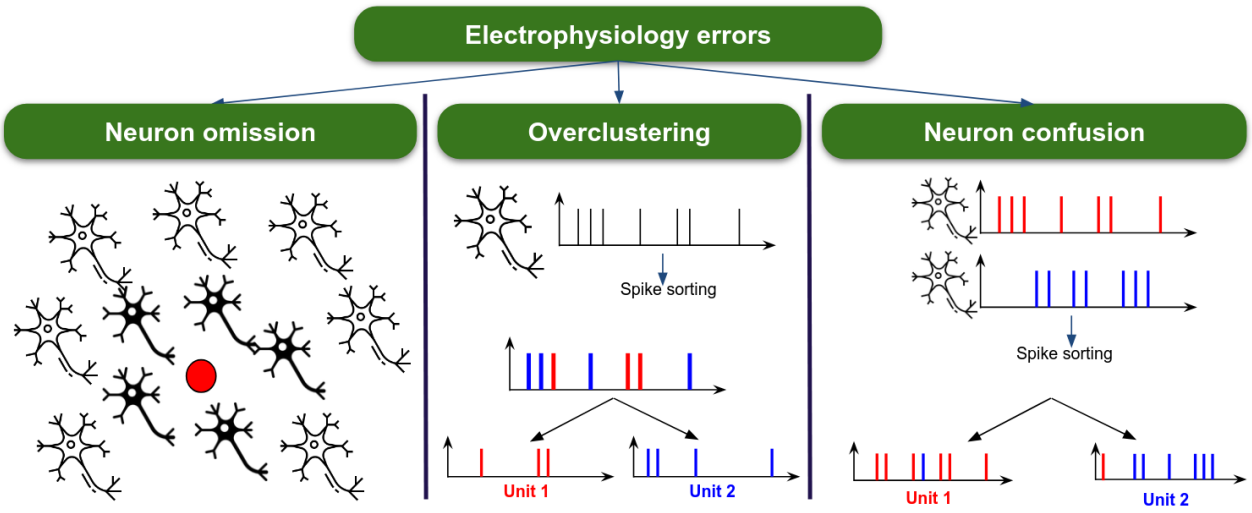

Figure 2.1: Main types of problems with extracellular electrophysiology recordings. Left: Sparsity of recorded neurons. **Middle:** Overclustering. **Right:** Wrong spike attribution.

data. However, it is important to acknowledge that each recording method introduces systematic errors and potential data corruption. Our goal is to model the most prominent of these effects. In this subsection, we will list and explain the most commonly encountered issues with extracellular electrophysiology recordings and demonstrate how to model them.

2.2.1 Sparsity of recorded neurons

As we stated before in the literature review, traditional electrophysiology techniques, such as single-unit recordings and multi-electrode arrays (MEAs), are limited in the number of neurons they can record simultaneously. Electrodes can only capture activity from neurons that are in close proximity to their tips (Fig. 2.1, left), meaning that only a tiny fraction of neurons within the immediate vicinity of each electrode are sampled, leaving the majority of the tissue unrecorded. While high-density electrodes offer better spatial resolution and can sample from more neurons, they still cover only a small portion of the overall neural population. Furthermore, the insertion of electrodes into brain tissue can cause damage, which restricts the number of electrodes that can be safely implanted, further limiting the scope of neuron sampling.

In addition to the limitations imposed by electrode coverage, there is the challenge of recording from ”silent” neurons. These are neurons that do not exhibit significant spontaneous or evoked action potentials within the timeframe or specific conditions of a typical recording session. Silent neurons may be active only under certain conditions not present during the recording, or they may have inherently low firing rates that are difficult to detect with standard techniques. The detection thresholds of electrophysiological equipment may not be sensitive enough to capture the very small or infrequent signals generated by these neurons, causing them to fall below the detection limit. Moreover, the few spikes produced by such neurons may not be sufficient to form distinct clusters during spike sorting, leading to these spikes being either discarded or incorrectly attributed to other neurons.

2.2.2 Spike sorting: wrong spike attribution

Another important issue in spike sorting occurs when a cluster predominantly composed of spikes from one neuron contains a small percentage of spikes originating from another neuron (Fig. 2.1, right). This misattribution can lead to significant errors in interpreting neural data, particularly when studying the firing patterns or functional roles of individual neurons.

Wrong spike attribution typically arises when the spike sorting algorithm fails to distingiush spikes from closely located neurons with similar waveform shapes. Factors such as noise, subtle variations in spike morphology, or overlapping spike trains can cause the algorithm to mistakenly

assign spikes from different neurons to the same cluster. This is particularly problematic in dense neural recordings where neurons are in close proximity and their electrical signals are difficult to disentangle. This issue can lead to incorrect conclusions about the neuron's role in a neural circuit, its response to stimuli, or its contribution to behavior. In extreme cases, the functional identity of the neuron may be misrepresented, potentially skewing the results of the entire study.

It is important to note that this type of confusion can only occur between neurons situated in close proximity to the same recording site. Each recording site has a limited range within which it can effectively detect signals. Therefore, confusion in spike attribution arises primarily among neurons within this range, as their signals are most likely to be similar and overlap on the recording equipment. Neurons located outside this region are too distant to contribute significantly to the signals detected by that particular site, making it less likely for their spikes to be confused with those of nearby neurons.

We model this effect as follows: Let's assume that a typical number of neurons surrounding a given electrode site is given by b. This value will depend of the type of tissue. Then, we will split all N neurons randomly into the groups of b neurons each. We assume that all of the neurons in the group have the same probability of attributing spike from a given neuron to another neuron in the same group. We call this probability p_c. Then, the matrix M_c that models the wrong spike attribution can be written as

$$P^{-1}\begin{pmatrix} \begin{matrix} 1-(b-1)p_c & p_c & \dots & p_c \\ p_c & 1-(b-1)p_c & \dots & p_c \\ \vdots & \vdots & \ddots & \vdots \\ p_c & p_c & \dots & 1-(b-1)p_c \end{matrix} & & & \\ & \begin{matrix} 1-(b-1)p_c & p_c & \dots & p_c \\ p_c & 1-(b-1)p_c & \dots & p_c \\ \vdots & \vdots & \ddots & \vdots \\ p_c & p_c & \dots & 1-(b-1)p_c \end{matrix} & & \\ & & \ddots & \end{pmatrix} P \tag{2.12}$$

where matrix P performs a random permutation of neurons to ensure that the groups of size b are selected at random.

2.2.3 Spike sorting: overclustering

Signal corruption in electrophysiology recordings often arises from challenges associated with spike sorting, a process that involves identifying neuronal spikes and clustering them based on similarities in spike shapes.

Overclustering occurs when a spike sorting algorithm mistakenly divides spikes from the same neuron into multiple clusters (Fig. 2.1, middle). This typically happens when the algorithm interprets slight variations in features like amplitude, waveform shape, or duration—often due to electrode noise or changes in recording conditions—as distinct neuron signals.

Underclustering, on the other hand, combines distinct signals from different neurons into a single cluster, resulting in a loss of resolution that obscures the unique firing patterns of individual neurons. This issue is particularly problematic in studies that require precise neuron-level information, such as neural coding.

While overclustering is not ideal, it is often considered a safer error because it preserves the distinct identities of spikes, even if they originate from the same neuron. Unlike underclustering, which can obscure important data, overclustering allows for potential correction through post-processing or manual cluster merging. For this reason, many algorithms err on the side of caution, preferring to overcluster rather than risk merging spikes from different neurons.

We model this effect as follows: Let f_s represent the fraction of the total recorded neurons whose signals are split into multiple clusters. Accordingly, we randomly select $f_s N$ neurons to

undergo this splitting process. For this, we use a random permutation matrix P to shuffle the order of the neurons, and take the first $f_s N$ out of them.

For each neuron i chosen for splitting, we assume its signal is divided into two clusters. (In practice, the signal from one neuron can be split into more than two clusters, but we model a simplified case.) The first cluster contains a fraction p_i of the original signal, while the second cluster contains the remaining fraction $1 - p_i$. The value of p_i is drawn from a uniform random distribution over the interval $[0, 1]$.

Then, this splitting can be written as a transformation of the initial vector e_i with a matrix M_s.

$$M_s = \left(\begin{array}{cccc|cccc} p_1 & 0 & \dots & 0 & & & & \\ 0 & p_2 & \dots & 0 & & & & \\ \vdots & \vdots & \ddots & \vdots & & 0 & & \\ 0 & 0 & \dots & p_{f_s N} & & & & \\ \hline & & & & 1 & 0 & \dots & 0 \\ & & & & 0 & 1 & \dots & 0 \\ & & 0 & & \vdots & \vdots & \ddots & \vdots \\ & & & & 0 & 0 & \dots & 1 \\ \hline 1-p_1 & 0 & \dots & 0 & & & & \\ 0 & 1-p_2 & \dots & 0 & & 0 & & \\ \vdots & \vdots & \ddots & \vdots & & & & \\ 0 & 0 & \dots & 1-p_{f_s N} & & & & \end{array}\right) \tag{2.13}$$

2.2.4 Data preprocessing: temporal smoothing

One of the most important preprocessing steps is temporal smoothing of the neural activity. Temporal smoothing serves several purposes, most notably:

1. Noise reduction: By smoothing the data over time, we can mitigate the effects of high-frequency noise, making the underlying signal more apparent. This is especially important in neural recordings where the inherent noise can obscure the true neural activity.

2. Continuous Representation: On the timescale of neural activity recording, spike duration is very short, appearing as discrete events. To facilitate analysis, these spike events are often converted into firing rates, providing a continuous representation of neural activity. Continuous data is generally easier to visualize, compare, and interpret across different neurons or experimental conditions.

Smoothing the data, while beneficial, introduces temporal correlations into the noise. This is a significant departure from the classical assumption made in spiked covariance model, where the noise in different time bins is assumed to be independent.

There are several methods to achieve temporal smoothing, each with its own advantages. The most used methods include

- Sliding Window (Moving Average): This method involves averaging the data over a window that moves across the time series. It's straightforward and effective for reducing noise, though it may also blur finer details of the neural activity.

- Kernel Smoothing: A more sophisticated approach, kernel smoothing involves convolving the neural data with a kernel function (such as a Gaussian). This method allows for more flexibility in how the data is smoothed and can preserve certain features of the signal better than simple averaging. Sliding window can be viewed as a particular case of kernel smoothing, where the kernel is taken as constant for a certain width.

We will model this effect by introducing the kernel G for the convolution.

2.2.5 Summary of the electrophysiology model

Therefore, the model for the electrophysiology can be written in a following way:

$$s_{i,t} = \sum_{\tau=0}^{T-1} G_{t-\tau} \left[\sum_k (x_\tau^{(k)} + \delta x_\tau^{(k)})(M_s M_c e)_i^{(k)} + z_{i,\tau} \right] \tag{2.14}$$

Parameter	Description	Additional information
N	Number of clusters detected by spike sorting	
T	Number of time bins in the recording	
K	Dimensionality of latent activity	
i	Cluster index	in range $1 \dots N$
t	Time bin index	in range $1 \dots T$
k	Index of a latent mode	in range $1 \dots K$
$e^{(k)}$	k-th mode of activity	Normalized: $\lvert e^{(k)}\rvert^2 = 1$
$x^{(k)}$	Temporal dynamics of k-th mode of activity	Trial-invariant
$\delta x^{(k)}$	Fluctuations of $x^{(k)}$	$\mathrm{Cov}\left(\delta x_t^{(k)}, x_{t'}^{(k')}\right) = (\xi^{(k)})^2 \delta_{k,k'} \Delta_{t,t'}$
z	Noise unrelated to latent processes	$\mathrm{Cov}\left(z_{i,t}, z_{j,t'}\right) = \sigma_i^2 \delta_{i,j} \delta_{t,t'}$
G	Convolution kernel used for temporal smoothing	

2.2.6 Large N, T regime and change of the notation

The model for calcium imaging data that we introduced earlier defines a data with T time bins and N neurons. Our calculation of the accuracy measures ρ and ϵ will be done for the limit $T, N \to \infty$, but their ratio $\frac{T}{N} = \alpha$ stays finite. Some parameters, in order to have non-negligible effects on the system in this limit, need to be scaled appropriately, so below we will introduce this scaling explicitly. We will also rename some other combinations of the variables introduced above for convenience.

- In the base model, vectors $e_i^{(k)}$ represent the modes of activity of the group of neurons that we wanted to record. However, the groups of neurons that are detected after the spike sorting are different. We introduce $\bar{e}_i$ as the modes of activity altered by the wrong spike attribution and overclustering:

$$M_s M_c e^{(k)} = \frac{|M_s M_c e^{(k)}|}{\sqrt{N}} \bar{e}^{(k)}, \quad |\bar{e}^{(k)}|^2 = N \tag{2.15}$$

- We will make a simplification and replace all expressions $|M_s M_c e^{(k)}|$ by an average value $\sqrt{\left\langle |M_s M_c e^{(k)}|^2 \right\rangle}$ over all possible realizations of the matrices M_c,M_s and the directions of vector $e^{(k)}$. Since the neurons for signal confusion and splitting are chosen randomly and independently, we approximate this expression further by

$$\left\langle |M_s M_c e^{(k)}|^2 \right\rangle \approx \frac{\left\langle |M_s e^{(k)}|^2 \right\rangle \left\langle |M_c e^{(k)}|^2 \right\rangle}{\left\langle |e^{(k)}|^2 \right\rangle} \tag{2.16}$$

  First, we can write the average value of $|M_c e^{(k)}|^2$:

$$\begin{aligned} \left\langle |M_c e^{(k)}|^2 \right\rangle &= \left\langle \sum_{i=1}^N \left[(1-(b-1)p_c)^2 (e_i^{(k)})^2 + \sum_{\text{j around same electrode as i}} p_c^2 (e_j^{(k)})^2 \right] \right\rangle = \\ &= \sum_{i=1}^N \left[(1-(b-1)p_c)^2 \left\langle (e_i^{(k)})^2 \right\rangle + \sum_{\text{j around same electrode as i}} p_c^2 \left\langle (e_j^{(k)})^2 \right\rangle \right] = \\ &= ((1-(b-1)p_c)^2 + (b-1)p_c^2) N \end{aligned} \tag{2.17}$$

Second, we write the average value of $|M_s e^{(k)}|^2$:

$$\left\langle |M_s e^{(k)}|^2 \right\rangle = \left\langle \sum_{\text{i not overclustered}} e_i^2 + \sum_{\text{j overclustered}} e_j^2 (p_j^2 + (1-p_j)^2) \right\rangle = \\ = \left\langle \sum_{i=1}^{N} e_i^2 \right\rangle + \left\langle \sum_{\text{j overclustered}} e_j^2 (2p_j^2 - 2p_j) \right\rangle = \left(1 - \frac{f_s}{3}\right) N \tag{2.18}$$

So, overall we get a factor of

$$\left\langle |M_s M_c e^{(k)}|^2 \right\rangle \approx ((1-(b-1)p_c)^2 + (b-1)p_c^2) \left(1 - \frac{f_s}{3}\right) N \tag{2.19}$$

- We can write explicitly how the kernel G acts on z. For this, we will redefine the notation: the z-term after the convolution with the kernel will be called $\bar{z}_{i,t}$.

$$\bar{z}_{i,t} := (G * z)_{i,t} \tag{2.20}$$

Before the temporal smoothing, we were assumming that the noise is uncorrelated in time. After the application of the kernel the covariance matrix of the noise can be calciulating in the following way:

$$\text{Cov}(\bar{z}_{i,t_1}, \bar{z}_{j,t_2}) = \text{Cov}\left(\sum_{\tau_1=0}^{T-1} G_{t_1-\tau_1} z_{i,\tau_1}, \sum_{\tau_2=0}^{T-1} G_{t_2-\tau_2} z_{j,\tau_2} \right) \tag{2.21}$$

Since the initial noise is uncorrelated in time, this can be rewritten as

$$\sum_{\tau=0}^{T-1} \text{Cov}\left(G_{t_1-\tau} z_{i,\tau}, G_{t_2-\tau} z_{j,\tau}\right) = \sum_{\tau=0}^{T-1} G_{t_1-\tau} G_{t_2-\tau} \delta_{ij} \sigma_i^2 \tag{2.22}$$

Thus, we see that the noise for different neurons has different variance

$$\text{Var}(\bar{z}_{i,t}) = \sum_{\tau=0}^{T-1} G_{t-\tau}^2 \sigma_i^2 \tag{2.23}$$

but the same correlation matrix in time

$$\text{Corr}(\bar{z}_{i,t_1}, \bar{z}_{i,t_2}) = \frac{\sum_{\tau=0}^{T-1} G_{t_1-\tau} G_{t_2-\tau}}{\sum_{\tau=0}^{T-1} G_\tau^2} \tag{2.24}$$

We will call the resulting correlation matrix Z_{t_1,t_2}, and denote variance for the noise of the neuron i as $\bar{\sigma}_i^2 N$, where i is the number of neuron. Scaling the variance with N allows to have non-negligible noise in the limit $T, N \to \infty$.

$$\text{Corr}(\bar{z}_{i,t_1}, \bar{z}_{j,t_2}) = \delta_{ij} Z_{t_1,t_2}, \quad \text{Var}(\bar{z}_{i,t}) = \bar{\sigma}_i^2 N \tag{2.25}$$

- In the same way, we can write how the kernel G acts on $\delta x^{(k)}$. Redefine the notation: the $\delta x_t^{(k)}$-term after the convolution with the kernel will be called $\delta \bar{x}_t^{(k)}$. We will also absorb the coefficient appearing due to the normalization of $\bar{e}^{(k)}$:

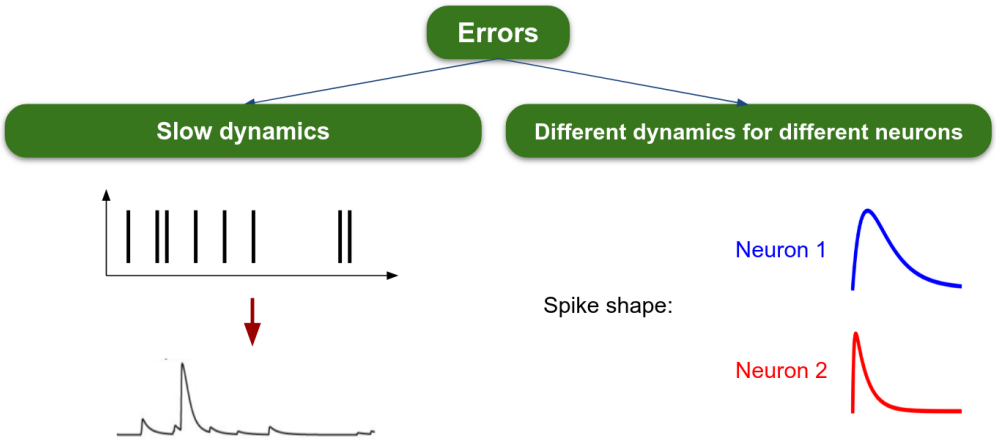

Figure 2.2: Data distortion by calcium imaging. Left: Slow dynamics of the calcium indicators. Right: Different rise and decay times for different neurons.

$$\delta\bar{x}_t^{(k)} := \frac{|M_s M_c e^{(k)}|}{\sqrt{N}}(G * \delta x^{(k)})_t \tag{2.26}$$

The covariance can be written as

$$\begin{aligned}\mathrm{Cov}\left(\delta\bar{x}_{t_1}^{(k_1)}, \delta\bar{x}_{t_2}^{(k_2)}\right) = \frac{|M_s M_c e^{(k)}|^2}{N}\mathrm{Cov}\left(\sum_{\tau_1=0}^{T-1} G_{t_1-\tau_1}\delta x_{\tau_1}^{(k_1)}, \sum_{\tau_2=0}^{T-1} G_{t_2-\tau_2}\delta x_{\tau_2}^{(k_2)}\right) = \\ = \delta_{k_1k_2}(\xi^{(k_1)})^2 \sum_{\tau_1=0}^{T-1} G_{t_1-\tau_1} \sum_{\tau_2=0}^{T-1} G_{t_2-\tau_2}\Delta_{\tau_1\tau_2}\end{aligned} \tag{2.27}$$

We will call the resulting correlation matrix $\bar{\Delta}_{t_1,t_2}$, and denote variance for $\delta\bar{x}^{(k)}$ as $(\bar{\xi}^{(k)})^2$:

$$\mathrm{Corr}(\delta\bar{x}_{t_1}^{(k_1)}, \delta\bar{x}_{t_2}^{(k_2)}) = \delta_{k_1k_2}\bar{\Delta}_{t_1,t_2}, \quad \mathrm{Var}(\delta\bar{x}_t^{(k)}) = (\bar{\xi}^{(k)})^2 \tag{2.28}$$

- The signal $x^{(k)}$ modified by the selected kernel G will be denoted as:

$$\bar{x}^{(k)} = \frac{|M_s M_c e^{(k)}|}{\sqrt{N}}(G * x^{(k)})_t \tag{2.29}$$

Here we absorb the factor (2.19) that emerged from acting with matrices M_s and M_c on the modes $e^{(k)}$.

Overall, in the new notation the model for exracellular electrophysiology can be written as

$$s_{i,t} = \sum_{k=1}^{K}(\bar{x}_t^{(k)} + \delta\bar{x}_t^{(k)})\bar{e}_t^{(k)} + \bar{z}_{i,t} \tag{2.30}$$

2.3 Calcium imaging model

As in the case with the model for electrophysiology recordings, this model will be built by modifying the base model of the data (2.5).

2.3.1 Double-exponential kernel for the fluorescence

As we stated before in the literature review, calcium imaging relies on indirect reporting of spikes through the fluorescence of indicators, which has slow dynamics (Fig. 2.2, left). For a given neuron, the fluorescence dynamics of an individual spike is commonly approximated by a double-exponential shape:

- Fast onset: This represents the rapid initial increase in fluorescence as calcium ions bind to the indicator, moving the system toward equilibrium. The time it takes for this initial rise is denoted as τ_r.
- Slow decay: This represents the slower, prolonged decline in fluorescence as calcium gradually unbinds from the indicator. The time associated with this decay is denoted as τ_d.

The shape of the fluorescence signal corresponding to a single spike can thus be described by the following equation:

$$F(t, \tau_r, \tau_d) = e^{-\frac{t}{\tau_d}} - e^{-\frac{t}{\tau_r}} \tag{2.31}$$

where t is the time since the beginning of the spike.

The indirect reporting of latent dynamics by fluorescence can thus be modeled by a convolution with the discretized version of the shape for the single spike:

$$\bar{x}_t^{(k)} = (F * x^{(k)})(t) = \sum_{\tau=0}^{T} x_\tau^{(k)} F(t - \tau, \tau_r, \tau_d) \tag{2.32}$$

2.3.2 Neuron-to-neuron kernel variability

The timing of calcium fluorescence signals varies among neurons (Fig. 2.2, right) due to several factors. One key reason is the inherent differences between neurons, such as the speed at which calcium is cleared from the cell or variations in the levels and types of calcium-binding proteins present. Additionally, differences in the expression levels of the calcium indicator itself can influence the fluorescence signal. If the indicator concentration is too high, it can buffer the calcium, changing the behavior of the calcium signals and affecting the rise and decay times of the fluorescence traces.

To model these variations, we define a family of discrete kernels $F_i(t, \tau_r, \tau_d)$, each kernel reflecting the shape of a single spike for a given neuron i. Then, we assume that for each neuron the rise and decay times (τ_r and τ_d) are sampled from a known joint distribution $p_{Ca}(\tau_r, \tau_d)$. For certain indicators and brain regions, these distributions have been measured (Wei et al. (2020)). Using this distribution, we can calculate the average shape of the fluorescence signal:

$$F(t) = \int_0^\infty d\tau_r \int_0^\infty d\tau_d \left[(e^{-\frac{t}{\tau_d}} - e^{-\frac{t}{\tau_r}}) p_{Ca}(\tau_r, \tau_d) \right] \tag{2.33}$$

Alternatively, we can write the kernel $F_i(t)$ as a sum of mean part $F(t)$ and a centered random variable $\delta F_i(t)$.

$$F_i(t) = F(t) + \delta F_i(t), \quad \langle \delta F_i(t) \rangle = 0 \tag{2.34}$$

To simplify the problem, instead of working in terms of random variables τ_r, τ_d we would like to directly work with the distribution of time bins of the kernels $\delta F_i(t)$. In order to be able to carry out the calculation, we would like to simplify this distribution by approximating it with Gaussian distribution of mean zero and variance Ξ.

$$\delta F_i(t) \sim \mathcal{N}(0, \Xi) \tag{2.35}$$

where Ξ can be found by knowing the distribution of rise and decay times:

$$\Xi(t_1.t_2) = \langle F(t_1,\tau_r,\tau_d)F(t_2,\tau_r,\tau_d)\rangle_{p_{Ca}} - \bar{F}(t_1)\bar{F}(t_2) =$$
$$= \int_0^\infty d\tau_r \int_0^\infty d\tau_d \left[(e^{-\frac{t_1}{\tau_d}} - e^{-\frac{t_1}{\tau_r}})(e^{-\frac{t_2}{\tau_d}} - e^{-\frac{t_2}{\tau_r}})p_{Ca}(\tau_r,\tau_d)\right] - \bar{F}(t_1)\bar{F}(t_2) \tag{2.36}$$

2.3.3 Data preprocessing: temporal smoothing

As in the case of extracellular electrophysiology, calciun imaging data is usually denoised by temporal smoothing. We will model it in the same way as in the case of electrophysiology, by introducing a convolution kernel G.

2.3.4 Summary of the calcium imaging model

The overall model for the calcium imaging case can be written as

$$s_{i,t} = \sum_{\tau_2} G_{t-\tau_2}\left[\sum_k \sum_{\tau_1} (F_{\tau_2-\tau_1} + \delta F_{i,\tau_2-\tau_1})\left(x_{\tau_1}^{(k)} + \delta x_{\tau_1}^{(k)}\right) e_i^{(k)} + z_{i,\tau_2}\right] \tag{2.37}$$

Parameter	Description	Additional information
N	Number of neurons	
T	Number of time bins in the recording	
K	Dimensionality of latent activity	
i	Cluster index	in range $1 \dots N$
t	Time bin index	in range $1 \dots T$
k	Index of a latent mode	in range $1 \dots K$
$e^{(k)}$	k-th mode of activity	Normalized: $\lvert e^{(k)}\rvert^2 = N$
$x^{(k)}$	Temporal dynamics of k-th mode of activity	Trial-invariant
$\delta x^{(k)}$	Fluctuations of $x^{(k)}$	$\mathrm{Cov}\left(\delta x_t^{(k)}, x_{t'}^{(k')}\right) = (\xi^{(k)})^2 \delta_{k,k'} \Delta_{t,t'}$
z	Noise unrelated to latent processes	$\mathrm{Cov}\left(z_{i,t}, z_{j,t'}\right) = \sigma_i^2 \delta_{i,j} \delta_{t,t'}$
G	Convolution kernel used for temporal smoothing	
$\bar{F}$	Average fluorescence profile for a single spike	
δF_i	Difference between $\bar{F}$ and the fluorescence profile for neuron i	$\delta F_i \sim \mathcal{N}(0, \Xi)$

2.3.5 Large N, T limit and change of the notation

As in the case with electrophysiology model, we need to adapt the model for the limit $T, N \to \infty$. We will also group and rename some of the combinations of the variables introduced above for convenience.

- Similarly to the case of the Electrophysiology model, we can introduce

$$\bar{z}_{i,t} = \sum_\tau G_{t-\tau} z_{i,\tau} \tag{2.38}$$

which has the correlation matrix

$$Z_{t_1,t_2} = \frac{\sum_{\tau=0}^{T-1} G_{t_1-\tau} G_{t_2-\tau}}{\sum_{\tau=0}^{T-1} G_{t-\tau}^2} \tag{2.39}$$

and variance

$$\bar{\sigma}_i^2 N = \sum_{\tau=1}^{T} G_\tau^2 \sigma_i^2 \tag{2.40}$$

- Since convolution is an associative operation, we can think of two consecutive convolution as of one convolution with another kernel. Let's introduce new notation for these kernels:

$$\bar{F}_t = \sum_\tau G_{t-\tau} F_\tau, \quad \delta\bar{F}_{i,t} = \sum_\tau G_{t-\tau}\delta F_{i,\tau} \tag{2.41}$$

We can see that as a sum of Gaussian variables with zero mean, $\delta\bar{F}_{i,t}$ is still a multivariate Gaussian variable with zero mean, and covariance

$$\begin{aligned}\mathrm{Cov}(\delta\bar{F}_{i,t_1}, \delta\bar{F}_{j,t_2}) &= \mathrm{Cov}(\sum_{\tau_1} G_{t_1-\tau_1}\delta F_{i,\tau_1}, \sum_{\tau_2} G_{t_2-\tau_2}\delta F_{j,\tau_2}) = \\ &= \delta_{ij}\sum_{\tau_1,\tau_2} G_{t_1-\tau_1}G_{t_2-\tau_2}\Xi_{\tau_1,\tau_2} = \delta_{ij}\bar{\Xi}_{t_1,t_2}N\end{aligned} \tag{2.42}$$

where N has been added to stay in the regime of non-negligible difference between kernels of different neurons as we consider the limit $N \to \infty$.

This way, the model can be rewritten as

$$s_{i,t} = \sum_k \sum_{\tau_1} (\bar{F}_{t-\tau_1} + \delta\bar{F}_{i,t-\tau_1}) \left(x_{\tau_1}^{(k)} + \delta x_{\tau_1}^{(k)}\right) \bar{e}_i^{(k)} + \bar{z}_{i,t} \tag{2.43}$$

- We can further simplify the notation by introducing $\bar{x}_t^{(k)} = \sum_\tau \bar{F}_\tau x_\tau^{(k)}$, $\delta\bar{x}_t^{(k)} = \sum_\tau \bar{F}_\tau \delta x_\tau^{(k)}$ - a result of application of the average fluorescence kernel to the signal and its fluctuations. The covariance matrix for $\delta\bar{x}_t^{(k)}$ can be written as

$$\begin{aligned}\mathrm{Cov}\left(\delta\bar{x}_{t_1}^{(k_1)}, \delta\bar{x}_{t_2}^{(k_2)}\right) &= \mathrm{Cov}\left(\sum_{\tau_1=0}^{T-1} \bar{F}_{t_1-\tau_1}\delta x_{\tau_1}^{(k_1)}, \sum_{\tau_2=0}^{T-1} \bar{F}_{t_2-\tau_2}\delta x_{\tau_2}^{(k_2)}\right) = \\ &= \delta_{k_1k_2}(\xi^{(k_1)})^2 \sum_{\tau_1=0}^{T-1} \bar{F}_{t_1-\tau_1} \sum_{\tau_2=0}^{T-1} \bar{F}_{t_2-\tau_2}\Delta_{\tau_1\tau_2}\end{aligned} \tag{2.44}$$

We will call the resulting correlation matrix $\bar{\Delta}_{t_1,t_2}$, and denote variance for $\delta\bar{x}^{(k)}$ as $(\bar{\xi}^{(k)})^2$:

$$\mathrm{Corr}(\delta\bar{x}_{t_1}^{(k_1)}, \delta\bar{x}_{t_2}^{(k_2)}) = \delta_{k_1k_2}\bar{\Delta}_{t_1,t_2}, \quad \mathrm{Var}(\delta\bar{x}_t^{(k)}) = (\bar{\xi}^{(k)})^2 \tag{2.45}$$

We will neglect double delta-term $\sum_\tau \delta\bar{F}_{i,\tau}\delta x_t^{(k)}$, assuming these fluctuations are small.

- Lastly, for coherence of notation with the electrophysiology model (2.30), we will rename $\bar{e}_i^{(k)} = e_i^{(k)}$.

Thus, after all of the simplifications of notation, we get the following model:

$$s_{i,t} = \sum_{k=1}^{K} (\bar{x}_t^{(k)} + \sum_\tau \delta\bar{F}_{i,t-\tau}x_\tau^{(k)} + \delta\bar{x}_t^{(k)})\bar{e}_i^{(k)} + \bar{z}_{i,t} \tag{2.46}$$

2.4 Outline of the calculation

2.4.1 Constructing the partition function

After the change of the notation, both models look almost exactly the same (2.30), (2.46). The only difference between them is the presence of the term containing δF for the calcium imaging case. This allows us to carry out the replica calculation for calcium imaging model, and then simply set Ξ=0 for the case of electrophysiology.

We will estimate the average value of ϵ and ρ with the replica method. For this, we will formulate our problem in the formalism of statistical mechanics. We consider a system where every state is represented by the set of K N-dimensional orthonormal vectors $\mathbf{v}$. This way, vector $v^{(k)}$ will describe the direction of k-th principal component. To define the energy of this system, we introduce the quadratic form

$$\begin{aligned}\frac{1}{N^2}\sum_{k=1}^{K}\sum_{i,j=1}^{N} v_i^{(k)} C_{ij} v_j^{(k)} &= \sum_{k=1}^{K}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} v_i^{(k)} s_{i,t}\right)^2 - \sum_{k=1}^{K}\left(\frac{1}{TN}\sum_{i,t} v_i^{(k)} s_{i,t}\right)^2 \\ &= \sum_{k=1}^{K}\left(\frac{1}{T}\sum_{t}(y_t^{(k)})^2\right) - \sum_{k=1}^{K}\left(\frac{1}{T}\sum_{t} y_t^{(k)}\right)^2 . \end{aligned} \tag{2.47}$$

This way, the maximum of this expression will be reached when the direction of all K vectors align with the directions that have the maximal variance in the data, so the first K principal components.

We therefore consider the following partition function:

$$\begin{aligned} Z(\{s_{i,t}\}) &= \int \prod_{k=1}^{K} d\mathbf{v}^{(k)} \int \prod_{k=1}^{K}\prod_{t=1}^{T} dy_t^{(k)} \prod_{k,t}\delta\left(\sum_i (v_i^{(k)})^2 - N\right) \prod_{k_1,k_1:k_1<k_2}\delta\left(\sum_i v_i^{(k_1)} v_i^{(k_2)}\right) \\ &\quad \prod_{k,t}\delta\left(y_t^{(k)} - \frac{1}{N}\sum_i v_i^{(k)} s_{i,t}\right) \times \exp\left(\beta T \frac{1}{N^2}\sum_{k=1}^{K}\sum_{i,j=1}^{N} v_i^{(k)} C_{ij} v_j^{(k)}\right) \end{aligned} \tag{2.48}$$

Then, we can explicitly introduce the definition of cosine similarity R, as well as source terms for accuracy measures ρ and ϵ:

$$\begin{aligned} Z(\{s_{i,t}\}) &= \int \prod_{k=1}^{K} d\mathbf{v}^{(k)} \int \prod_{k=1}^{K}\prod_{t=1}^{T} dy_t^{(k)} \prod_{k,t}\delta\left(\sum_i (v_i^{(k)})^2 - N\right) \prod_{k_1,k_1:k_1<k_2}\delta\left(\sum_i v_i^{(k_1)} v_i^{(k_2)}\right) \\ &\quad \prod_{k,t}\delta\left(y_t^{(k)} - \frac{1}{N}\sum_i v_i^{(k)} s_{i,t}\right) \int \prod_{k_1,k_2=1}^{K} dR^{(k_1,k_2)} \delta\left(R^{(k_1,k_2)} - \frac{1}{N}\sum_i v_i^{(k_1)} \bar{e}_i^{(k_2)}\right) \\ &\times \exp\left(\beta T \frac{1}{N^2}\sum_{k=1}^{K}\sum_{i,j=1}^{N} v_i^{(k)} C_{ij} v_j^{(k)}\right) \exp\left(\sum_{k_1,k_2=1}^{K}\sum_{i=1}^{N} \eta_i^{(k_1,k_2)} \frac{1}{2}(v_i^{(k_1)} - \bar{e}_i^{(k_1)})(v_i^{(k_2)} - \bar{e}_i^{(k_2)})\right) \\ &\times \exp\left(\sum_{k_1,k_2=1}^{K} \gamma^{(k_1,k_2)} \sum_{t=1}^{T} \left(\bar{x}_t^{(k_1)} - y_t^{(k_1)}\right)\left(\bar{x}_t^{(k_2)} - y_t^{(k_2)}\right)\right) \end{aligned} \tag{2.49}$$

Finally, using (2.47), we arrive at

$$
\begin{aligned}
Z\big(\{s_{i,t}\}\big) &= \int \prod_{k=1}^{K} d\mathbf{v}^{(k)} \int \prod_{k=1}^{K}\prod_{t=1}^{T} dy_t^{(k)} \prod_{k,t} \delta\left(\sum_i (v_i^{(k)})^2 - N\right) \prod_{k_1,k_1:k_1<k_2} \delta\left(\sum_i v_i^{(k_1)} v_i^{(k_2)}\right) \\
&\quad \prod_{k,t} \delta\left(y_t^{(k)} - \frac{1}{N}\sum_i v_i^{(k)} s_{i,t}\right) \int \prod_{k_1,k_2=1}^{K} dR^{(k_1,k_2)} \delta\left(R^{(k_1,k_2)} - \frac{1}{N}\sum_i v_i^{(k_1)} \bar{e}_i^{(k_2)}\right) \\
&\times \exp\left(\beta \sum_{k=1}^{K}\left(\sum_t (y_t^{(k)})^2\right) - \beta \frac{1}{T}\sum_{k=1}^{K}\left(\sum_t y_t^{(k)}\right)^2\right) \\
&\times \exp\left(\sum_{k_1,k_2,i} \eta_i^{(k_1,k_2)} \frac{1}{2}(v_i^{(k_1)} - \bar{e}_i^{(k_1)})(v_i^{(k_2)} - \bar{e}_i^{(k_2)}) + \sum_{k_1,k_2,t} \gamma^{(k_1,k_2)}\big(\bar{x}_t^{(k_1)} - y_t^{(k_1)}\big)\big(\bar{x}_t^{(k_2)} - y_t^{(k_2)}\big)\right)
\end{aligned}
\tag{2.50}
$$

2.4.2 Introduction of many replicas

Now, following the outline of the calculation in (1.5.5), we introduce n different replicas:

$$
\begin{aligned}
Z^n\big(\{s_{i,t}\}\big) &= \int \prod_{a=1}^{n}\prod_{k=1}^{K} d\mathbf{v}_a^{(k)} \int \prod_{a=1}^{n}\prod_{k=1}^{K}\prod_{t=1}^{T} \left[dy_{a,t}^{(k)} \delta\left(\sum_i (v_{a,i}^{(k)})^2 - N\right) \delta\left(y_{a,t}^{(k)} - \frac{1}{N}\sum_i v_{a,i}^{(k)} s_{i,t}\right)\right] \\
&\times \prod_{a,k_1<k_2} \delta\left(\sum_i v_{a,i}^{(k_1)} v_{a,i}^{(k_2)}\right) \int \prod_{a,k_1,k_2} dR_a^{(k_1,k_2)} \delta\left(R_a^{(k_1,k_2)} - \frac{1}{N}\sum_i v_{a,i}^{(k_1)} \bar{e}_i^{(k_2)}\right) \times \\
&\times \exp\left(\beta \sum_{a=1}^{n}\sum_{k=1}^{K}\left(\sum_t (y_{a,t}^{(k)})^2\right) - \beta \frac{1}{T}\sum_{a=1}^{n}\sum_{k=1}^{K}\left(\sum_t y_{a,t}^{(k)}\right)^2\right) \times \\
&\times \exp\left(\sum_{a=1}^{n}\sum_{k_1,k_2,i} \eta_i^{(k_1,k_2)} \frac{1}{2}(v_{a,i}^{(k_1)} - \bar{e}_i^{(k_1)})(v_{a,i}^{(k_2)} - \bar{e}_i^{(k_2)})\right) \\
&\times \exp\left(\sum_{a=1}^{n}\sum_{k_1,k_2,t} \gamma^{(k_1,k_2)}\big(\bar{x}_t^{(k_1)} - y_{a,t}^{(k_1)}\big)\big(\bar{x}_t^{(k_2)} - y_{a,t}^{(k_2)}\big)\right)
\end{aligned}
\tag{2.51}
$$

Next step: expressing all of the delta-functions as integrals.

$$
\begin{aligned}
Z^n(\{s_{i,t}\}) &= \int \prod_{a,k} d\mathbf{v}_a^{(k)} \prod_{a,k,t} dy_{a,t}^{(k)} \frac{d\hat{y}_{a,t}^{(k)}}{2\pi} \int \prod_{a,k_1,k_2} dR_a^{(k_1,k_2)} \frac{d\hat{R}_a^{(k_1,k_2)}}{2\pi N} \frac{d\hat{u}_a^{(k_1,k_2)}}{2\pi} \\
&\times \exp\left(\sum_{a,k_1<k_2} \hat{u}_a^{(k_1,k_2)} \left(\sum_i v_i^{(k_1)} v_i^{(k_2)} \right) \right) \exp\left(i \sum_{a,k} \hat{u}_a^{(k,k)} \left(\sum_i (v_{a,i}^{(k)})^2 - N \right) \right) \\
&\times \exp\left(i \sum_{a,k,t} \hat{y}_{a,t}^{(k)} \left(y_{a,t}^{(k)} - \frac{1}{N} \sum_i v_{a,i}^{(k)} s_{i,t} \right) \right) \exp\left(iN \sum_{a,k_1,k_2} R_a^{(k_1,k_2)} \left(R_a^{(k_1,k_2)} - \frac{1}{N} \sum_i v_{a,i}^{(k_1)} \bar{e}_i^{(k_2)} \right) \right) \\
&\times \exp\left(\beta \sum_{a=1}^{n} \sum_{k=1}^{K} \left(\sum_t (y_{a,t}^{(k)})^2 \right) - \beta \frac{1}{T} \sum_{a=1}^{n} \sum_{k=1}^{K} \left(\sum_t y_{a,t}^{(k)} \right)^2 \right) \times \\
&\times \exp\left(\sum_{a,k_1,k_2,i} \eta_i^{(k_1,k_2)} \frac{1}{2} (v_{a,i}^{(k_1)} - \bar{e}_i^{(k_1)})(v_{a,i}^{(k_2)} - \bar{e}_i^{(k_2)}) \right) \\
&\times \exp\left(\sum_{a,k_1,k_2,t} \gamma^{(k_1,k_2)} \left(\bar{x}_t^{(k_1)} - y_{a,t}^{(k_1)} \right) \left(\bar{x}_t^{(k_2)} - y_{a,t}^{(k_2)} \right) \right)
\end{aligned}
\tag{2.5}
$$

Now, we can express $s_{i,t}$ using (2.46), and average over all sources of disorder in the system, starting from $\bar{z}_{i,t}$:

$$
\langle Z^n(s_{i,t}) \rangle_{\bar{z}} = \int \frac{1}{(\sqrt{\det \bar{Z}})^N} \prod_i \frac{1}{\sqrt{\sigma_i^2 N}^T} \prod_t \frac{d\bar{z}_{i,t}}{\sqrt{2\pi}} \exp\left(- \sum_i \frac{\sum_{t_1,t_2} \bar{z}_{i,t_1} (\bar{Z}^{-1})_{t_1,t_2} \bar{z}_{i,t_2}}{2\bar{\sigma}_i^2 N} \right) Z^n(s_{i,t}) \tag{2.53}
$$

Then, average over $\delta \bar{F}_{i,t}$:

$$
\langle Z^n(s_{i,t}) \rangle_{\bar{z},\bar{F}} = \int \frac{1}{\sqrt{\det \bar{\Xi}}^N} \prod_{i,t} \frac{d\bar{F}_{i,t}}{\sqrt{2\pi}} \exp\left(- \sum_i \frac{\sum_{t_1,t_2} \delta \bar{F}_{i,t_1} (\bar{\Xi}^{-1})_{t_1,t_2} \delta \bar{F}_{i,t_2}}{2N} \right) \langle Z^n(s_{i,t}) \rangle_{\bar{z}} \tag{2.54}
$$

and finally, averaging over $\delta \bar{x}_t^{(k)}$:

$$
\langle Z^n(s_{i,t}) \rangle = \int \frac{1}{\sqrt{\det \bar{\Delta}}^K} \prod_k \frac{1}{(\bar{\xi}^{(k)})^T} \prod_t \frac{d\delta \bar{x}_{i,t}}{\sqrt{2\pi}} \exp\left(- \sum_k \frac{\sum_{t_1,t_2} \delta \bar{x}_{t_1}^{(k)} (\bar{\Delta}^{-1})_{t_1,t_2} \delta \bar{x}_{t_2}^{(k)}}{2(\bar{\xi}^{(k)})^2} \right) \langle Z^n(s_{i,t}) \rangle_{\bar{z},\bar{F}} \tag{2.55}
$$

After performing these Gaussian integrals, we will get

$$
\begin{aligned}
\langle Z^n(s_{i,t})\rangle \;=\; & \int \prod_{a,k}\left[d\mathbf{v}_a^{(k)}\prod_{t=1}^{T}dy_{a,t}^{(k)}\frac{d\hat{y}_{a,t}^{(k)}}{2\pi}\right]\int\prod_{a,k_1,k_2}\left[dR_a^{(k_1,k_2)}\frac{d\hat{R}_a^{(k_1,k_2)}}{2\pi N}\frac{d\hat{u}_a^{(k_1,k_2)}}{2\pi}\right] \\
\times\; & \exp\left(-\sum_i\frac{\bar{\sigma}_i^2\sum\limits_{a,b,k_1,k_2,t_1,t_2}v_{a,i}^{(k_1)}v_{b,i}^{(k_2)}\hat{y}_{a,t_1}^{(k_1)}\bar{Z}_{t_1,t_2}\hat{y}_{b,t_2}^{(k_2)}}{2N}\right) \\
\times\; & \exp\left(i\sum_{a,k_1<k_2}\hat{u}_a^{(k_1,k_2)}\left(\sum_i v_{a,i}^{(k_1)}v_{a,i}^{(k_2)}\right)+i\sum_{a,k}\hat{u}_a^{(k,k)}\left(\sum_i(v_{a,i}^{(k)})^2-N\right)\right) \\
\times\; & \exp\left(-N\sum_{i,\tau,\tau'}\frac{\left(\sum\limits_{k_1,k_1'}\frac{1}{N}\bar{e}_i^{(k_1)}\sum\limits_a v_{a,i}^{(k_1')}\sum\limits_t\hat{y}_{a,t}^{(k_1')}x_{t-\tau}^{(k_1)}\right)\bar{\Xi}_{\tau,\tau'}\left(\sum\limits_{k_2,k_2'}\frac{1}{N}\bar{e}_i^{(k_2)}\sum\limits_b v_{b,i}^{(k_2')}\sum\limits_{t'}\hat{y}_{b,t'}^{(k_2')}x_{t'-\tau'}^{(k_2)}\right)}{2}\right) \\
\times\; & \exp\left(-\sum_k\frac{(\bar{\xi}^{(k)})^2\sum\limits_{t,t',k_1,k_2}\sum_{a,b}\hat{y}_{a,t}^{(k_1)}R_a^{(k_1,k)}\Delta_{t,t'}\hat{y}_{b,t'}^{(k_2)}R_b^{(k_2,k)}}{2}\right) \\
\times\; & \exp\left(i\sum_{a,k_1,t}\hat{y}_{a,t}^{(k_1)}\left(y_{a,t}^{(k_1)}-\sum_{k_2=1}^{K}R_a^{(k_1,k_2)}\bar{x}_t^{(k_2)}\right)+iN\sum_{a,k_1,k_2}\hat{R}_a^{(k_1,k_2)}\left(R_a^{(k_1,k_2)}-\frac{1}{N}\sum_i v_{a,i}^{(k_1)}\bar{e}_i^{(k_2)}\right)\right)\times \\
\times\; & \exp\left(\beta\sum_{a,k}\left(\sum_t(y_{a,t}^{(k)})^2\right)-\beta\frac{1}{T}\sum_{a,k}\left(\sum_t y_{a,t}^{(k)}\right)^2\right)\times \\
\times\; & \exp\left(\sum_{a,k_1,k_2,i}\eta_i^{(k_1,k_2)}\frac{1}{2}(v_{a,i}^{(k_1)}-\bar{e}_i^{(k_1)})(v_{a,i}^{(k_2)}-\bar{e}_i^{(k_2)})\right) \\
\times\; & \exp\left(\sum_{a,k_1,k_2,t}\gamma^{(k_1,k_2)}\left(\bar{x}_t^{(k_1)}-y_{a,t}^{(k_1)}\right)\left(\bar{x}_t^{(k_2)}-y_{a,t}^{(k_2)}\right)\right)
\end{aligned}
\tag{2.56}
$$

Once again, we can simplify the notation by introducing

$$
\mathcal{X}_{t,t'}^{(k_1,k_2)}=\sum_{\tau,\tau'}x_{t-\tau}^{(k_1)}\bar{\Xi}_{\tau,\tau'}x_{t'-\tau'}^{(k_2)} \tag{2.57}
$$

Following the outline of the calculation (1.5.5), we introduce the overlaps:

$$
\begin{cases}
q_{ab}^{(k_1,k_2,k_1',k_2')}=\frac{1}{N}\sum\limits_i\bar{e}_i^{(k_1)}\bar{e}_i^{(k_2)}v_{a,i}^{(k_1')}v_{b,i}^{(k_2')}, & k_1\leq k_2,\ a\leq b \\
m_{ab}^{(k,k')}=\frac{1}{N}\sum\limits_i\bar{\sigma}_i^2v_{a,i}^{(k)}v_{b,i}^{(k')}, & a\leq b
\end{cases}
\tag{2.58}
$$

by adding an additional integration over delta-function:

$$
\langle Z^n(s_{i,t})\rangle=\int\prod_{a\leq b,k_1\leq k_2,k_1',k_2'}dq_{ab}^{(k_1,k_2,k_1',k_2')}\delta\left(q_{ab}^{(k_1,k_2,k_1',k_2')}-\frac{1}{N}\sum_i\bar{e}_i^{(k_1)}\bar{e}_i^{(k_2)}v_{a,i}^{(k_1')}v_{b,i}^{(k_2')}\right)\langle Z^n(s_{i,t})\rangle \tag{2.59}
$$

(and likewise for $m_{ab}^{(k,k')}$).

2.4.3 Replica-symmetric ansatz

We can replace the delta-function above by integration of a complex exponent, and introduce $\hat{q}_{ab}^{(d_1,d_2,d_1',d_2')}$ and $\hat{m}_{ab}^{(k,k')}$. Following the outline of the calculation (1.5.5) we can apply the replica-symmetric ansatz:

$$\begin{cases} q_{ab}^{(d_1,d_2,d_1',d_2')} = q^{(d_1,d_2,d_1',d_2')},\ \hat{q}_{ab}^{(d_1,d_2,d_1',d_2')} = \hat{q}^{(d_1,d_2,d_1',d_2')} & \text{for } a \neq b \\ q_{aa}^{(d_1,d_2,d_1',d_2')} = r^{(d_1,d_2,d_1',d_2')},\ \hat{q}_{aa}^{(d_1,d_2,d_1',d_2')} = \hat{r}^{(d_1,d_2,d_1',d_2')} & \\ m_{ab}^{(k,k')} = m^{(k,k')},\ \hat{m}_{ab}^{(k,k')} = \hat{m}^{(k,k')} & \text{for } a \neq b \\ m_{aa}^{(k,k')} = l^{(k,k')},\ \hat{m}_{aa}^{(k,k')} = \hat{l}^{(k,k')} & \\ R_a^{(k,k')} = R^{(k,k')} & \\ \hat{R}_a^{(k,k')} = \hat{R}^{(k,k')} & \\ \hat{u}_a^{(k,k')} = \hat{u}^{(k,k')} & \end{cases} \tag{2.60}$$

We can consider that $y_{a,t}^{(k)}$ - elements of one big vector y in a tensor product of spaces of principal components (K-dimensional space), of time bins (T-dimensional space), and of replicas (n-dimensional space). Similarly, we can consider $\hat{y}_{a,t}^{(k)}$ as elements of vector $\hat{y}$. We can see $m^{(k,k')}$ as elements of a matrix M acting in the space of principal components, $\hat{m}^{(k,k')}$ - as elements of $\hat{M}$, $l^{(k,k')}$ - as elements of L, $\hat{l}^{(k,k')}$ - as elements of $\hat{L}$. We also introduce the matrix

$$\hat{U}_{k_1,k_2} = \hat{u}^{(k_1,k_2)}(1 + \delta_{k_1,k_2}) \tag{2.61}$$

We will write the identity matrices as Id_K, Id_T, and Id_n for component, time bin, and replica spaces, respectively. Similarly, we use J_K, J_T, and J_n to denote square matrices of ones. For vectors, we will denote 1_K, 1_T, 1_n as vectors consisting of ones. We use $\text{diag}_K(...)$ to denote a matrix that is diagonal in the space of components, with the elements of the argument on its diagonal. Additionally, we denote the tensor product of matrices with $\otimes$.

We use a circle to denote contraction on all indices; for example:

$$\hat{r} \circ r = \sum_{k_1,k_1',k_2,k_2'} \hat{r}^{(k_1,k_2,k_1',k_2')} r^{(k_1,k_2,k_1',k_2')} \tag{2.62}$$

and a dot for contracting the first two indices:

$$(q \cdot \mathcal{X})_{t,t'}^{(k_1',k_2')} = \sum_{k_1,k_2} q^{(k_1,k_2,k_1',k_2')} \mathcal{X}_{t,t'}^{(k_1,k_2)} \tag{2.63}$$

Then, the entire expression can be rewritten as

$$
\begin{aligned}
\langle Z^n(s_{i,t})\rangle \;=\; & \int dv \frac{dy\hat{y}}{(2\pi)^{nKT}} \frac{dRd\hat{R}}{(2\pi N)^{K^2}} \frac{dLd\hat{L}}{(2\pi)^{K^2}} \frac{d\hat{U}}{(2\pi)^{K^2}} \frac{dMd\hat{M}}{(2\pi)^{K^2}} \\
& \int \prod_{k_1,k_1',k_2,k_2'} \frac{dq^{(k_1,k_2,k_1',k_2')} d\hat{q}^{(k_1,k_2,k_1',k_2')}}{2\pi} \frac{dr^{(k_1,k_2,k_1',k_2')} d\hat{r}^{(k_1,k_2,k_1',k_2')}}{2\pi} \\
\times \;& \exp\left(iNn\mathrm{Tr}(\hat{L}L^T) - iv^T(\hat{L}\otimes Id_n \otimes \mathrm{diag}_N(\bar{\sigma}^2))v + \frac{iNn(n-1)}{2} Tr(\hat{M}M^T) \right) \\
\times \;& \exp\left(\frac{i}{2} v^T(\hat{M}\otimes(J_n - Id_n)\otimes \mathrm{diag}_N(\bar{\sigma}^2))v - \frac{\hat{y}^T(L\otimes Z\otimes Id_n)\hat{y}}{2} - \frac{\hat{y}^T(M\otimes Z\otimes(J_n - Id_n))\hat{y}}{2} \right) \\
\times \;& \exp\left(iNn(\hat{r}\circ r) - \sum_{k_1,k_1',k_2,k_2',a} i\hat{r}^{(k_1,k_2,k_1',k_2')} \sum_i \bar{e}_i^{(k_1)} \bar{e}_i^{(k_2)} v_{a,i}^{(k_1')} v_{a,i}^{(k_2')} \right) \\
\times \;& \exp\left(iN\frac{n(n-1)}{2}(\hat{q}\circ q) - \sum_{k_1,k_1',k_2,k_2',a<b} i\hat{q}^{(k_1,k_2,k_1',k_2')} \left(\sum_i \bar{e}_i^{(k_1)} \bar{e}_i^{(k_2)} v_{a,i}^{(k_1')} v_{b,i}^{(k_2')} \right) \right) \\
\times \;& \exp\left(-\frac{\hat{y}^T((q\cdot\mathcal{X})\otimes(J_n - Id_n))\hat{y}}{2} - \frac{\hat{y}^T((r\cdot\mathcal{X})\otimes Id_n)\hat{y}}{2} \right) \\
\times \;& \exp\left(-\frac{iNn}{2}\mathrm{Tr}\hat{U} + \frac{i}{2} v^T(\hat{U}\otimes Id_n \otimes Id_N)v - \frac{\hat{y}^T(R\mathrm{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\otimes J_n)\hat{y}}{2} \right) \\
\times \;& \exp\left(i\hat{y}^T y - i((R\otimes Id_T\otimes Id_n)(\bar{x}\otimes 1_n))^T\hat{y} + iN\mathrm{Tr}\hat{R}^T R - iv^T(\hat{R}\otimes Id_n\otimes Id_N)(e\otimes\mathbb{1}_n) \right) \times \\
\times \;& \exp\left(\beta y^T \left(Id_{K,n} \otimes \left(Id_T - \frac{1}{T} J_T \right) \right) y + \frac{1}{2}(v - \bar{e}\otimes 1_n)^T(\mathrm{diag}_N(\eta_i)\otimes Id_n)(v - \bar{e}\otimes 1_n) \right) \\
\times \;& \exp\left((y - \bar{x}\otimes 1_n)^T(\gamma\otimes Id_T\otimes Id_n)(y - \bar{x}\otimes 1_n) \right)
\end{aligned}
\tag{2.64}
$$

Now, we can continue integrating over the variables. First, integrating with respect to $\hat{y}$. We can see that this is just a simple Gaussian integral with the quadratic term

$$
\mathcal{A} = ((L-M)\otimes Z + (r-q)\cdot\mathcal{X})\otimes Id_n + \left(M\otimes Z + q\cdot\mathcal{X} + R\mathrm{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\right)\otimes J_n \tag{2.65}
$$

and a linear term $i(y - (R\otimes Id_n\otimes Id_T)(\bar{x}\otimes\mathbb{1}_n))$.

After this, we can integrate with respect to y. The integral will also be Gaussian, with quadratic part

$$
\mathcal{B} = \mathcal{A}^{-1} - 2\beta Id_K\otimes Id_n\otimes Id_T + 2\beta Id_K\otimes Id_n\otimes\frac{1}{T}J_T - 2\gamma\otimes Id_n\otimes Id_T \tag{2.66}
$$

and a linear part $(\mathcal{A}^{-1}(R\otimes Id_{n,T}) - 2\gamma\otimes Id_{n,T})(\bar{x}\otimes\mathbb{1}_n)$

Then, after integration

$$\begin{aligned}
\langle Z^n \rangle &= \int \frac{d\hat{U}dv}{(2\pi)^{K^2}} \frac{d\hat{R}dR}{(2\pi N)^{K^2}} \frac{dMd\hat{M}}{(2\pi)^{K^2}} \frac{dLd\hat{L}}{(2\pi)^{K^2}} \frac{1}{\sqrt{\det \mathcal{A} \det \mathcal{B}}} \exp\left(iNnTr(\hat{L}L^T) + \frac{iNn(n-1)}{2} Tr(\hat{M}M^T) \right) \\
&\int \prod_{k_1,k_1',k_2,k_2'} \frac{dq^{(k_1,k_2,k_1',k_2')} d\hat{q}^{(k_1,k_2,k_1',k_2')}}{2\pi} \frac{dr^{(k_1,k_2,k_1',k_2')} d\hat{r}^{(k_1,k_2,k_1',k_2')}}{2\pi} \\
&\times \exp\left(\frac{(\bar{x}^T \otimes \mathbb{1}_n^T)(\mathcal{A}^{-1}(R \otimes Id_{n,T}) - 2\gamma \otimes Id_{n,T})^T \mathcal{B}^{-1} (\mathcal{A}^{-1}(R \otimes Id_{n,T}) - 2\gamma \otimes Id_{n,T})(\bar{x} \otimes \mathbb{1}_n)}{2} \right) \\
&\times \exp\left(-\frac{i}{2} v^T (\hat{M} \otimes (J_n - Id_n) \otimes \mathrm{diag}_N(\bar{\sigma}^2))v - iv^T(\hat{L} \otimes Id_n \otimes \mathrm{diag}_N(\bar{\sigma}^2))v \right) \\
&\times \exp\left(iNn(\hat{r} \circ r) - \sum_{k_1,k_1',k_2,k_2',a} i\hat{r}^{(k_1,k_2,k_1',k_2')} \sum_i \bar{e}_i^{(k_1)} \bar{e}_i^{(k_2)} v_{a,i}^{(k_1')} v_{a,i}^{(k_2')} \right) \\
&\times \exp\left(iN\frac{n(n-1)}{2}(\hat{q} \circ q) - \sum_{k_1,k_1',k_2,k_2',a<b} i\hat{q}^{(k_1,k_2,k_1',k_2')} \left(\sum_i \bar{e}_i^{(k_1)} \bar{e}_i^{(k_2)} v_{a,i}^{(k_1')} v_{b,i}^{(k_2')} \right) \right) \\
&\times \exp\left(iNn\mathrm{Tr}(\hat{R}R^T) - iv^T(\hat{R} \otimes Id_n \otimes Id_N)(\bar{e} \otimes \mathbb{1}_n) - \frac{iNn}{2}\mathrm{Tr}\hat{U} + \frac{i}{2} v^T(\hat{U} \otimes Id_n \otimes Id_N)v \right) \\
&\times \exp\left(-\frac{(\bar{x} \otimes \mathbb{1}_n)^T (R^T \otimes Id_n \otimes Id_T)\mathcal{A}^{-1}(R \otimes Id_n \otimes Id_T)(\bar{x} \otimes \mathbb{1}_n)}{2} \right) \\
&\times \exp\left(\frac{1}{2}(v - \bar{e} \otimes 1_n)^T (\mathrm{diag}_N(\eta_i) \otimes Id_n)(v - \bar{e} \otimes 1_n) + (\bar{x} \otimes 1_n)^T (\gamma \otimes Id_T \otimes Id_n)(\bar{x} \otimes 1_n) \right)
\end{aligned} \tag{2.67}$$

Now let's take the integral w.r.t. $v_{a,i}$. Introduce

$$\begin{cases}
\tilde{q} = i\hat{q} \\
\tilde{r} = i\hat{r} \\
\tilde{M} = i\hat{M} \\
\tilde{L} = i\hat{L} \\
\tilde{R} = \frac{i}{\beta}\hat{R} \\
\tilde{U} = -i\hat{U}
\end{cases} \tag{2.68}$$

Remembering the notation for contraction of the first two indices, we write

$$(\tilde{q} \cdot \bar{e}_i \bar{e}_i^T)^{(k_1',k_2')} = \sum_{k_1,k_2} \tilde{q}^{(k_1,k_2,k_1',k_2')} \bar{e}_i^{(k_1)} \bar{e}_i^{(k_2)} \tag{2.69}$$

Then, we see that the integral over v is Gaussian once again, with quadratic term

$$\mathcal{C}_i = (\bar{\sigma}_i^2(2\tilde{L} - \tilde{M}) + \tilde{U} + (2\tilde{r} - \tilde{q}) \cdot \bar{e}_i \bar{e}_i^T) \otimes Id_n + (\bar{\sigma}_i^2 \tilde{M} + \tilde{q} \cdot \bar{e}_i \bar{e}_i^T) \otimes J_n - \eta_i \otimes Id_n \tag{2.70}$$

and integrating, we get

$$
\begin{aligned}
\langle Z^n \rangle &= \int \frac{dqd\tilde{q}}{(2\pi)^{K^4}} \frac{drd\tilde{r}}{(2\pi)^{K^4}} \int \frac{d\tilde{U}}{(2\pi)^{K^2}} \frac{d\tilde{R}dR}{(2\pi N)^{K^2}} \frac{dMd\tilde{M}}{(2\pi)^{K^2}} \frac{dLd\tilde{L}}{(2\pi)^{K^2}} \frac{1}{\sqrt{\det \mathcal{A} \det \mathcal{B} \prod_i \det \mathcal{C}_i}} \exp\left(NnTr(\tilde{L}L^T)\right) \\
&\times \exp\left(\frac{Nn}{2}\mathrm{Tr}\tilde{U} + Nn(\tilde{r} \circ r) + \frac{Nn(n-1)}{2}(\tilde{q} \circ q) + Nn\mathrm{Tr}(\tilde{R}R^T) + \frac{Nn(n-1)}{2}Tr(\tilde{M}M^T)\right) \\
&\times \exp\left(\frac{(\bar{x}^T \otimes \mathbb{1}_n^T)(\mathcal{A}^{-1}(R \otimes Id_{n,T}) - 2\gamma \otimes Id_{n,T})^T \mathcal{B}^{-1}(\mathcal{A}^{-1}(R \otimes Id_{n,T}) - 2\gamma \otimes Id_{n,T})(\bar{x} \otimes \mathbb{1}_n)}{2}\right) \\
&\times \exp\left(-\frac{(\bar{x} \otimes \mathbb{1}_n)^T (R^T \otimes Id_n \otimes Id_T)\mathcal{A}^{-1}(R \otimes Id_n \otimes Id_T)(\bar{x} \otimes \mathbb{1}_n)}{2}\right) \\
&\times \exp\left(\frac{1}{2}(\bar{e} \otimes 1_n)^T (\mathrm{diag}_N(\eta_i) \otimes Id_n)(\bar{e} \otimes 1_n) + (\bar{x} \otimes 1_n)^T (\gamma \otimes Id_T \otimes Id_n)(\bar{x} \otimes 1_n)\right) \\
&\times \prod_i \left[\exp\left(\frac{1}{2}(\bar{e}_i \otimes \mathbb{1}_n)^T ((\frac{\tilde{\eta}_i + \tilde{\eta}_i^T}{2} + \tilde{R}) \otimes Id_n)^T \mathcal{C}_i^{-1} ((\frac{\tilde{\eta}_i + \tilde{\eta}_i^T}{2} + \tilde{R}) \otimes Id_n)(\bar{e}_i \otimes \mathbb{1}_n)\right)\right] \qquad (2.71)
\end{aligned}
$$

We assume that as $\beta \to \infty$, the following combinations of variables go to a finite limit:

$$
\begin{cases}
v = \beta(r - q) \\
W = \beta(L - M) \\
\tilde{v} = \frac{1}{\beta}(2\tilde{r} - \tilde{q}) \\
\tilde{W} = \frac{1}{\beta}(2\tilde{L} - \tilde{M}) \\
U = \frac{1}{\beta}\tilde{U} \\
p = \frac{1}{\beta^2}\tilde{q} \\
S = \frac{1}{\beta^2}\tilde{M} \\
\tilde{\gamma} = \frac{1}{\beta}\gamma \\
\tilde{\eta} = \frac{1}{\beta}\eta
\end{cases} \qquad (2.72)
$$

Now we can take $n \to 0$ and $\beta \to \infty$ keeping only leading orders.

$$
\begin{aligned}
&-\frac{(\bar{x} \otimes \mathbb{1}_n)^T (R^T \otimes Id_n \otimes Id_T)\mathcal{A}^{-1}(R \otimes Id_n \otimes Id_T)(\bar{x} \otimes \mathbb{1}_n)}{2} = \\
&= -\frac{\beta n \bar{x}^T (R^T \otimes Id_T) \left(W \otimes Z + v \cdot \mathcal{X}\right)^{-1} (R \otimes Id_T)\bar{x}}{2} + O(n^2) \qquad (2.73)
\end{aligned}
$$

$$
\begin{aligned}
&\frac{(\bar{x}^T \otimes \mathbb{1}_n^T)(\mathcal{A}^{-1}(R \otimes Id_{n,T}) - 2\gamma \otimes Id_{n,T})^T \mathcal{B}^{-1}(\mathcal{A}^{-1}(R \otimes Id_{n,T}) - 2\gamma \otimes Id_{n,T})(\bar{x} \otimes \mathbb{1}_n)}{2} = \\
&\frac{n\beta}{2}\bar{x}^T \left((R^T \otimes Id_T)\left(W \otimes Z + v \cdot \mathcal{X}\right)^{-1} - 2\tilde{\gamma}^T \otimes Id_T\right) \\
&\left(\left(W \otimes Z + v \cdot \mathcal{X}\right)^{-1} - 2Id_{K,T} + 2Id_K \otimes \frac{1}{T}J_T - 2\tilde{\gamma} \otimes Id_T\right)^{-1} \\
&\left(\left(W \otimes Z + v \cdot \mathcal{X}\right)^{-1} (R \otimes Id_T) - 2\gamma \otimes Id_T\right)\bar{x} + O(n^2) \qquad (2.74)
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\beta^2}{2}(\bar{e}_i \otimes \mathbb{1}_n)^T ((\frac{\tilde{\eta}_i + \tilde{\eta}_i^T}{2} + \tilde{R}) \otimes Id_n)^T \mathcal{C}_i^{-1} ((\frac{\tilde{\eta}_i + \tilde{\eta}_i^T}{2} + \tilde{R}) \otimes Id_n)(\bar{e}_i \otimes \mathbb{1}_n) = \\
&= \frac{n\beta}{2}\bar{e}_i^T (\frac{\tilde{\eta}_i + \tilde{\eta}_i^T}{2} + \tilde{R})^T (\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T - \tilde{\eta}_i)^{-1} (\frac{\tilde{\eta}_i + \tilde{\eta}_i^T}{2} + \tilde{R})\bar{e}_i + O(n^2) \qquad (2.75)
\end{aligned}
$$

Now we simplify the logarithms:

$$
\begin{aligned}
-\frac{1}{2}\ln\left[\det\mathcal{A}\det\mathcal{B}\right] = -\frac{1}{2}\ln\det\Bigg[Id_{K,n,T} - 2\left(W\otimes Z + v\cdot\mathcal{X}\right)\left(\tilde{\gamma}\otimes Id_T + Id_K\otimes\left(Id_T - \frac{1}{T}J_T\right)\right)\otimes Id_n \\
-2\beta\left(\left(L-\frac{W}{\beta}\right)\otimes Z + \left(r-\frac{v}{\beta}\right)\cdot\mathcal{X} + R\mathrm{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\right)\left(\gamma\otimes Id_T - Id_K\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)\otimes J_n\Bigg] = \\
= n\beta\mathrm{Tr}\Bigg[\left(Id_{K,T} - 2\left(W\otimes Z + v\cdot\mathcal{X}\right)\left(\tilde{\gamma}\otimes Id_T + Id_K\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)\right)^{-1} \\
\left(L\otimes Z + r\cdot\mathcal{X} + R\mathrm{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\right)\left(\tilde{\gamma}\otimes Id_T + Id_K\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)\Bigg] + O(n^2)
\end{aligned}
\tag{2.76}
$$

$$
-\frac{1}{2}\ln\left(\prod_i \det\mathcal{C}_i\right) = -\frac{n\beta}{2}\sum_i \mathrm{Tr}\left[(\bar{\sigma}_i^2\tilde{W} + U + \tilde{v}\cdot\bar{e}_i\bar{e}_i^T - \tilde{\eta}_i)^{-1}(\bar{\sigma}_i^2 S + p\cdot\bar{e}_i\bar{e}_i^T)\right] + O(n^2) \tag{2.77}
$$

Finally, combining everything together, we can write

$$
\begin{aligned}
\langle Z^n\rangle &= \int \frac{dvd\tilde{v}}{(2\pi)^{K^4}}\frac{dpdr}{(2\pi)^{K^4}}\frac{dU}{(2\pi)^{K^2}}\frac{d\tilde{R}dR}{(2\pi N)^{K^2}}\frac{dWd\tilde{W}dSdL}{(2\pi)^{2K^2}} \\
&\times \exp\Bigg(n\beta\mathrm{Tr}\Bigg[\left(Id_{K,T} - 2\left(W\otimes Z + v\cdot\mathcal{X}\right)\left(\tilde{\gamma}\otimes Id_T + Id_K\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)\right)^{-1} \\
&\quad \left(L\otimes Z + r\cdot\mathcal{X} + R\mathrm{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\right)\left(\tilde{\gamma}\otimes Id_T + Id_K\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)\Bigg]\Bigg) \\
&\times \exp\left(-\frac{n\beta}{2}\sum_i \mathrm{Tr}\left[(\bar{\sigma}_i^2\tilde{W} + U + \tilde{v}\cdot\bar{e}_i\bar{e}_i^T - \tilde{\eta}_i)^{-1}(\bar{\sigma}_i^2 S + p\cdot\bar{e}_i\bar{e}_i^T)\right]\right) \\
&\times \exp\left(\frac{\beta Nn}{2}TrU + \frac{\beta Nn}{2}(\tilde{v}\circ r) + \frac{\beta Nn}{2}(p\circ v) + \beta Nn\mathrm{Tr}(\tilde{R}R^T) + \frac{\beta Nn}{2}Tr(\tilde{W}L^T) + \frac{\beta Nn}{2}Tr(SW^T)\right) \\
&\times \exp\Bigg(\frac{n\beta}{2}\bar{x}^T\Big(\left(R\otimes Id_T\right)^T\left(W\otimes Z + v\cdot\mathcal{X}\right)^{-1} - 2\tilde{\gamma}^T\otimes Id_T\Big) \\
&\quad \left(\left(W\otimes Z + v\cdot\mathcal{X}\right)^{-1} - 2Id_{K,T} + 2Id_K\otimes\frac{1}{T}J_T - 2\tilde{\gamma}\otimes Id_T\right)^{-1} \\
&\quad \Big(\left(W\otimes Z + v\cdot\mathcal{X}\right)^{-1}\left(R\otimes Id_T\right) - 2\tilde{\gamma}\otimes Id_T\Big)\,\bar{x}\Bigg) \\
&\times \exp\left(-\frac{\beta n\bar{x}^T(R^T\otimes Id_T)\left(W\otimes Z + v\cdot\mathcal{X}\right)^{-1}(R\otimes Id_T)\bar{x}}{2}\right) \\
&\times \exp\left(\frac{\beta}{2}(\bar{e}\otimes 1_n)^T(\mathrm{diag}_N(\tilde{\eta}_i)\otimes Id_n)(\bar{e}\otimes 1_n) + \beta(\bar{x}\otimes 1_n)^T(\tilde{\gamma}\otimes Id_T\otimes Id_n)(\bar{x}\otimes 1_n)\right) \\
&\times \prod_i\left[\exp\left(\frac{n\beta}{2}\bar{e}_i^T(\frac{\tilde{\eta}_i+\tilde{\eta}_i^T}{2} + \tilde{R})^T(\bar{\sigma}_i^2\tilde{W} + U + \tilde{v}\cdot\bar{e}_i\bar{e}_i^T - \tilde{\eta}_i)^{-1}(\frac{\tilde{\eta}_i+\tilde{\eta}_i^T}{2} + \tilde{R})\bar{e}_i\right)\right]
\end{aligned}
\tag{2.78}
$$

2.4.4 Saddle point approximation

We will use the approximation for the integrals in $\langle Z^n\rangle$ valid for large β. The key idea is to evaluate the integrated expression at its optimal point, which we approximate as the stationary point of the free energy I. Then, combining (1.11) and (1.12), we can write

$$
\begin{aligned}
I &= \underset{v,\tilde{v},p,r,U,\tilde{R},R,W,\tilde{W},S,L}{\text{optimum}} \left(-\frac{1}{2N}\sum_i \text{Tr}\left[(\bar{\sigma}_i^2\tilde{W} + U + \tilde{v}\cdot \bar{e}_i\bar{e}_i^T - \tilde{\eta}_i)^{-1}(\bar{\sigma}_i^2 S + p\cdot \bar{e}_i\bar{e}_i^T)\right]\right. \\
&+ \frac{1}{N}\text{Tr}\left[\left(Id_{K,T} - 2\left(W\otimes Z + v\cdot\mathcal{X}\right)\left(\tilde{\gamma}\otimes Id_T + Id_{K,T}\otimes\left(Id_T - \frac{1}{T}J_T\right)\right)\right)^{-1}\right. \\
&\quad \left.\left(L\otimes Z + r\cdot\mathcal{X} + R\text{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\right)\left(\tilde{\gamma}\otimes Id_T + Id_K\otimes\left(Id_T - \frac{1}{T}J_T\right)\right)\right] \\
&+ \frac{1}{2}\text{Tr}U + \frac{1}{2}(\tilde{v}\circ r) + \frac{1}{2}(p\circ v) + \text{Tr}(\tilde{R}R^T) + \frac{1}{2}Tr(\tilde{W}L^T) + \frac{1}{2}Tr(SW^T) \\
&+ \frac{1}{2N}\bar{x}^T\left((R\otimes Id_T)^T(W\otimes Z + v\cdot\mathcal{X})^{-1} - 2\tilde{\gamma}^T\otimes Id_T\right) \\
&\quad \left((W\otimes Z + v\cdot\mathcal{X})^{-1} - 2Id_{K,T} + 2Id_K\otimes\frac{1}{T}J_T - 2\tilde{\gamma}\otimes Id_T\right)^{-1} \\
&\quad \left((W\otimes Z + v\cdot\mathcal{X})^{-1}(R\otimes Id_T) - 2\tilde{\gamma}\otimes Id_T\right)\bar{x} \\
&- \frac{1}{2N}\bar{x}^T(R^T\otimes Id_T)(W\otimes Z + v\cdot\mathcal{X})^{-1}(R\otimes Id_T)\bar{x} \\
&+ \frac{1}{2N}\bar{e}^T\text{diag}_N(\tilde{\eta}_i)\bar{e} + \frac{1}{N}\bar{x}^T(\tilde{\gamma}\otimes Id_T)\bar{x} \\
&+ \left.\sum_i \frac{1}{2N}\bar{e}_i^T(\frac{\tilde{\eta}_i + \tilde{\eta}_i^T}{2} + \tilde{R})^T(\bar{\sigma}_i^2\tilde{W} + U + \tilde{v}\cdot\bar{e}_i\bar{e}_i^T - \tilde{\eta}_i)^{-1}(\frac{\tilde{\eta}_i + \tilde{\eta}_i^T}{2} + \tilde{R})\bar{e}_i\right)
\end{aligned}
\tag{2.79}
$$

Then, we can recover ϵ by writing:

$$
\begin{aligned}
\left.\frac{\partial I}{\partial\tilde{\gamma}}\right|_{\tilde{\gamma},\tilde{\eta}=0} &= -\frac{2}{N}Tr_T\left[\left(Id_{K,T} - 2\left(W\otimes Z\left(Id_T - \frac{1}{T}J_T\right) + v\cdot\mathcal{X}\right)\right)^{-1}(R\otimes Id_T)\bar{x}\bar{x}^T\right] \\
&+ \frac{1}{N}Tr_T\left[\left(Id_{K,T} - 2\left(W\otimes Z\left(Id_T - \frac{1}{T}J_T\right) + v\cdot\mathcal{X}\right)\right)^{-1}(R\otimes Id_T)\bar{x}\bar{x}^T(R\otimes Id_T)^T\right. \\
&\quad \left.\left(Id_{K,T} - 2\left(W\otimes Z\left(Id_T - \frac{1}{T}J_T\right) + v\cdot\mathcal{X}\right)\right)^{-1}\right] \\
&+ \frac{1}{N}\text{Tr}_T\left[\left(L\otimes Z + r\cdot\mathcal{X} + R\text{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\right)\right. \\
&\quad \left.\left(Id_{K,T} - 2\left(W\otimes Z\left(Id_T - \frac{1}{T}J_T\right) + v\cdot\mathcal{X}\right)\right)^{-1}\right] \\
&+ \frac{2}{N}\text{Tr}_T\left[(W\otimes Z + v\cdot\mathcal{X})\left(Id_{K,T} - 2\left(W\otimes Z\left(Id_T - \frac{1}{T}J_T\right) + v\cdot\mathcal{X}\right)\right)^{-1}\right. \\
&\quad \left(L\otimes Z\left(Id_T - \frac{1}{T}J_T\right) + r\cdot\mathcal{X} + R\text{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right)\right) \\
&\quad \left.\left(Id_{K,T} - 2\left(W\otimes Z\left(Id_T - \frac{1}{T}J_T\right) + v\cdot\mathcal{X}\right)\right)^{-1}\right] + \frac{1}{N}\text{Tr}_T\bar{x}\bar{x}^T
\end{aligned}
\tag{2.80}
$$

In the limit $T, N \to \infty$, $T/N = \alpha$ this expression can be simplified:

$$\begin{aligned}
\left.\frac{\partial I}{\partial \tilde{\gamma}}\right|_{\tilde{\gamma},\tilde{\eta}=0} &= \frac{1}{N}\text{Tr}_T \bar{x}\bar{x}^T - \frac{2}{N} Tr_T \left[\left(Id_{K,T} - 2\left(W \otimes Z + v \cdot \mathcal{X}\right)\right)^{-1} \left(R \otimes Id_T\right)\bar{x}\bar{x}^T \right] \\
&+ \frac{1}{N} Tr_T \Big[\left(Id_{K,T} - 2\left(W \otimes Z + v \cdot \mathcal{X}\right)\right)^{-1} \\
&\quad \left(L \otimes Z + r \cdot \mathcal{X} + (R \otimes Id_T)(\text{diag}(\bar{\xi}^2) \otimes \bar{\Delta}(Id_T - \frac{1}{T} J_T) + \bar{x}\bar{x}^T)(R^T \otimes Id_T) \right) \\
&\quad \left(Id_{K,T} - 2\left(W \otimes Z + v \cdot \mathcal{X}\right)\right)^{-1} \Big] \\
&+ \frac{1}{N}\text{Tr}_T \left[\left(L \otimes Z + r \cdot \mathcal{X} + R\text{diag}(\bar{\xi}^2) R^T \otimes \bar{\Delta} \frac{1}{T} J_T \right) \left(Id_{K,T} - 2\left(W \otimes Z + v \cdot \mathcal{X}\right)\right)^{-1} \right]
\end{aligned} \tag{2.81}$$

Similarly, we can recover ρ by writing

$$\begin{aligned}
\left.\frac{\partial I}{\partial \tilde{\eta}_i}\right|_{\tilde{\gamma},\tilde{\eta}=0} &= \left.\frac{\partial}{\partial \tilde{\eta}}\right|_{\tilde{\gamma},\tilde{\eta}=0} \left[-\frac{1}{2N}\text{Tr}\left[(\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T - \tilde{\eta}_i)^{-1} (\bar{\sigma}_i^2 S + p \cdot \bar{e}_i \bar{e}_i^T) \right] + \right. \\
&+ \left. \frac{1}{2N} \bar{e}_i^T \tilde{\eta}_i \bar{e}_i + \frac{1}{2N} \bar{e}_i^T \left(\frac{\tilde{\eta}_i + \tilde{\eta}_i^T}{2} + \tilde{R} \right)^T (\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T - \tilde{\eta}_i)^{-1} \left(\frac{\tilde{\eta}_i + \tilde{\eta}_i^T}{2} + \tilde{R} \right) \bar{e}_i \right]
\end{aligned} \tag{2.82}$$

$$\begin{aligned}
\left.\frac{\partial I}{\partial \tilde{\eta}_i}\right|_{\tilde{\gamma},\tilde{\eta}=0} &= -\frac{1}{2N} (\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T)^{-1} (\bar{\sigma}_i^2 S + p \cdot \bar{e}_i \bar{e}_i^T - \tilde{R} \bar{e}_i \bar{e}_i^T \tilde{R}^T) (\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T)^{-1} + \\
&+ \frac{1}{2N} \left(\bar{e}_i \bar{e}_i^T \right) + \frac{1}{N} (\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T)^{-1} \left(\frac{\tilde{R} + \tilde{R}^T}{2} \right) \bar{e}_i \bar{e}_i^T
\end{aligned} \tag{2.83}$$

We can verify that this expression for ρ, when summed for all neurons i, obeys the property (2.10). The proof is given in the Appendix B.

To get the solution for all order parameters, we set $\tilde{\gamma} = 0$, $\tilde{\eta} = 0$.

$$\begin{aligned}
I &= \underset{v,\tilde{v},p,r,U,\tilde{R},R,W,\tilde{W},S,L}{\text{optimum}} \left(-\frac{1}{2N} \sum_i \text{Tr}\left[(\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T)^{-1} (\bar{\sigma}_i^2 S + p \cdot \bar{e}_i \bar{e}_i^T) \right] \right. \\
&+ \frac{1}{N}\text{Tr}\left[\left(Id_{K,T} - 2\left(W \otimes Z + v \cdot \mathcal{X}\right) \left(Id_K \otimes \left(Id_T - \frac{1}{T} J_T \right) \right) \right)^{-1} \right. \\
&\quad \left. \left(L \otimes Z + r \cdot \mathcal{X} + R\text{diag}(\bar{\xi}^2) R^T \otimes \bar{\Delta} \right) \left(Id_K \otimes \left(Id_T - \frac{1}{T} J_T \right) \right) \right] \\
&+ \frac{1}{2}\text{Tr}U + \frac{1}{2}(\tilde{v} \circ r) + \frac{1}{2}(p \circ v) + \text{Tr}(\tilde{R} R^T) + \frac{1}{2} Tr(\tilde{W} L^T) + \frac{1}{2} Tr(S W^T) \\
&+ \frac{1}{2N} \bar{x}^T \left((R \otimes Id_T)^T \left(W \otimes Z + v \cdot \mathcal{X} \right)^{-1} \right) \\
&\quad \left(Id_{K,T} - 2\left(W \otimes Z + v \cdot \mathcal{X}\right) \left(Id_{K,T} - \frac{1}{T} Id_K \otimes J_T \right) \right)^{-1} (R \otimes Id_T) \bar{x} \\
&- \frac{1}{2N} \bar{x}^T (R \otimes Id_T)^T \left(W \otimes Z + v \cdot \mathcal{X} \right)^{-1} (R \otimes Id_T) \bar{x} \\
&+ \left. \sum_i \frac{1}{2N} \bar{e}_i^T \tilde{R}^T (\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T)^{-1} \tilde{R} \bar{e}_i \right)
\end{aligned} \tag{2.84}$$

We can simplify this expression in the limit $T, N \to \infty$, $T/N = \alpha$:

$$
\begin{aligned}
I &= \underset{v,\tilde{v},p,r,U,\tilde{R},R,W,\tilde{W},S,L}{\text{optimum}} \Bigg(-\frac{1}{2N}\sum_i \text{Tr}\left[(\bar{\sigma}_i^2\tilde{W} + U + \tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}(\bar{\sigma}_i^2 S + p\cdot\bar{e}_i\bar{e}_i^T - \tilde{R}\bar{e}_i\bar{e}_i^T\tilde{R}^T)\right] \\
&+ \frac{1}{N}\text{Tr}\Big[(Id_{K,T} - 2\,(W\otimes Z + v\cdot\mathcal{X}))^{-1} \\
&\quad \left(L\otimes Z + r\cdot\mathcal{X} + (R\otimes Id_T)(\text{diag}(\bar{\xi}^2)\otimes\bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right) + \bar{x}\bar{x}^T)(R\otimes Id_T)^T\right)\Big] \\
&+ \frac{1}{2}\text{Tr}U + \frac{1}{2}(\tilde{v}\circ r) + \frac{1}{2}(p\circ v) + \text{Tr}(\tilde{R}R^T) + \frac{1}{2}Tr(\tilde{W}L^T) + \frac{1}{2}Tr(SW^T)\Bigg)
\end{aligned} \tag{2.85}
$$

To solve for the order parameters, we perform a saddle point calculation by taking the derivatives of the free energy with respect to each order parameter and finding where these derivatives vanish:

$$
\frac{\partial I}{\partial v} = 0, \quad \frac{\partial I}{\partial \tilde{v}} = 0, \quad \dots \tag{2.86}
$$

However, it's important to note that we do not necessarily maximize the free energy with respect to all order parameters. For some, we actually need to minimize the free energy. The reasons for this sign flip in certain parameters stem from two aspects of our calculation:

1. **Complex Integration Over Certain Parameters.** Some of our variables, *e.g.* Lagrange multipliers conjugated to the "physical" order parameters, are integrated over the imaginary axis, and not the real axis. In standard optimization problems, extremization usually occurs over real variables, where one clearly distinguishes between maximization and minimization based on the sign of the second derivative. In the complex plane, the stationary points are still determined by setting the first derivative to zero, but the interpretation of extremization is different. In this case, a point where the first derivative vanishes might correspond to neither a strict maximum nor a minimum in the usual sense, but rather a saddle point in the complex domain. As a result, what might appear as a maximization in a real parameter space could effectively become a minimization along a complex contour, leading to the observed sign flip for these parameters.

2. **Replica limit $n \to 0$.** In this case, the sign flip is not due to the nature of the order parameters themselves but rather due to the analytic continuation procedure of the replica trick. In the calculation, we worked with n copies (replicas) of the system, before going to the limit $n \to 0$. Certain terms in the free energy are proportional to $\frac{n(n-1)}{2}$, corresponding to interactions between pairs of replicas (see Eq. (2.64)). As the limit is taken, the sign of such terms flips. This is because, as $n \to 0$, the factor $\frac{n(n-1)}{2}$ becomes negative. This flip in sign affects the optimization conditions, turning what would have been a maximization into a minimization (or vice versa). Thus, we will have to minimize the free energy with respect to these order parameters to properly account for the negative contributions that emerge due to the $n \to 0$ limit of the replica trick.

The free energy I depends on a large number of order parameters, and we can think of this problem as an optimization in a $7K^2 + 4K^4$-dimensional space (since $U, \tilde{R}, R, W, \tilde{W}, S$ and L are matrices in the space of size K, and $v, \tilde{v}, p$ and r are 4-index tensors in the same space). In this formulation, all the order parameters are collected into a single long vector ω. To find the saddle point of the free energy, we use Newton's method. The key steps of this method are:

1. **Gradient Calculation.** At each iteration, we compute the gradient ∇I of the free energy with respect to the order parameters. This gradient gives us the direction of the steepest ascent (or descent) for each parameter.

2. **Second derivatives calculation.** We then compute the Hessian matrix H, which is the matrix of second derivatives of the free energy with respect to each pair of order parameters. The Hessian encodes information about the local curvature of the free energy landscape.

3. **Update Step:** Once we have the gradient and Hessian, Newton's method updates the order parameters according to the rule:

$$\omega_{\text{new}} = \omega_{\text{old}} - H^{-1}\nabla I \tag{2.87}$$

This update step works by using both the gradient (which tells us the direction of steepest change) and the Hessian (which adjusts for the curvature of the landscape). By multiplying the gradient by the inverse of the Hessian, we adjust our step size according to the local curvature—taking smaller steps in directions with steep curvature and larger steps where the curvature is flatter. This allows us to move efficiently toward the saddle point.

There are several reasons why Newton's method is particularly well-suited to our situation:

- **No need to specify maximization or minimization directions:** In our saddle point problem, some order parameters need to be maximized, while others need to be minimized. Newton's method solely focuses on finding where the gradient is zero, and it doesn't discriminate between the types of stationary points - minimum, maximum or saddle point. As long as the algorithm is initialized near a saddle point and the Hessian matrix allows for movement towards the saddle, Newton's method can naturally converge to it.
- **Efficient handling of multi-dimensional problems:** Different directions may behave very differently. Some parameters might change the free energy sharply, while others might have a much smaller effect. Newton's method adjusts for this automatically because of the Hessian, which captures the curvature along each direction. This allows the method to take larger steps in more flat directions and smaller steps in more steep directions.
- **Fast convergence:** Since Newton's method uses both the gradient and the Hessian, it often converges to the solution much faster than methods that only rely on the gradient, such as the method of steepest descent.

Starting point for the convergence procedure.

Newton's method requires a starting point. We will use a simple case of $\bar{\sigma}_i = 0, \Xi = 0$. In this case, the (2.85) can be rewritten as

$$\begin{aligned} I &= \underset{v,\tilde{v},p,r,U,\tilde{R},R,W,\tilde{W},S,L}{\text{optimum}} \Bigg(-\frac{1}{2N}\sum_i \text{Tr}\left[(U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}(p\cdot\bar{e}_i\bar{e}_i^T - \tilde{R}\bar{e}_i\bar{e}_i^T\tilde{R}^T)\right] \\ &+ \frac{1}{N}\text{Tr}\Big[(Id_{K,T} - 2W\otimes Z)^{-1} \\ &\quad \left(L\otimes Z + (R\otimes Id_T)(\text{diag}(\bar{\xi}^2)\otimes\bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right) + \bar{x}\bar{x}^T)(R\otimes Id_T)^T\right)\Big] \\ &+ \frac{1}{2}\text{Tr}U + \frac{1}{2}(\tilde{v}\circ r) + \frac{1}{2}(p\circ v) + \text{Tr}(\tilde{R}R^T) + \frac{1}{2}\text{Tr}(\tilde{W}L^T) + \frac{1}{2}\text{Tr}(SW^T)\Bigg) \end{aligned} \tag{2.88}$$

In this case, setting derivatives of I with respect to v, r, S and $\tilde{W}$ will yield

$$\begin{cases} \tilde{v} = 0 \\ p = 0 \\ W = 0 \\ L = 0 \end{cases} \tag{2.89}$$

simplifying our expression for I further:

$$I = \underset{U,\tilde{R},R}{\text{optimum}} \left(\frac{1}{N}\text{Tr}\left[R\text{Tr}_T(\text{diag}(\bar{\xi}^2)\otimes\bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right) + \bar{x}\bar{x}^T)R^T\right] + \frac{1}{2}\text{Tr}U + \text{Tr}(\tilde{R}R^T) + \frac{1}{2}\text{Tr}\tilde{R}^TU^{-1}\tilde{R}\right)$$

Since the matrix $\frac{1}{T}\text{Tr}_T(\text{diag}(\bar{\xi}^2) \otimes \bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right) + \bar{x}\bar{x}^T)$ is real and symmetric, it can be diagonalized:

$$\frac{1}{T}\text{Tr}_T(\text{diag}(\bar{\xi}^2) \otimes \bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right) + \bar{x}\bar{x}^T) = PDP^T \tag{2.90}$$

where P is orthogonal matrix and D is diagonal. Then,

$$\begin{cases} \frac{\partial I}{\partial U} = \frac{1}{2}I - \frac{1}{2}U^{-1}\tilde{R}\tilde{R}^T U^{-1} \\ \frac{\partial I}{\partial \tilde{R}} = R + U^{-1}\tilde{R} \\ \frac{\partial I}{\partial R} = \frac{2T}{N}RPDP^T + \tilde{R} \end{cases} \tag{2.91}$$

to which we find the solution:

$$\begin{cases} R = P^T \\ \tilde{R} = -2\alpha DP^T \\ U = 2\alpha D \end{cases} \tag{2.92}$$

By substituting this into (2.88) and taking the derivatives with respect to $W, L, \tilde{v}$ and p, we can find the rest of the order parameters:

$$\begin{cases} v = \frac{1}{2\alpha}Id_K \otimes D^{-1} \\ \tilde{v} = 0 \\ p = 0 \\ r = \frac{1}{N}\sum_i \bar{e}_i\bar{e}_i^T \otimes \left(U^{-1}(\tilde{R}\bar{e}_i\bar{e}_i^T\tilde{R}^T)U^{-1}\right) = \frac{1}{N}\sum_i \bar{e}_i\bar{e}_i^T \otimes \left(P^T\bar{e}_i\bar{e}_i^T P\right) \\ U = 2\alpha D \\ \tilde{R} = -2\alpha DP^T \\ R = P^T \\ W = 0 \\ \tilde{W} = -\frac{2}{N}\text{Tr}ZId_K = -2\alpha Id_K \\ S = -\frac{4\alpha}{T}Tr_T\left((P^T \otimes Z)(\text{diag}(\bar{\xi^2}) \otimes \bar{\Delta}(Id_T - \frac{1}{T}J_T) + \bar{x}\bar{x}^T)(P \otimes Id_T)\right) \\ L = 0 \end{cases} \tag{2.93}$$

Therefore, for $\bar{\sigma}_i = 0$ and $\bar{\Xi} = 0$ we have the exact solution. To find the solution for the given values of $\bar{\sigma}_i$ and $\bar{\Xi}$, we will slowly interpolate between the zero and desired value of these parameters.

- First, we will gradually introduce the variability in the calcium kernel by increasing $\bar{\Xi}$.
- Then, we will introduce the noise $\bar{z}_{i,t}$ by increasing $\bar{\sigma}_i$.

2.5 Results on the synthetic data.

2.5.1 Electrophysiology

Generating the synthetic data

To test our prediction, we explore a range of values for the parameters of the extracellular electrophysiology model (2.2.5), varying one parameter at a time while holding the others constant. For each set of parameter values, we generate synthetic data using the model (2.14). We then perform PCA on the synthetic data, calculate the two accuracy measures ρ and ϵ, and compare the results with the predictions for ρ and ϵ made based on the parameters used for data generation. This allows us to see how the accuracy of the PCA depends on any given parameter.

- We generate synthetic dataset with two-dimensional latent activity by setting the temporal profile of the modes in the following way:

$$\begin{cases} x_t^{(1)} = a\left(\frac{1}{\sqrt{2}}\sin\left(\frac{2\pi t}{T}\right) + \frac{1}{\sqrt{2}}\sin\left(\frac{4\pi t}{T}\right)\right) \\ x_t^{(2)} = b\left(-\frac{1}{\sqrt{2}}\sin\left(\frac{2\pi t}{T}\right) + \frac{1}{\sqrt{2}}\sin\left(\frac{4\pi t}{T}\right)\right) \end{cases} \tag{2.94}$$

  where coefficients a and b fix a specific variance for $x^{(1)}$ and $x^{(2)}$.

- The modes $e^{(1)}$ and $e^{(2)}$ are selected as two orthogonal vectors from a uniform distribution over N-dimensional normalized vectors.

- For the trial-to-trial noise $\delta x^{(k)}$ that we can also call "directional noise", we select the temporal correlation matrix Δ (see 2.14) as having Gaussian correlation in time, with the correlation width of τ_ξ. The resulting correlation matrix is written as

$$\Delta_{t_1,t_2} = e^{-\frac{-(t_1-t_2)^2}{2\tau_\xi^2}} \tag{2.95}$$

- The variances $(\xi^{(k)})^2$ of directional noise δx are selected as proportional to the variances of the signal components $x^{(k)}$, so that

$$\frac{(\xi^{(1)})^2}{\mathrm{Var}(x^{(1)})} = \frac{(\xi^{(2)})^2}{\mathrm{Var}(x^{(2)})} \tag{2.96}$$

- We select variance of the noise z to be equal for all neurons i, so $\sigma_i = \sigma$.

- The kernel G used for data smoothing is selected as Gaussian with width τ_z.

Results without smoothing kernel

First, we will see how both accuracy measures ρ_i and ϵ depend on the parameters of the system if no smoothing kernel has been applied.

As it can be seen from the figures for ϵ^2 (Fig, 2.3) and ρ (Fig. 2.4), overall the outcome of the PCA on the data generated with electrophysiology model can be separated into two phases: the phase where we observe the partial recovery of the right direction, and the phase where this does not happen at all. This corresponds to the phenomenon of the retarded learning and the phase transition found for the spiked covariance model.

Dependence of the accuracy parameters on the number of neurons.

As expected, the accuracy improves as the number of recorded neurons increases, reflected by the decrease in both ρ and ϵ. However, even with an infinitely large population of neurons, achieving perfect accuracy is impossible. The underlying intuition behind this limitation lies in the increasing dimensionality of the problem. As the number of recorded neurons grows, the dimensional space in which we must identify the correct neural patterns or directions also expands. This higher dimensionality introduces more complexity, making it increasingly difficult to find the right solution.

Dependence of the accuracy parameters on the noise strength.

We see that when the amplitude σ of measurement noise z is set to zero, the principal components are recovered perfectly, as seen from $\langle\rho\rangle = 1$, despite the presence of the trial-to-trial variability noise δx. Despite the perfect recovery of the principal components in this limit, the shape of the neural trajectory is not recovered perfectly, since the trial-to-trial variability noise δx is still present. As noise strength σ increases, the individual components will exhibit a phase transition: both $\left\langle\rho^{(k,k)}\right\rangle$ and $\left\langle\epsilon^{(k,k)}\right\rangle$ will have a phase transition at the same time, with $\left\langle\rho^{(k,k)}\right\rangle$ becoming one, indicating complete loss of the direction recovery, and ϵ exhibiting a cusp.

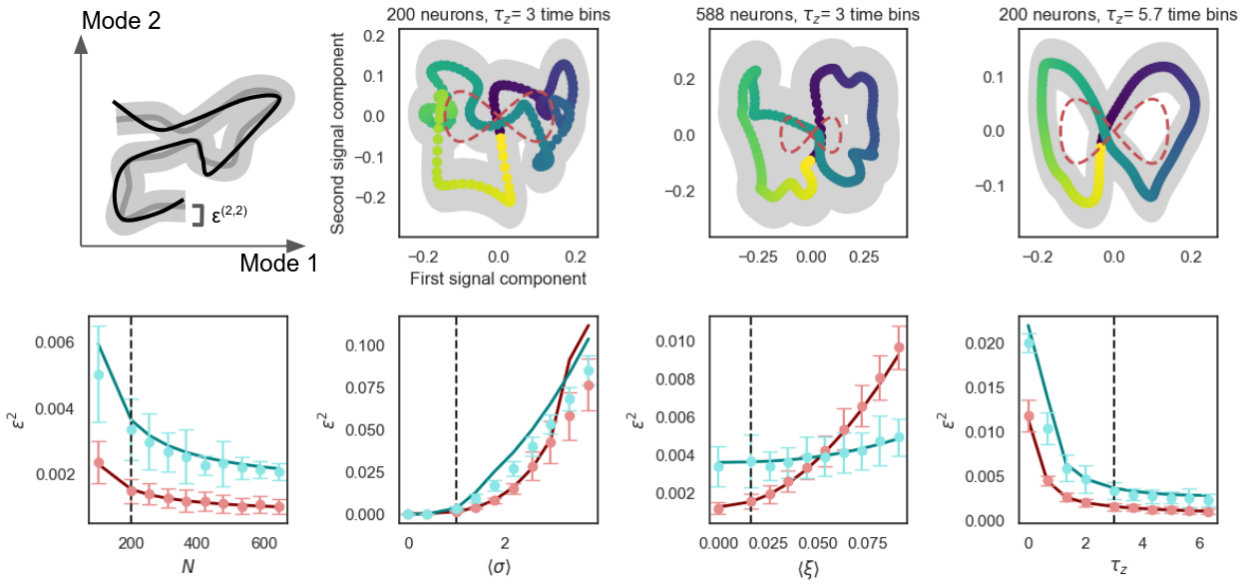

Figure 2.3: Accuracy measure ϵ as a function of different parameters of the dataset done for the case of extracellular electrophysiology. **Top left:** Definition of ϵ as the mean distance between true (black) and inferred (grey) neural dynamics. **Rest of the top row:** Neural trajectories inferred from synthetic data with the two-dimensional 8-shaped latent dynamics. The quality of the inference depends of the parameters of the dataset. The inferred trajectories are shown as a gradient from the first time bin (purple) to the yellow (final time bin).The true neural dynamics is shown by dashed red line. **Bottow row:** Comparison of the prediction and the results of PCA performed on synthetic data with the two-dimensional 8-shaped latent dynamics. $\epsilon^{(1,1)}$ is shown in red, $\epsilon^{(2,2)}$ in cyan. Theoretical predictions are shown as solid lines, average values obtained from generated data are shown with error bars. Across all plots, dataset parameters are varied around a reference point, indicated by a dashed line: $T = 200$ time bins, $N = 200$ neurons, $Var(x^{(1)}) = 0.01$, $Var(x^{(2)}) = 0.002$, $\sigma_i = 1$, $\tau_z = 3$, and $\langle \xi \rangle = 0.017$.

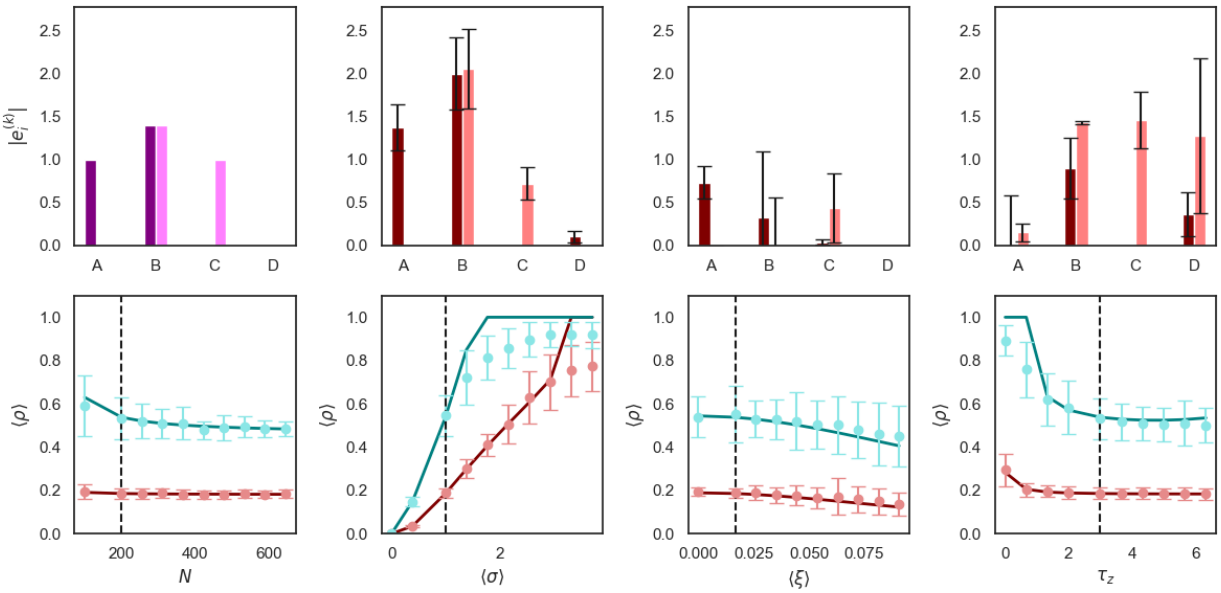

Figure 2.4: Accuracy measure ρ as a function of different parameters of the dataset done for the case of extracellular electrophysiology. **Top left:** Components of the activity modes $e_i^{(k)}$ for four different neurons A,B,C,D in the synthetic data. The first component $e^{(1)}$ is shown in dark purple. The second component $e^{(2)}$ is shown in pink. **Rest of the top row:** Examples of $v_i^{(k)}$ for the same neurons A,B,C,D for synthetic data with different number of recorded neurons and smoothing kernel width. The first component $v^{(1)}$ is given in dark red, the second component $v^{(2)}$ is sigen in light red. **Bottow row:** Comparison of the prediction and the results of PCA performed on synthetic data with the two-dimensional 8-shaped latent dynamics. $\rho^{(1,1)}$ is shown in red, $\rho^{(2,2)}$ in cyan. Theoretical predictions are shown as solid lines, average values obtained from generated data are shown with error bars. Across all plots, dataset parameters are varied around a reference point, indicated by a dashed line: $T = 200$ time bins, $N = 200$ neurons, $Var(x^{(1)}) = 0.01$, $Var(x^{(2)}) = 0.002$, $\sigma_i = 1$, $\tau_z = 3$, and $\langle \xi \rangle = 0.017$.

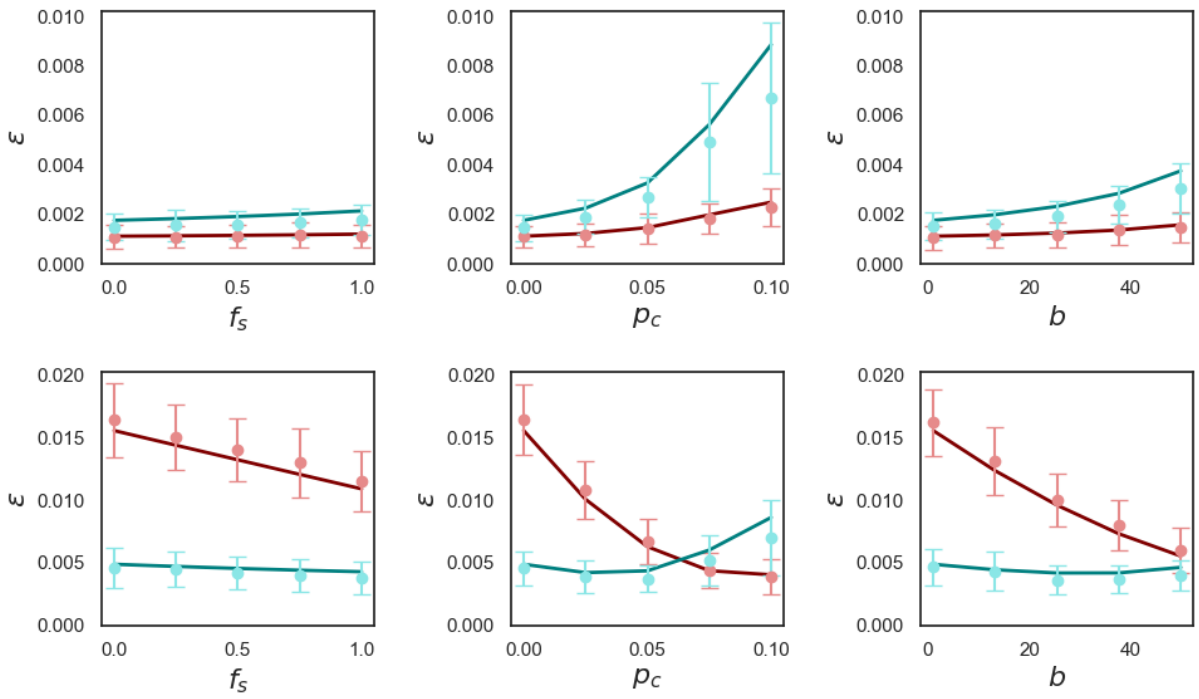

Figure 2.5: Changes in accuracy induced by the errors in the spike sorting. **Left column:** ϵ as a function of the fraction of neurons that were split into two clusters. Top plot shows the behavior of ϵ in case of $\xi^{(1)} = \xi^{(2)} = 0$, $\text{Var}(x^{(1)}) = 0.2$, $\text{Var}(x^{(2)}) = 0.04$. Bottom plot shows the case of $\xi^{(1)} = \xi^{(2)} = 0.15$. **Middle column:** For the same values of ξ and $\text{Var}(x)$, we plot ϵ as a function of the fraction p_c of signal confused between two neurons around the same recording site. **Right column:** For the same values of ξ and $\text{Var}(x)$, we plot ϵ as a function of the number b of neurons around the same recording site.

Dependence of the accuracy parameters on the trial-to-trial variability.

With the increase of the variance ξ of the trial-to-trial variability δx, the inferred neural trajectory becomes more noisy, resulting in the increase of ϵ. However, the presence of such noise will on average help with the recovery of principal components, as indicated by decrease in $\langle \rho \rangle$. This is because of the following: while $\delta x^{(k)}$ is a noise term, it occurs along the direction $e^{(k)}$, same as the signal $x_t^{(k)}$ (see (2.14)). This means that the noise $\delta x^{(k)}$ contributes to increasing the variance along the direction $e^{(k)}$, making it easier to detect.

Dependence of the accuracy parameters on the width of kernel used for smoothing.

Initially, as the width of the kernel increases, the accuracy improves. This makes sense because using a wider kernel effectively reduces the noise in the data. Noise reduction can help enhance the ability to extract meaningful latent dynamics or principal components from the data, improving the performance of PCA.

At this stage, more data points are averaged together, which increases the signal-to-noise ratio (SNR). As the kernel smooths the data, it becomes easier to identify the true underlying patterns in neural activity.

However, there is a limit to how much the kernel width can be increased before it starts to negatively impact the analysis. When the kernel becomes too wide, the smoothing begins to have a detrimental effect: the true structure of the neural dynamics becomes "blurred" or flattened because the kernel is no longer simply reducing noise—it is also smoothing out the actual signal. As a result, fast variations in the latent dynamics are suppressed. At this point, accuracy starts to decline, as the excessive smoothing averages out important signal features, making it harder to accurately capture the true underlying neural dynamics.

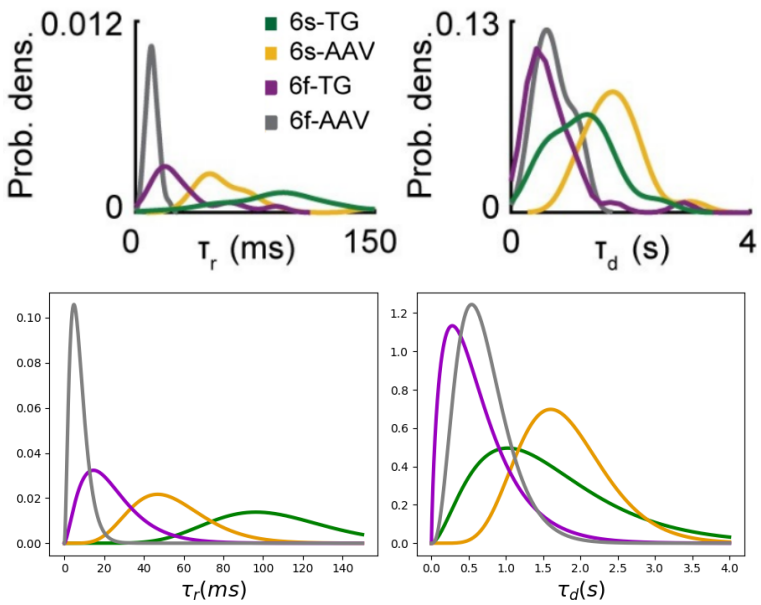

Figure 2.6: Top row: Rise and decay times for calcium indicators GCaMP6s and GCaMP6f measured for two different modes of delivery in the tissue: transduction with adeno-associated virus (AAV) and transgenic expression (TG). Adapted from Wei et al. (2020). Bottom row: Gamma-distributions for the rise and decay times taken for the generation of synthetic calcium imaging data.

Spike sorting-induced effects

It is easy to imagine that accuracy tends to decrease as the spike sorting becomes worse, such as with the increase of the number of split neurons, or with larger provability of confusing the signal between the neurons around the same electrode. This idea is intuitive because splitting and confusing neural activity effectively scales the original signal. This effect can be explicitly observed in equation (2.29), where the scaling of the signal becomes apparent. In the specific case where the directional noise δx is absent, we observe a clear, monotonic increase in the error terms $\epsilon^{(1,1)}$ and $\epsilon^{(2,2)}$ (Fig. 2.5).

However, when the variance ξ of the directional noise δx is not zero, the situation becomes more complex. In this case, signal confusion and signal splitting not only reduce the effective variance of the signal itself, but they also reduce the effective variance of the directional noise δx. As we saw from Fig. 2.4, δx may be beneficial for the recovery of the principal components, However, as we saw from Fig. 2.3, δx also decreases the accuracy of the reconstruction of the shape x_t of the neural trajectory. Therefore, effective scaling of δx by the spike sorting induced effects may result in non-monotonic behaviour, as seen in the bottom row of Fig. 2.5.

2.5.2 Calcium imaging: Results

To generate synthetic neural population recordings that mimic realistic calcium imaging data, it is important to choose appropriate values for the rise and decay times of the fluorescence signals and to understand how much these values typically fluctuate from neuron to neuron.

To estimate the general order of magnitude for rise and decay times, as well as their variability, we use the distributions reported by Wei et al. (2020) for different variants of the widely used calcium indicator GCaMP6, expressed in the left anterolateral motor cortex of mice. The authors provide rise and decay times for GCaMP6s and GCaMP6f under two modes of delivery: transduction via adeno-associated virus and transgenic expression (Fig. 2.6, top row).

For generating the kernels in our synthetic data, we approximate the distribution of rise and decay times using Gamma distributions. We capture the general shape by selecting means and variances that are in the same order of magnitude as the measured distributions. This level of approximation is sufficient for our purposes, as we aim to test the prediction of accuracy for any

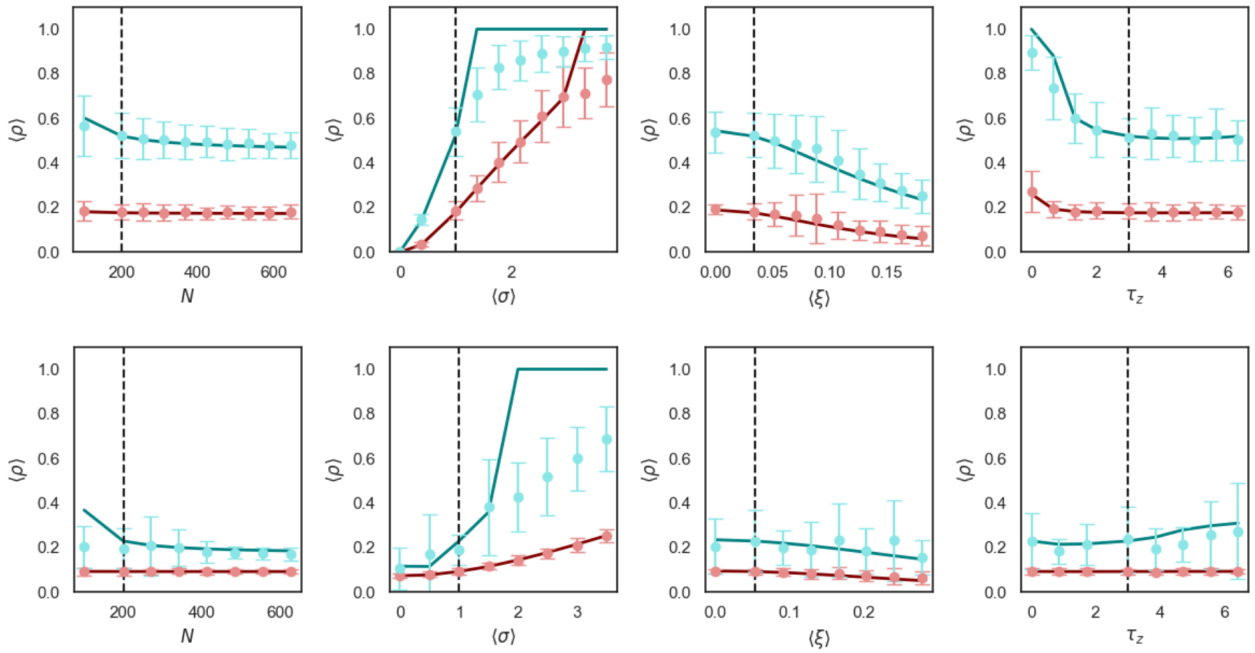

Figure 2.7: Comparison of the effects of the slow fluorescent dynamics on the accuracy measure ρ. Top row: Average rise and decay times are selected similar to "6s-AAV" case. Bottow row: Same for the "6s-TG" case.

given parameters of rise and decay times.

Dependence of the accuracy parameters on the variations in the shape of the calcium imaging kernel.

The presence of slow fluorescence kernel does not affect the overall trends in the dependencies of ϵ and ρ on the different parameters of the model: as can be expected, accuracy still increases with the number of neurons, decreases with the noise strength σ, and benefits from the wide smoothing kernel. The directional noise δx will stay beneficial for the recovery of the right directions of the principal components, but make the accuracy of the recovery of the shape of the neural trajectory worse.

However, the difference in rise and decay time of the fluorescence produce several effects not observed for the model of electrophysiology data. Notably, even in the absence of the noise z, (so $\sigma_i = 0$), the recovery of the direction of the principal component is not perfect, as seen by $\langle\rho\rangle > 0$ (Fig 2.7). The effect is more pronounced for the large variations in the rise and decay times.

The presence of the kernel also introduces off-diagonal terms, absent for the extracellular electrophysiology case.

2.6 Multiple Trial Data

Most real-world experiments consist of multiple trials. We have seen before the importance of the trial-to-trial variability in our analysis. However, until now, the fact that the dataset were consisting in an ensemble of trials was just taken into account in the quantity $\delta x^{(k)}$, which represents the fluctuations of the signal across trials. In this section, we want to precise how we can analyze datasets that consist of many trials. In particular, in the literature, these experiments are typically processed in one of two ways: either by averaging across trials to produce a single matrix of trial-averaged firing rates, or by concatenating the data from multiple trials into a single, extended time series matrix. We aim thus at adapting our analysis to these different types of analysis.

2.6.1 Modeling Trial Variability

When considering multiple trials, it is important to differentiate between the components of the data that remain consistent across trials and those that vary from one trial to the next. Specifically:

- **Latent Dynamics Stability:** The signal components representing latent neural dynamics, consisting of the neural modes $e_i^{(k)}$ and their temporal profiles of activation $x_t^{(k)}$ are assumed to reflect underlying neural processes that remain stable across trials. Therefore, these components will be identical for each trial.

- **Trial-Dependent Noise:** The noise terms $z_{i,t}$ and $\delta x_t^{(k)}$ play distinct roles in the model: $z_{i,t}$ captures the measurement noise, while $\delta x_t^{(k)}$ models the variability in the latent neural dynamics across trials. It is natural to assume that both sources of noise will vary from trial to trial, reflecting the inherent differences in both the experimental conditions and the underlying neural activity.

- **Measurement Artifacts Consistency:** We assume that any systematic distortions introduced by the recording method (whether electrophysiology or calcium imaging) are consistent across trials. For instance, if a particular neuron is overclustered in one trial, it will be overclustered in all trials. Similarly, for calcium imaging, the rise and decay times of the fluorescence signal for a given neuron are expected to remain the same across trials. These assumptions ensure that trial-to-trial variability is driven primarily by noise, rather than changes in the measurement process itself.

2.6.2 Trial-Averaged Data

When data from multiple trials are averaged, the firing rates for each neuron are computed across all trials, resulting in a single matrix that represents the trial-averaged activity. This averaging process reduces the influence of noise and thus requires an adjustment to the noise terms in our model.

For a dataset consisting of M trials, the trial-averaged data, with N observed neurons and T time bins per trial, can still be described using the previously introduced models. However, the variances of the noise terms must be scaled by a factor of $1/M$, reflecting the reduction in noise magnitude due to averaging over multiple trials. Specifically, the noise terms scale as:

$$\xi^{(k)}_{(M \text{ trials})} = \frac{1}{\sqrt{M}} \xi^{(k)}_{(\text{single trial})}, \quad \sigma_{(M \text{ trials}),i} = \frac{1}{\sqrt{M}} \sigma_{(\text{single trial}),i} \tag{2.97}$$

The effect of this noise reduction on the accuracy of recovering the latent dynamics can be both positive and negative. While decreasing the variance σ^2 of $z_{i,t}$ (the measurement noise) is always beneficial—as illustrated in Fig. 2.3, Fig. 2.4, and Fig. 2.7—reducing the variance $(\xi^{(k)})^2$ of $\delta x^{(k)}$ (the variability in latent dynamics) may impair the recovery of the principal component $e^{(k)}$.

In such situations, averaging across trials can result in a mean activity with a much smaller amplitude compared to an individual trial. This represents a common scenario in real data, where high trial-to-trial variability causes the trial-averaged data to "flatten" the latent activity, making it harder to accurately recover the underlying neural dynamics.

2.6.3 Trial-Concatenated Data

An alternative approach is to concatenate data from all trials into one long time series. This method involves stacking the time bins from multiple trials, creating a single data matrix with $T \times M$ time bins. In this case, each trial's data is appended to the previous trial's data, without any averaging.

For trial-concatenated data, the noise variances remain unchanged because no averaging has been performed. However, the length of the time series increases, as the number of time bins is now the product of the number of trials and the number of time bins per trial ($T \times M$). The signal

components $x_t^{(k)}$ that represent the latent dynamics of the system are repeated across each trial, reflecting the assumption that these dynamics remain consistent across trials.

This approach is particularly suitable for cases where there is high trial-to-trial variability in the latent dynamics, as it preserves the differences in latent activity across trials, avoiding the issue of averaging out variability that can occur in trial-averaged data. However, unlike in the trial-averaged approach, the noise level σ does not decrease in the concatenated data, as no averaging has been applied. This means that while the variability between trials is preserved, the noise level remains constant, which may still impact the accuracy of recovering the latent dynamics.

CHAPTER 3

Application to synthetic data: parameter inference and design

In this chapter, we will apply the above theoretical results in a statistically consistent setting with controlled synthetic data. The setting is realistic in the sense that we have access to data, but not to the characteristic parameters of their distributions. Those must be inferred from the data, as we explain below. Once inference is done, we can use our calilbrated theoretical model to predict the expected performance in situations differing from the data set, for instance for larger number of trials or of recorded neurons.

3.1 Procedure for parameter inference

We saw above that the prediction of accuracy relies on knowing many parameters of the data. While some of them, like the number of neurons, number of time bins and the shape of the kernel used for smoothing are known, some other parameters, such as signal-to-noise ratio, the shape of the latent activity itself, and the temporal correlations in the noise must be somehow inferred (approximated) from the data before doing the prediction.

- **Shape of the latent activity of $\bar{x}_t^{(k)}$.** As we see from the expression of I in (2.85), the saddle point equations will not depend on the variances var$(x^{(k)})$ of the components only, but on all the points $x_t^{(k)}$. We will approximate their values with $y_t^{(k)}$, potentially scaled to obtain right signal to noise ratio (the procedure will be described below).

- **Composition of neural assemblies $e_i^{(k)}$.** We will take the best available approximation for $e_i^{(k)}$, which is the inferred principal component $v_i^{(k)}$.

- **Inferring the Temporal Correlations in the Trial-to-Trial Variability Term δx.** In the previously introduced model, we assumed that the trial-to-trial variability, $\delta\bar{x}$, exhibits some degree of temporal correlation. To simplify the inference of these correlations, we further assume that they can be described by a characteristic temporal correlation timescale $\tau_{\delta x}$. As a simple model, we take these correlations to be Gaussian in time with the associated covariance kernel:

$$\bar{\Delta}_{t_1,t_2} = \exp\left(-\frac{(t_1-t_2)^2}{2\tau_{\delta x}^2}\right) \tag{3.1}$$

  The general idea is as follows: We begin by taking the trial-averaged temporal profile of the k-th principal component, $\bar{y}_t^{(k)}$ (also known as the score), and subtract it from the projection of a single trial onto the corresponding inferred principal component $v^{(k)}$. If the direction of the principal component is inferred well ($v_i^{(k)} \approx e_i^{(k)}$), then the difference should approximately be equal to

$$\delta\bar{y}_t^{(k)} = \delta\bar{F}x_t^{(k)} + \delta\bar{x}_t^{(k)} + \frac{1}{N}\sum_i v_i^{(k)}\bar{z}_{i,t} \tag{3.2}$$

  For the case of extracellular electrophysiology recordings, the term $\delta\bar{F}x_t^{(k)}$ is absent, and we end up with $\delta\bar{y}_t^{(k)}$ being the mixture of two types of noise, $\delta\bar{x}^{(k)}$ and $\bar{z}$.

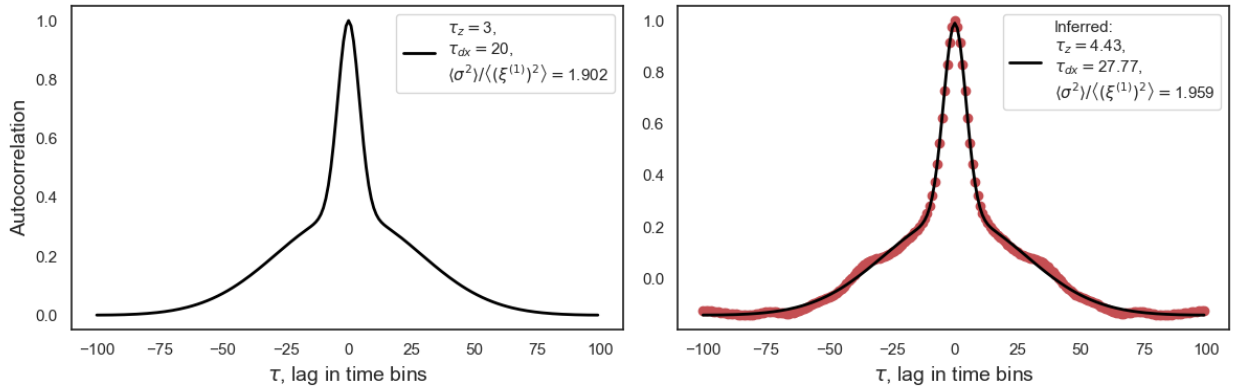

Figure 3.1: Inference of the noise temporal correlation times and amplitudes for electrophysiology data. Left: Autocorrelation imposed by the model, with correlation time τ_z for the noise term z, and correlation time $\tau_{\delta x}$ for the directional term δ. Right: Empirical autocorrelation inferred from 100 trials of generated synthetic data.

First, we can calculate $\delta\bar{y}_t^{(k)}$ independently for every trial available. Then, we calculate the temporal autocorrelation. For a single trial, the autocorrelation function $C_{\delta\bar{y}}(\tau)$ at lag τ is calculated as

$$C_{\delta\bar{y}}(\tau) = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau}\delta\bar{y}_t^{(k)}\delta\bar{y}_{t+\tau}^{(k)}, \tag{3.3}$$

where T is the total length of the time series. Then, we can the average of resulting autocorrelation over all available trials.

The resulting autocorrelation should resemble the sum of two Gaussians: one arising from $\bar{z}_{i,t}$ and the other from $\delta\bar{x}_t^{(k)}$ (Fig. 3.1):

$$\langle C_{\delta\bar{y}}(\tau)\rangle \sim \langle\bar{\sigma}_i^2\rangle \exp\left(-\frac{\tau^2}{2\tau_z^2}\right) + (\bar{\xi}^{(k)})^2 \exp\left(-\frac{\tau^2}{2\tau_{\delta x}^2}\right) \tag{3.4}$$

Then, we can find back the values of $\tau_{\delta x}$, and the ratio $\frac{\langle\bar{\sigma_i}^2\rangle}{(\xi^{(k)})^2}$ by fitting a double exponential curve

$$a_1 \exp\left(-\frac{\tau^2}{2b_1^2}\right) + a_2 \exp\left(-\frac{\tau^2}{2b_2^2}\right) + c \tag{3.5}$$

where c is an additional constant used to accommodate for the finite-size effects of the system. We fit the coefficients a_1, a_2, b_1, b_2, c by using the method of least squares, consisting of minimizing the sum of the squared difference between the double exponential and the observed mean autocorrelaton. Then, assuming that $b_1 < b_2$, the inferred value of b_1 will approximate τ_z, and the inferred value of b_2 will give τ_{δ_x}. In this case a_1/a_2 will be approximately equal to $\frac{\langle\sigma_i^2\rangle}{(\xi^{(k)})^2}$.

In the case of the calcium imaging, the additional term $\delta\bar{F}x_t^{(k)}$ contribures to the correlation as

$$\frac{1}{T-\tau}\sum_{t=1}^{T-\tau}\delta\bar{F}x_t^{(k)}\delta\bar{F}x_{t+\tau}^{(k)} = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau}\sum_{t_1=0}^{T-1}\sum_{t_2=0}^{T-1}\delta\bar{F}_{t_1}\delta\bar{F}_{t_2}x_{t-t_1}^{(k)}x_{t+\tau-t_2}^{(k)} \tag{3.6}$$

which, after averaging on trials, can be rewritten using the covariance matrix $\bar{\Xi}$ (2.46):

$$\left\langle\frac{1}{T-\tau}\sum_{t=1}^{T-\tau}\delta\bar{F}x_t^{(k)}\delta\bar{F}x_{t+\tau}^{(k)}\right\rangle = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau}\sum_{t_1=0}^{T-1}\sum_{t_2=0}^{T-1}\bar{\Xi}_{t_1,t_2}x_{t-t_1}^{(k)}x_{t+\tau-t_2}^{(k)} \tag{3.7}$$

Finally, we will approximate $x_t^{(k)}$ in this expression by the scores $y_k^{(k)}$. Then, we replace the variational expression in 3.5 with the following one to carry out the fit,

$$a_1 \exp\left(-\frac{\tau^2}{2b_1^2}\right) + a_2 \exp\left(-\frac{\tau^2}{2b_2^2}\right) + \frac{a_3}{T-\tau}\sum_{t=1}^{T-\tau}\sum_{t_1=0}^{T-1}\sum_{t_2=0}^{T-1}\bar{\Xi}_{t_1,t_2} y_{t-t_1}^{(k)} y_{t+\tau-t_2}^{(k)} + c \tag{3.8}$$

- **Inference of the Signal-to-Noise Ratio.** To infer the signal-to-noise ratio, we will predict the average top eigenvalues

$$\lambda^{(k)} = \left\langle \frac{1}{N}\sum_{i,j} v_i^{(k)} C_{ij} v_j^{(k)} \right\rangle \tag{3.9}$$

and compare them to the top eigenvalues $\mu^{(k)}$ obtained by applying PCA to the dataset. Replica calculation described above can be easily modified for finding $\lambda^{(k)}$. For that, we simply need to add a source term to the partition function (2.48):

$$\begin{aligned} Z(\{s_{i,t}\}) \;=\; & \int \prod_k d\mathbf{v}^{(k)} \int \prod_{k,t} dy_t^{(k)} \prod_{k,t} \delta\left(\sum_i (v_i^{(k)})^2 - N\right) \prod_{k_1,k_1:k_1<k_2} \delta\left(\sum_i v_i^{(k_1)} v_i^{(k_2)}\right) \\ & \prod_{k,t} \delta\left(y_t^{(k)} - \frac{1}{N}\sum_i v_i^{(k)} s_{i,t}\right) \times \exp\left(\beta T \frac{1}{N^2}\sum_{k=1}^{K}\sum_{i,j=1}^{N} v_i^{(k)} C_{ij} v_j^{(k)}\right) \\ \times \;& \exp\left(\phi^{(k)} \beta \frac{1}{N}\sum_{i,j=1}^{N} v_i^{(k)} C_{ij} v_j^{(k)}\right) \end{aligned} \tag{3.10}$$

and carry out the calculation as before (see Appendix for details). The outcome is the following analytical expression for $\lambda(k)$:

$$\begin{aligned} \lambda^{(k)} = \frac{N}{T}\frac{\partial I}{\partial \theta^{(k)}}\bigg|_{\theta^{(k)}=0} = \frac{1}{T}\mathrm{Tr}_T \Big[(Id_{K,T} - 2\,(W \otimes Z + v \cdot \mathcal{X}))^{-1} \\ \Big(L \otimes Z + r \cdot \mathcal{X} + (R \otimes Id_T)(\mathrm{diag}(\bar{\xi}^2) \otimes \bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right) + \tilde{x}\tilde{x}^T)(R \otimes Id_T)^T \Big) \\ (Id_{K,T} - 2\,(W \otimes Z + v \cdot \mathcal{X}))^{-1} \Big]^{(k,k)} \end{aligned} \tag{3.11}$$

We can then minimize the difference between $\mu^{(k)}$ and $\lambda^{(k)}$ to determine $\mathrm{var}(x)$. The inference process follows these steps:

1. **Initial Assumption:** We start by assuming that the variance of the principal components, $\mathrm{Var}(\bar{x}_t^{(k)})$, equals $\mu^{(k)}$. While this assumption is not entirely accurate (as the projection on the principal components always contains some noise), it provides a useful starting point for further calculations.
2. **Approximating Noise Variance:** To refine this initial estimate, we approximate the noise variance as

$$\mathrm{var}(s_{i,t} - \sum_k \bar{x}_{i,t}^{(k)} e_i^{(k)}) \tag{3.12}$$

where $s_{i,t}$ represents the observed data, $\bar{x}_{i,t}^{(k)}$ is the projection onto the k-th principal component, and $e_i^{(k)}$ is the corresponding eigenvector. Using this expression, we infer the relative variances of $\bar{z}_{i,t}$ (the true signal) and $\delta\bar{x}_t^{(k)}$ (the noise).

3. **Predicting Eigenvalues:** Using the inferred variances, we predict the eigenvalues $\lambda^{(k)}$ for each $k = 1, \ldots, K$. The goal is to minimize the difference between $\lambda^{(k)}$ and $\mu^{(k)}$. To achieve this, we compute the gradient of the squared difference, $(\lambda^{(k)} - \mu^{(k)})^2$, with respect to the variance of the signal components, $Var(x^{(k)})$.

4. **Updating Variances:** Finally, we update the estimate of the variance $Var(\bar{x}^{(k)})$ iteratively, adjusting it to minimize the difference between the predicted and observed eigenvalues.

3.2 Benchmarking

In this section, we apply our inference method to synthetic data for the sake of validation. For simplicity, we will demonstrate the performance of the method applied to the synthetic electrophysiology data, where trial-to-trial variability $\delta x^{(k)}$ is excluded.

We generate the data using the same parameters as the ones used for Fig. 2.3 and Fig. 2.4, aside from $\xi^{(k)}$, which we set to zero. As in the previous case, we will vary one parameter at a time, keeping the rest fixed, to see how the reliability of the inference changes with the varied parameter.

The benchmarking follows these steps:

- We generate multiple realizations of synthetic neural data for each specific set of parameters (see 2.14). For each realization, we perform PCA, and using the inferred principal components and scores, we calculate empirical values for both accuracy measures ρ and ϵ.

- We then apply the inference procedure described above to each synthetic data realization, recovering the signal-to-noise ratio, noise correlation times, and the latent neural dynamics.

- Using the inferred parameters from each realization, we compute predictions for the expected average values of ρ and σ. By comparing these predictions with the empirical values, we assess the accuracy of our inference method for the synthetic data.

Fig. 3.2 illustrates the performance of our inference method in estimating the accuracy measures ρ and ϵ for synthetic electrophysiology data. The figure presents results for two principal components, shown in red and cyan, with error bars representing empirically calculated values from multiple realizations of the synthetic dataset.

We see that there is some variability in the predicted accuracy depending of the particular realisation of the data, which decreases with the increase of accuracy. This is likely due to the fact that our procedure for the inference of the parameters relies on finding the signal-to-noise ratio that minimizes the difference between the eigenvalues of the empirical covariance matrix and the predicted average values for these eigenvalues. However, due to the presence of noise, the eigenvalues of the empirical covariance matrix can depend from one data realization to the other.

The figure confirms that our inference method performs effectively, especially when the correlation parameter ρ is small, indicating that the direction is well inferred (we indirectly make this assumption when approximate the shape of $x_t^{(k)}$ with $y_t^{(k)}$, and $e_i^{(k)}$ with $v_i^{(k)}$). Under these conditions, the prediction accuracy improves, showing that lower noise in the directional inference allows for more reliable parameter estimation. The first principal component is inferred with higher accuracy than the second, which comes from the fact that first component captures more signal variance, and thus is less susceptible to noise interference.

Figure 3.3 shows how the errors on the predicted observables change depending on the parameter varied in the data. We observe that the errors decay with the number N of neurons and increase with the noise in the data as expected. The scalings of $\Delta\langle\rho\rangle$ and $\Delta\epsilon$ seem to be compatible with a $1/\sqrt{N}$ algebraic dependence.

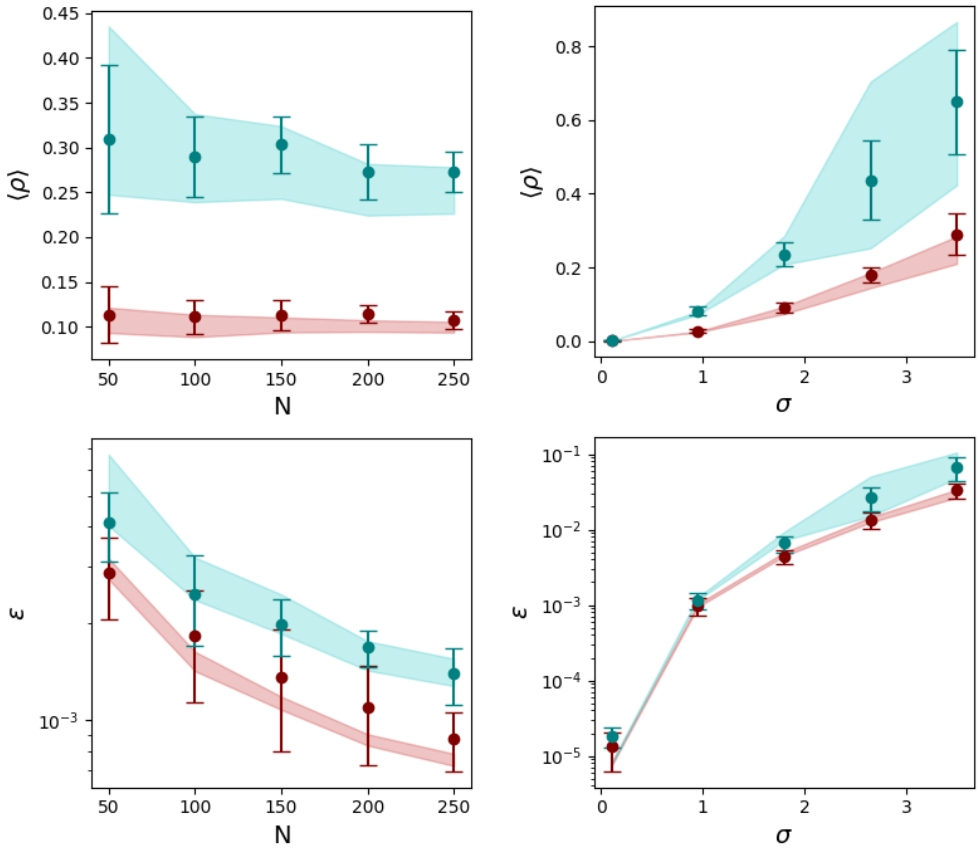

Figure 3.2: Inference of accuracy parameters ρ and ϵ on the synthetic data. Here, we show the inference for two principal components (shown in red - 1st component - and cyan - 2nd component). The error bars show empirically calculated values for each parameter based on repeated data realizations. For theoretical predictions, ρ and ϵ were calculated for each data realization separately; the plot shows one standard deviation around the mean (shaded areas).

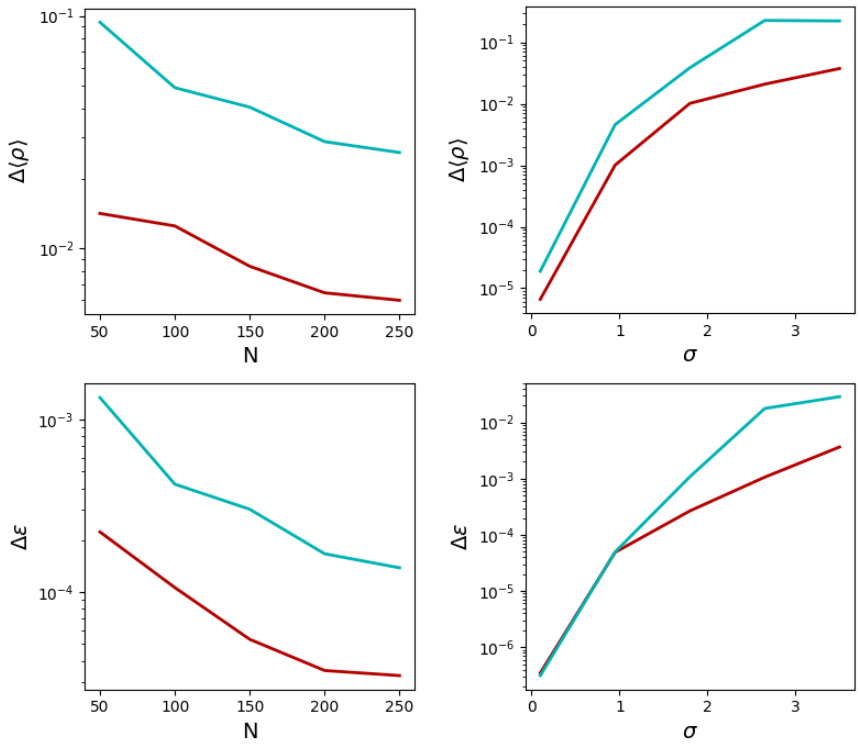

Figure 3.3: Error in the inference of accuracy parameters ρ and ϵ on the synthetic data. Here, we show the error in the inference for two principal components (shown in red and cyan) based on repeated data realizations.

3.3 Experimental design

In the previous sections, we showed how to calculate the accuracy of PCA for a model of neural data that is designed to reflect key features of real neural recordings. By applying an approach based on statistical mechanics, we were able to derive the average accuracy of PCA across all possible datasets generated from a given set of parameters. This self-averaging property entails we do not need the actual dataset to make predictions about PCA accuracy: knowledge the parameters defining the data distribution is sufficient.

As a result, we can calculate the average accuracy for a hypothetical dataset — even if we do not have the actual data available. This ability is very useful for experimental design. Suppose we have a small preliminary dataset; if we assume that this sample is representative of the larger population, we can extract the necessary data parameters from it to use in our model for predicting accuracy.

By adjusting any of these parameters within the model, we can predict the average accuracy for a hypothetical dataset that corresponds to new values of those parameters. For example, we might want to know how increasing the number of neurons affects the accuracy of PCA. Starting with recordings from just 50 neurons, we can use our model to predict what the average accuracy would be if we had recordings from 500 neurons. This approach allows us to make informed predictions about experimental outcomes without needing to collect large amounts of new data upfront.

The pipeline for the prediction of the accuracy is given in the Fig.3.4:

- **Dataset preprocessing.** Before any preprocessing, the dataset consists of many individual trials, each containing a matrix of firing activity of the recorded group of neurons. To calculate the covariance matrix of the activity of the neurons, the data from different trials needs to be combined into one matrix, either by trial-averaging or trial-concatenation (see Section 2.6). After this, the activity is convoluted with a kernel to reduce the noise.

- **Inference of low-dimensional structure with PCA.** The covariance matrix of the prepro-

cessed data is used to find the top principal components $v^{(k)}$ and the scores $y^{(k)}$.

- **Inference of the latent parameters.** As described in the Section 3.1, the projections of individual trials on the inferred principal components $v^{(k)}$ can be used to infer the correlation time of the direction noise δx, as well as to infer the ratio of amplitudes between the two types of noise $\delta x^{(k)}$ and z. We approximate $e^{(k)}$ with $v^{(k)}$, and $x^{(k)}$ with a scaled version of $y^{(k)}$, where the scaling is done by matching the predicted top eigenvalues to the real top eigenvalues of the covariance matrix of the data.

- **Modification of dataset parameters for a larger dataset size.** To be able to make the prediction for a larger dataset size, some of the inferred parameters must be modified.

  – If the prediction is done for a dataset with larger number of neurons, we have to provide the mode coefficients $e_i^{(k)}$ and noise σ_i for all of the new neurons i. This can be done by sampling them from the existing values. To preserve the statistics of $e_i^{(k)}$, the new values for k-th mode are being sampled only from the existing values of the same mode. After sampling, we normalize the new $e^{(k)}$ to ensure they satisfy the requirement of the model of the data used for prediction, in which $|e^{(k)}|^2 = N$.

  – If the prediction is done for a trial-averaged dataset with larger number of trials, the noise amplitudes σ_i and $\xi^{(k)}$ must be scaled, reflecting the fact that trial-averaging reduces the noise (see Section 2.6). If the available dataset has M_{old} trials, and we want to predict the accuracy on the hypothetical dataset with M_{new} trials, both σ_i and $\xi^{(k)}$ must be scaled as $\sigma_{i,new} = \sqrt{\frac{M_{\text{old}}}{M_{\text{new}}}}\sigma_i$ and $\xi^{(k)}_{new} = \sqrt{\frac{M_{\text{old}}}{M_{\text{new}}}}\xi^{(k)}$.

  – If the prediction is done for a trial-concatenated dataset with larger number of trials, the concatenated matrix of the neural activity will become longer. This means that the inferred $x_t^{(k)}$, that already consisted of M_{old} repetitions of the same signal, will need to be expanded further to have a total length of $M_{new} \times T$, so M_{new} repetitions of the signal overall. To do this, we will sample the $x_t^{(k)}$ of the lacking trials from the inferred values.

- **Prediction of the accuracy.** After the modification, the new set of parameters can be used to find the accuracy measures ρ and ϵ by finding the optimum of the free energy, as described in (2.4.4).

Before trying it on the real data, we will test these predictions on the synthetic data. As in the case of the parameter inference, this was only tested on the electrophysiology model of the data, and only without the directional noise term δx. For this we will define all the model parameters (see 2.14), generate the synthetic dataset, perform all the steps described above, and empirically validate the accuracy prediction by generating many realizations of the dataset of the larger size and calculating the accuracy parameters ϵ and ρ empirically.

Fig. 3.5 shows an example of such prediction done on a synthetic dataset with two-dimensional latent activity. We show the prediction for the larger number of neurons and larger number of trials in the case when the dataset is trial-averaged. We test the predictions based on the datasets of different sizes. The figure confirms our intuitive expectations: the more data is available, the more precise predictions are. Our model makes this intuition quantitative, allowing us to estimate the number of trials or neurons needed to achieve the desired precision, for example to test a hypothesis on the nature of trajectories or the mixed selectivity of neurons.

As was the case for the inference of ρ and ϵ (see Fig. 3.2), we observe a variability in the predicted accuracy coming from the fluctuations in the top eigenvalues of the empirical covariance matrix. We see that this error decreases with increased number of available neurons, since it effectively increases overall signal-to-noise ratio, and large signal-to-noise ratio corresponds to smaller fluctuations in the top eigenvalue. Trial-averaging also increases signal-to-noise ratio, resulting in a similar decrease of the fluctuation. Trial-concatenation does not change signal-to-noise ratio, but provide more independent observation points, which also contributes to the decrease of the fluctuation.

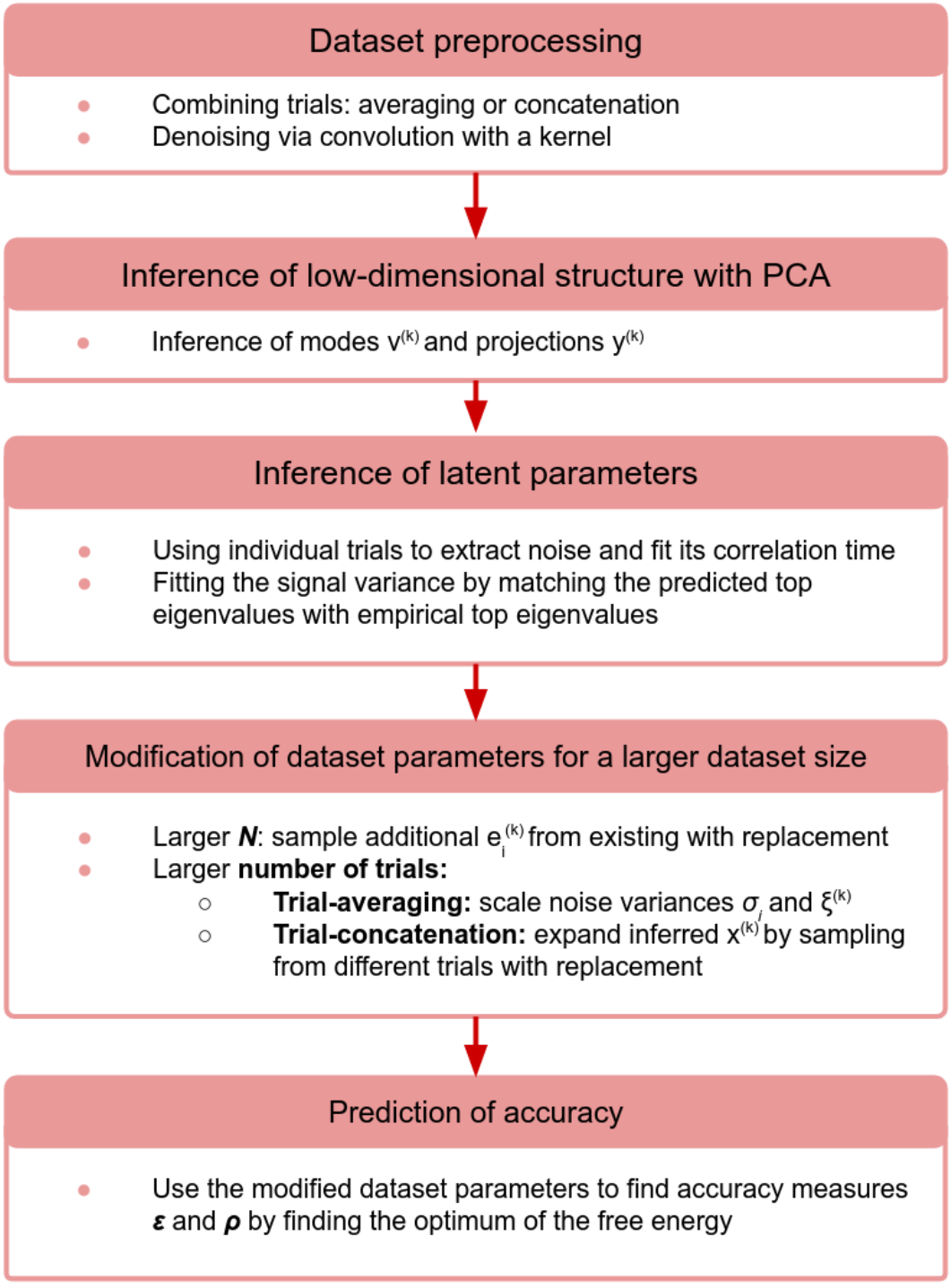

Figure 3.4: General pipeline for prediction of hypothetical experimental results based on a limited dataset.

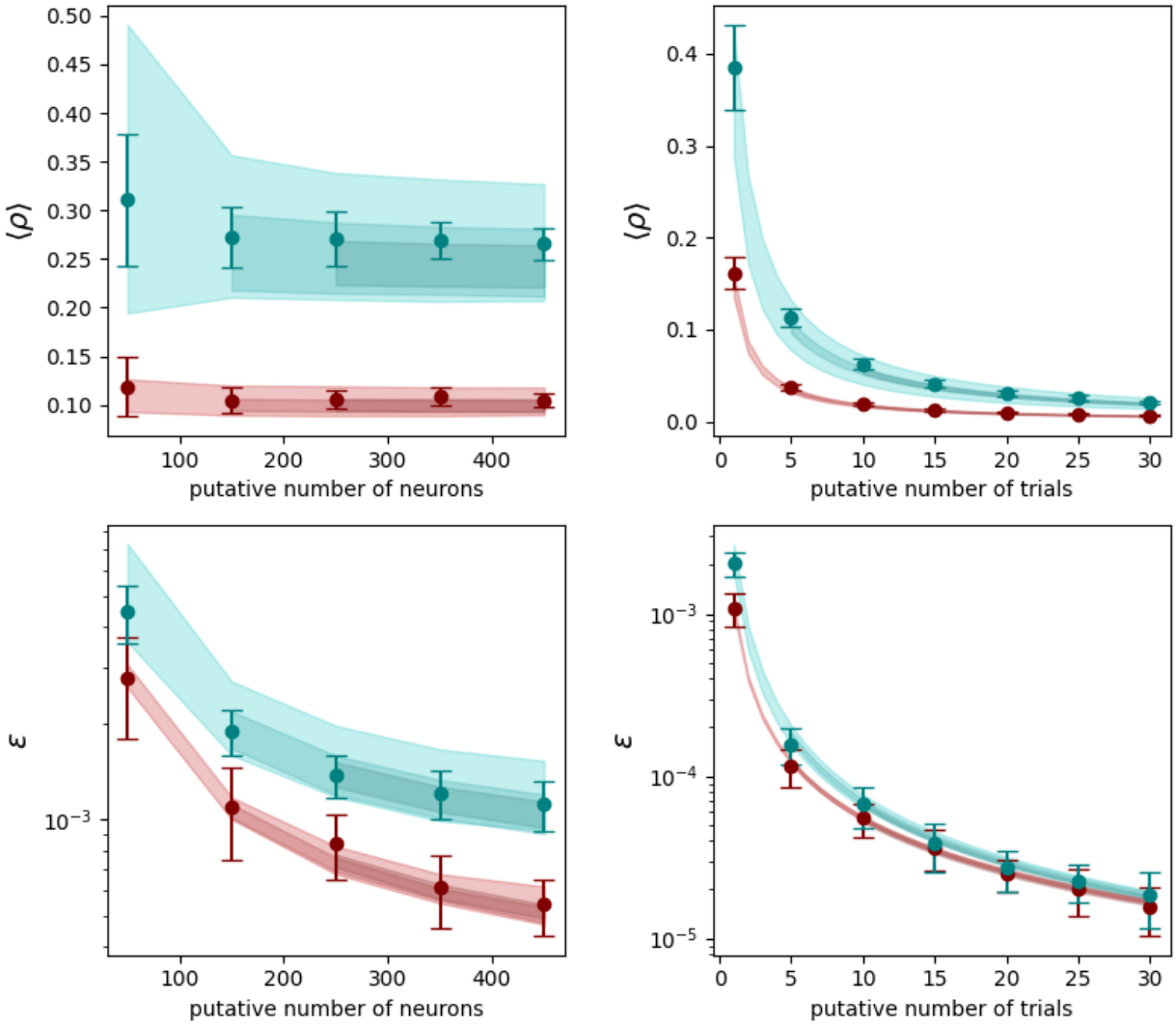

Figure 3.5: Inference and prediction of the accuracy parameters ρ and ϵ as functions of the number of available neurons and of trials. Left column: prediction for hypothetical datasets with larger number of neurons. Parameters necessary for the estimation of σ and ρ were independently inferred from the datasets containing 50, 150 or 250 neurons. Then the inferred parameters were used for predicting the accuracy of potential larger dataset sizes (up to 500 neurons). The three predictions are shown as shaded areas, each area showing one standard deviation around the mean of many predictions done on independent realizations of the data. The real values of ρ and ϵ for these realisations of data are shown in error bars. Right column: same for the different number of trials, combined with trial-averaging. The prediction were done on data containing 1, 5 and 10 trials. Same color code as in Fig. 3.2.

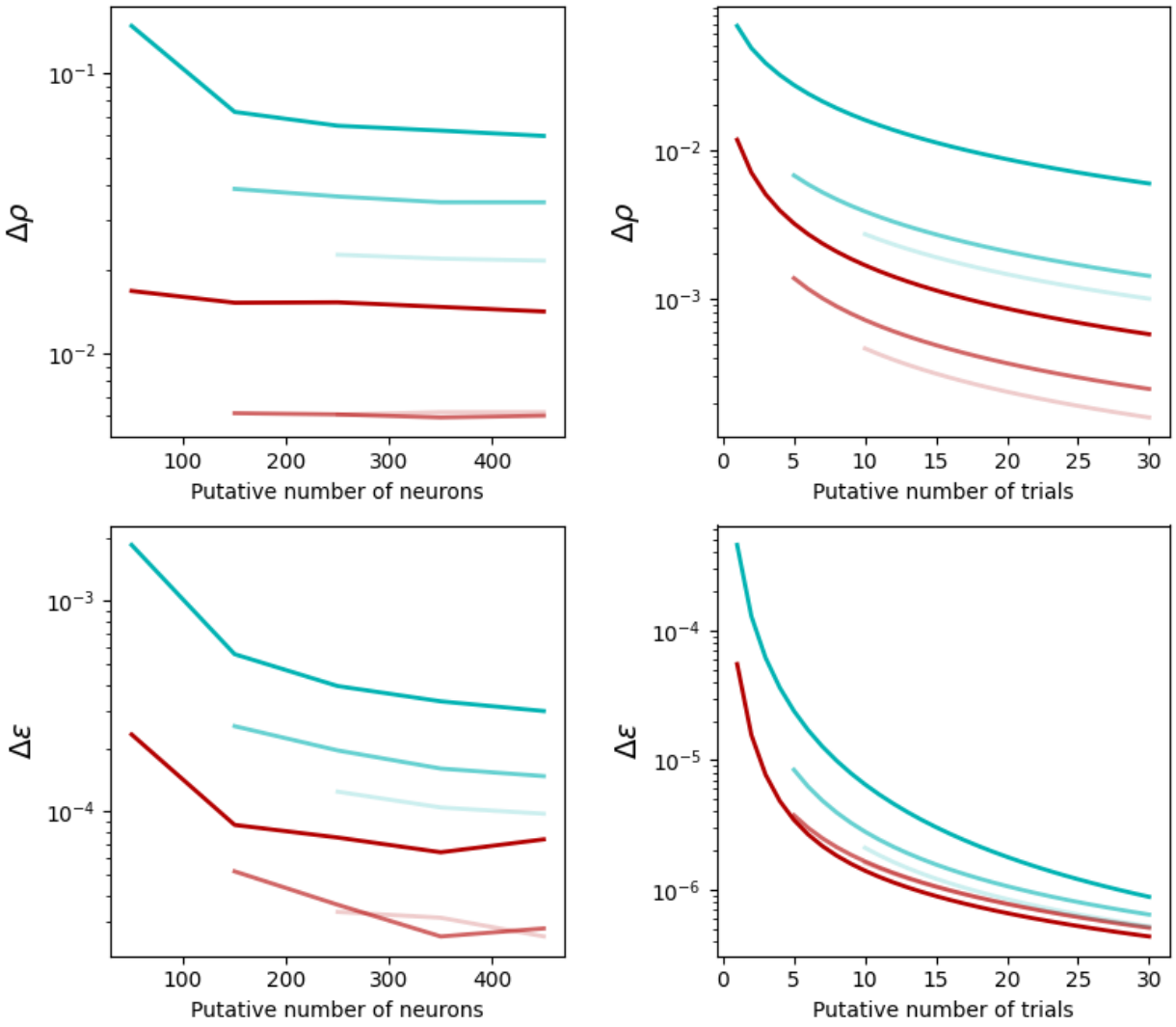

Figure 3.6: Difference in predicted measures ρ and ϵ done on different realizations of the data with two-dimensional latent dynamics. Left column: predictions done for hypothetical datasets with larger number of neurons. Predictions were done on synthetic datasets containing 50, 150, and 200 neurons are shown with dark, medium and light colours respectively. Right column: predictions done for hypothetical datasets with larger number of trials. Predictions were done on synthetic datasets containing 1, 5, and 10 trials. Same color code as in Fig. 3.2.

The errors on the inference of the two accuracy measures ρ and ϵ are shown in Fig. 3.6. For both trial-averaging and trial-concatenation cases, adding more data improves the prediction accuracy. However, we observe that the error saturates with the increase of the number of neurons, while it keeps decaying when the number of trials is increased. As stated before, this is likely connected to the fact that our inference procedure seeks to match the predicted eigenvalues with the empirical eigenvalues. With the increase of the number of trials, we approach the situation of zero noise in the data and perfect direction recovery, and therefore, no fluctuations in the top eigenvalue. However, as, we saw before, the increase of number of neurons does not lead to the perfect direction recovery, and therefore, the error of prediction of ρ and σ will not go to zero, but stabilize on a nonzero value instead.

CHAPTER 4

Experimental design and inference on real data

In the previous chapter, we implemented and validated on synthetic data a procedure for model inference; in turn, the inferred model allowed us to make predictions for extended settings. In this chapter, we show how the same approach can be applied to extracellular electrophysiology datasets, and make possible to carry out specific predictions for larger number of neurons and of trials.

4.1 Subsampling of real data.

To test the quality of inference and prediction procedures, we resort to the technique of subsampling. We take a part of the dataset, infer the model parameters on these training data, and then do a prediction for a hypothetical larger sized dataset, still within the size of the original data available to us. Then, we compare the results with some proxy for the ground-truth values. How do we derive this proxy without access to the true modes $e_i^{(k)}$ and their temporal profiles $x_t^{(k)}$?

Here we describe how to get the empirical value of the accuracy using subsampling technique. Let's say that the total dataset has N neurons and M trials. Then, for an average subset wit $N_1 < N$ neurons and $M_1 < M/2$ trials we can obtain the empirical values of both accuracy parameters in the following way:

- Among the initial N neurons we select a random subset of N_1 neurons. Among M trials, we select two non-overlapping subsets of M_1 trials. This results in two separate small non-overlapping datasets.
- For each of the two new sub-datasets, we carry out PCA analysis. We call $v^{(k),1}$ the inferred principal components for the first dataset, and $v^{(k),2}$ their counterparts for the second dataset. The inferred scores are denoted by, respectively, $y^{(k),1}$ and $y^{(k),2}$.
- By using the definition of the cosine similarity (see 2.7), we can show that the average value of $\frac{1}{N}\sum_i v_i^{(k),1} v_i^{(k),2}$ is $\sum_{k'}(R^{(k,k')})^2$.

$$\left\langle \frac{1}{N}\sum_i v_i^{(k),1} v_i^{(k),2} \right\rangle = \left\langle \frac{1}{N}\sum_i (v_i^{(k),1} - \sum_{k'} R^{(k,k')} e_i^{(k')} + \sum_{k'} R^{(k,k')} e_i^{(k')}) \right.$$
$$\left. (v_i^{(k),2} - \sum_{k'} R^{(k,k')} e_i^{(k')} + \sum_{k'} R^{(k,k')} e_i^{(k')}) \right\rangle =$$
$$= \left\langle \frac{1}{N}\sum_i \sum_{k',k''} (v_i^{(k),1} - R^{(k,k')} e_i^{(k')}) R^{(k,k'')} e_i^{(k'')} \right\rangle + \left\langle \frac{1}{N}\sum_i \sum_{k',k''} R^{(k,k')} e_i^{(k')} (v_i^{(k),2} - R^{(k,k'')} e_i^{(k'')}) \right\rangle +$$
$$+ \sum_{k'} (R^{(k,k')})^2 \frac{1}{N}\sum_i (e_i^{(k')})^2 + \left\langle \frac{1}{N}\sum_i (v_i^{(k),2} - R^{(k,k)} e_i^{(k)})(v_i^{(k),2} - R^{(k,k)} e_i^{(k)}) \right\rangle \quad (4.1)$$

Since the average of the cosine similarity of $v^{(k)}$ and $e_i^{(k')}$ is $R^{(k,k')}$, the first two terms are zero. The last term vanish as well, since $v_i^{(k),1}$ and $v_i^{(k),2}$ are independent. Then,

$$\left\langle \frac{1}{N}\sum_i v_i^{(k),1} v_i^{(k),2} \right\rangle = \sum_{k'} (R^{(k,k')})^2 \frac{1}{N}\sum_i (e_i^{(k')})^2 = \sum_{k'} (R^{(k,k')})^2 \tag{4.2}$$

And knowing the definition of ρ_i, we can find $\left\langle \rho^{(k,k)} \right\rangle_i = 1 - R^{(k,k)}$ (see 2.8).

It can be equally shown that the average value of $\frac{1}{T}\sum_t (y_t^{(k),1} - y_t^{(k),2})^2$ is $2\epsilon^{(k,k)}$:

$$\left\langle \frac{1}{T}\sum_t (y_t^{(k),1} - y_t^{(k),2})^2 \right\rangle = \left\langle \frac{1}{T}\sum_t (y_t^{(k),1} - \sum_{k'} x_t^{(k')} + \sum_{k'} x_t^{(k')} - y_t^{(k),2})^2 \right\rangle =$$

$$= \sum_{k'} \left\langle \frac{1}{T}\sum_t (y_t^{(k),1} - x_t^{(k')})^2 \right\rangle + \left\langle \frac{1}{T}\sum_t (y_t^{(k),2} - x_t^{(k)})^2 \right\rangle - 2\left\langle \frac{1}{T}\sum_t (y_t^{(k),1} - x_t^{(k)})(y_t^{(k),2} - x_t^{(k)}) \right\rangle \tag{4.3}$$

Each of the first two terms equals to $\epsilon^{(k,k)}$, while the last term vanishes, since $y^{(k),1}$ and $y^{(k),2}$ are obtained from independent samples of data. Thus, we can estimate $\epsilon^{(k,k)}$ as

$$\epsilon^{(k,k)} \approx \frac{1}{2T}(y_t^{(k),1} - y_t^{(k),2})^2 \tag{4.4}$$

4.2 Application I: center-out reach task in monkeys

In their 2023 study, Safaie et al. (2023) compared latent neural dynamics during motor planning and execution across different animals of the same species. One of their experiments involved monkeys performing an eight-target instructed-delay center-out reaching task using a manipulandum controlled by a hand (Fig. 4.1, left). The monkeys controlled a cursor on a screen, moving it toward one of eight targets after a variable delay and an auditory cue. While the monkeys performed the task, the neural activity was recorded from the motor cortex using extracellular electrophysiology through 96-channel Utah electrode arrays.

To understand the underlying neural dynamics, the authors used PCA to reduce the complexity of the neural data and estimate the primary patterns of neural activity. They then applied Canonical Correlation Analysis (CCA) to align these neural activity patterns across different monkeys, allowing them to compare the obtained low-dimensional representations.

Despite the natural differences in brain circuitry from one monkey to another, the neural activity patterns were highly similar across monkeys when they performed the same task (Fig. 4.1, right). The authors suggest that there are common neural constraints, or shared mechanisms, that allow similar patterns of neural dynamics to emerge in different individuals. This discovery of preserved latent dynamics across individuals has important implications for both basic research and practical applications. One potential application is in the field of brain-machine interfaces: by aligning the neural activity of one individual with that of another, it may be possible to train a movement decoder on one person and then transfer it to others with minimal retraining. This could significantly reduce the time and effort required to adapt brain-controlled devices to new users.

Change in the accuracy with the dataset size

The quality of alignment of the two neural trajectories obtained from different animals will depend on how well both of these trajectories were inferred. Better alignment may improve the accuracy of movement decoding, and as a result, improve the control of the prosthetic. As we saw from the analysis of synthetic datasets in the previous chapter, there are several ways of increasing the accuracy of the inference of the latent dynamics. One simple way to do so is to record more trials.

Our prediction may be useful in this case. By estimating the errors ρ and ϵ of inference of the neural trajectories and predicting how they will decrease with additional trials, we can determine

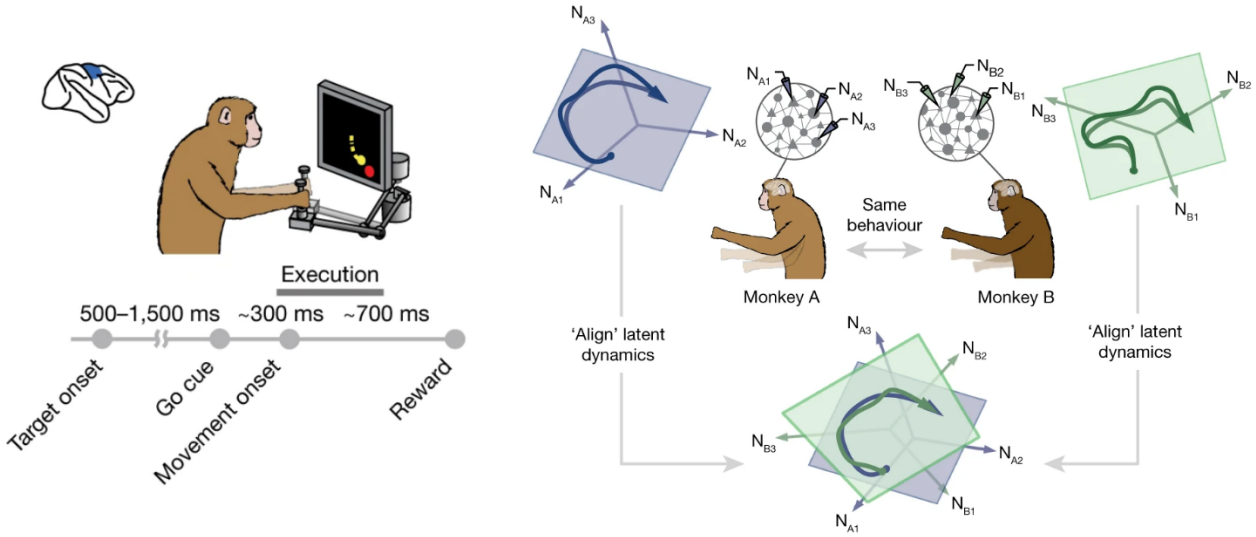

Figure 4.1: Scheme and main results of the work of Safaie et al. (2023). Left: Sheme of the experiment. A monkey controls a cursor to one of eight possible targets using a manipulandum. Right: low-dimensional representations of activity extracted from different animals performing the same task can be aligned. Figures taken from Safaie et al. (2023).

whether collecting more data would significantly improve alignment accuracy — and, if so, how many additional trials would be optimal. In practice, we take a small part of the original dataset (in this case, smaller number of trials that were recorded in total), infer our model, make predictions, and then compare with the empirical values that we obtain via the subsampling procedure described above.

Dataset: Gallego-Carracedo et al. (2022), used in Safaie et al. (2023)

Parameter	Value
Animal	Rhesus macaques
Area of recording	motor cortex, premotor cortex, primary somatosensory cortex
Number of animals	4
Trial types	Eight: one type per possible target
T	depends on the animal; around 100 time bins (trial length around 2s)
N	depends on the animal; between 29 and 86 for M1, 76 for PMd, 23-32 for S1
M	depends on the animal and recording session; up to 1038 for all targets ($\sim$ 130 per target)

To test our prediction, we take the data from one of the recording sessions for one monkey. For our analysis, we took two out of eight possible trial types, consisting of reaching two opposite targets. We take the same number of trials for both targets, perform trial-averaging within the trial of each type, and then concatenate the resulting averaged activity into one matrix with twice as many time bins. Next, we perform the PCA on the resulting data matrix. We predict the accuracy of the the first two principal components, assuming that the dynamics is two-dimensional, reflecting the fact that the monkey performs a movement in the two-dimensional space.

Figure 4.2 shows the difference between the predictions for $\langle \rho \rangle$ done assuming that there is no directional noise $\delta x^{(k)}$ (that is, no trial-to-trial variability of the signal) and the predictions when $\delta x^{(k)}$ is inferred with the procedure described in section 3.1. Example plots of the directional noise $\delta x^{(k)}$, which mainly represents trial-to-trial variability (see section 3.1 for details)[1], can be seen on Figure 4.3. In both cases, we see that the prediction is biased: our theoretical estimate of the error is always higher than the real error. However, the bias is smaller in the case when we assume the presence of directional noise δx and infer its variance.

[1]This may be particularly informative for experiments in which trial-dependent noise may gradually change its statistics, for example in the case of learning.

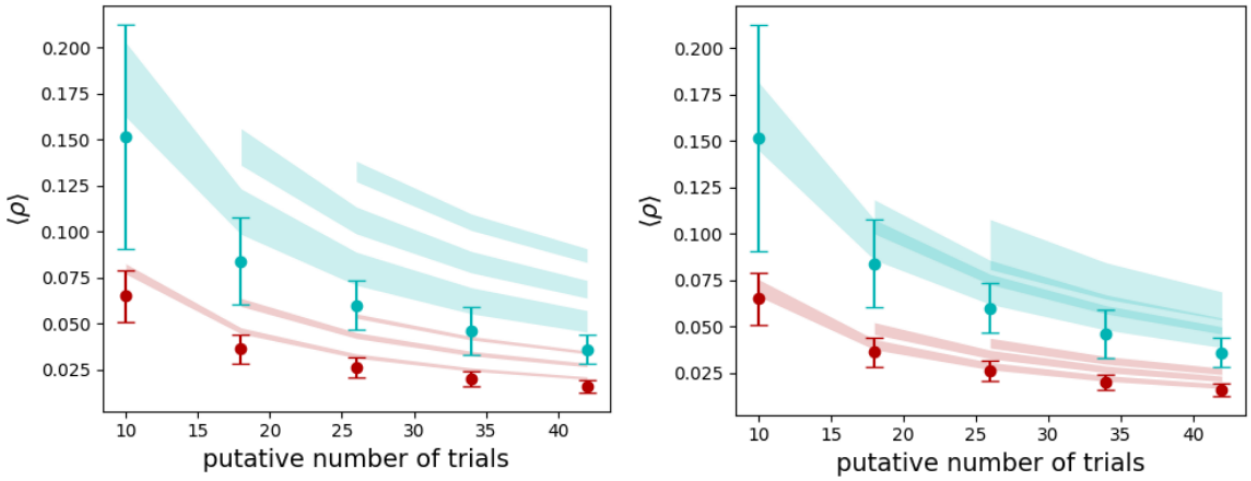

Figure 4.2: Difference between the predictions done with and without the inference of directional noise δx. Left: Prediction of $\langle \rho \rangle$ for datasets with larger number of trials assuming that there is no directional noise δx. Right: same prediction done with the inference of δx. The first principal component is shown in red, the second - in teal. Shaded areas show theoretical prediction, error bars - empirical values. Predictions are done based on 10, 18 and 26 trials.

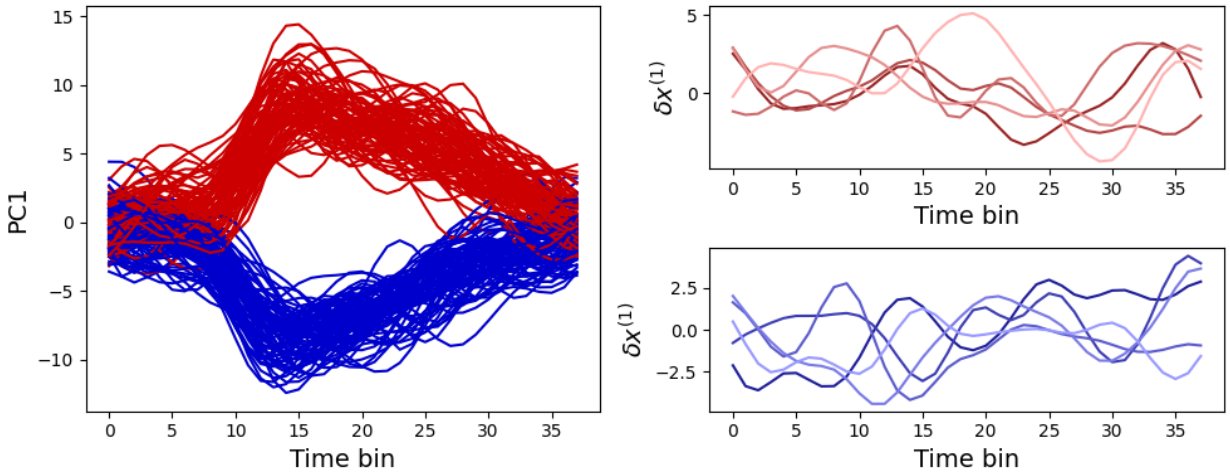

Figure 4.3: Extraction of trial-to-trial variability in principal components. Left: projections of individual trials on the first principal component. The two types of trials with opposite targets are shown in red and blue. Right: Examples of variability δx between trials.

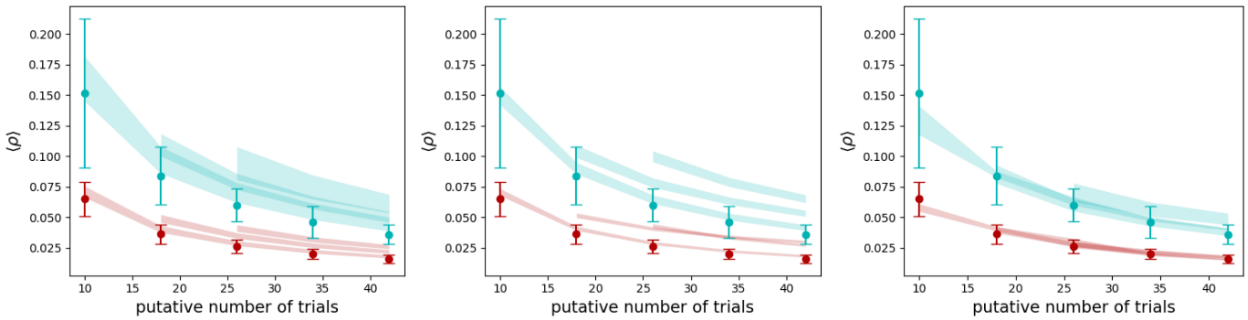

Figure 4.4: Prediction of accuracy of the first two principal components done assuming different dimensionality of the latent dynamics. Left: prediction done assuming the latent dynamics is two-dimensional. Middle: same assuming three-dimensional dynamics. Right: same assuming four-dimensional dynamics. The first principal component is shown in red, the second - in teal. Shaded areas show theoretical prediction, error bars - empirical values. Predictions are done based on 10, 18 and 26 trials.

Figure 4.4 shows the prediction of accuracy done for increasing the number of trials in the dataset for the trial-averaged case, in the same way as we did for the synthetic data in the previous chapter (see Fig. 3.5).

- **Left Plot**: The prediction was performed assuming that the underlying latent dynamics is two-dimensional. The results indicate a biased prediction: the estimated error for direction recovery (ρ) was consistently larger than the error obtained using a subsampling procedure described above. While increasing the number of trials reduces the variability of predictions obtained with different subsamples, the bias remains.

  The reason for this bias lies in the underestimation of the dimensionality of the underlying neural activity. Since the inference procedure estimates the signal-to-noise ratio by scaling the scores $y_t^{(k)}$ obtained from PCA and subsequently inferring the noise variance using the total variance of each individual neuron's activity, an incorrect assumption about the number of latent dimensions leads to a systematic overestimation of the noise variance. This results in a bias that cannot be eliminated by simply adding more trials.

- **Middle and Right Plots**: To address this limitation, predictions were also made assuming higher-dimensional latent dynamics:

  - The middle plot shows the prediction of accuracy of the first two principal components assuming a three-dimensional latent space.
  - The right plot shows the prediction assuming a four-dimensional latent space.

  These results clearly demonstrate that increasing the assumed latent dimensionality reduces the prediction bias for this dataset. The decrease in the bias indicates that higher latent dimensionality allows for better modeling of the complex underlying neural dynamics. The bias does not completely disappear, showing that me may want to continue increasing the number of considered dimensions.

A natural question arises from this analysis: How do we determine the correct dimensionality of the latent dynamics? One possible approach is to make multiple predictions of $\langle\rho\rangle$, gradually increasing the number of principal components considered each time. At some point, we may reach a situation where the latest added principal component does not correspond to any specific latent mode, but instead represents purely random noise. This case would be indicated by a prediction value where the average recovery error $\langle\rho\rangle$ reaches 1, implying that the added component no longer captures any meaningful structure.

4.3 Application II: Tactile delayed response task in mice

In their 2020 study, Wei et al. (2020) examined differences in neural population dynamics recorded via calcium imaging and electrophysiology in actively behaving mice. The mice engaged in a tactile delayed-response task, in which they used their whiskers to identify the position of a pole (either anterior or posterior) and then responded to an auditory cue by licking a designated left or right water spout to obtain a reward (Fig. 4.5, left). Neural activity was tracked in the left anterior lateral motor cortex (ALM), a brain region involved in decision-making processes, using silicon probes for electrophysiology and calcium imaging with three different calcium indicators (GCaMP6s, GCaMP6s-TG, and GCaMP6f) delivered by either viral gene transfer and transgenic expression.

To study neural population activity, the authors analyzed single-cell selectivity and applied PCA to reduce data dimensionality and identify the primary patterns of neural variance in each recording type. They showed that there were significant differences in the results of analysis done on neural dynamics recorded with calcium imaging and electrophysiology, including fewer multiphasic neurons in the case of calcium imaging and delayed, temporally dispersed patterns of activity. The authors proposed that the inherent transformations in calcium imaging, such as lower temporal resolution and nonlinear signal properties, contributed to these differences.

Dataset: Wei et al. (2020) (electrophysology recordings)

Parameter	Value
Animal	Mice
Area of recording	left anterior lateral motor cortex (ALM)
Number of animals	19
Trial types	4: successful trials for left and right licking, error trials for left and right licking
T	77 (trial length around 5s)
N	720
M	Different for different neurons, up to 117 trials for a given neuron

To test our approach on this dataset and make sure we have a large amount of trials to subsample from, among all 720 neurons recorded with extracellular electrophysiology, we select all that were recorded in at least 50 trials with correct lick of the left spout, and at least 50 trials with the correct lick of the right spout. We end up with 263 neurons. For each of these 263 neurons we take 50 successful trials for left lick and 50 successful trials for right lick. Each trial is binned with the bin width of 67 ms and aligned on the movement onset.

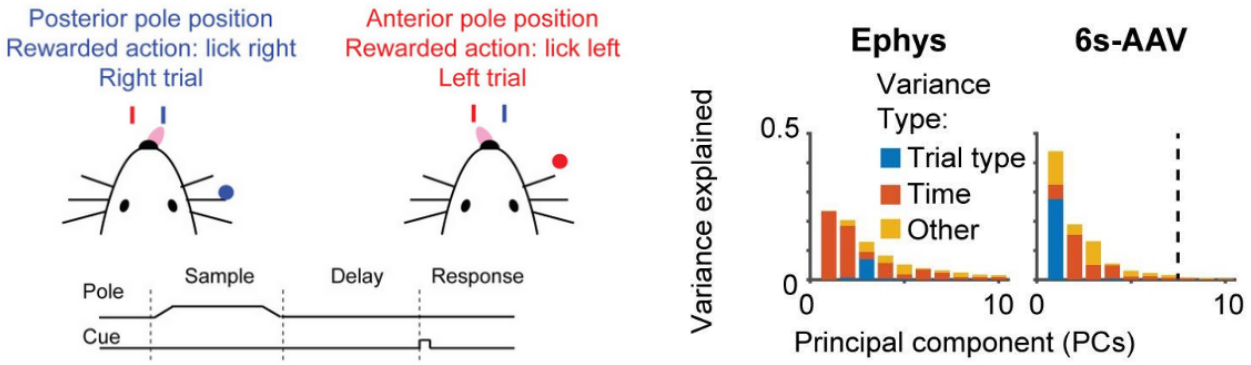

Figure 4.5: Tactile delayed response task studied by Wei et al. (2020). Left: Scheme of the experiment. Mice had to report the position of the pole by choosing either left or right water spout to receive a reward. Right: Types of variance present in first ten principal component for electrophysiology ("Ephys") and calcium imaging done with adeno-associated virus expressing GCaMP6s ("6s-AAV"). Figures taken from Wei et al. (2020).

Tuning of individual neurons

In their analysis, the authors looked at the sources of variance for each principal component:

- "Time" variance: the variance that comes from dynamics in time, regardless of the type of trial (lick-left vs lick-right);
- "Trial type" variance: the variance that comes from the difference in the neural trajectories for the two types of trials.

The authors proposed to calculate the fractions of each type of variance explained by a given component. In our notation, where the score of the (k)-th principal component is denoted by $y_t^{(k)}$ (see 2.6), their definition can be written in the following way:

$$EV^{(k)}_{\text{time}} = \frac{\langle \langle y_t^{(k)} \rangle^2_{\text{trial type}} \rangle_{\text{time}}}{\langle (y_t^{(k)})^2 \rangle_{\text{time,trial type}}}, \qquad EV^{(k)}_{\text{trial type}} = \frac{\langle \langle y_t^{(k)} \rangle^2_{\text{time}} \rangle_{\text{trial type}}}{\langle (y_t^{(k)})^2 \rangle_{\text{time,trial type}}} \tag{4.5}$$

In the case of temporal dynamics variance, the authors first averaged the principal component score across trial types. This gave an average response profile over time, capturing how neural activity evolves over time regardless of the specific trial type. They then calculated the variance of this temporal profile. This value reflected the portion of the component's variance explained by time-dependent changes that were consistent across trials.

For trial-type variance, the authors reversed this process. They first averaged the principal component score over time, yielding a single value for each trial type. This represented the mean response level for each trial type, independent of when it occurs during the trial. They then calculated the variance of these mean values across trial types, which reflected the component's variance explained by differences between trial types.

In both cases, the result was normalized by the total variance explained by the k-th principal component. This normalization allowed each fraction to be interpreted as the proportion of the component's variance explained by either temporal dynamics or trial-type differences.

The authors found notable differences in the sources of variance for electrophysiology and calcium imaging data: in the electrophysiology data, the first two principal components predominantly captured temporal dynamics, reflecting more of the time variance, while, in the calcium imaging data, the first principal component was more strongly influenced by trial type variance. In the case of electrophysiology recordings, the trial type variance was mostly present in the third principal component (Figure 4.5, right).

Are there neurons that simultaneously encode the trial type and the temporal dynamics for the trial? Identifying such neurons would reveal a form of mixed selectivity, where individual neurons respond to multiple types of information simultaneously. We can try answering to this question by searching for neurons that have a high participation coefficient $e_i^{(k)}$ for both the mode that mostly encode temporal variance and the mode that mostly encode trial-type variance. As we see from Figure 4.5 (right), in the case of electrophysiology recordings we could search for neurons that contribute significantly to both the time-related first component and the trial-related third component.

However, simply identifying neurons that appear in both components does not necessarily confirm genuine mixed selectivity. These neurons could exhibit high participation in multiple components due to noise, rather than true encoding of both time and trial-type information.

This is where our parameter ρ, defined in (2.8) can be helpful, as it gives the average squared difference between $e_i^{(k)}$ and the $v_i^{(k)}$ that we infer with PCA. If our prediction estimates that ρ is large, it could indicate that we could reject high participation in a given component by chance. Being able to predict how ρ changes when we increase the dataset size can tell us which dataset is large enough to have enough accuracy on estimating $v_i^{(k)}$.

To find mixed selectivity neurons, we select 10 random trials, perform PCA, and take the neurons that have high participation in both first and third PCs. An example of two such neurons is given in Figure 4.6 (left). We predict the average distance between the $v_i^{(k)}$ and $e_i^{(k)}$ as $\sqrt{\rho^{(k,k)_i}}$, and

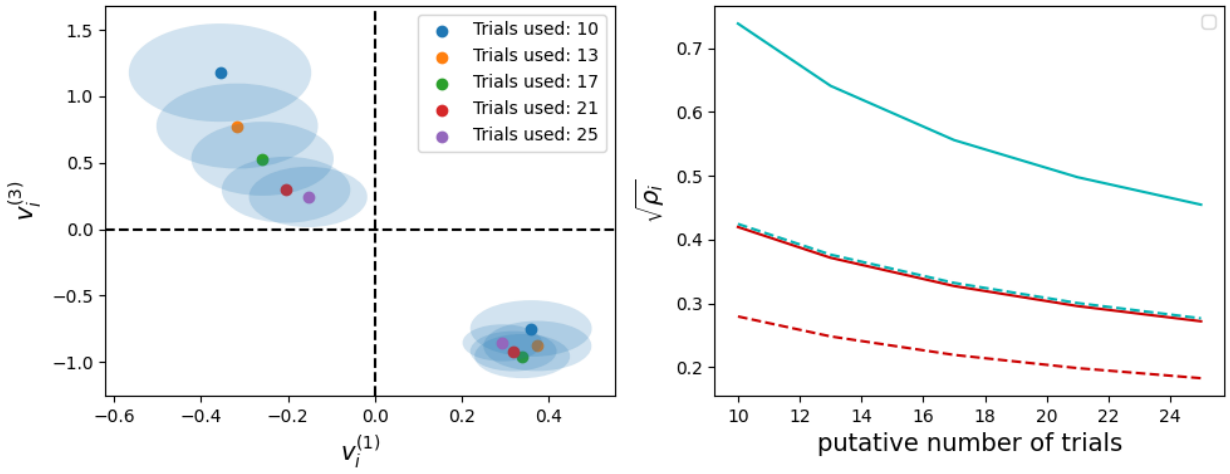

Figure 4.6: Two neurons with apparent mixed selectivity. Left: $v_i^{(k)}$ for two selected neurons for different number of available trials. Shaded area indicates predicted average distance to $e_i^{(k)}$. Right: prediction of $\sqrt{\rho^{(k,k)_i}}$ done on a given 10 trials for the two selected neurons. Prediction for the selective neuron is shown with a dashed line, for non-selective - with a solid line. For both neurons, the first principal component is shown in red, the third - in teal.

show it as a shaded area in blue. To see how $v_i^{(k)}$ will drift when more data is available, we add more trials to the ones we already used, and redo the PCA. As can be observed on the figure, despite the fact that for 10 initially available trials both of the selected neurons seemed to have mixed selectivity, one of them turned out to be non-selective when the number of trials was increased, while the other stayed selective.

Change in the accuracy with the dataset size

From the definitions of variance explained by different sources (eq. 4.5), it is clear that the interpretation of principal component scores $y_t^{(k)}$ — whether they predominantly capture temporal dynamics or trial-type variance — depends heavily on the accuracy with which these scores are inferred. Noise in the data results in the inaccuracies in the score estimation process and can lead to misinterpretations. For instance, apparent temporal dynamics in a component may simply reflect noise, leading us to incorrectly conclude that the main activity encoded by neurons is time-dependent. Similarly, one component might appear to overrepresent temporal dynamics while another underrepresents it. This is particularly relevant in the context of the comparison of calcium imaging and electrophysiology, where the authors observe a discrepancy. The two recording methods seem to attribute trial-type variance to different principal components, which could stem from differences in score inference accuracy between methods.

If there is a way to understand how the error in inferring the score $y_t^{(k)}$ translates into errors in estimating the distribution of variance sources for a given principal component, then knowing the average squared distance ϵ (Eq. 2.11) between the inferred scores $y_t^{(k)}$ and the true dynamics $x_t^{(k)}$ provides a way to judge the reliability of the observed distribution of variance sources. This would clarify whether observed differences in the variance distributions across principal components derived from electrophysiology and calcium imaging recordings reflect true methodological distinctions or are instead artifacts from score inference errors.

Additionally, similarly to what we saw above for the experiment conducted by Safaie et al. (2023), predicting how this error might change with additional data—such as more neurons or trials—can help with experimental planning. This approach would allow researchers to ensure their datasets are large enough to achieve a specific level of precision in estimating the sources of variance.

Figure 4.7 shows the prediction of ϵ done on a small subset of available trials (either 10, 13

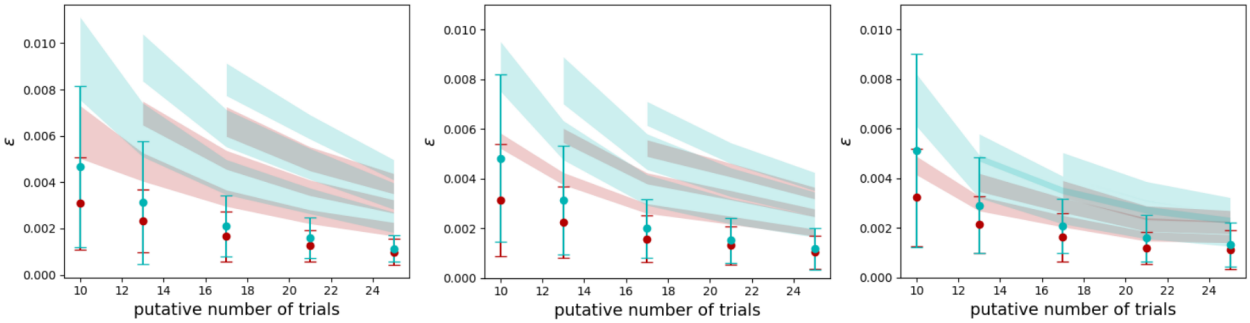

Figure 4.7: Prediction of the accuracy measure ϵ for the first principal component of the electrophysiology recordings done by Wei et al. (2020). Prediction is done based on 10, 13 and 17 trials. Left: inference done assuming the true latent dynamics is two-dimensional. Middle: same, but assuming the dynamics is is three-dimensional. Right: same, but assuming the dynamics is is four-dimensional.

or 17) for the datasets with potentially larger number of trials. As in the case of the prediction done on the dataset from the work of Safaie et al. (2023), we see that the prediction is biased, systematically overestimating the error for both principal components, and that the bias in the prediction that decreases with the increase of the assumed dimensionality of the true underlying latent dynamics.

CHAPTER 5

Discussion

5.1 Summary

Neural systems, despite their high-dimensional structure, often exhibit coherent patterns of activity across neurons that are linked to specific cognitive or motor processes. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), simplify the analysis of these systems by revealing collective neural patterns encoding different brain functions like decision-making and motor control. However, assessing these representations accurately is challenging due to noise, variability, and differences in recording techniques.

We introduced a theoretical framework to evaluate PCA accuracy in neural data, validated on synthetic data and applied to real electrophysiology recordings. We began with a model that reflects the properties of real neural recordings, accounting for noise and trial-to-trial variability. Validation on synthetic data demonstrated that model predictions reliably estimate accuracy across different data conditions, offering a foundation for robust experimental planning. Applying the model to real data further highlighted how predictions can guide efficient data collection strategies, and help determine minimal neuron counts and trial numbers for reliable experimental outcomes.

Accurate Modeling of Neural Data

The first contribution of this work lies in the development of a model that captures essential characteristics of neural data. Standard spiked covariance models do not account for factors such as trial-invariant patterns and neuron-specific noise, making it hard to transfer analytical results of PCA from theory to real neural recordings.

To address these points, we designed a model that incorporates trial-invariant neural modes, trial-specific variability, and unique noise characteristics at the neuron level, providing a more faithful representation of underlying neural dynamics. We also tailored the model to accommodate for the two most popular recording techniques: electrophysiology and calcium imaging.

Each technique presents unique challenges that influence data quality and structure. For electrophysiology, we addressed issues such as wrong spike attribution, neuron overclustering, and recording sparcity. For calcium imaging, we modeled the slower signal dynamics using a double-exponential kernel. Given that neurons exhibit varying fluorescence dynamics, we introduced fluctuations in rise and decay times of this kernel across neurons. By working within a Gaussian framework, we assumed that these fluctuations in fluorescence dynamics, arising from the neuron-specific variations, are approximately normally distributed. This approach allowed us to keep the key features of variability using just the mean and variance of the kernel distribution, significantly simplifying the mathematical treatment and making the calculation analytically tractable.

Statistical Tools for Accuracy Prediction

To analyze and predict the accuracy of PCA within this model, we used tools from random matrix theory and the statistical physics of disordered systems. The replica method is particularly well-suited for handling systems with quenched disorder, such as neural data. By using this approach, we found the expected values of two distinct accuracy measures.

Accuracy Measures: ρ_i and ϵ

The two accuracy measures, ρ_i and ϵ, offer complementary perspectives on PCA accuracy. Neuron-specific alignment (ρ_i) quantifies how well PCA captures the latent modes for each individual neuron. This measure is valuable for understanding how each neuron contributes to neural assemblies and for identifying neurons with mixed selectivity, which may participate in multiple functional groups.

In contrast, trajectory reconstruction error (ϵ) represents the mean squared distance between the true and inferred low-dimensional neural trajectories, providing a global measure of accuracy of PCA in capturing the overall shape of neural dynamics across the entire population.

Together, ρ_i and ϵ provide different ways to quantify the accuracy of a given low-dimensional representation obtained with PCA, supporting analyses of both specific neuron contributions and the structure of low-dimensional neural dynamics.

Validation on Synthetic Data

The next phase involved validating the outcome of our calculation on synthetic data generated with the model above. This procedure allowed us to verify our predictions across different values of dataset parameters, such as neuron count, trial number, and signal-to-noise ratio, simulating a range of neural recording conditions. This validation demonstrated that the model consistently predicts both accuracy measures, ρ_i and ϵ.

Furthermore, we showed it was possible to infer the parameters defining the model from data. Once these are inferred, we used the model to make predictions for hypothetical larger datasets, illustrating how data size and quality influence dimensionality reduction outcomes. Such predictive insights are valuable for planning data collection in neuroscience, guiding decisions on the necessary scale of trials or neurons for achieving accurate PCA results.

Application to Real Neural Recordings

In the last chapter, we applied our procedure to real neural recordings, specifically extracellular electrophysiology data from studies focused on motor control and decision-making. This practical application illustrated the model's utility in predicting how additional data (e.g., more neurons or trials) affects PCA accuracy and can affect the analysis and conclusion that were drawn from the low-dimensional representation obtained with PCA.

By predicting the impact of dataset size on the accuracy of PCA, this model allows to tailor data collection strategies to their specific experimental needs, potentially reducing the need for excessive data collection without sacrificing result quality. Additionally, the model's adaptability for various recording techniques, including calcium imaging, suggests broad applicability across neuroscience studies.

5.2 Perspectives

Testing Accuracy Predictions on Synthetic and Real Calcium Imaging Data

Although the calcium imaging model has been benchmarked, we have not yet tested the expansion of the accuracy predictions to larger dataset sizes. This testing should be conducted on both synthetic and real calcium imaging data. For synthetic data, the approach would mirror the process used for electrophysiology: by systematically varying parameters such as neuron count, trial number, and signal-to-noise ratio, we can evaluate the model's ability to predict PCA outcomes when scaled to hypothetical larger datasets.

The next step involves validating these predictions on real calcium imaging datasets. Applying the model to real data will allow us to confirm that its predictions align with observed PCA performance, thereby providing empirical validation. This phase would involve comparing predicted versus observed accuracy (via subsampling) across a variety of calcium imaging datasets, ensuring that the model remains reliable and effective in real-world scenarios.

Investigating Directional Noise ($\delta x^{(k)}$) Variability Across Trials

Our model enables the extraction of a directional noise component, δx, which provides an way to examine trial-specific variations in neural dynamics. Other techniques, such as tensor PCA, also address trial-to-trial variability by capturing patterns across multiple modes (e.g., neurons, time points, and trials) simultaneously. While tensor PCA is effective for reducing dimensionality in complex neural datasets, it generally lacks the capability to isolate noise at the level of individual trials. Instead, it provides an overall compressed representation, blending trial-specific noise into a collective pattern across all data modes.

In contrast, our approach with δx allows for examining noise specific to each individual trial, offering insights into trial-to-trial fluctuations that tensor PCA cannot isolate. This capability is particularly valuable for observing changes in neural dynamics over time, such as those that may occur during learning or adaptation. By isolating directional noise on a trial-by-trial basis, our approach provides a more detailed perspective on how neural activity evolves across trials, complementing the broader analysis offered by tensor PCA.

Additionally, interpreting tensor PCA can be challenging due to the non-uniqueness of its decomposition. This non-uniqueness arises from the complexity of fitting multiple interacting modes (e.g., neurons, time points, and trials), where each mode contributes overlapping patterns of variability. Non-uniqueness in the decomposition can make it difficult to disentangle trial-based fluctuations from variations due to other modes, such as neuron-specific or time-specific patterns.

In contrast, our approach relies on the unique decomposition provided by PCA when applied to matrices that are either trial-averaged or trial-concatenated. This setup allows us to examine trial-specific noise free from confounding effects across modes.

In the near future, we plan to apply our model to a wider range of experimental datasets to better understand the ways in which δx encodes trial variability. By examining different contexts, we hope to demonstrate that δx may contain information about neural response dynamics, potentially encoding aspects of the behavioral context, such as current task demands, environmental cues, or internal states like attention or motivation. Additionally, δx may reflect learning-related changes, adapting to varying conditions or internal states across trials. An example of such learning-induced changes is the phenomenon of representational drift in the hippocampus studied by Khatib et al. (2023) in mice, specifically focusing on whether it is driven more by the passage of time or by active experience. Among the main results the authors show that over time, the number of active place cells (cells associated with specific spatial locations) decreased, but each cell gradually stored more spatial information. Representational drift occurred gradually across sessions, as shown in both single-cell and population vector correlations. This drift appeared across the entire environment rather than in specific areas of the track, indicating a context-wide effect.

Studying the Sensitivity of Analysis to Spike Sorting

Another future direction involves examining the necessity of spike sorting when applying PCA to electrophysiological neural data. Spike sorting, a preprocessing step that identifies and isolates spikes from individual neurons, may or may not be essential for interpreting low-dimensional representations derived through PCA. By applying the model to real data both with and without spike sorting, we can assess whether this step significantly affects the conclusions drawn from PCA analysis. If spike sorting proves unnecessary for certain analyses, this finding could simplify preprocessing workflows, making data processing more efficient.

Studying Mixed Selectivity

For years, neuroscience research focused primarily on understanding how neurons respond to specific stimuli, usually in a linear way—where a neuron might activate when it detects a particular feature, like color or movement. However, complex behaviors and decision-making capabilities suggest a presence of more intricate mechanism, known as non-linear mixed selectivity. It involves neurons that respond not to single features but to combinations of features in complex, non-linear ways.

While historically nonlinear mixed selectivity was underexplored, recent studies are beginning to highlight its significance in cognitive functions. For instance, Rigotti et al. (2013) demonstrated that neurons in the prefrontal cortex exhibit diverse nonlinear mixed selectivity and encode multiple aspects of a task, creating a high-dimensional representation of information. This allows the brain to decode task aspects even when single neurons don't exhibit direct selectivity to those aspects. The authors also show that the dimensionality of neural responses is predictive of performance. During correct trials, the high-dimensional encoding remains intact, while errors correlate with a decrease in dimensionality.

Similarly, Ledergerber et al. (2021) found that subicular neurons exhibit strong mixed selectivity by integrating multiple navigational variables, such as position, head direction, and speed. This integration was more prominent during goal-directed navigation tasks compared to random foraging, suggesting that mixed selectivity in the subiculum is dynamically modulated to meet task demands.

Our tool can help with estimating the statistical significance of the analysis done in such studies. Analyzing the weights associated with each neuron can reveal important information about mixed selectivity, as they indicate how individual neurons contribute to the neural assemblies represented in each principal component. Using the model to analyze such data would allow us to evaluate how well PCA captures these overlapping representations and identify neurons that contribute significantly to mixed selectivity.

Developing a Tool for Experimental Planning

To support experimental neuroscientists in planning data collection, a useful future direction would be to create a user-friendly tool or software package based on this model. This tool would provide a practical interface where users can input parameters—such as neuron count, trial number, and anticipated noise levels—and receive guidance on experimental design to achieve desired PCA accuracy levels.

Appendix

Theoretical prediction for the top eigenvalues

since $v^{(k)}$ is k-th top eigenvector, its eigenvalue $\lambda^{(k)}$ can be written as

$$\frac{1}{N}\sum_{i,j=1}^{N} v_i^{(k)} C_{ij} v_j^{(k)} \tag{A.1}$$

Then, to find its theoretically predicted value, we can use the partition function introduced in (2.48), but with added source term:

$$\begin{aligned} Z(\{s_{i,t}\}) &= \int \prod_{k=1}^{K} d\mathbf{v}^{(k)} \int \prod_{k=1}^{K}\prod_{t=1}^{T} dy_t^{(k)} \prod_{k,t} \delta\left(\sum_i (v_i^{(k)})^2 - N\right) \prod_{k_1,k_1:k_1<k_2} \delta\left(\sum_i v_i^{(k_1)} v_i^{(k_2)}\right) \\ &\quad \prod_{k,t} \delta\left(y_t^{(k)} - \frac{1}{N}\sum_i v_i^{(k)} s_{i,t}\right) \times \exp\left(\beta T \frac{1}{N^2} \sum_{k=1}^{K}\sum_{i,j=1}^{N} v_i^{(k)} C_{ij} v_j^{(k)}\right) \\ &\times \exp\left(\beta T \frac{1}{N^2} \sum_{k=1}^{K} \theta^{(k)} \sum_{i,j=1}^{N} v_i^{(k)} C_{ij} v_j^{(k)}\right) \end{aligned} \tag{A.2}$$

Then, after introducing R and using (2.47), we get

$$\begin{aligned} Z(\{s_{i,t}\}) &= \int \prod_{k=1}^{K} d\mathbf{v}^{(k)} \int \prod_{k=1}^{K}\prod_{t=1}^{T} dy_t^{(k)} \prod_{k,t} \delta\left(\sum_i (v_i^{(k)})^2 - N\right) \prod_{k_1,k_1:k_1<k_2} \delta\left(\sum_i v_i^{(k_1)} v_i^{(k_2)}\right) \\ &\quad \prod_{k,t} \delta\left(y_t^{(k)} - \frac{1}{N}\sum_i v_i^{(k)} s_{i,t}\right) \int \prod_{k_1,k_2=1}^{K} dR^{(k_1,k_2)} \delta\left(R^{(k_1,k_2)} - \frac{1}{N}\sum_i v_i^{(k_1)} \bar{e}_i^{(k_2)}\right) \\ &\times \exp\left(\beta \sum_{k=1}^{K} (1+\theta^{(k)}) \left(\sum_t (y_t^{(k)})^2\right) - \beta \frac{1}{T} \sum_{k=1}^{K} (1+\theta^{(k)}) \left(\sum_t y_t^{(k)}\right)^2\right) \end{aligned} \tag{A.3}$$

Following the outline of the calculation(1.5.5), we introduce n different replicas:

$$\begin{aligned} Z^n(\{s_{i,t}\}) &= \int \prod_{a=1}^{n}\prod_{k=1}^{K} d\mathbf{v}_a^{(k)} \int \prod_{a=1}^{n}\prod_{k=1}^{K}\prod_{t=1}^{T} \left[dy_{a,t}^{(k)} \delta\left(\sum_i (v_{a,i}^{(k)})^2 - N\right) \delta\left(y_{a,t}^{(k)} - \frac{1}{N}\sum_i v_{a,i}^{(k)} s_{i,t}\right)\right] \\ &\times \prod_{a,k_1<k_2} \delta\left(\sum_i v_{a,i}^{(k_1)} v_{a,i}^{(k_2)}\right) \int \prod_{a,k_1,k_2} dR_a^{(k_1,k_2)} \delta\left(R_a^{(k_1,k_2)} - \frac{1}{N}\sum_i v_{a,i}^{(k_1)} \bar{e}_i^{(k_2)}\right) \times \\ &\times \exp\left(\beta \sum_{a=1}^{n}\sum_{k=1}^{K} (1+\theta^{(k)}) \left(\sum_t (y_{a,t}^{(k)})^2\right) - \beta \frac{1}{T} \sum_{a=1}^{n}\sum_{k=1}^{K} (1+\theta^{(k)}) \left(\sum_t y_{a,t}^{(k)}\right)^2\right) \end{aligned} \tag{A.4}$$

Next step: expressing all of the delta-functions as integrals.

$$
\begin{aligned}
Z^n\big(\{s_{i,t}\}\big) &= \int \prod_{a,k} d\mathbf{v}_a^{(k)} \prod_{a,k,t} dy_{a,t}^{(k)} \frac{d\hat{y}_{a,t}^{(k)}}{2\pi} \int \prod_{a,k_1,k_2} dR_a^{(k_1,k_2)} \frac{d\hat{R}_a^{(k_1,k_2)}}{2\pi N} \frac{d\hat{u}_a^{(k_1,k_2)}}{2\pi} \\
&\times \exp\left(i \sum_{a,k_1<k_2} \hat{u}_a^{(k_1,k_2)} \left(\sum_i v_i^{(k_1)} v_i^{(k_2)} \right) \right) \exp\left(i \sum_{a,k} \hat{u}_a^{(k,k)} \left(\sum_i (v_{a,i}^{(k)})^2 - N \right) \right) \\
&\times \exp\left(i \sum_{a,k,t} \hat{y}_{a,t}^{(k)} \left(y_{a,t}^{(k)} - \frac{1}{N} \sum_i v_{a,i}^{(k)} s_{i,t} \right) \right) \\
&\times \exp\left(iN \sum_{a,k_1,k_2} R_a^{(k_1,k_2)} \left(R_a^{(k_1,k_2)} - \frac{1}{N} \sum_i v_{a,i}^{(k_1)} \bar{e}_i^{(k_2)} \right) \right) \times \\
&\times \exp\left(\beta \sum_{a=1}^{n} \sum_{k=1}^{K} (1+\theta^{(k)}) \left(\sum_t (y_{a,t}^{(k)})^2 \right) - \beta \frac{1}{T} \sum_{a=1}^{n} \sum_{k=1}^{K} (1+\theta^{(k)}) \left(\sum_t y_{a,t}^{(k)} \right)^2 \right) \qquad \text{(A.5)}
\end{aligned}
$$

Now, we can express $s_{i,t}$ using (2.46), and average over all sources of disorder in the system, starting from $\bar{z}_{i,t}$:

$$
\langle Z^n(s_{i,t}) \rangle_{\bar{z}} = \int \frac{1}{(\sqrt{\det \bar{Z}})^N} \prod_i \frac{1}{\sqrt{\sigma_i^2 N^T}} \prod_t \frac{d\bar{z}_{i,t}}{\sqrt{2\pi}} \exp\left(- \sum_i \frac{\sum_{t_1,t_2} \bar{z}_{i,t_1} (\bar{Z}^{-1})_{t_1,t_2} \bar{z}_{i,t_2}}{2\bar{\sigma}_i^2 N} \right) Z^n(s_{i,t}) \qquad \text{(A.6)}
$$

Then, average over $\delta\bar{F}_{i,t}$:

$$
\langle Z^n(s_{i,t}) \rangle_{\bar{z},\bar{F}} = \int \frac{1}{\sqrt{\det \bar{\Xi}^N}} \prod_{i,t} \frac{d\bar{F}_{i,t}}{\sqrt{2\pi}} \exp\left(- \sum_i \frac{\sum_{t_1,t_2} \delta\bar{F}_{i,t_1} (\bar{\Xi}^{-1})_{t_1,t_2} \delta\bar{F}_{i,t_2}}{2N} \right) \langle Z^n(s_{i,t}) \rangle_{\bar{z}} \qquad \text{(A.7)}
$$

and finally, averaging over $\delta\bar{x}_t^{(k)}$:

$$
\langle Z^n(s_{i,t}) \rangle = \int \frac{1}{\sqrt{\det \bar{\Delta}^K}} \prod_k \frac{1}{(\bar{\xi}^{(k)})^T} \prod_t \frac{d\delta\bar{x}_{i,t}}{\sqrt{2\pi}} \exp\left(- \sum_k \frac{\sum_{t_1,t_2} \delta\bar{x}_{i,t_1} (\bar{\Delta}^{-1})_{t_1,t_2} \delta\bar{x}_{i,t_2}}{2(\bar{\xi}^{(k)})^2} \right) \langle Z^n(s_{i,t}) \rangle_{\bar{z},\bar{F}} \qquad \text{(A.8)}
$$

After performing these Gaussian integrals, we will get

$$
\begin{aligned}
\langle Z^n(s_{i,t})\rangle \;=\; & \int \prod_{a,k}\left[d\mathbf{v}_a^{(k)}\prod_{t=1}^{T}dy_{a,t}^{(k)}\frac{d\hat{y}_{a,t}^{(k)}}{2\pi}\right]\int\prod_{a,k_1,k_2}\left[dR_a^{(k_1,k_2)}\frac{d\hat{R}_a^{(k_1,k_2)}}{2\pi N}\frac{d\hat{u}_a^{(k_1,k_2)}}{2\pi}\right]\\
\times\;& \exp\left(-\sum_i\frac{\bar{\sigma}_i^2\sum\limits_{a,b,k_1,k_2,t_1,t_2}v_{a,i}^{(k_1)}v_{b,i}^{(k_2)}\hat{y}_{a,t_1}^{(k_1)}\bar{Z}_{t_1,t_2}\hat{y}_{b,t_2}^{(k_2)}}{2N}\right)\\
\times\;& \exp\left(i\sum_{a,k_1<k_2}\hat{u}_a^{(k_1,k_2)}\left(\sum_i v_{a,i}^{(k_1)}v_{a,i}^{(k_2)}\right)+i\sum_{a,k}\hat{u}_a^{(k,k)}\left(\sum_i(v_{a,i}^{(k)})^2-N\right)\right)\\
\times\;& \exp\left(-N\sum_{i,\tau,\tau'}\frac{\left(\sum\limits_{k_1,k_1'}\frac{1}{N}\bar{e}_i^{(k_1)}\sum\limits_a v_{a,i}^{(k_1')}\sum\limits_t\hat{y}_{a,t}^{(k_1')}x_{t-\tau}^{(k_1)}\right)\bar{\Xi}_{\tau,\tau'}\left(\sum\limits_{k_2,k_2'}\frac{1}{N}\bar{e}_i^{(k_2)}\sum\limits_b v_{b,i}^{(k_2')}\sum\limits_{t'}\hat{y}_{b,t'}^{(k_2')}x_{t'-\tau'}^{(k_2)}\right)}{2}\right)\\
\times\;& \exp\left(-\sum_k\frac{(\bar{\xi}^{(k)})^2\sum\limits_{t,t',k_1,k_2}\sum_{a,b}\hat{y}_{a,t}^{(k_1)}R_a^{(k_1,k)}\Delta_{t,t'}\hat{y}_{b,t'}^{(k_2)}R_b^{(k_2,k)}}{2}\right)\\
\times\;& \exp\left(i\sum_{a,k_1,t}\hat{y}_{a,t}^{(k_1)}\left(y_{a,t}^{(k_1)}-\sum_{k_2=1}^{K}R_a^{(k_1,k_2)}\bar{x}_t^{(k_2)}\right)+iN\sum_{a,k_1,k_2}\hat{R}_a^{(k_1,k_2)}\left(R_a^{(k_1,k_2)}-\frac{1}{N}\sum_i v_{a,i}^{(k_1)}\bar{e}_i^{(k_2)}\right)\right)\\
\times\;& \exp\left(\beta\sum_{a,k}(1+\theta^{(k)})\left(\sum_t(y_{a,t}^{(k)})^2\right)-\beta\frac{1}{T}\sum_{a,k}(1+\theta^{(k)})\left(\sum_t y_{a,t}^{(k)}\right)^2\right)
\end{aligned}
\tag{A.9}
$$

We simplify the notation by introducing

$$
\mathcal{X}_{t,t'}^{(k_1,k_2)}=\sum_{\tau,\tau'}x_{t-\tau}^{(k_1)}\bar{\Xi}_{\tau,\tau'}x_{t'-\tau'}^{(k_2)} \tag{A.10}
$$

Following the outline of the calculation (1.5.5), we introduce the overlaps in the same way as we did in the calculation in the main text:

$$
\begin{cases}
q_{ab}^{(k_1,k_2,k_1',k_2')}=\frac{1}{N}\sum\limits_i\bar{e}_i^{(k_1)}\bar{e}_i^{(k_2)}v_{a,i}^{(k_1')}v_{b,i}^{(k_2')}, & k_1\le k_2,\ a\le b\\
m_{ab}^{(k,k')}=\frac{1}{N}\bar{\sigma}_i^2v_{a,i}^{(k)}v_{b,i}^{(k')}, & a\le b
\end{cases}
\tag{A.11}
$$

by adding an additional integration over delta-function:

$$
\langle Z^n(s_{i,t})\rangle=\int\prod_{a\le b,k_1\le k_2,k_1',k_2'}dq_{ab}^{(k_1,k_2,k_1',k_2')}\delta\left(q_{ab}^{(k_1,k_2,k_1',k_2')}-\frac{1}{N}\sum_i\bar{e}_i^{(k_1)}\bar{e}_i^{(k_2)}v_{a,i}^{(k_1')}v_{b,i}^{(k_2')}\right)\langle Z^n(s_{i,t})\rangle \tag{A.12}
$$

(and likewise for $m_{ab}^{(k,k')}$).

We can replace the delta-function above by integration of a complex exponent, and introduce $\hat{q}_{ab}^{(d_1,d_2,d_1',d_2')}$ and $\hat{m}_{ab}^{(k,k')}$. Following the outline of the calculation (1.5.5) we can apply the replica-symmetric ansatz as we did in (2.60).

We can consider that $y_{a,t}^{(k)}$ - elements of one big vector y in a tensor product of spaces of principal components (K-dimensional space), of time bins (T-dimensional space), and of replicas (n-dimensional space). Similarly, we can consider $\hat{y}_{a,t}^{(k)}$ as elements of vector $\hat{y}$. We can see $m^{(k,k')}$ as

elements of a matrix M acting in the space of principal components, $\hat{m}^{(k,k')}$ - as elements of $\hat{M}$, $l^{(k,k')}$ - as elements of L, $\hat{l}^{(k,k')}$ - as elements of $\hat{L}$. We also introduce the matrix

$$\hat{U}_{k_1,k_2} = \hat{u}^{(k_1,k_2)}(1+\delta_{k_1,k_2}) \tag{A.13}$$

We will write the identity matrices as Id_K, Id_T and Id_n for component, time bin and replica spaces respectively. Similarly, we use J_K, J_T and J_n to denote matrices of ones. For vectors, we will denote 1_K,1_T,1_n as vectors consisting of ones. We use $\text{diag}_K(...)$ to denote a matrix that is diagonal in space of components, and has elements of the argument on its diagonal.

We use a circle to denote contraction on all indices, for example

$$\hat{r}\circ r = \sum_{k_1,k_1',k_2,k_2'} \hat{r}^{(k_1,k_2,k_1',k_2')} r^{(k_1,k_2,k_1',k_2')} \tag{A.14}$$

and dot for contracting first two indices:

$$(q\cdot\mathcal{X})^{(k_1',k_2')}_{t,t'} = \sum_{k_1,k_2} q^{(k_1,k_2,k_1',k_2')}\mathcal{X}^{(k_1,k_2)}_{t,t'} \tag{A.15}$$

Then, the entire expression can be rewritten as

$$\begin{aligned}
\langle Z^n(s_{i,t})\rangle \;=\; & \int dv \frac{dy\hat{y}}{(2\pi)^{nKT}} \frac{dRd\hat{R}}{(2\pi N)^{K^2}} \frac{dLd\hat{L}}{(2\pi)^{K^2}} \frac{d\hat{U}}{(2\pi)^{K^2}} \frac{dMd\hat{M}}{(2\pi)^{K^2}} \\
& \int \prod_{k_1,k_1',k_2,k_2'} \frac{dq^{(k_1,k_2,k_1',k_2')} d\hat{q}^{(k_1,k_2,k_1',k_2')}}{2\pi} \frac{dr^{(k_1,k_2,k_1',k_2')} d\hat{r}^{(k_1,k_2,k_1',k_2')}}{2\pi} \\
\times\; & \exp\left(iNn\text{Tr}(\hat{L}L^T) - iv^T(\hat{L}\otimes Id_n\otimes \text{diag}_N(\bar{\sigma}^2))v + \frac{iNn(n-1)}{2} Tr(\hat{M}M^T)\right) \\
\times\; & \exp\left(\frac{i}{2} v^T(\hat{M}\otimes(J_n - Id_n)\otimes \text{diag}_N(\bar{\sigma}^2))v - \frac{\hat{y}^T(L\otimes Z\otimes Id_n)\hat{y}}{2} - \frac{\hat{y}^T(M\otimes Z\otimes(J_n-Id_n))\hat{y}}{2}\right) \\
\times\; & \exp\left(iNn(\hat{r}\circ r) - \sum_{k_1,k_1',k_2,k_2',a} i\hat{r}^{(k_1,k_2,k_1',k_2')} \sum_i \bar{e}_i^{(k_1)}\bar{e}_i^{(k_2)} v_{a,i}^{(k_1')} v_{a,i}^{(k_2')}\right) \\
\times\; & \exp\left(iN\frac{n(n-1)}{2}(\hat{q}\circ q) - \sum_{k_1,k_1',k_2,k_2',a<b} i\hat{q}^{(k_1,k_2,k_1',k_2')}\left(\sum_i \bar{e}_i^{(k_1)}\bar{e}_i^{(k_2)} v_{a,i}^{(k_1')} v_{b,i}^{(k_2')}\right)\right) \\
\times\; & \exp\left(-\frac{\hat{y}^T((q\cdot\mathcal{X})\otimes(J_n-Id_n))\hat{y}}{2} - \frac{\hat{y}^T((r\cdot\mathcal{X})\otimes Id_n)\hat{y}}{2}\right) \\
\times\; & \exp\left(-\frac{iNn}{2}\text{Tr}\hat{U} + \frac{i}{2} v^T(\hat{U}\otimes Id_n\otimes Id_N)v - \frac{\hat{y}^T(R\text{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\otimes J_n)\hat{y}}{2}\right) \\
\times\; & \exp\left(i\hat{y}^T y - i((R\otimes Id_T\otimes Id_n)(\tilde{x}\otimes 1_n))^T\hat{y} + iN\text{Tr}\hat{R}^T R - iv^T(\hat{R}\otimes Id_n\otimes Id_N)(e\otimes\mathbb{1}_n)\right)\times \\
\times\; & \exp\left(\beta y^T\left(\text{diag}(1+\theta^{(k)})\otimes\left(Id_T - \frac{1}{T}J_T\right)\otimes Id_n\right)y\right)
\end{aligned} \tag{A.16}$$

Now, we can continue integrating over the variables. First, integrating with respect to $\hat{y}$. We can see that this is just a simple Gaussian integral with the quadratic term

$$\mathcal{A} = ((L-M)\otimes Z + (r-q)\cdot\mathcal{X})\otimes Id_n + \left(M\otimes Z + q\cdot\mathcal{X} + R\text{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\right)\otimes J_n \tag{A.17}$$

and a linear term $i(y-(R\otimes Id_n\otimes Id_T)(\bar{x}\otimes\mathbb{1}_n))$.

After this, we can integrate with respect to y. The integral will also be Gaussian, with quadratic part

$$\mathcal{B} = \mathcal{A}^{-1} - 2\beta \text{diag}(1+\theta^{(k)}) \otimes Id_n \otimes Id_T + 2\beta \text{diag}(1+\theta^{(k)}) \otimes Id_n \otimes \frac{1}{T} J_T \tag{A.18}$$

and a linear part $(\mathcal{A}^{-1}(R \otimes Id_{n,T}))(\bar{x} \otimes \mathbb{1}_n)$
Then, after integration

$$\begin{aligned}
\langle Z^n \rangle &= \int \frac{d\hat{U}dv}{(2\pi)^{K^2}} \frac{d\hat{R}dR}{(2\pi N)^{K^2}} \frac{dMd\hat{M}}{(2\pi)^{K^2}} \frac{dLd\hat{L}}{(2\pi)^{K^2}} \frac{1}{\sqrt{\det \mathcal{A} \det \mathcal{B}}} \exp\left(iNnTr(\hat{L}L^T) + \frac{iNn(n-1)}{2} Tr(\hat{M}M^T) \right) \\
&\int \prod_{k_1,k_1',k_2,k_2'} \frac{dq^{(k_1,k_2,k_1',k_2')} d\hat{q}^{(k_1,k_2,k_1',k_2')}}{2\pi} \frac{dr^{(k_1,k_2,k_1',k_2')} d\hat{r}^{(k_1,k_2,k_1',k_2')}}{2\pi} \\
&\times \exp\left(\frac{(\tilde{x}^T \otimes \mathbb{1}_n^T)(\mathcal{A}^{-1}(R \otimes Id_{n,T}))^T \mathcal{B}^{-1} (\mathcal{A}^{-1}(R \otimes Id_{n,T}))(\bar{x} \otimes \mathbb{1}_n)}{2} \right) \\
&\times \exp\left(-\frac{i}{2} v^T (\hat{M} \otimes (J_n - Id_n) \otimes \text{diag}_N(\bar{\sigma}^2)) v - i v^T (\hat{L} \otimes Id_n \otimes \text{diag}_N(\bar{\sigma}^2)) v \right) \\
&\times \exp\left(iNn(\hat{r} \circ r) - \sum_{k_1,k_1',k_2,k_2',a} i\hat{r}^{(k_1,k_2,k_1',k_2')} \sum_i \bar{e}_i^{(k_1)} \bar{e}_i^{(k_2)} v_{a,i}^{(k_1')} v_{a,i}^{(k_2')} \right) \\
&\times \exp\left(iN \frac{n(n-1)}{2} (\hat{q} \circ q) - \sum_{k_1,k_1',k_2,k_2',a<b} i\hat{q}^{(k_1,k_2,k_1',k_2')} \left(\sum_i \bar{e}_i^{(k_1)} \bar{e}_i^{(k_2)} v_{a,i}^{(k_1')} v_{b,i}^{(k_2')} \right) \right) \\
&\times \exp\left(iNn \text{Tr}(\hat{R} R^T) - i v^T (\hat{R} \otimes Id_n \otimes Id_N)(\bar{e} \otimes \mathbb{1}_n) - \frac{iNn}{2} \text{Tr}\hat{U} + \frac{i}{2} v^T (\hat{U} \otimes Id_n \otimes Id_N) v \right) \\
&\times \exp\left(-\frac{(\tilde{x} \otimes \mathbb{1}_n)^T (R^T \otimes Id_n \otimes Id_T) \mathcal{A}^{-1} (R \otimes Id_n \otimes Id_T)(\tilde{x} \otimes \mathbb{1}_n)}{2} \right)
\end{aligned} \tag{A.19}$$

Now let's take the integral w.r.t. $v_{a,i}$. Introduce

$$\begin{cases}
\tilde{q} = i\hat{q} \\
\tilde{r} = i\hat{r} \\
\tilde{M} = i\hat{M} \\
\tilde{L} = i\hat{L} \\
\tilde{R} = \frac{i}{\beta}\hat{R} \\
\tilde{U} = -i\hat{U}
\end{cases} \tag{A.20}$$

Remembering the notation for contraction of the first two indices, we write

$$(\tilde{q} \cdot \bar{e}_i \bar{e}_i^T)^{(k_1',k_2')} = \sum_{k_1,k_2} \tilde{q}^{(k_1,k_2,k_1',k_2')} \bar{e}_i^{(k_1)} \bar{e}_i^{(k_2)} \tag{A.21}$$

Then, we see that the integral over v is Gaussian once again, with quadratic term

$$\mathcal{C}_i = (\bar{\sigma}_i^2 (2\tilde{L} - \tilde{M}) + \tilde{U} + (2\tilde{r} - \tilde{q}) \cdot \bar{e}_i \bar{e}_i^T) \otimes Id_n + (\bar{\sigma}_i^2 \tilde{M} + \tilde{q} \cdot \bar{e}_i \bar{e}_i^T) \otimes J_n \tag{A.22}$$

and integrating, we get

$$
\begin{aligned}
\langle Z^n \rangle &= \int \frac{dqd\tilde{q}}{(2\pi)^{K^4}} \frac{drd\tilde{r}}{(2\pi)^{K^4}} \int \frac{d\tilde{U}}{(2\pi)^{K^2}} \frac{d\tilde{R}dR}{(2\pi N)^{K^2}} \frac{dMd\tilde{M}}{(2\pi)^{K^2}} \frac{dLd\tilde{L}}{(2\pi)^{K^2}} \frac{1}{\sqrt{\det \mathcal{A} \det \mathcal{B} \prod_i \det \mathcal{C}_i}} \exp\left(NnTr(\tilde{L}L^T)\right) \\
&\times \exp\left(\frac{Nn}{2}\mathrm{Tr}\tilde{U} + Nn(\tilde{r}\circ r) + \frac{Nn(n-1)}{2}(\tilde{q}\circ q) + Nn\mathrm{Tr}(\tilde{R}R^T) + \frac{Nn(n-1)}{2}Tr(\tilde{M}M^T)\right) \\
&\times \exp\left(\frac{(\tilde{x}^T \otimes \mathbb{1}_n^T)(\mathcal{A}^{-1}(R\otimes Id_{n,T}) - 2\gamma \otimes Id_{n,T})^T \mathcal{B}^{-1}(\mathcal{A}^{-1}(R\otimes Id_{n,T}) - 2\gamma \otimes Id_{n,T})(\bar{x}\otimes \mathbb{1}_n)}{2}\right) \\
&\times \exp\left(-\frac{(\tilde{x}\otimes \mathbb{1}_n)^T (R^T \otimes Id_n \otimes Id_T)\mathcal{A}^{-1}(R\otimes Id_n \otimes Id_T)(\tilde{x}\otimes \mathbb{1}_n)}{2}\right) \\
&\times \prod_i \left[\exp\left(\frac{1}{2}(\bar{e}_i \otimes \mathbb{1}_n)^T (\tilde{R}\otimes Id_n)^T \mathcal{C}_i^{-1}(\tilde{R}\otimes Id_n)(\bar{e}_i \otimes \mathbb{1}_n)\right)\right]
\end{aligned} \tag{A.23}
$$

We assume that as $\beta \to \infty$, the following combinations of variables go to a finite limit:

$$
\begin{cases}
v = \beta(r-q) \\
W = \beta(L-M) \\
\tilde{v} = \frac{1}{\beta}(2\tilde{r} - \tilde{q}) \\
\tilde{W} = \frac{1}{\beta}(2\tilde{L} - \tilde{M}) \\
U = \frac{1}{\beta}\tilde{U} \\
p = \frac{1}{\beta^2}\tilde{q} \\
S = \frac{1}{\beta^2}\tilde{M} \\
\tilde{\gamma} = \frac{1}{\beta}\gamma \\
\tilde{\eta} = \frac{1}{\beta}\eta
\end{cases} \tag{A.24}
$$

Now we can take $n \to 0$ and $\beta \to \infty$ keeping only leading orders.

$$
\begin{aligned}
&-\frac{(\tilde{x}\otimes \mathbb{1}_n)^T (R^T \otimes Id_n \otimes Id_T)\mathcal{A}^{-1}(R\otimes Id_n \otimes Id_T)(\tilde{x}\otimes \mathbb{1}_n)}{2} = \\
&= -\frac{\beta n \tilde{x}^T (R^T \otimes Id_T)\left(W\otimes Z + v\cdot \mathcal{X}\right)^{-1}(R\otimes Id_T)\tilde{x}}{2} + O(n^2)
\end{aligned} \tag{A.25}
$$

$$
\begin{aligned}
&\frac{(\tilde{x}^T \otimes \mathbb{1}_n^T)(\mathcal{A}^{-1}(R\otimes Id_{n,T}))^T \mathcal{B}^{-1}(\mathcal{A}^{-1}(R\otimes Id_{n,T}))(\bar{x}\otimes \mathbb{1}_n)}{2} = \\
&\frac{n\beta}{2}\tilde{x}^T \left((R^T \otimes Id_T)\left(W\otimes Z + v\cdot \mathcal{X}\right)^{-1}\right)\left(\left(W\otimes Z + v\cdot \mathcal{X}\right)^{-1} - 2\mathrm{diag}(1+\theta^{(k)}) \otimes \left(Id_T - \frac{1}{T}J_T\right)\right)^{-1} \\
&\left(\left(W\otimes Z + v\cdot \mathcal{X}\right)^{-1}(R\otimes Id_T)\right)\tilde{x} + O(n^2)
\end{aligned} \tag{A.26}
$$

$$
\frac{\beta^2}{2}(\bar{e}_i \otimes \mathbb{1}_n)^T (\tilde{R}\otimes Id_n)^T \mathcal{C}_i^{-1}(\tilde{R}\otimes Id_n)(\bar{e}_i \otimes \mathbb{1}_n) = \frac{n\beta}{2}\bar{e}_i^T \tilde{R}^T (\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v}\cdot \bar{e}_i \bar{e}_i^T)^{-1}\tilde{R}\bar{e}_i + O(n^2) \tag{A.27}
$$

Now we simplify the logarithms:

$$-\frac{1}{2}\ln\left[\det\mathcal{A}\det\mathcal{B}\right]=-\frac{1}{2}\ln\det\left[Id_{K,n,T}-2\left(W\otimes Z+v\cdot\mathcal{X}\right)\left(\text{diag}(1+\theta^{(k)})\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)\otimes Id_n\right.$$
$$\left.-2\beta\left(\left(L-\frac{W}{\beta}\right)\otimes Z+\frac{r-\frac{v}{\beta}}{N}\cdot\mathcal{X}+R\text{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\right)\left(-\text{diag}(1+\theta^{(k)})\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)\otimes J_n\right]=$$
$$=n\beta\text{Tr}\left[\left(Id_{K,T}-2\left(W\otimes Z+v\cdot\mathcal{X}\right)\left(\text{diag}(1+\theta^{(k)})\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)\right)^{-1}\right.$$
$$\left.\left(L\otimes Z+r\cdot\mathcal{X}+R\text{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\right)\left(\text{diag}(1+\theta^{(k)})\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)\right]+O(n^2) \tag{A.28}$$

$$-\frac{1}{2}\ln\left(\prod_i\det\mathcal{C}_i\right)=-\frac{n\beta}{2}\sum_i\text{Tr}\left[(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}(\bar{\sigma}_i^2S+p\cdot\bar{e}_i\bar{e}_i^T)\right]+O(n^2) \tag{A.29}$$

Finally, combining everything together, we can write

$$\begin{aligned}
\langle Z^n\rangle &= \int\frac{dvd\tilde{v}}{(2\pi)^{K^4}}\frac{dpdr}{(2\pi)^{K^4}}\frac{dU}{(2\pi)^{K^2}}\frac{d\tilde{R}dR}{(2\pi N)^{K^2}}\frac{dWd\tilde{W}dSdL}{(2\pi)^{2K^2}}\\
&\times\ \exp\left(n\beta\text{Tr}\left[\left(Id_{K,T}-2\left(W\otimes Z+v\cdot\mathcal{X}\right)\left(\text{diag}(1+\theta^{(k)})\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)\right)^{-1}\right.\right.\\
&\quad\left.\left.\left(L\otimes Z+r\cdot\mathcal{X}+R\text{diag}(\bar{\xi}^2)R^T\otimes\bar{\Delta}\right)\left(\text{diag}(1+\theta^{(k)})\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)\right]\right)\\
&\times\ \exp\left(-\frac{n\beta}{2}\sum_i\text{Tr}\left[(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}(\bar{\sigma}_i^2S+p\cdot\bar{e}_i\bar{e}_i^T)\right]\right)\\
&\times\ \exp\left(\frac{\beta Nn}{2}TrU+\frac{\beta Nn}{2}(\tilde{v}\circ r)+\frac{\beta Nn}{2}(p\circ v)+\beta Nn\text{Tr}(\tilde{R}R^T)+\frac{\beta Nn}{2}Tr(\tilde{W}L^T)+\frac{\beta Nn}{2}Tr(SW^T)\right)\\
&\times\ \exp\left(\frac{n\beta}{2}\tilde{x}^T\left((R\otimes Id_T)^T\left(W\otimes Z+v\cdot\mathcal{X}\right)^{-1}\right)\right.\\
&\quad\left(\left(W\otimes Z+v\cdot\mathcal{X}\right)^{-1}-2\text{diag}(1+\theta^{(k)})\otimes\left(Id_T-\frac{1}{T}J_T\right)\right)^{-1}\\
&\quad\left.\left(\left(W\otimes Z+v\cdot\mathcal{X}\right)^{-1}(R\otimes Id_T)\right)\tilde{x}\right)\\
&\times\ \exp\left(-\frac{\beta n\tilde{x}^T(R^T\otimes Id_T)\left(W\otimes Z+v\cdot\mathcal{X}\right)^{-1}(R\otimes Id_T)\tilde{x}}{2}\right)\\
&\times\ \prod_i\left[\exp\left(\frac{n\beta}{2}\bar{e}_i^T\tilde{R}^T(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}\tilde{R}\bar{e}_i\right)\right]
\end{aligned} \tag{A.30}$$

Going to the limit of $T,N\to\infty$, $T/N=\alpha$, we can simplify this expression:

$$
\begin{aligned}
\langle Z^n \rangle &= \int \frac{dvd\tilde{v}}{(2\pi)^{K^4}} \frac{dpdr}{(2\pi)^{K^4}} \frac{dU}{(2\pi)^{K^2}} \frac{d\tilde{R}dR}{(2\pi N)^{K^2}} \frac{dWd\tilde{W}dSdL}{(2\pi)^{2K^2}} \\
&\times \exp\left(n\beta \mathrm{Tr}\left[\left(Id_{K,T} - 2\left(W \otimes Z + v \cdot \mathcal{X}\right)\left(\mathrm{diag}(1+\theta^{(k)}) \otimes Id_T\right)\right)^{-1} \right.\right. \\
&\quad \left.\left. \left(L \otimes Z + r \cdot \mathcal{X} + R\mathrm{diag}(\bar{\xi}^2)R^T \otimes \bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right)\right)\left(\mathrm{diag}(1+\theta^{(k)}) \otimes Id_T\right)\right]\right) \\
&\times \exp\left(-\frac{n\beta}{2} \sum_i \mathrm{Tr}\left[(\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T)^{-1} (\bar{\sigma}_i^2 S + p \cdot \bar{e}_i \bar{e}_i^T) \right] \right) \\
&\times \exp\left(\frac{\beta Nn}{2} TrU + \frac{\beta Nn}{2}(\tilde{v} \circ r) + \frac{\beta Nn}{2}(p \circ v) + \beta Nn \mathrm{Tr}(\tilde{R}R^T) + \frac{\beta Nn}{2} Tr(\tilde{W}L^T) + \frac{\beta Nn}{2} Tr(SW^T) \right) \\
&\times \exp\left(\frac{n\beta}{2} \tilde{x}^T \left((R \otimes Id_T)^T \left(W \otimes Z + v \cdot \mathcal{X}\right)^{-1} \right) \left((W \otimes Z + v \cdot \mathcal{X})^{-1} - 2\mathrm{diag}(1+\theta^{(k)}) \otimes Id_T \right)^{-1} \right. \\
&\quad \left. \left((W \otimes Z + v \cdot \mathcal{X})^{-1} (R \otimes Id_T) \right) \tilde{x} \right) \\
&\times \exp\left(-\frac{\beta n \tilde{x}^T (R^T \otimes Id_T) \left(W \otimes Z + v \cdot \mathcal{X}\right)^{-1} (R \otimes Id_T) \tilde{x}}{2} \right) \\
&\times \prod_i \left[\exp\left(\frac{n\beta}{2} \bar{e}_i^T \tilde{R}^T (\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T)^{-1} \tilde{R} \bar{e}_i \right) \right]
\end{aligned} \tag{A.31}
$$

As in the case for the calculation presented in the main text, we will use the approximation for the integrals in $\langle Z^n \rangle$ valid for large β, replacing the integrated expression by its value at its optimal point, which we approximate as the stationary point of the free energy I. Then,

$$
\begin{aligned}
I &= \underset{v,\tilde{v},p,r,U,\tilde{R},R,W,\tilde{W},S,L}{\mathrm{optimum}} \left(-\frac{1}{2N} \sum_i \mathrm{Tr}\left[(\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T)^{-1} (\bar{\sigma}_i^2 S + p \cdot \bar{e}_i \bar{e}_i^T) \right] \right. \\
&+ \frac{1}{N} \mathrm{Tr}\left[\left(Id_{K,T} - 2\left(W \otimes Z + v \cdot \mathcal{X}\right)\left(\mathrm{diag}(1+\theta^{(k)}) \otimes Id_T\right)\right)^{-1} \right. \\
&\quad \left. \left(L \otimes Z + r \cdot \mathcal{X} + R\mathrm{diag}(\bar{\xi}^2)R^T \otimes \bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right)\right)\left(\mathrm{diag}(1+\theta^{(k)}) \otimes Id_T\right)\right] \\
&+ \frac{1}{2}\mathrm{Tr}U + \frac{1}{2}(\tilde{v} \circ r) + \frac{1}{2}(p \circ v) + \mathrm{Tr}(\tilde{R}R^T) + \frac{1}{2} Tr(\tilde{W}L^T) + \frac{1}{2} Tr(SW^T) \\
&+ \frac{1}{2N} \tilde{x}^T \left((R \otimes Id_T)^T \left(W \otimes Z + v \cdot \mathcal{X}\right)^{-1} \right) \\
&\quad \left((W \otimes Z + v \cdot \mathcal{X})^{-1} - 2\mathrm{diag}(1+\theta^{(k)}) \otimes Id_T \right)^{-1} \left((W \otimes Z + v \cdot \mathcal{X})^{-1} (R \otimes Id_T) \right) \tilde{x} \\
&- \frac{\tilde{x}^T (R^T \otimes Id_T) \left(W \otimes Z + v \cdot \mathcal{X}\right)^{-1} (R \otimes Id_T) \tilde{x}}{2N} \\
&+ \left. \sum_i \frac{1}{2N} \bar{e}_i^T \tilde{R}^T (\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v} \cdot \bar{e}_i \bar{e}_i^T)^{-1} \tilde{R} \bar{e}_i \right)
\end{aligned} \tag{A.32}
$$

Or, simplified,

$$
\begin{aligned}
I &= \underset{v,\tilde{v},p,r,U,\tilde{R},R,W,\tilde{W},S,L}{\text{optimum}} \Bigg(-\frac{1}{2N}\sum_i \text{Tr}\left[(\bar{\sigma}_i^2\tilde{W} + U + \tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}(\bar{\sigma}_i^2 S + p\cdot\bar{e}_i\bar{e}_i^T - \tilde{R}\bar{e}_i\bar{e}_i^T\tilde{R}^T)\right] \\
&+ \frac{1}{N}\text{Tr}\Bigg[\Big(Id_{K,T} - 2\left(W\otimes Z + v\cdot\mathcal{X}\right)\Big(\text{diag}(1+\theta^{(k)})\otimes Id_T\Big)\Big)^{-1} \\
&\quad \Bigg(L\otimes Z + r\cdot\mathcal{X} + (R\otimes Id_T)(\text{diag}(\bar{\xi}^2)\otimes\bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right) + \tilde{x}\tilde{x}^T)(R\otimes Id_T)^T\Bigg) \\
&\quad \Big(\text{diag}(1+\theta^{(k)})\otimes Id_T\Big)\Bigg] \\
&+ \frac{1}{2}\text{Tr}U + \frac{1}{2}(\tilde{v}\circ r) + \frac{1}{2}(p\circ v) + \text{Tr}(\tilde{R}R^T) + \frac{1}{2}Tr(\tilde{W}L^T) + \frac{1}{2}Tr(SW^T)
\end{aligned}
\tag{A.33}
$$

Setting $\theta^{(k)} = 0$ will give us (2.85), for which the procedure for finding the optimal point is described in the main text. Then, we find the theoretical prediction for the top eigenvalue as

$$
\begin{aligned}
\lambda^{(k)} = \frac{N}{T}\frac{\partial I}{\partial\theta^{(k)}}\Big|_{\theta^{(k)}=0} &= \frac{1}{T}\text{Tr}_T\Big[\left(Id_{K,T} - 2\left(W\otimes Z + v\cdot\mathcal{X}\right)\right)^{-1} \\
&\Bigg(L\otimes Z + r\cdot\mathcal{X} + (R\otimes Id_T)(\text{diag}(\bar{\xi}^2)\otimes\bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right) + \tilde{x}\tilde{x}^T)(R\otimes Id_T)^T\Bigg)\Bigg]^{(k,k)} + \\
&+ \frac{2}{T}\text{Tr}_T\Big[\left(Id_{K,T} - 2\left(W\otimes Z + v\cdot\mathcal{X}\right)\right)^{-1} \\
&\Bigg(L\otimes Z + r\cdot\mathcal{X} + (R\otimes Id_T)(\text{diag}(\bar{\xi}^2)\otimes\bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right) + \tilde{x}\tilde{x}^T)(R\otimes Id_T)^T\Bigg) \\
&\left(Id_{K,T} - 2\left(W\otimes Z + v\cdot\mathcal{X}\right)\right)^{-1}\left(W\otimes Z + v\cdot\mathcal{X}\right)\Big]^{(k,k)}
\end{aligned}
\tag{A.34}
$$

which after simplification gives

$$
\begin{aligned}
\lambda^{(k)} = \frac{N}{T}\frac{\partial I}{\partial\theta^{(k)}}\Big|_{\theta^{(k)}=0} &= \frac{1}{T}\text{Tr}_T\Big[\left(Id_{K,T} - 2\left(W\otimes Z + v\cdot\mathcal{X}\right)\right)^{-1} \\
&\Bigg(L\otimes Z + r\cdot\mathcal{X} + (R\otimes Id_T)(\text{diag}(\bar{\xi}^2)\otimes\bar{\Delta}\left(Id_T - \frac{1}{T}J_T\right) + \tilde{x}\tilde{x}^T)(R\otimes Id_T)^T\Bigg) \\
&\left(Id_{K,T} - 2\left(W\otimes Z + v\cdot\mathcal{X}\right)\right)^{-1}\Big]^{(k,k)}
\end{aligned}
\tag{A.35}
$$

Connection between ρ and R

In the main text, we obtained the following expression for ρ_i (2.83):

$$\begin{aligned}\left.\frac{\partial I}{\partial \tilde{\eta}_i}\right|_{\tilde{\gamma},\tilde{\eta}=0} &= -\frac{1}{2N}(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}(\bar{\sigma}_i^2 S+p\cdot\bar{e}_i\bar{e}_i^T-\tilde{R}\bar{e}_i\bar{e}_i^T\tilde{R}^T)(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}+\\ &+ \frac{1}{2N}\left(\bar{e}_i\bar{e}_i^T\right)+\frac{1}{N}(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}\left(\frac{\tilde{R}+\tilde{R}^T}{2}\right)\bar{e}_i\bar{e}_i^T \end{aligned} \tag{B.1}$$

If we sum the result for different i, we get

$$\begin{aligned}\sum_i\left.\frac{\partial I}{\partial \tilde{\eta}_i}\right|_{\tilde{\gamma},\tilde{\eta}=0} &= -\frac{1}{2N}\sum_i(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}(\bar{\sigma}_i^2 S+p\cdot\bar{e}_i\bar{e}_i^T-\tilde{R}\bar{e}_i\bar{e}_i^T\tilde{R}^T)(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}+\\ &+ \underbrace{\sum_i\frac{1}{2N}\left(\bar{e}_i\bar{e}_i^T\right)}_{\frac{1}{2}Id_K}+\frac{1}{N}\sum_i(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}\left(\frac{\tilde{R}+\tilde{R}^T}{2}\right)\bar{e}_i\bar{e}_i^T \end{aligned} \tag{B.2}$$

Using the expression obtained for I (2.85)

$$\begin{aligned} I &= \underset{v,\tilde{v},p,r,U,\tilde{R},R,W,\tilde{W},S,L}{\text{optimum}}\left(-\frac{1}{2N}\sum_i \text{Tr}\left[(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}(\bar{\sigma}_i^2 S+p\cdot\bar{e}_i\bar{e}_i^T-\tilde{R}\bar{e}_i\bar{e}_i^T\tilde{R}^T)\right]\right.\\ &+ \frac{1}{N}\text{Tr}\left[\left(Id_{K,T}-2\left(W\otimes Z+v\cdot\mathcal{X}\right)\right)^{-1}\right.\\ &\left.\left.\left(L\otimes Z+r\cdot\mathcal{X}+(R\otimes Id_T)(\text{diag}(\bar{\xi}^2)\otimes\bar{\Delta}\left(Id_T-\frac{1}{T}J_T\right)+\bar{x}\bar{x}^T)(R\otimes Id_T)^T\right)\right]\right.\\ &+ \left.\frac{1}{2}\text{Tr}U+\frac{1}{2}(\tilde{v}\circ r)+\frac{1}{2}(p\circ v)+\text{Tr}(\tilde{R}R^T)+\frac{1}{2}Tr(\tilde{W}L^T)+\frac{1}{2}Tr(SW^T)\right) \end{aligned} \tag{B.3}$$

we take the derivative of I w.r.t. U to get one one the conditions for finding the optimum of I:

$$\frac{\partial I}{\partial U}=0=\frac{1}{2}Id_K+\frac{1}{2N}\sum_i(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}(\bar{\sigma}_i^2 S+p\cdot\bar{e}_i\bar{e}_i^T-\tilde{R}\bar{e}_i\bar{e}_i^T\tilde{R}^T)(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1} \tag{B.4}$$

which means that the expression (B.5) further simplifies to

$$\sum_i\left.\frac{\partial I}{\partial \tilde{\eta}_i}\right|_{\tilde{\gamma},\tilde{\eta}=0} = Id_K+\frac{1}{N}\sum_i(\bar{\sigma}_i^2\tilde{W}+U+\tilde{v}\cdot\bar{e}_i\bar{e}_i^T)^{-1}\left(\frac{\tilde{R}+\tilde{R}^T}{2}\right)\bar{e}_i\bar{e}_i^T \tag{B.5}$$

Similarly, the derivative of I w.r.t. $\tilde{R}$ gives another condition for finding the optimum of I:

$$\frac{\partial I}{\partial \tilde{R}} = 0 = R + \frac{1}{N}\sum_i (\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v}\cdot \bar{e}_i \bar{e}_i^T)^{-1} \tilde{R} \bar{e}_i \bar{e}_i^T \tag{B.6}$$

meaning that

$$R = -\frac{1}{N}\sum_i (\bar{\sigma}_i^2 \tilde{W} + U + \tilde{v}\cdot \bar{e}_i \bar{e}_i^T)^{-1} \tilde{R} \bar{e}_i \bar{e}_i^T \tag{B.7}$$

and thus, by comparing B.5 to with the expression above, we get

$$\sum_i \frac{\partial I}{\partial \tilde{\eta}_i}\bigg|_{\tilde{\gamma},\tilde{\eta}=0} = Id_K - \frac{R + R^T}{2} \tag{B.8}$$

Bibliography

Advani, M., Lahiri, S., and Ganguli, S. (2013). Statistical mechanics of complex neural systems and high dimensional data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03014.

Akemann, W., Wolf, S., Villette, V., Mathieu, B., Tangara, A., Fodor, J., Ventalon, C., Léger, J.-F., Dieudonné, S., and Bourdieu, L. (2022). Fast optical recording of neuronal activity by three-dimensional custom-access serial holography. *Nat. Methods*, 19(1):100–110.

Altan, E., Solla, S. A., Miller, L. E., and Perreault, E. J. (2021). Estimating the dimensionality of the manifold underlying multi-electrode neural recordings. *PLOS Computational Biology*, 17(11):e1008591.

Baik, J., Arous, G. B., and Peche, S. (2004). Phase transition of the largest eigenvalue for non-null complex sample covariance matrices. *The Annals of Probability*.

Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *J. Physiol.*, 119(1):69–88.

Biehl, M. and Mietzner, A. (1994). Statistical mechanics of unsupervised structure recognition. *Journal of Physics A: Mathematical and General*, 27(6):1885–1897.

Bramlett, H., Green, E., and Dietrich, W. (1997). Hippocampally dependent and independent chronic spatial navigational deficits following parasagittal fluid percussion brain injury in the rat. *Brain Research*, 762(1–2):195–202.

Brette, R. and Destexhe, A. (2012). *Intracellular recording*, page 44–91. Cambridge University Press.

Briggman, K. L., Abarbanel, H. D. I., and Kristan, W. B. (2005). Optical imaging of neuronal populations during decision-making. *Science*, 307(5711):896–901.

Bumstead, J. R. (2018). Designing a large field-of-view two-photon microscope using optical invariant analysis. *Neurophotonics*, 5(02):1.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276.

Cheng, S. and Sabes, P. N. (2006). Modeling sensorimotor learning with linear dynamical systems. *Neural Computation*, 18(4):760–793.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, 487(7405):51–56.

Cunningham, J. P. and Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509.

Dana, H., Mohar, B., Sun, Y., Narayan, S., Gordus, A., Hasseman, J. P., Tsegaye, G., Holt, G. T., Hu, A., Walpita, D., Patel, R., Macklin, J. J., Bargmann, C. I., Ahrens, M. B., Schreiter, E. R., Jayaraman, V., Looger, L. L., Svoboda, K., and Kim, D. S. (2016). Sensitive red protein calcium indicators for imaging neural activity. *Elife*, 5.

De Nó, R. L. (1938). Analysis of the activity of the chains of internuncial neurons. *Journal of Neurophysiology*, 1(3):207–244.

Edwards, S. F. and Anderson, P. W. (1975). Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965–974.

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press.

Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M., and Cunningham, J. P. (2016). Reorganization between preparatory and movement population responses in motor cortex. *Nature Communications*, 7(1).

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379.

Fries, P. and Maris, E. (2022). What to do if n is two? *Journal of Cognitive Neuroscience*, 34(7):1114–1118.

Gallant, J. L., Connor, C. E., and Van Essen, D. C. (1998). Neural activity in areas v1, v2 and v4 during free viewing of natural scenes compared to controlled viewing. *NeuroReport*, 9(9):2153–2158.

Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A., and Miller, L. E. (2020). Long-term stability of cortical population dynamics underlying consistent behavior. *Nature Neuroscience*, 23(2):260–270.

Gallego, J. A., Perich, M. G., Miller, L. E., and Solla, S. A. (2017). Neural manifolds for the control of movement. *Neuron*, 94(5):978–984.

Gallego, J. A., Perich, M. G., Naufel, S. N., Ethier, C., Solla, S. A., and Miller, L. E. (2018). Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nat. Commun.*, 9(1):4233.

Gallego-Carracedo, C., Perich, M. G., Chowdhury, R. H., Miller, L. E., and Gallego, J. (2022). Local field potentials reflect cortical population dynamics in a region-specific and frequency-dependent manner. *eLife*, 11:e73155.

Gardner, E. and Derrida, B. (1988). Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271–284.

Garner, J. P. (2005). Stereotypies and other abnormal repetitive behaviors: potential impact on validity, reliability, and replicability of scientific outcomes. *ILAR J.*, 46(2):106–117.

Gong, Y., Huang, C., Li, J. Z., Grewe, B. F., Zhang, Y., Eismann, S., and Schnitzer, M. J. (2015). High-speed recording of neural spikes in awake mice and flies with a fluorescent voltage sensor. *Science*, 350(6266):1361–1366.

Grewe, B. F., Helmchen, F., and Kampa, B. M. (2013). *Two-Photon Imaging of Neuronal Network Dynamics in Neocortex*, page 133–150. Humana Press.

Grienberger, C., Giovannucci, A., Zeiger, W., and Portera-Cailliau, C. (2022). Two-photon calcium imaging of neuronal activity. *Nature Reviews Methods Primers*, 2(1).

Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory*. Wiley.

Henze, D. A., Borhegyi, Z., Csicsvari, J., Mamiya, A., Harris, K. D., and Buzsáki, G. (2000). Intracellular features predicted by extracellular recordings in the hippocampus in vivo. *Journal of Neurophysiology*, 84(1):390–400.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117(4):500–544.

Hölscher and Munk (2008). *Information Processing by Neuronal Populations*. Cambridge University Press.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160(1):106–154.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):e124.

Izenman, A. J. (2013). *Linear Discriminant Analysis*, page 237–280. Springer New York.

Jacobs, E. A., Steinmetz, N. A., Peters, A. J., Carandini, M., and Harris, K. D. (2020). Cortical state fluctuations during sensory decision making. *Current Biology*, 30(24):4944–4955.e7.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.*, 29(2):295–327.

Joliffe, I. (2002). *Principal Component Analysis*. Springer-Verlag.

Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., Lee, A. K., Anastassiou, C. A., Andrei, A., Aydın, u., Barbic, M., Blanche, T. J., Bonin, V., Couto, J., Dutta, B., Gratiy, S. L., Gutnisky, D. A., Häusser, M., Karsh, B., Ledochowitsch, P., Lopez, C. M., Mitelut, C., Musa, S., Okun, M., Pachitariu, M., Putzeys, J., Rich, P. D., Rossant, C., Sun, W.-l., Svoboda, K., Carandini, M., Harris, K. D., Koch, C., O'Keefe, J., and Harris, T. D. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236.

Katona, G., Szalay, G., Maák, P., Kaszás, A., Veress, M., Hillier, D., Chiovini, B., Vizi, E. S., Roska, B., and Rózsa, B. (2012). Fast two-photon in vivo imaging with three-dimensional random-access scanning in large tissue volumes. *Nature Methods*, 9(2):201–208.

Kaufman, M. T., Churchland, M. M., Ryu, S. I., and Shenoy, K. V. (2014). Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience*, 17(3):440–448.

Kayser, C., Montemurro, M. A., Logothetis, N. K., and Panzeri, S. (2009). Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron*, 61(4):597–608.

Kelly, M. and Woodbury, D. J. (2003). *Advantages and disadvantages of patch clamping versus using BLM*, page 699–721. Elsevier.

Khatib, D., Ratzon, A., Sellevoll, M., Barak, O., Morris, G., and Derdikman, D. (2023). Active experience, not time, determines within-day representational drift in dorsal ca1. *Neuron*, 111(15):2348–2356.e4.

Knöpfel, T. and Song, C. (2019). Optical voltage imaging in neurons: moving from technology development to practical tool. 20(12):719–727.

Kondo, M., Kobayashi, K., Ohkura, M., Nakai, J., and Matsuzaki, M. (2017). Two-photon calcium imaging of the medial prefrontal cortex and hippocampus without cortical invasion. *eLife*, 6:e26839.

Kremer, Y., Léger, J.-F., Lapole, R., Honnorat, N., Candela, Y., Dieudonné, S., and Bourdieu, L. (2008). A spatio-temporally compensated acousto-optic scanner for two-photon microscopy providing large field of view. *Optics Express*, 16(14):10066.

Ledergerber, D., Battistin, C., Blackstad, J. S., Gardner, R. J., Witter, M. P., Moser, M.-B., Roudi, Y., and Moser, E. I. (2021). Task-dependent mixed selectivity in the subiculum. *Cell Reports*, 35(8):109175.

Machens, C. K., Romo, R., and Brody, C. D. (2010). Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *The Journal of Neuroscience*, 30(1):350–360.

Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction.

Morcos, A. S. and Harvey, C. D. (2016). History-dependent variability in population dynamics during evidence accumulation in cortex. *Nature Neuroscience*, 19(12):1672–1681.

Nadella, K. M. N. S., Roš, H., Baragli, C., Griffiths, V. A., Konstantinou, G., Koimtzis, T., Evans, G. J., Kirkby, P. A., and Silver, R. A. (2016). Random-access scanning microscopy for 3d imaging in awake behaving animals. *Nature Methods*, 13(12):1001–1004.

Nechyporuk-Zloy, V., editor (2022). *Principles of light microscopy: From basic to advanced*. Springer International Publishing, Cham, Switzerland, 1 edition.

Nicolas-Alonso, L. F. and Gomez-Gil, J. (2012). Brain computer interfaces, a review. *Sensors*, 12(2):1211–1279.

Noguchi, A., Ikegaya, Y., and Matsumoto, N. (2021). In vivo whole-cell patch-clamp methods:

Recent technical progress and future perspectives. *Sensors*, 21(4):1448.

Ohki, K., Chung, S., Ch'ng, Y. H., Kara, P., and Reid, R. C. (2005). Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature*, 433(7026):597–603.

Otsu, Y., Marcaggi, P., Feltz, A., Isope, P., Kollo, M., Nusser, Z., Mathieu, B., Kano, M., Tsujita, M., Sakimura, K., and Dieudonné, S. (2014). Activity-dependent gating of calcium spikes by a-type k+ channels controls climbing fiber signaling in purkinje cell dendrites. *Neuron*, 84(1):137–151.

Panichello, M. F. and Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*, 592(7855):601–605.

Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E., and Fellin, T. (2017). Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron*, 93(3):491–507.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642.

Pellegrino, A., Stein, H., and Cayco-Gajic, N. A. (2024). Dimensionality reduction beyond neural subspaces with slice tensor component analysis. *Nature Neuroscience*, 27(6):1199–1210.

Peron, S. P., Freeman, J., Iyer, V., Guo, C., and Svoboda, K. (2015). A cellular resolution map of barrel cortex activity during tactile behavior. *Neuron*, 86(3):783–799.

Pettersen, K. H., Lindén, H., Dale, A. M., and Einevoll, G. T. (2012). *Extracellular spikes and CSD*, page 92–135. Cambridge University Press.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999.

Raposo, D., Kaufman, M. T., and Churchland, A. K. (2014). A category-free neural population supports evolving demands during decision-making. *Nature Neuroscience*, 17(12):1784–1792.

Reimann, P. and Van den Broeck, C. (1996). Learning by examples from a nonuniform distribution. *Physical Review E*, 53(4):3989–3998.

Rey, H. G., Pedreira, C., and Quian Quiroga, R. (2015). Past, present and future of spike sorting techniques. *Brain Research Bulletin*, 119:106–117.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Russo, A. A., Khajeh, R., Bittner, S. R., Perkins, S. M., Cunningham, J. P., Abbott, L. F., and Churchland, M. M. (2020). Neural trajectories in the supplementary motor area and motor cortex exhibit distinct geometries, compatible with different classes of computation. *Neuron*, 107(4):745–758.e6.

Sadtler, P. T., Quick, K. M., Golub, M. D., Chase, S. M., Ryu, S. I., Tyler-Kabara, E. C., Yu, B. M., and Batista, A. P. (2014). Neural constraints on learning. *Nature*, 512(7515):423–426.

Safaie, M., Chang, J. C., Park, J., Miller, L. E., Dudman, J. T., Perich, M. G., and Gallego, J. A. (2023). Preserved neural dynamics across animals performing similar behaviour. *Nature*, 623(7988):765–771.

Saha, D., Leong, K., Li, C., Peterson, S., Siegel, G., and Raman, B. (2013). A spatiotemporal coding mechanism for background-invariant odor recognition. *Nature Neuroscience*, 16(12):1830–1839.

Salomé, R., Kremer, Y., Dieudonné, S., Léger, J.-F., Krichevsky, O., Wyart, C., Chatenay, D., and Bourdieu, L. (2006). Ultrafast random-access scanning in two-photon microscopy using acousto-optic deflectors. *J. Neurosci. Methods*, 154(1-2):161–174.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

Schwartz, A., Kettner, R., and Georgopoulos, A. (1988). Primate motor cortex and free arm movements to visual targets in three- dimensional space. i. relations between single cell discharge and direction of movement. *The Journal of Neuroscience*, 8(8):2913–2927.

Sherrington, D. and Kirkpatrick, S. (1975). Solvable model of a spin-glass. *Physical Review Letters*,

35(26):1792–1796.

Shinn, M. (2023). Phantom oscillations in principal component analysis. *Proceedings of the National Academy of Sciences*, 120(48).

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., Chen, S., Colonell, J., Gardner, R. J., Karsh, B., Kloosterman, F., Kostadinov, D., Mora-Lopez, C., O'Callaghan, J., Park, J., Putzeys, J., Sauerbrei, B., van Daal, R. J. J., Vollan, A. Z., Wang, S., Welkenhuysen, M., Ye, Z., Dudman, J. T., Dutta, B., Hantman, A. W., Harris, K. D., Lee, A. K., Moser, E. I., O'Keefe, J., Renart, A., Svoboda, K., Häusser, M., Haesler, S., Carandini, M., and Harris, T. D. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539).

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., and Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365.

Svoboda, K. and Yasuda, R. (2006). Principles of two-photon excitation microscopy and its applications to neuroscience. *Neuron*, 50(6):823–839.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Theumann, W. K. and Köberle, R. (1990). *Neural Networks and Spin Glasses: Proceedings of the STATPHYS 17 Workshop*. WORLD SCIENTIFIC.

Tischbirek, C., Birkner, A., Jia, H., Sakmann, B., and Konnerth, A. (2015). Deep two-photon brain imaging with a red-shifted fluorometric ca2+ indicator. 112(36):11377–11382.

Titus, D. J., Wilson, N. M., Freund, J. E., Carballosa, M. M., Sikah, K. E., Furones, C., Dietrich, W. D., Gurney, M. E., and Atkins, C. M. (2016). Chronic cognitive dysfunction after traumatic brain injury is improved with a phosphodiesterase 4b inhibitor. *Journal of Neuroscience*, 36(27):7095–7108.

Törnqvist, E., Annas, A., Granath, B., Jalkesten, E., Cotgreave, I., and Öberg, M. (2014). Strategic focus on 3R principles reveals major reductions in the use of animals in pharmaceutical toxicity testing. *PLoS One*, 9(7):e101638.

Villette, V., Chavarha, M., Dimov, I. K., Bradley, J., Pradhan, L., Mathieu, B., Evans, S. W., Chamberland, S., Shi, D., Yang, R., Kim, B. B., Ayon, A., Jalil, A., St-Pierre, F., Schnitzer, M. J., Bi, G., Toth, K., Ding, J., Dieudonné, S., and Lin, M. Z. (2019). Ultrafast two-photon imaging of a high-gain voltage indicator in awake behaving mice. *Cell*, 179(7):1590–1608.e23.

Vinje, W. E. and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276.

Wang, T., Chen, Y., and Cui, H. (2022). From parametric representation to dynamical system: Shifting views of the motor cortex in motor control. *Neuroscience Bulletin*, 38(7):796–808.

Wasmuht, D. F., Spaak, E., Buschman, T. J., Miller, E. K., and Stokes, M. G. (2018). Intrinsic neuronal dynamics predict distinct functional roles during working memory. *Nature Communications*, 9(1).

Watkin, T. L. H. and Nadal, J. P. (1994). Optimal unsupervised learning. *Journal of Physics A: Mathematical and General*, 27(6):1899–1915.

Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use t-sne effectively. *Distill*.

Wei, Z., Lin, B.-J., Chen, T.-W., Daie, K., Svoboda, K., and Druckmann, S. (2020). A comparison of neuronal population dynamics measured with calcium imaging and electrophysiology. *PLOS Computational Biology*, 16(9):e1008198.

Widrow, B. (1962). Self-organizing systems 1962.

Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V., and Sahani, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102(1):614–635.

Yu, Y., McTavish, T. S., Hines, M. L., Shepherd, G. M., Valenti, C., and Migliore, M. (2013). Sparse distributed representation of odors in a large-scale olfactory bulb circuit. *PLoS Computational Biology*, 9(3):e1003014.

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, 16(8):487–497.

Acknowledgments

I would like to first extend my deepest gratitude to my supervisors, Rémi Monasson and Laurent Bourdieu, for their guidance and support over these past few years. Thanks to both of you, I had the privilege of working in an environment that was both encouraging and intellectually stimulating. I thank you for your understanding, non-judgmental attitude, and especially your patience. Knowing that you believe in me and my potential, shown by offering me the chance to continue our work together beyond this PhD, means a lot to me.

My sincere thanks go to each member of my thesis jury: to Ada Altieri and Fleur Zeldenrust, and especially to Brice Bathellier and David Dean for agreeing to be my thesis reviewers.

I'd like to thank my wonderful colleagues, whose support, friendship, and shared wisdom made this PhD journey so much more enjoyable. Over the years, I was lucky to work side by side with Emanuele, Jorge, Sebastian, Vito, Simona, Eugenio, Andrea, Cyril, Mauro, Francesco, Thomas, Leo, Hugo, Massimilliano, Anthony, Alice, Cathie, Jean-François, Sebastien, Caio, Vladimir, Josephine, Anaïs, Walther, and Louis-Victor. Each of you has played a part in making these years an unforgettable experience, full of learning, growth, and friendship. Thank you for being such an important part of this journey.

A special thanks to Francesco and Mauro for the productive discussions on the mathematical aspects of this work. And double thanks to Francesco, Alice and Caio for their emotional support, especially at the end of this tough journey.

I would like to extend my deepest thanks to my friends outside the lab, whose companionship and support have been a vital part of my journey. I feel incredibly lucky to have met such wonderful people here in France, who have been there to share in the highs and help carry me through the lows. A special thanks to Giulia and Federico, whom I met during my masters studies here at ENS, and who have been constant sources of encouragement, joy and music. I am also especially grateful to Adrien for his kindness, humor, and support.

I would like to thank my family back in Russia, whose unwavering support has been a constant source of strength throughout my PhD journey. Being able to see you during these years was increasingly challenging due to many reasons. Despite the distance, your encouragement, love, and belief in me were deeply felt and kept me grounded. Thank you for being there for me. I am also incredibly grateful to the family of my partner, Guillaume, especially his parents, Adeline and Nicolas, who welcomed me with open arms and made me feel like part of their own.

To Guillaume, who not only tolerated my late nights and thesis rants but somehow managed to make me laugh through it all – thank you. Now, as you approach the end of your own PhD, I'm doing my best to return the favor, and I hope I can be as steady a support for you as you've been for me.

RÉSUMÉ

Les techniques d'enregistrement multi-électrodes ou d'imagerie calcique permettent désormais de mesurer l'activité de centaines, voire de milliers de neurones dans une région du cerveau. Une hypothèse courante est que cette activité de haute dimension est intégrée dans une variété de basse dimension, et décrit une trajectoire qui renseigne, par exemple, sur les calculs sensoriels ou moteurs en cours. L'analyse en composantes principales, parmi d'autres méthodes de reconstruction de variété, est couramment utilisée pour inférer le sous-espace neuronal pertinent et visualiser les trajectoires correspondantes.
Un enjeu important est la fiabilité de ce processus de reconstruction de trajectoire. La durée limitée d'enregistrement, la taille de la population enregistrée, la présence de bruit neuronal dynamique, ainsi que les limitations intrinsèques des méthodes d'enregistrement, telles que les erreurs ou les biais dans l'identification des spikes à partir de signaux de tension ou de fluorescence, peuvent considérablement affecter les trajectoires reconstruites, et notre compréhension des processus computationnels sous-jacents.
Dans notre travail, nous présentons une caractérisation systématique de ces effets sur l'inférence de trajectoire en basse dimension. Nous obtenons des expressions analytiques pour l'erreur de reconstruction en utilisant les outils de la physique statistique des systèmes désordonnés, et établissons des diagrammes de phase localisant les régions où l'erreur est maîtrisée, et celles où la reconstruction est impossible. Nos résultats sont confirmés par des simulations numériques étendues. Nous montrons ensuite, sur divers enregistrements existants dans le cortex visuel, l'hippocampe et le cortex préfrontal, comment nos résultats peuvent être utilisés pour caractériser l'erreur attendue sur les trajectoires obtenues à partir de jeux de données réels.

MOTS CLÉS

Physique statistique, Neuroscience, Conception expérimentale, Analyse en composantes principales, Réduction de la dimensionnalité, Modélisation

ABSTRACT

Multi-electrode or calcium-imaging techniques now make possible to record the activity of hundreds or thousands of neurons in a brain area. A common assumption is that this high-dimensional activity is embedded on a low-dimensional manifold, and describes a trajectory informative about the ongoing e.g. sensory or motor computations. Principal component analysis, among other manifold-reconstruction methods, is routinely used to infer the relevant neural subspace and visualize the corresponding trajectories.
An important issue is the reliability of this trajectory reconstruction process. The limited recording time, the size of the recorded population, the presence of dynamical neural noise, as well as intrinsic limitations of the recording methods, such as errors or biases in spike identification from voltage or fluorescence signals may considerably affect the reconstructed trajectories, and our understanding of the underlying computational processes.
In our work, we present a systematic characterization of these effects on low-dimensional trajectory inference. We derive analytical expressions for the reconstruction error using the tools of the statistical physics of disordered systems, and derive phase diagrams locating regions in which the error is under control, or in which reconstruction is not possible. Our results are confirmed by extensive numerical simulations. We then show, on various existing recordings in the visual cortex, the hippocampus, and in the prefrontal cortex, how our results can be used to characterize the expected error on trajectory obtained from real datasets.

KEYWORDS

Statistical physics, Neuroscience, Experimental design, Principal Component Analysis, Dimensionality reduction, Modeling