Université de recherche Paris Sciences et Lettres
PSL Research University

École Doctorale de Dauphine – ED 543

Synthèse de Textures Dynamiques pour l'Étude de la Vision en Psychophysique et en Electrophysiologie

Dynamic Textures Synthesis for Probing Vision in Psychophysics and Electrophysiology

THÉSE DE DOCTORAT

Pour l'obtention du titre de

Docteur en Sciences
Spécialité Mathématiques Appliquées

Présentée par

Jonathan Vacher

Soutenue publiquement le 18 janvier 2017 devant le jury composé de

Frédéric	Chavane	CNRS, INT, Aix-Marseille Université	Examinateur
Laurent	Cohen	CNRS, CEREMADE, Université Paris Dauphine	Président
Cyril	Monier	CNRS, UNIC	Directeur
Jean Michel	Morel	CMLA, École Normale Supérieure de Paris-Saclay	Rapporteur
Peter	Neri	LSP, École Normale Supérieure	Examinateur
Gabriel	Peyré	CNRS, DMA, École Normale Supérieure	Directeur
Bertrand	Thirion	INRIA	Rapporteur

Résumé

Le but de cette thèse est de proposer une modélisation mathématique des stimulations visuelles afin d'analyser finement des données expérimentales en psychophysique et en electrophysiologie. Plus précisément, afin de pouvoir exploiter des techniques d'analyse de données issues des statistiques Bayésiennes et de l'apprentissage automatique, il est nécessaire de développer un ensemble de stimulations qui doivent être dynamiques, stochastiques et d'une complexité paramétrée. Il s'agit d'un problème important afin de comprendre la capacité du système visuel à intégrer et discriminer differents stimuli. En particulier, les mesures effectuées à de multiples échelles (neurone, population de neurones, cognition) nous permette d'étudier les sensibilités particulières des neurones, leur organisation fonctionnelle et leur impact sur la prise de décision. Dans ce but, nous proposons un ensemble de contributions théoriques, numériques et expérimentales, organisées autour de trois axes principaux : (1) un modèle de synthése de textures dynamiques Gaussiennes spécialement paramètrée pour l'étude de la vision; (2) un modèle d'observateur Bayésien rendant compte du biais positif induit par fréquence spatiale sur la perception de la vitesse; (3) l'utilisation de méthodes d'apprentissage automatique pour l'analyse de données obtenues en imagerie optique par colorant potentiométrique et au cours d'enregistrements extra-cellulaires. Ce travail, au carrefour des neurosciences, de la psychophysique et des mathématiques, est le fruit de plusieurs collaborations interdisciplinaires.

Mots-clés: stimulation visuelle, synthèse de textures, inférence Bayésienne inverse, discrimination de vitesse, apprentissage supervisé, imagerie optique par colorant potentiométrique, enregistrement extra-cellulaire, cartes d'orientations, selectivité à l'orientation, neurones.

Abstract

The goal of this thesis is to propose a mathematical model of visual stimulations in order to finely analyze experimental data in psychophysics and electrophysiology. More precisely, it is necessary to develop a set of dynamic, stochastic and parametric stimulations in order to exploit data analysis techniques from Bayesian statistics and machine learning. This problem is important to understand the visual system capacity to integrate and discriminate between stimuli. In particular, the measures performed at different scales (neurons, neural population, cognition) allow to study the particular sensitivities of neurons, their functional organization and their impact on decision making. To this purpose, we propose a set of theoretical, numerical and experimental contributions organized around three principal axes: (1) a Gaussian dynamic texture synthesis model specially crafted to probe vision; (2) a Bayesian observer model that accounts for the positive effect of spatial frequency over speed perception; (3) the use of machine learning techniques to analyze voltage sensitive dye optical imaging and extracellular data. This work, at the crossroads of neurosciences, psychophysics and mathematics is the fruit of several interdisciplinary collaborations.

Keywords: visual stimulation, texture synthesis, inverse Bayesian inference, speed discrimination, supervised learning, voltage sensitive dye optical imaging, extracellular recordings, orientation maps, orientation selectivity, neurons.

Acknowledgments

Cette thèse est le fruit de mon travail effectué au cours des trois dernières années. Celui-ci n'aurait pu être conduit ni même aboutir sans un ensemble de personnes que je souhaite remercier ici.

Tout d'abord, je remercie mon directeur de thèse Gabriel Peyré qui m'a proposé ce sujet entre mathématiques et neurosciences correspondant parfaitement à mes attentes. Tout au long de cette thèse Gabriel s'est montré enthousiaste et disponible, il a parfaitement su me transmettre le bagage nécessaire du chercheur (discussions et écriture scientifiques, présentations orales, collaborations, ...). Qu'attendre de plus d'un directeur de thèse ? Tout cela fera pour moi référence. Je remercie également Yves Frégnac sans qui je ne n'aurais pas eu la chance d'être au plus prés des neurosciences expérimentales et qui m'a toujours poussé à renforcer mes connaissances dans ce domaine. Je remercie enfin Cyril Monier, mon co-directeur de thèse, qui s'est toujours montré débordant d'idées et avec qui nous avons pu dépasser les difficultés de communication qu'il peut y avoir entre un mathématicien et un neuroscientifique.

Ce travail pluridisciplinaire m'a amené à travailler avec de nombreuses personnes. Mes voyages à l'INT à Marseille m'ont permis de travailler avec Laurent Perrinet avec qui nous avons enrichi le modèle des Motion Clouds qu'il avait introduit quelques années plus tôt. Ce fût un réel plaisir de partager repas et café sur la terrasse ensoleillée de l'INT. De même, j'ai appris à l'INT au coté d'Andrew Meso, comment conduire une expérience de psychophysique. Nos rencontres fûrent toujours agréables et intéressantes tant sur le plan scientifique qu'humain. Enfin, j'ai pû avoir quelques discussions avec Frédèric Chavane qui fûrent également très utiles pour mener les analyses des données obtenues en imagerie optique.

Bien entendu, c'est à l'UNIC que j'ai eu le plus d'échanges ! Merci à Luc Foubert pour son accueil et ses discussions toujours intéressantes. Merci à Yannick Passarelli pour sa bonne humeur, son envie de déconner et sa motivation à travailler avec moi. Merci à Margot pour ses remarques scientifiques très utiles et nos débats sans fin au sujet des notebooks. Un grand merci à Gérard Sadoc qui a codé en CUDA notre algorithme de synthèse de textures dynamiques dans le logiciel Elphy. Merci à Isabelle Férézou pour ce partage de données infructueux et son intérêt pour mon travail. Merci à Jan Antolik pour nos discussions toujours fructueuses. Merci à tous les autres: Valérie, Benoit, Pauline, Xoana, Evan, Matias, Christophe et j'en oublie...

Merci également aux nombreux thésards du CEREMADE ! En particulier, merci à mes "frères" et "sœur" de thèse , Nicolas, Hugo, Samuel, Quentin, Lénaïc, Aude, Paul pour les discussions, parfois débiles, parfois scientifiques. Les autres: Maxime (mon cher co-bureau), Luca, Raphaël et Fang.

Merci à tous mes amis: Kenan et Swann pour leur intérêt et leurs questions, Adrien, Simon, Ron, Quentin, Pierre, Vincent, Anaïs Élise, Alice. Merci aux Toulousaings: Alex, Fab, Rafi, Muriel, Lisa, Le Kéh, ... Merci également au groupe Zorionak Fermat pour le spam pendant les matchs de Rugby! Merci particulièrement à Manon qui m'a supporté et soutenu quoiqu'il arrive et qui a su être présente dans les moments difficiles !

Enfin je veux remercier mes parents, mon frère et le reste de ma famille pour leur soutien sans faille tout au long de mes études qui s'achèvent enfin !

Table of Contents

Introduction

1 Outline

Overview. The goal of this PhD is to develop a mathematically sound framework to perform both stochastic visual stimulation and statistic data analysis, for psychophysics and electrophysiology of the visual brain.

Scientific Context. This work is a collaboration between the team of Gabriel Peyré initially at the research center CEREMADE (Université Paris-Dauphine, France) and now at DMA (École Normale Supérieure de Paris), and the visual neuroscience team of Yves Frégnac at the UNIC lab (Gif-sur-Yvette, France). As a strong interdisciplinary project, my work was supervised by Gabriel Peyré in mathematics and Cyril Monier in neurosciences. I have also developed a strong collaboration in psychophysics of vision with Laurent Perrinet from INT (Marseille, France) and Andrew Meso initially at INT and now at the CCNRC (Bournemouth University, Bournemouth, UK) .

This collaboration led me to work in various fields in addition to my initial mathematical background. The target readers of this manuscript are mathematicians, neuroscientists and psychophysicists. In addition to promote such interactions, I intend to show how they benefit to experimental neurosciences and psychophysics. In particular, when studying vision, it is important to go trough more complex, mathematically grounded visual stimulation models. Increasing the complexity of these models allows us to build innovative experimental protocols and the mathematical description provides a clear understanding of the stimuli. Such an understanding is necessary in order to be able to make sense of the collected data. Moreover, the increasing amount of available data stimulates the development of machine learning based techniques that experimental neurosciences should benefit from.

Manuscript Organization. This manuscript is organized into six chapters among which three principal parts can be distinguished:

- *Chapter I: description of the Motion Cloud (MC) visual stimulation model.* In Chapter I, we detail the mathematical description of a dynamic texture model specially crafted to probe vision. This work was done in collaboration with Laurent Perrinet and Andrew Meso at INT. Gérard Sadoc (UNIC) implemented our algorithm into the Elphy software used at UNIC for stimulation and recording. This constitutes our first contribution. Initially developed by Leon [169] as a spatio-temporal

Gaussian field, we embed this model in a general framework which enables us to give three equivalent formulations. On the one hand, these three formulations provide a biologically plausible justification of the model; on the other hand, they provide a real-time synthesis algorithm.

- *Chapters II and III: a Bayesian approach to discrimination tasks in psychophysics and an application to the study of motion perception.* In Chapter II, we describe a mathematical formulation of an ideal Bayesian observer model and we show how to use it to analyze data in psychophysics. We tackle the question of inverse Bayesian inference *ie* knowing decisions made using a Bayesian model, we intend to infer the likelihood and prior. In Chapter III, we use this approach to explain the results obtained in a two-alternate forced choice (2AFC) speed discrimination experiment using MC stimulations. This work was done in collaboration with Laurent Perrinet and Andrew Meso at INT and constitutes our second sets of contributions. We confront our model to real data and successfully describe the positive effect of spatial frequency over speed perception using a bi-variate prior.

- *Chapters IV, V and VI: the use of machine learning techniques to analyze electrophysiological data.* In Chapter IV, we recall the basics of supervised learning and define a new classification error measure. Then, in Chapters V and VI, we make use of supervised learning to analyze Voltage Sensitive Dye optical Imaging (VSDi) data and Extracellular Recordings (ER). This work was done in collaboration with Cyril Monier, Luc Foubert, Yannick Passarelli and Margot Larroche from the vision team at UNIC. Unfortunately, experiments using MCs and VSDi were not fruitful, however we obtained some interesting results using standard grating stimuli. In contrast, the results obtained by using MCs and ER are promising. Supervised learning appears relevant to analyze the spatio-temporal dynamic of the VSDi signal, and this allows us to provide a methodology to analyze different types of protocols and to conclude on a simple model of VSDi signal. We conduct similar analysis on ER, and we show that neural populations contain enough information to discriminate between stimuli that differ with regards to parameters of the MC stimulations (orientation and spatial frequency bandwidth). We conclude by showing that these findings are compatible with a simple neural computational model.

2 Problem under Study and Previous Works

2.1 Probing the Visual System: from Stimulation to Data Analysis

The visual system can be viewed as a machine that receives external inputs (the photons that reach the eyes) and produces internal outputs (electrical signal in the brain). In order to understand how this machine transforms inputs into outputs we need to have a clear understanding of both inputs and outputs. The inputs are physically well understood and we have been able to build some optical instruments to capture visible light from our environment. However, we still have a poor understanding of natural images as we have no low dimensional mathematical model to account for the complexity of natural images. Indeed, the vast majority of the litterature focuses on establishing statistical properties or on identifying key features of natural images [86, 163, 195, 206]. Furthermore, the problem of dynamical modeling is much less studied, see for instance the previous work of Dong [44]. Even if the design of generic natural model is out of reach, the statistical modeling of texture was succesfully applied to the problem of texture synthesis (see Section 1.1). We mention here the work of Portilla *et al.* [149] who design an algorithm based on physiology and psychophysics. We also highlight the work of Galerne *et al.* [63] who settle the theoretical basis of our dynamic Motion Cloud model. Concerning the outputs, the biophysical mechanisms at stake in neurons and synapses are now well understood. While experimental instruments and techniques have progressed a lot, recording very large assemblies of single neurons is impossible. Therefore, we only have partial samples of brain outputs, and experimental techniques open small windows at different scales on these outputs. Probing the visual system thus corresponds to understanding the relation between inputs that have an enormous complexity and outputs that are not fully understood and are only partially accessible [53, 176].

The question of the class of stimulation that should be used to probe the visual system is a long standing debate. This debate opposes artificial stimulation [164] to natural stimulation [68]. For the last 60 years, artificial stimuli have been most commonly used. Generally, they consist of moving bars, dots, sine waves or noise. These artificial stimuli are well controlled and allow to test particular features present in natural images such as oriented edges, spatial and temporal frequencies, and movement directions. They play an essential role in the concept of tuning curve which represents the firing activity of a neuron as a function of some stimulus parameter. Noise as a stochastic stimulus appears fundamental to the estimation of receptive fields using Spike Triggered Average (STA) and Spike Triggered Covariance (STC) methods [177]. Natural stimuli are less used but the improvement of optical instruments and the spreading of images in the numerical environment make them more and more attractive to be used as stimuli. Natural stimulation is mainly motivated by the idea that our brain has adapted to its environment through long term evolution mechanisms [175]. Therefore, by using natural stimulation we prevent unnatural bias that can emerge from artificial stimulation [140]. However, choosing naturalness raises the issue of the high statistical complexity of natural images. Handling this complexity requires to make simple hypotheses about the specific features of natural images that elicit neural responses. Therefore, there is a risk to miss some multifactorial features that explain the responses. Finally, the fundamental question lies in the gap that separates artificial and natural stimulation. There are ways to fill the gap on whether it consists in building more complex parametric model [149, 58] or deteriorating natural images [160, 115]. Our work is rooted in these attempts by trying to increase the complexity of artificial and stochastic stimulations, yet keeping a reasonable number of parameters.

In experimental neurosciences the recording techniques are multiple: intracellular and extracellular recordings (ER), electroencephalography, functional magnetic resonance imaging (fMRI), two-photons imaging, voltage sensitive dye optical imaging (VSDi), *etc.* These measuring instruments have specific advantages and drawbacks and are generally associated to different spatio-temporal scales. For example, fMRI records signal at the scale of the entire brain with a low temporal resolution while intracellular recording measures the electric activity of a single neurons at a high temporal resolution. Under such constraints it becomes difficult to embed datasets from different techniques in a common framework. In fact, for each scale there exist an adapted mathematical framework. For instance, Hodgkin-Huxley modeled the signal of a single neuron; when moving to a neural population it becomes more adapted to use mean-fields combined with dynamical systems theory [50, 31] or build a neural network. In psychophysics, data collection is restricted; it consists in asking an observer if she detects or discriminates different stimuli. In such a way, we are able to measure detection or discrimination thresholds and bias at cognition scale. Along this work, we try to make sense of data collected with different techniques at different scales under similar stimulation.

2.2 Bayesian Modeling Visual Motion Perception

A normative explanation for the function of perception is to infer relevant unknown real world parameters Q from the sensory input, with respect to a generative model [74]. Equipped with a prior about the distribution of the parameter (typically learned and which reflects some knowledge about the surrounding world) the modeling representation that emerges corresponds to the *Bayesian brain* hypothesis [103, 46, 34, 102]. This assumes that when given some sensory information S, the brain takes a decision using the posterior distribution of the parameter given the sensory information, which, by Bayes theorem, can be obtained as :

$$\mathbb{P}_{Q|S}(q|s) = \frac{\mathbb{P}_{S|Q}(s|q)\mathbb{P}_Q(q)}{\mathbb{P}_S(s)}. \tag{2.1}$$

where $\mathbb{P}_{S|Q}$ is the likelihood and $\mathbb{P}_Q$ represents prior knowledge. This hypothesis is well illustrated with the case of motion perception [211]. This work uses a Gaussian parameterization of the generative model and a unimodal (Gaussian) prior in order to estimate perceived speed v when observing a visual input I. However, such a Bayesian hypothesis – based on the formalization of unimodal Gaussian prior and likelihood functions for instance – does not always fit psychophysical results [209, 79]. As such, a major challenge is to refine the definition of generative models so that they are consistent with a larger set of experimental results.

The estimation problem inherent to perception can be somehow be aleviated by defining an adequate generative model. The simplest generative model to describe visual motion is probably the luminance conservation equation [3]. It states that luminance $I(x,t)$ for $(x,t) \in \mathbb{R}^2 \times \mathbb{R}$ is approximately conserved along trajectories defined as integral lines of a vector field $v(x,t) \in \mathbb{R}^2 \times \mathbb{R}$. The corresponding generative model defines random fields as solutions to the stochastic partial differential equation (sPDE),

$$\langle v, \nabla I \rangle + \frac{\partial I}{\partial t} = W, \tag{2.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product in $\mathbb{R}^2$, ∇I is the spatial gradient of I. To match the spatial scale or frequency statistics of natural scenes (*ie* 1/f amplitude fall-off) or of some alternative category of textures, the driving term W is usually defined as a stationary colored Gaussian noise corresponding to the average localized spatio-temporal correlation (which we refer to as spatio-temporal coupling), and is parameterized by a covariance matrix Σ, while the field is usually a constant vector $v(x,t) = v_0$ accounting for a full-field translation with constant speed.

Ultimately, the application of this generative model is essential for probing the visual system, for instance for one seeking to understand how observers might detect motion in a scene. Indeed, as shown by [135, 211], the negative log-likelihood of the probability distribution of the solutions I to the luminance conservation equation (2.2), on some space-time observation domain $\Omega \times [0, T]$, for some hypothesized constant speed $v(x, t) = v_0$, is proportional to the value of the motion-energy model [3]

$$\int_\Omega \int_0^T |\langle v_0, \nabla(K \star I)(x, t)\rangle + \frac{\partial(K \star I)}{\partial t}(x, t)|^2 \mathrm{d}t \, \mathrm{d}x \tag{2.3}$$

where K is the whitening filter corresponding to the inverse square root of Σ, and $\star$ is the convolution operator. Using a prior knowledge about the expected distribution of motions (preference for slow speeds, for instance), a Bayesian formalization can be applied to this inference problem [210, 211]. One of the purposes of this dissertation is to refine this class of dynamic stochastic models to perform motion estimation using energy models associated to the stimulation.

2.3 Data Analysis in Electrophysiology

Voltage Sensitive Dye Optical Imaging The VSDi technique is a promising recording modality for the cortical activity at meso-scale. It consists in staining the cortical surface with some voltage sensitive dye and filming this surface [75]. In presence of electrical activity and light, the dye re-emits light. It is therefore possible to identify some areas of activity under different stimulations. However, the signal is known to be corrupted by many artifacts, which leads many people to tackle this question [159, 216, 156]. These drawbacks have not prevented experimentalists to reproduce results obtained in intrinsic optical imaging. They use oriented drifting gratings as stimuli that elicit responses in different areas of the primary visual cortex (see *eg* [174]). This reveals the existence of orientation maps in the primary cortex of many mammals: the surface of the primary visual cortex is clustered in different domains where neurons share the same orientation tuning. The main interest of VSDi is its high temporal resolution. The paper by Sharon [174] highlights the increasing-decreasing dynamic of the difference between preferred and orthogonal orientation responses. Among other data set, we analyze data that is particularly related to the work of Chavane *et al.* [32] where VSDi is used to understand the lateral spread of orientation selectivity by comparing responses evoked by local/full-field and center/surround stimuli. We also tackle the problem of the transient dynamic due to changes in motion direction as in the paper of Wu *et al.* [212] where it is shown that cortical dynamic combined with population coding is well suited to encode these changes. Let us remark that only a few previous works make use of supervised machine learning techniques to analyze VSDi signals [7, 8, 24]. One of the purposes of this dissertation is to design some machine learning based analyses of VSDi data.

Extracellular Recordings The ER technique is one of the oldest recording modalities. It consists in recording the electrical activity of cells in the small area surrounding an electrode. Today, it is possible to record multiple neurons at the same time over large and deep volumes of cortex [36, 47]. In ER the signal is twofold: the high-frequency component (400Hz to few thousands) which corresponds to spikes of single neurons and is known as multiunit activity (MUA), and the low-frequency component (cut off at about 300Hz) which represents an average activity of multiple neurons and is known as local field potential (LFP). In this dissertation, we focus on the high-frequency component corresponding to the neurons' spiking activity. The spiking activity of one neuron allows to compute their tuning curve and receptive field [87]. A tuning curve is the simple representation of the spiking activity as a function of one stimulation parameter. It should be understood as a sensitivity curve. A receptive field is the region of the visual field in which a stimulus modifies its firing rate. Different reverse correlation techniques exist to estimate receptive fields [98, 40]. Also known as Spike-Triggered Analysis (STA) and Spike-Triggered Covariance (STC), they constitute the standard methods. Let us also single out the paper of Park *et al.* [145] who perform a Bayesian estimation of receptive fields and provide impressive results. Based on these concepts, different neural network models are built. They generally consist in a first linear step followed by a non-linear thresholding step that is then converted to spiking activity by using a Poisson distribution [172]. Such models are also known as Linear/Non-linear Poisson spiking models (LNP). Few papers make use of standard supervised learning techniques. In particular, Hung *et al.* [90] use a kernel linear regression to classify the responses of IT neurons to different stimuli and make stimulus predictions knowing the neurons' response. More recently Yamins *et al.* [215] use a classifier to classify the responses of neurons and make predictions about the stimulus. Finally, to our knowledge, the paper of Goris *et al.*[71] is the only one to use MC-like stimulations in ER to study the origin of tuning diversity in the visual cortex. One of our goals is to perform some machine learning based analyses of ER data obtained under MC stimulations that we subsequently compare to analysis of synthetic data using a LNP model.

3 Contributions

3.1 Chapter I: A Model of Visual Stimulation

In Chapter I, we seek to reach a better understanding of human perception by improving generative models for dynamic texture synthesis. From that perspective, we motivate the generation of visual stimulation within a stationary Gaussian dynamic texture model I.2.1. We develop the proposed model by extending, mathematically detailing and robustly testing previously introduced dynamic noise textures [169, 178, 194] coined “Motion Clouds” (MC or MCs). Our main contribution is a complete axiomatic derivation of the model (see Section I.2). We detail three equivalent formulations of the Gaussian dynamic texture model. First, the composite dynamic textures are constructed by the random aggregation of warped patterns (“textons”), which can then be viewed as 3D Gaussian fields. Third, these textures are cast as solutions to a stochastic partial differential equation (sPDE). A second contribution is the mathematical proof of the equivalence between the two first formulations and a class of linear sPDEs (see Section I.3). This shows that our model is a generalization of the well-known luminance conservation Equation (2.2). This sPDE formulation has two chief advantages: it allows for a real-time synthesis using an AR recurrence and allows one to recast the log-likelihood of the model as a generalization of the classical motion energy model, which in turn is crucial to allow for Bayesian modeling of perceptual biases. Finally, we provide the source code[1] of this model of dynamic textures as an open source software development and reproducible research, which is crucial to advance the state of the art of real time stimulation for neurosciences. Some additional examples of MCs and texture synthesis from examples are available online[2].

3.2 Chapter II: Probabilistic and Bayesian Approach in Psychophysics

In Chapter II, first, we briefly review the Bayesian approaches in neurosciences and psychophysics. Then, we introduce the problem of “inverse Bayesian inverse”. We formalize the concept of observer’s internal representations in a probabilistic model. In this model, we are able to formulate the “Bayesian brain” hypothesis: our brain estimates external parameters as if they have generated their sensory representations. Then, we are able to define in mathematical terms the notion of psychometric curve obtained in a two-alternate forced choice (2AFC) experiment. This general definition, combined with our ideal Bayesian observer model appears intractable in absence of specific assumptions. We thus exemplify the psychometric curve by making simplifying assumptions on the likelihood and prior and we give numerical ex-

[1] `http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/motion_clouds`

[2] `https://jonathanvacher.github.io/mc_examples.html`

amples (see online[3]). Finally, we provide an algorithm which allows to solve the inverse Bayesian inference problem. The algorithm is illustrated in the following chapter.

3.3 Chapter III: Effect of Spatial Frequency over Speed Perception

In Chapter III, we present and analyze the result of psychophysical experiments that we performed to probe speed perception in humans using zoom-like changes in MCs spatial frequency content. We simplify the general Bayesian model developed in Chapter II by assuming a Gaussian likelihood and a Laplacian prior (see Section III.2.2). As the MC model allows for the derivation of a local motion-energy model, we use it to estimate speed in the experimental stimuli. We then validate the fitting process of the model using synthesized data in Section III.4.2. The human data replicates previous findings that relative perceived speed is positively biased by spatial frequency increments. By comparing the estimated variances of likelihoods to the distribution of the motion-energy model estimates of speed, we show that they are not compatible (see Section III.3). The effect cannot be fully accounted for by previous models, but the current prior acting on the spatio-temporal likelihoods has proved necessary in accounting for the perceptual bias (see Section III.5). We provide an online[4] example of data synthesis and analysis.

3.4 Chapter IV: Supervised Classification

Chapter IV is addressed to readers that are not familiar with supervised classification. From a mathematical point of view this Chapter provides very few contributions. We review in Section III.2 the different approaches to supervised classification. We give some useful and sometimes original examples to the different supervised classification approaches. In Section III.5, we give precise definitions of the different tools we use in the following chapters. In summary, this Chapter is closer to a graduate course in machine learning than to a contribution to research. However, within an interdisciplinary study, we feel it is necessary to set up the general problem of supervised classification and to introduce the different algorithms as particular cases of a common

[3] `http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/bayesian_observer/`

[4] `http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/bayesian_observer/`

framework before we apply them in Chapters V and VI. The goal is to introduce these tools to experimental neuroscientists and psychophysicists so they can properly understand the subsequent analysis we perform. Finally, having a better understanding of supervised data analysis will be useful for experimentalists to better craft their experiments. We provide the source code[5] of Examples 6 and 7 that illustrate Section 2.

3.5 Chapter V: Analysis of VSDi Data

Chapter V presents the visual system organization and its intracortical connectivity. Then, we review the VSDi technique and its processing. We also summarize the different machine learning approaches used to analyze functional Magnetic Resonance Imaging (fMRI) and VSDi data. The first major contribution of this chapter is an automatic method to select the number of components of the Principal Component Analysis (PCA) based on the classification performances (see Section V.4.1). The second main contribution is a methodology of local space-time analysis of classification performances, which enables to identify the most predictive pixels and to precisely quantify the temporal dynamic (see Section V.4.2). The third major contribution is the definition of a simple and efficient model of the VSDi signal obtained using oriented stimuli (see Section V.5). In addition, we make several minor biological contributions related to the experimental protocols that we analyze. In particular, we find that activation of neural populations is faster when stimulated after a blank than when stimulated after a first stimulus (see Section V.4.3.2). Moreover, the simple proposed model supports the role of lateral connections for a neural population to handle an abrupt change of stimulus orientation (see V.5.3). We provide an online[6] example of data synthesis using the proposed model. Moreover, additional Figures are also available online[7].

3.6 Chapter VI: Analysis of ER Data

Chapter VI introduces the Extracellular Recording (ER) technique, the standard processings that are applied to this signal and the few machine learning approaches that have been used to analyze this type of data. The first major contribution results from the use of MCs as stimuli. We find that small

[5] http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/examples_classif/

[6] http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/model_vsd/

[7] https://jonathanvacher.github.io/chapV-supp.html

neural populations (few dozens of neurons) contain enough information to discriminate between homogeneously and heterogeneously oriented stimuli. Such populations also contain enough information to discriminate between stimuli with narrow and broad spatial frequency bands (see Section VI.4). The second major contribution is a methodology of temporal analysis of prediction performances. We find that neural populations systematically have better classification performances than any single neurons when stimulated by MCs (see Section VI.5). However, when stimulated with natural movies, some neurons that provides classification performances that are similar to these of the entire population. The third major contribution is a simple Linear/Non-linear Poisson (LNP) spiking neurons model that generates synthetic data (see Section VI.6). When generated with MCs, the synthetic data provides results that are similar to the these obtained on the experimental recordings. We provide an online[8] example of data synthesis using the proposed model and MCs.

[8] http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/lnp_spiking_neurons/

⋆ I ⋆

Dynamic and Stochastic Models for Visual Stimulation

In this chapter, we give a mathematical description of a dynamic texture model specifically crafted to probe vision. It is first derived in a set of axiomatic steps constrained by biological plausibility. We then detail contributions by detailing three equivalent formulations of the Gaussian dynamic texture model. First, the composite dynamic textures are constructed by the random aggregation of warped patterns, which can be viewed as 3D Gaussian fields. Second, these textures are cast as solutions to a stochastic partial differential equation (sPDE). This essential step enables real time, on-the-fly, texture synthesis using time-discretized auto-regressive processes. Finally, we use the stochastic differential equation formulation from which the parameters are inferred from texture examples, in order to perform synthesis.

Contents

1 Introduction

1.1 Dynamic Texture Synthesis.

The model defined in (2.2) is quite simplistic with respect to the complexity of natural scenes. It is therefore useful here to discuss solutions to generative model problems previously proposed by texture synthesis methods in the computer vision and computer graphics community. Indeed, the literature on the subject of static textures synthesis is abundant (*eg* [208]). Of particular interest for us is the work of Galerne et al. [64, 63], which proposes a stationary Gaussian model restricted to static textures and provides an equivalent generative model based on Poisson shot noise. Realistic dynamic texture models are however less studied, and the most prominent method is the non-parametric Gaussian auto-regressive (AR) framework of Doretto [45], which has been thoroughly explored [213, 218, 35, 54, 91, 1]. These works generally consists in finding an appropriate low-dimensional feature space in which an AR process models the dynamics. Many of these approaches focus on the feature space where decomposition such as Singular Value Decomposition (SVD) and its Higher Order version (HOSVD) [45, 35] has shown their efficiency. In [1], the feature space is the Fourier frequency, and the AR recursion is carried over independently over each frequency, which defines the space-time stationary processes. A similar approach is used in [213] to compute the average of several dynamic texture models. Properties of these AR models are studied by Hyndman [91] where they found that higher order AR processes are able to capture perceptible temporal features. A different approach aims at learning the manifold structure of a given dynamic texture [110] while yet another deals with motion statistics [157]. All these works have in common the will to reproducing the natural spatio-temporal behavior of dynamic textures with rigorous mathematical tools. In addition, our concern is to design a dynamic texture model that is precisely parametrized for experimental purposes in visual neurosciences and psychophysics.

1.2 Stochastic Differential Equations (sODE and sPDE).

Stochastic Ordinary differential equation (sODE) and their higher dimensional counter-parts, stochastic partial differential equation (sPDE) can be viewed as continuous-time versions of these 1-D or higher dimensional auto-regressive (AR) models. Conversely, AR processes can therefore also be used to compute numerical solutions to these sPDE using finite difference approximations of time derivatives. Informally, these equations can be understood as partial differential equations perturbed by a random noise. The theoretical and numerical study of these sDE is of fundamental interest in fields as diverse as physics and chemistry [196], finance [49] or neuroscience [56]. They allow the dynamic study of complex, irregular and random phenomena such as particle interactions, stocks' or savings' prices, or ensembles of neurons. In psychophysics, sODE have been used to model decision making tasks in which the stochastic variable represents some accumulation of knowledge until the decision is taken, thus providing detailed information about predicted response times [181]. In imaging sciences, sPDE with sparse non-Gaussian driving noise has been proposed as model of natural signals and images [192]. As described above, the simple motion energy model (2.3) can similarly be demonstrated to rely on the sPDE (2.2) stochastic model of visual sensory input. This has not previously been presented in a formal way in the literature. One key goal of the current work is to comprehensively describe a parametric family of Gaussian sPDEs which generalize the modeling of moving images (and the corresponding synthesis of visual stimulation) and thus allow for a fine-grained systematic exploration of visual neurosciences and psychophysics.

1.3 Contributions

In this chapter, we attempt to engender a better understanding of human perception by improving generative models for dynamic texture synthesis. From that perspective, we motivate the generation of optimal visual stimulation within a stationary Gaussian dynamic texture model. We develop our current model by extending, mathematically detailing and robustly testing previously introduced dynamic noise textures [169, 178, 194] coined "Motion Clouds" (MC or MCs). Our first contribution is a complete axiomatic derivation of the model, seen as a shot noise aggregation of dynamically warped "textons". Within our generative model, the parameters correspond to average spatial and temporal transformations (*ie* zoom, orientation and translation speed) and associated standard deviations of random fluctuations, as illustrated in Figure 2.1, with respect to external (objects) and internal (observers) movements. A second contribution is the explicit demonstration of the equivalence between this model and a class of linear sPDEs. This shows that our model is a generalization of the well-known luminance conservation equation 2.2. This sPDE formulation has two chief advantages: it allows for a real-time synthesis using an AR recurrence and allows one to recast the log-likelihood of the model as a generalization of the classical motion energy model, which in turn is crucial to allow for Bayesian modeling of perceptual biases. Finally, we provide the source code[1] of this model of dynamic textures as an open source software development and reproducible research, which is crucial to advance the state of the art of real time stimulation for neurosciences. Some additional examples of MCs and texture synthesis from examples are available online[2].

2 Axiomatic Construction of the Dynamic Textures

Efficient dynamic textures to probe visual perception should be naturalistic yet low-dimensional parametric stochastic models. They should embed meaningful physical parameters (such as the effect of head rotations or whole-field

[1] `http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/motion_clouds`

[2] `https://jonathanvacher.github.io/mc_examples.html`

scene movements, see Figure 2.1) into the local or global dependencies (e.g. the covariance) of the random field. In the luminance conservation model (2.2), the generative model is parameterized by a spatio-temporal coupling encoded in the covariance Σ of the driving noise and the motion flow v_0. This localized space-time coupling (e.g. the covariance if one restricts its attention to Gaussian fields) is essential as it quantifies the extent of the spatial integration area as well as the integration dynamics. This is an important issue in neuroscience when considering the implementation of spatio-temporal integration mechanisms from very small to very large scales i.e. going from local to global visual features [162, 17, 42]. In particular, this is crucial to understand the modular sensitivity within the different lower visual areas. In primates for instance, this is evident in the range of spatio-temporal scales of selectivity for generally smaller features observed in the Primary Visual Cortex (V1) and in contrast, ascending the processing hierarchy, for larger features in Middle Temple area (MT). By varying the frequency bandwidth of such dynamic textures, distinct mechanisms for perception and action have been identified in humans [178]. Our goal here is to develop a principled, axiomatic definition of these dynamic textures.

2.1 From Shot Noise to Motion Clouds

We propose a mathematically-sound derivation of a general parametric model of dynamic textures. This model is defined by aggregation, through summation, of a basic spatial "texton" template $g(x)$. The summation reflects a transparency hypothesis, which has been adopted for instance in [64]. While one could argue that this hypothesis is overly simplistic and does not model occlusions or edges, it leads to a tractable framework of stationary Gaussian textures, which has proved useful to model static micro-textures [64] and dynamic natural phenomena [213]. The simplicity of this framework allows for a fine tuning of frequency-based (Fourier) parameterization, which is desirable for the interpretation of psychophysical experiments with respect to underlying spatio-temporal neural sensitivity.

We define a random field as

$$I_\lambda(x,t) \overset{\text{def.}}{=} \frac{1}{\sqrt{\lambda}} \sum_{p\in\mathbb{N}} g(\varphi_{A_p}(x - X_p - V_p t)) \tag{2.1}$$

where $\varphi_a : \mathbb{R}^2 \to \mathbb{R}^2$ is a planar warping parameterized by a finite dimensional vector a. The parameters $(X_p, V_p, A_p)_{p\in\mathbb{N}}$ are independent and identically distributed random vectors. They account for the variability in the position of objects or observers and their speed, thus mimicking natural motions in an

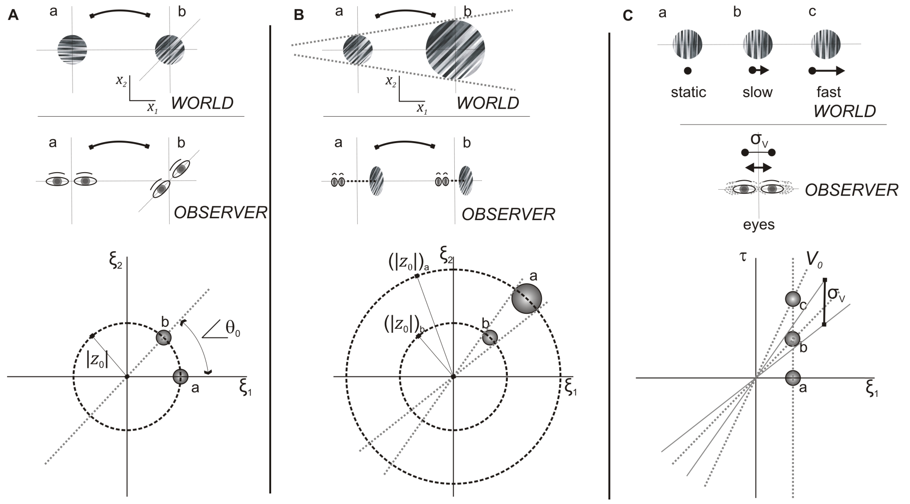


Figure 2.1: *Parameterization of the class of Motion Clouds stimuli.* The illustration relates the parametric changes in MC with real world (top row) and observer (second row) movements. **(A)** Orientation changes resulting in scene rotation are parameterized through θ as shown in the bottom row where a horizontal a and obliquely oriented b MC are compared. **(B)** Zoom movements, either from scene looming or observer movements in depth, are characterized by scale changes reflected by a scale or frequency term z shown for a larger or closer object b compared to more distant a. **(C)** Translational movements in the scene characterized by V using the same formulation for static (a) slow (b) and fast moving MC, with the variability in these speeds quantified by σ_V. (ξ and τ) in the third row are the spatial and temporal frequency scale parameters. The development of this formulation is detailed in the text.

ambient scene. The set of translations $(X_p)_{p\in\mathbb{N}}$ is a 2-D Poisson point process of intensity $\lambda > 0$. This means that, defining for any measurable A, $C(A) = \sharp\{p\ ;\ X_p \in A\}$, one has that $C(A)$ has a Poisson distribution with mean $\lambda|A|$ (where $|A|$ is the measure of A) and $C(A)$ is independent of $C(B)$ if $A \cap B = \emptyset$.

Intuitively, this model (2.1) corresponds to a dense mixing of stereotyped, static, textons as in [64]. The originality is two-fold. First, the components of this mixing are derived from the texton by visual transformations φ_{A_p} which may correspond to arbitrary transformations such as zooms or rotations, illustrated in Figure 1. Second, we explicitly model the motion (position X_p and speed V_p) of each individual texton.

In the following, we denote $\mathbb{P}_A$ the common distribution of the i.i.d. $(A_p)_p$, and we denote $\mathbb{P}_V$ the distribution in $\mathbb{R}^2$ of the speed vectors $(V_p)_p$. Section 2.3 instantiates this model and proposes canonical choices for these variabilities.

The following result shows that the model (2.1) converges for high point density $\lambda \to +\infty$ to a stationary Gaussian field and gives the parameterization of the covariance. Its proof follows from a specialization of [63, Theorem 3.1] to our setting.

Proposition 1. *I_λ is stationary with bounded second order moments. Its covariance is $\Sigma(x, t, x', t') = \gamma(x - x', t - t')$ where γ satisfies*

$$\forall\, (x, t) \in \mathbb{R}^3, \quad \gamma(x, t) = \iiint_{\mathbb{R}^2} c_g(\varphi_a(x - \nu t))\mathbb{P}_V(\nu)\mathbb{P}_A(a)\mathrm{d}\nu\mathrm{d}a \tag{2.2}$$

where $c_g = g \star \bar{g}$ is the auto-correlation of g. When $\lambda \to +\infty$, it converges (in the sense of finite dimensional distributions) toward a stationary Gaussian field I of zero mean and covariance Σ.

This proposition enables us to give a precise definition of a MC.

Definition 1. *A Motion Cloud (MC) is a stationary Gaussian field whose covariance is given by* (2.2)*.*

Note that, following [65], the convergence result of Proposition 1 could be used in practice to simulate a Motion Cloud I using a high but finite value of λ in order to generate a realization of I_λ. We do not use this approach, and rather rely on the sPDE characterization proved in Section 3, which is well tailored for an accurate and computationally efficient dynamic synthesis.

2.2 Toward “Motion Clouds” for Experimental Purposes

The previous Section provides a theoretical definition of MC 1 that is characterized by $c_g, \varphi_a, \mathbb{P}_A$ and $\mathbb{P}_V$. The high dimension of these parameters has to be reduced for experimental purposes, therefore it is essential to specify these parameters to have a better control of the covariance γ. We further study this model in the specific case where the warpings φ_a are rotations and scalings (see Figure 2.1). They account for the characteristic orientations and sizes (or spatial scales) in a scene with respect to the observer. We thus set

$$\forall\, a = (\theta, z) \in [-\pi, \pi) \times \mathbb{R}_+^*, \quad \varphi_a(x) \overset{\text{def.}}{=} zR_{-\theta}(x),$$

where R_θ is the planar rotation of angle θ. We now give some physical and biological motivation underlying our particular choice for the distributions of the parameters. We assume that the distributions $\mathbb{P}_Z$ and $\mathbb{P}_\Theta$ of spatial scales z and orientations θ, respectively (see Figure 1), are independent and have densities, thus considering

$$\forall\, a = (\theta, z) \in [-\pi, \pi) \times \mathbb{R}_+^*, \quad \mathbb{P}_A(a) = \mathbb{P}_Z(z)\,\mathbb{P}_\Theta(\theta).$$

The speed vector ν is assumed to be randomly fluctuating around a central speed $v_0 \in \mathbb{R}^2$, so that

$$\forall\, \nu \in \mathbb{R}^2, \quad \mathbb{P}_V(\nu) = \mathbb{P}_{\|V-v_0\|}(\|\nu - v_0\|). \tag{2.3}$$

In order to obtain "optimal" responses to the stimulation (as advocated by [217]) and based on the structure of a standard receptive field of V1, it makes sense to define the texton to be equal to an oriented Gabor which acts as the generic atom

$$g_\sigma(x) = \frac{1}{2\pi} \cos\left(\langle x,\, \xi_0\rangle\right) e^{-\frac{\sigma^2}{2}\|x\|^2} \tag{2.4}$$

where σ is the inverse standard deviation and $\xi_0 \in \mathbb{R}^2$ is the spatial frequency. Since the orientation and scale of the texton is handled by the (θ, z) parameters, we can impose without loss of generality the normalization $\xi_0 = (1, 0)$. In the special case where $\sigma \to 0$, g_σ is a grating of frequency ξ_0, and the image I is a dense mixture of drifting gratings, whose power-spectrum has a closed form expression detailed in Proposition 2. It is fully parameterized by the distributions $(\mathbb{P}_Z, \mathbb{P}_\Theta, \mathbb{P}_V)$ and the central frequency and speed (ξ_0, v_0). Note that it is possible to consider any arbitrary textons g, which would give rise to more complicated parameterizations for the power spectrum $\hat{g}$, but we decided here to stick to the simple asymptotic case of gratings.

Proposition 2. *Consider the texton g_σ , when $\sigma \to 0$, the Gaussian field $I_\sigma(x,t)$ defined in Proposition 1 converges toward a stationary Gaussian field of covariance having the power-spectrum*

$$\forall\,(\xi, \tau) \in \mathbb{R}^2 \times \mathbb{R},\ \hat{\gamma}(\xi, \tau) = \frac{\mathbb{P}_Z\left(\|\xi\|\right)}{\|\xi\|^2} \mathbb{P}_\Theta\left(\angle\xi\right) \mathcal{L}(\mathbb{P}_{\|V-v_0\|}) \left(-\frac{\tau + \langle v_0,\, \xi\rangle}{\|\xi\|}\right), \tag{2.5}$$

where the linear transform $\mathcal{L}$ is such that

$$\forall\, u \in \mathbb{R}, \quad \mathcal{L}(f)(u) \overset{\text{def.}}{=} \int_{-\pi}^{\pi} f(-u/\cos(\varphi))\mathrm{d}\varphi.$$

and $\xi = (\|\xi\| \cos(\angle\xi), \|\xi\| \sin(\angle\xi))$.

Proof. We recall the expression (2.2) of the covariance

$$\forall\,(x,t) \in \mathbb{R}^3, \quad \gamma(x,t) = \iiint_{\mathbb{R}^2} c_{g_\sigma}(\varphi_a(x-\nu t))\mathbb{P}_V(\nu)\mathbb{P}_A(a)\mathrm{d}\nu\mathrm{d}a \qquad (2.6)$$

We denote $(\theta,\varphi,z,r) \in \Gamma = [-\pi,\pi)^2 \times \mathbb{R}_+^2$ the set of parameters. Denoting $h(x,t) = c_{g_\sigma}(zR_\theta(x-\nu t))$, one has, in the sense of distributions (taking the Fourier transform with respect to (x,t))

$$\hat{h}(\xi,\tau) = z^{-2}\hat{g}_\sigma(z^{-1}R_\theta(\xi))^2\delta_{\mathcal{Q}}(\nu) \quad \text{where} \quad \mathcal{Q} = \left\{\nu \in \mathbb{R}^2\ ;\ \tau + \langle \xi,\, \nu\rangle = 0\right\}.$$

Taking the Fourier transform of (2.6) and using this computation, the result is that $\hat{\gamma}(\xi,\tau)$ is equal to

$$\int_\Gamma \frac{|\hat{g}_\sigma\,(z^{-1}R_\theta(\xi))\,|^2}{z^2}\delta_{\mathcal{Q}}(v_0 + r(\cos(\varphi),\sin(\varphi)))\mathbb{P}_\Theta(\theta)\mathbb{P}_Z(z)\mathbb{P}_{\|V-v_0\|}(r)\ \mathrm{d}\theta\mathrm{d}z\mathrm{d}r\mathrm{d}\varphi.$$

Therefore when $\sigma \to 0$, one has in the sense of distributions

$$|\hat{g}_\sigma\left(z^{-1}R_\theta(\xi)\right)|^2 \to \delta_{\mathcal{B}}(\theta,z) \quad \text{where} \quad \mathcal{B} = \left\{(\theta,z)\ ;\ z^{-1}R_\theta(\xi) = \xi_0\right\}.$$

Observing that $\delta_{\mathcal{Q}}(\nu)\delta_{\mathcal{B}}(\theta,z) = \delta_{\mathcal{C}}(\theta,z,r)$ where

$$\mathcal{C} = \left\{(\theta,z,r)\ ;\ z = \|\xi\|,\ \theta = \angle\xi,\ r = -\frac{\tau}{\|\xi\|\cos(\angle\xi-\varphi)} - \frac{\|v_0\|\cos(\angle\xi - \angle v_0)}{\cos(\angle\xi-\varphi)}\right\}$$

one obtains the desired formula. □

Remark 1. *Note that the envelope of $\hat{\gamma}$ as defined in* (2.5) *is constrained to lie within a cone in the spatio-temporal domain with the apex at zero. This is an important and novel contribution when compared to a classical Gabor. In particular, the bandwidth is then constant around the speed plane or the orientation line with respect to spatial frequency. Basing the generation of the textures on all possible translations, rotations and zooms, we thus provide a principled approach to show that bandwidth should be parametrically scaled with spatial frequency to provide a better model of moving textures.*

2.3 Biologically-inspired Parameter Distributions

We now give meaningful specialization for the probability distributions $\mathbb{P}_Z$, $\mathbb{P}_\Theta$, and $\mathbb{P}_{\|V-v_0\|}$, which are inspired by some known scaling properties of the visual transformations relevant to dynamic scene perception.

Parameterization of $\mathbb{P}_Z$. First, small, centered, linear movements of the observer along the axis of view (orthogonal to the plane of the scene) generate centered planar zooms of the image. From the linear modeling of the observer's displacement and the subsequent multiplicative nature of zoom, scaling should follow a Weber-Fechner law stating that subjective sensation when quantified is proportional to the logarithm of stimulus intensity. Thus, we choose the scaling z drawn from a log-normal distribution $\mathbb{P}_Z$, defined in (2.7). The bandwidth σ_Z quantifies the variance in the amplitude of zooms of individual textons relative to the characteristic scale z_0. We thus define

$$\mathbb{P}_Z(z) \propto \frac{\tilde{z}_0}{z} \exp\left(- \frac{\ln\left(\frac{z}{\tilde{z}_0}\right)^2}{2\ln\left(1+\tilde{\sigma}_Z^2\right)} \right), \tag{2.7}$$

where $\propto$ means that we ignored the normalizing constant.

In practice, the parameters $(\tilde{z}_0, \tilde{\sigma}_Z)$ are not convenient to manipulate because they have no "physical meaning". Instead, we use another, more intuitive, parametrization using mode and variance (z_0, σ_Z)

$$z_0 \overset{\text{def.}}{=} \operatorname{argmax}_z \mathbb{P}_Z(z) \quad \text{and} \quad \sigma_Z^2 \overset{\text{def.}}{=} \mathbb{E}(Z^2) - \mathbb{E}(Z)^2.$$

Once (z_0, σ_Z) are fixed, it is easy to compute the corresponding $(\tilde{z}_0, \tilde{\sigma}_Z)$ to plug into expression (2.7), simply by solving a polynomial equation (2.8), as detailed in the following proposition.

Proposition 3. *One has*

$$z_0 = \frac{\tilde{z}_0}{1+\tilde{\sigma}_Z^2} \quad \textit{and} \quad \sigma_Z^2 = \tilde{z}_0 \tilde{\sigma}_Z^2 (1+\tilde{\sigma}_Z^2).$$

Such formula can be inverted by finding the unique positive root of

$$\tilde{\sigma}_Z^2(1+\tilde{\sigma}_Z^2)^2 - \frac{\sigma_Z^2}{z_0} = 0 \quad \textit{and} \quad \tilde{z}_0 = z_0(1+\tilde{\sigma}_Z^2). \tag{2.8}$$

Proof. The primary relations are established using standard calculations from the probability density function $\mathbb{P}_Z$ [97]. The relations (2.8) follow standard arithmetic.

□

Parametrization of $\mathbb{P}_Z$ by mode and octave bandwidth Differences in perception are often more relevant in a log domain, therefore it is useful to parametrize $\mathbb{P}_Z$ by its mode z_0 and octave bandwidth B_Z which is defined by

$$B_Z \overset{\text{def.}}{=} \frac{\ln\left(\frac{z_+}{z_-}\right)}{\ln(2)}$$

where (z_-, z_+) are respectively the successive half-power cutoff frequencies, that is, which verify $\mathbb{P}_Z(z_-) = \mathbb{P}_Z(z_+) = \frac{\mathbb{P}_Z(z_0)}{2}$ with $z_- \leq z_+$.

Proposition 4. *One has*

$$B_Z = \sqrt{\frac{8\ln(1+\tilde{\sigma}_Z^2)}{\ln(2)}} \quad \textit{and conversely} \quad \tilde{\sigma}_Z = \sqrt{\exp\left(\frac{\ln(2)}{8}B_Z^2\right) - 1}. \tag{2.9}$$

Proof. Using the fact that $\mathbb{P}_Z(z_-) = \mathbb{P}_Z(z_+) = \frac{\mathbb{P}_Z(z_0)}{2}$, one shows that $X_+ = \ln\left(\frac{z_+}{z_0}\right)$ and $X_- = \ln\left(\frac{z_-}{z_0}\right)$ are the two roots of the following polynomial (with $X_- \leq X_+$).

$$Q(X) = X^2 + 2\ln(1+\tilde{\sigma}_Z^2)X - 2\ln(2)\ln(1+\tilde{\sigma}_Z^2) + \frac{1}{2}\ln(1+\tilde{\sigma}_Z^2)^2$$

This allows to compute B_Z. □

Through Proposition 4 it is possible to obtain the parametrization of bandwidth prevalent in manipulations used in visual psychophysics experiments.

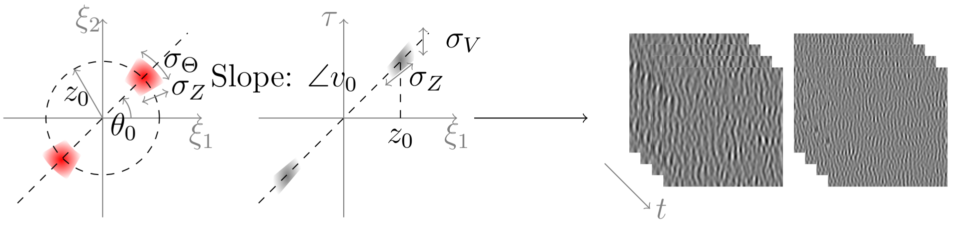
Parameterization of $\mathbb{P}_\Theta$. Similarly, the texture is perturbed by variations in the global angle θ of the scene: for instance, the head of the observer may roll slightly around its normal position. The von-Mises distribution – as a good approximation of the warped Gaussian distribution around the unit circle – is an adapted choice for the distribution of θ with mean θ_0 and bandwidth σ_Θ,

$$\mathbb{P}_\Theta(\theta) \propto e^{\frac{\cos(2(\theta-\theta_0))}{4\sigma_\Theta^2}} \tag{2.10}$$

Parameterization of $\mathbb{P}_{\|V-v_0\|}$. We may similarly consider that the position of the observer is variable in time. On first order approximation, movements perpendicular to the axis of view dominate, generating random perturbations to the global translation v_0 of the image at speed $\nu - v_0 \in \mathbb{R}^2$. These perturbations are for instance described by a Gaussian random walk: take for instance tremors, which are constantly jittering, small ($\leqslant$ 1 deg) movements of the eye. This justifies the choice of a radial distribution (2.3) for $\mathbb{P}_V$. This radial distribution $\mathbb{P}_{\|V-v_0\|}$ is thus selected as a bell-shaped function of width σ_V, and we choose here a Gaussian function for simplicity

$$\mathbb{P}_{\|V-v_0\|}(r) \propto e^{-\frac{r^2}{2\sigma_V^2}}. \tag{2.11}$$

Note that, as detailed in Section 3.2 a slightly different bell-function (with a more complicated expression) should be used to obtain an exact equivalence with the sPDE discretization.



Two different projections of $\|\xi\|^2\hat{\gamma}(\xi,\tau)$ in Fourier space

MC of two different spatial frequencies z_0

Figure 2.2: Graphical representation of the covariance γ (left) — note the cone-like shape of the envelopes– and an example of synthesized dynamics for narrow-band and broad-band Motion Clouds (right).

Putting everything together. Plugging these expressions (2.7), (2.10) and (2.11) into the definition (2.5) of the power spectrum of the motion cloud, one obtains a parameterization which shares similarities with the one originally introduced in [178].

The following table recaps the parameters of the biologically-inpired MC models. It is composed of the central parameters (v_0) for the speed, (θ_0) for orientation and (z_0) for the frequency modulus, as well as corresponding "dispersion" parameters ($\sigma_V, \sigma_\Theta, B_Z$) which account for the typical deviation around these centers.

	Speed	Freq. orient.	Freq. amplitude
(mean, dispersion)	(v_0, σ_V)	$(\theta_0, \sigma_\Theta)$	(z_0, B_Z)

Figure 2.2 shows graphically the influence of these parameters on the shape of the MC power spectrum $\hat{\gamma}$.

We show in Figure 2.3 two examples of such stimuli for different spatial frequency bandwidths. In particular, by tuning this bandwidth, in previous studies it has been possible to dissociate its respective role in action and perception [178]. Using this formulation to extend the study of visual perception to other dimensions, such as orientation or speed bandwidths, should provide a means to systematically titrate their respective role in motion integration and obtain essential novel data.

3 sPDE Formulation and Synthesis Algorithm

In this section, we show that the MC model (Definition 1) can equally be described as the stationary solution of a stochastic partial differential equation (sPDE). This sPDE formulation is important since we aim to deal with

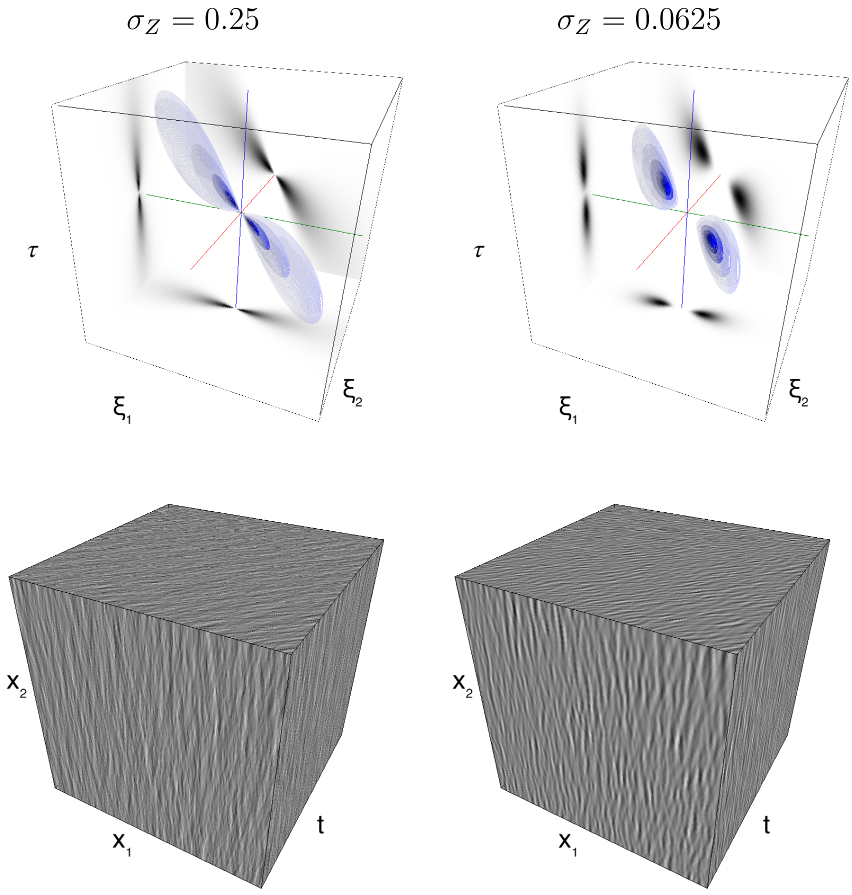


Figure 2.3: Comparison of the broadband (left) vs. a narrowband (right) stimulus. Two instances (left and right columns) of two motions clouds having the same parameters except the frequency bandwidths σ_Z, which were different. The top column displays iso-surfaces of $\hat{\gamma}$ in the form of enclosing volumes at different energy values with respect to the peak amplitude of the Fourier spectrum. The bottom column shows an isometric view of the faces of a movie cube, which is a realization of the random field I. The first frame of the movie lies on the $(x_1, x_2, t = 0)$ spatial plane. The Motion Cloud with the broadest bandwidth is often thought to best represent stereotyped natural stimuli, since, it similarly contains a broad range of frequency components.

dynamic stimulation, which should be described by a causal equation which is local in time. This is crucial for numerical simulations, since this allows us to perform real-time synthesis of stimuli using an auto-regressive time discretization. This is a significant departure from previous Fourier-based implementation of dynamic stimulation [169, 178]. Moreover, this is also important to simplify the application of MC inside a Bayesian model of psychophysical experiments (see Chapters II and III). In particular, the derivation of an equivalent sPDE model exploits a spectral formulation of MCs as Gaussian Random fields. The full proof along with the synthesis algorithm follows.

To be mathematically correct, all the sPDE in this article are written in the sense of generalized stochastic processes (GSP) which are to stochastic processes what generalized functions are to functions. This allows the consideration of linear transformations of stochastic processes like differentiation or Fourier transforms as for generalized functions. We refer to [193] for a recent use of GSP and to [69] for the foundation of the theory. The connection between GSP and stochastic processes has been described by previous work [122]

3.1 Dynamic Textures as Solutions of sPDE

In the following, we first restrict our attention to the case $v_0 = 0$ in order to define a simple sPDE, and then detail the general case.

sPDE without global translation, $v_0 = 0$. We first give the definition of a sPDE cloud I making use of another cloud I_0 without translation speed.

Definition 2. *For a given stationary spatial covariance σ_w, 2-D spatial filters (α, β) and a translation speed $v_0 \in \mathbb{R}^2$, a sPDE cloud is defined as*

$$I(x,t) \overset{\text{def.}}{=} I_0(x - v_0 t, t). \tag{3.1}$$

where I_0 is a stationary Gaussian field satisfying for all (x,t)

$$\mathcal{D}(I_0) = \frac{\partial W}{\partial t} \quad \textit{where} \quad \mathcal{D}(I_0) \overset{\text{def.}}{=} \frac{\partial^2 I_0}{\partial t^2} + \alpha \star \frac{\partial I_0}{\partial t} + \beta \star I_0 \tag{3.2}$$

where the driving noise $\frac{\partial W}{\partial t}$ is white in time (i.e. corresponding to the temporal derivative of a Brownian motion in time) and has the spatial stationary covariance σ_W and $\star$ is the spatial convolution operator.

The random field I_0 solving (3.2) thus corresponds to a sPDE cloud with no translation speed, $v_0 = 0$. The filters (α, β) parameterizing this sPDE cloud aim at enforcing an additional correlation in time of the model. Section 3.2 explains how to choose $(\alpha, \beta, \sigma_W)$ so that these sPDE clouds, which are stationary solutions of (3.2), have the power spectrum given in (2.5) (in the case that $v_0 = 0$), i.e. are motion clouds.

Defining a causal equation that is local in time is crucial for numerical simulation (as explained in Section 3.3) but also to simplify the application of MC inside a Bayesian model of psychophysical experiments (see Section 3).

The sPDE equation (3.2) corresponds to a set of independent stochastic ODEs over the spatial Fourier domain, which reads, for each frequency ξ,

$$\forall\, t \in \mathbb{R}, \quad \frac{\partial^2 \hat{I}_0(\xi,t)}{\partial t^2} + \hat{\alpha}(\xi)\frac{\partial \hat{I}_0(\xi,t)}{\partial t} + \hat{\beta}(\xi)\hat{I}_0(\xi,t) = \hat{\sigma}_W(\xi)\frac{\hat{W}(\xi,t)}{\partial t} \tag{3.3}$$

where $\hat{I}_0(\xi,t)$ denotes the Fourier transform with respect to the spatial variable x only. The Fourier transform of the stationary spatial covariance $\hat{\sigma}_W(\xi)^2$ is the spatial power spectrum of $\frac{\partial W}{\partial t}$ and $\hat{W}(\xi,t+\delta t)-\hat{W}(\xi,t) \sim \mathcal{CN}(0,\delta t)$ where $\mathcal{CN}(0,\delta t)$ denotes the complex normal distribution of variance δt *ie* $\hat{W}(\xi,t+\delta t)-\hat{W}(\xi,t)$ is a white noise in space and time. While the equation (3.3) should hold for all time $t \in \mathbb{R}$, the construction of stationary solutions (hence sPDE clouds) of this equation is obtained by solving the sODE (3.3) forward for time $t > t_0$ with arbitrary boundary conditions at time $t = t_0$, and letting $t_0 \to -\infty$. This is consistent with the numerical scheme detailed in Section 3.3.

While it is beyond the scope of this paper to study theoretically the equation (3.2), one can show the existence and uniqueness results of stationary solutions for this class of sPDE under stability conditions on the filers (α,β) (see for instance [192, 25]) that are automatically satisfied for the particular case of Section 3.2.

Theorem 1. *If $(\hat{\alpha},\hat{\beta})$ are non-negative and $\frac{\hat{\sigma}_W^2}{\hat{\alpha}\hat{\beta}} \in L^1$, then Equation (3.2) has a unique causal and stationary solution, i.e. it defines uniquely the distribution of a sPDE cloud.*

Proof. Consider (3.3), the Fourier transform of (3.2) which has causal and stationary solutions (see the general case of Levy-driven sPDE, Theorem 3.3 in [25]). Hence $\frac{\hat{\sigma}_W}{\hat{\alpha}\hat{\beta}} \in L^1$, these solutions have an integrable spatial power spectrum. Then, one could take their inverse Fourier transform and get the solution which is unique by construction. □

Remark 2. *There are different ways to define uniqueness of solution for sPDE. In Theorem 1, uniqueness has to be understood in terms of sample path, meaning that two solutions $(X,\tilde{X})$ of Equation (3.2) verifies $\mathbb{P}(\forall t \in \mathbb{R}, \quad X_t = \tilde{X}_t) = 1$. This notion of uniqueness is strong and it implies uniqueness in distribution meaning that X and $\tilde{X}$ have the same law.*

sPDE with global translation. The easiest way to define and synthesize a sPDE cloud I with non-zero translation speed v_0 is to first define I_0

solving (3.3) and then translating it with constant speed using (3.1). An alternative way is to derive the sPDE satisfied by I, as detailed in the following proposition. This is useful to define motion energy in Section 3.

Proposition 5. *The MCs noted I with parameters $(\alpha, \beta, \sigma_w)$ and translation speed v_0 are the stationary solutions of the sPDE*

$$\mathcal{D}(I) + \langle \mathcal{G}(I),\, v_0 \rangle + \langle \mathcal{H}(I) v_0,\, v_0 \rangle = \frac{\partial W}{\partial t} \tag{3.4}$$

where $\mathcal{D}$ is defined in (3.2)*, $\partial_x^2 I$ is the Hessian of I (second order spatial derivative), where*

$$\mathcal{G}(I) \overset{\text{def.}}{=} \alpha \star \nabla_x I + 2\partial_t \nabla_x I \quad \textit{and} \quad \mathcal{H}(I) \overset{\text{def.}}{=} \nabla_x^2 I. \tag{3.5}$$

Proof. This follows by computing the derivative in time of the warping equation (3.1), denoting $y \overset{\text{def.}}{=} x + v_0 t$

$$\begin{aligned}
&\partial_t I_0(x,t) = \partial_t I(y,t) + \langle \nabla I(y,t),\, v_0 \rangle, \\
&\partial_t^2 I_0(x,t) = \partial_t^2 I(y,t) + 2\langle \partial_t \nabla I(y,t),\, v_0 \rangle + \langle \partial_x^2 I(y,t) v_0,\, v_0 \rangle
\end{aligned}$$

and plugging this into (3.2) after remarking that the distribution of $\frac{\partial W}{\partial t}(x,t)$ is the same as the distribution of $\frac{\partial W}{\partial t}(x - v_0 t, t)$. □

3.2 Equivalence between the spectral and sPDE formulations

Since both MCs and sPDE clouds are obtained by a uniform translation with speed v_0 of a motionless cloud, we can restrict without loss of generality our analysis to the case $v_0 = 0$.

In order to relate MCs to sPDE clouds, equation (3.3) makes explicit that the functions $(\hat{\alpha}(\xi), \hat{\beta}(\xi))$ should be chosen in order for the temporal covariance of the resulting process to be equal (or at least to approximate well) the temporal covariance appearing in (2.5). This covariance should be localized around 0 and be non-oscillating. It thus makes sense to constrain $(\hat{\alpha}(\xi), \hat{\beta}(\xi))$ for the corresponding ODE (3.3) to be critically damped, which corresponds to imposing the following relationship

$$\forall\, \xi, \quad \hat{\alpha}(\xi) = \frac{2}{\hat{\nu}(\xi)} \quad \text{and} \quad \hat{\beta}(\xi) = \frac{1}{\hat{\nu}^2(\xi)}$$

for some relaxation step size $\hat{\nu}(\xi)$. The model is thus solely parameterized by the noise variance $\hat{\sigma}_W(\xi)$ and the characteristic time $\hat{\nu}(\xi)$.

The following proposition shows that the sPDE cloud model (3.2) and the motion cloud model (2.5) are identical for an appropriate choice of function $\mathbb{P}_{\|V-v_0\|}$.

Proposition 6. *When considering*

$$\forall\, r > 0, \quad \mathbb{P}_{\|V-v_0\|}(r) = \mathcal{L}^{-1}(h)(r/\sigma_V) \quad \textit{where} \quad h(u) = (1+u^2)^{-2} \tag{3.6}$$

where $\mathcal{L}$ is defined in (2.5)*, equation* (3.2) *admits a solution I which is a stationary Gaussian field with power spectrum* (2.5) *when setting*

$$\hat{\sigma}_W^2(\xi) = \frac{4}{\hat{\nu}(\xi)^3\|\xi\|^2}\mathbb{P}_Z(\|\xi\|)\mathbb{P}_\Theta(\angle\xi), \quad \textit{and} \quad \hat{\nu}(\xi) = \frac{1}{\sigma_V\|\xi\|}. \tag{3.7}$$

Proof. For this proof, we denote I^{MC} the motion cloud defined by (2.5), and I a stationary solution of the sPDE defined by (3.2) which exists according to Theorem 1 because $\hat{\sigma}_W^2\hat{\nu}^3 \in L^1$, indeed $\mathbb{P}_Z$ and $\mathbb{P}_\Theta$ are probability distributions and $\xi \mapsto \frac{1}{\|\xi\|^2}$ does not change the continuity at 0. We aim to show that under the specification (3.7), they have the same covariance. This is equivalent to showing that $I_0^{\text{MC}}(x,t) = I^{\text{MC}}(x+ct,t)$ has the same covariance as I_0. For any fixed ξ, equation (3.3) admits a unique stationary solution $\hat{I}_0(\xi,\cdot)$ (Theorem 1) which is a stationary Gaussian process of zero mean and with a covariance which is $\hat{\sigma}_W^2(\xi) r\star\bar{r}$ where r is the impulse response (i.e. taking formally $a=\delta$) of the ODE $r'' + 2r'/u + r/u^2 = a$ where we denoted $u = \hat{\nu}(\xi)$. This impulse response can be shown to be $r(t) = te^{-t/u}\mathbb{1}_{\mathbb{R}^+}(t)$. The covariance of $\hat{I}_0(\xi,\cdot)$ is thus, after some computation, equal to $\hat{\sigma}_W^2(\xi) r\star\bar{r} = \hat{\sigma}_W^2(\xi)h(\cdot/u)$ where $h(t) = \frac{u^3}{4}(1+|t|)e^{-|t|}$. Taking the Fourier transform of this equality, the power spectrum $\hat{\gamma}_0$ of I_0 thus reads

$$\hat{\gamma}_0(\xi,\tau) = \frac{1}{4}\hat{\sigma}_W^2(\xi)\hat{\nu}(\xi)^3\tilde{h}(\hat{\nu}(\xi)\tau) \quad \text{where} \quad \tilde{h}(s) = \frac{1}{(1+s^2)^2} \tag{3.8}$$

and where it should be noted that this function h is the same as the one introduced in (3.6). The covariance γ^{MC} of I^{MC} and γ_0^{MC} of I_0^{MC} are related by the relation

$$\hat{\gamma}_0^{\text{MC}}(\xi,\tau) = \hat{\gamma}^{\text{MC}}(\xi,\tau - \langle \xi,\, v_0\rangle) = \frac{1}{\|\xi\|^2}\mathbb{P}_Z(\|\xi\|)\mathbb{P}_\Theta\left(\angle\xi\right)\hat{h}\left(-\frac{\tau}{\sigma_V\|\xi\|}\right). \tag{3.9}$$

where we used the expression (2.5) for $\hat{\gamma}^{\text{MC}}$ and the value of $\mathcal{L}(\mathbb{P}_{\|V-v_0\|})$ given by (3.6). Condition (3.7) guarantees that expression (3.8) and (3.9) coincide, and thus $\hat{\gamma}_0 = \hat{\gamma}_0^{\text{MC}}$. □

Expression for $\mathbb{P}_{\|V-v_0\|}$. Equation (3.6) states that in order to obtain a perfect equivalence between the MC defined by (2.5) and by (3.2), the function $\mathcal{L}^{-1}(h)$ has to be well-defined. It means we need to compute the inverse of the transform of the linear operator $\mathcal{L}$

$$\forall u \in \mathbb{R}, \quad \mathcal{L}(f)(u) = 2\int_0^{\pi/2} f(-u/\cos(\varphi))\mathrm{d}\varphi.$$

to the function h. The following proposition gives a closed-form expression for this function, and shows in particular that it is a function in $L^1(\mathbb{R})$, i.e. it has a finite integral, which can be normalized to unity to define a density distribution. Figure 3.1 shows a graphical display of that distribution.

Proposition 7. *One has*

$$\mathcal{L}^{-1}(h)(u) = \frac{2-u^2}{\pi(1+u^2)^2} - \frac{u^2(u^2+4)(\log(u)-\log(\sqrt{u^2+1}+1))}{\pi(u^2+1)^{5/2}}.$$

In particular, one has

$$\mathcal{L}^{-1}(h)(0) = \frac{2}{\pi} \quad \textit{and} \quad \mathcal{L}^{-1}(h)(u) \sim \frac{1}{2\pi u^3} \quad \textit{when} \quad u \to +\infty.$$

Proof. The variable substitution $x = \cos(\varphi)$ can be used to rewrite (3.2) as

$$\forall u \in \mathbb{R}, \quad \mathcal{L}(h)(u) = 2\int_0^1 h\left(-\frac{u}{x}\right)\frac{x}{\sqrt{1-x^2}}\frac{\mathrm{d}x}{x}.$$

In such a form, we recognize a Mellin convolution which could be inverted by the use of Mellin convolution table [137]. □

3.3 AR(2) Discretization of the sPDE

Most previous works for Gaussian texture synthesis (such as [64] for static and [169, 178] for dynamic textures) have used a global Fourier-based approach and the explicit power spectrum expression (2.5). The main drawbacks of such an approach are: (i) it introduces an artificial periodicity in time and thus can only be used to synthesize a finite number of frames; (ii) these frames must be synthesized at once, before the stimulation, which prevents real-time synthesis; (iii) the discrete computational grid may introduce artifacts, in particular when one of the included frequencies is of the order of the discretization step or when a bandwidth is to small.

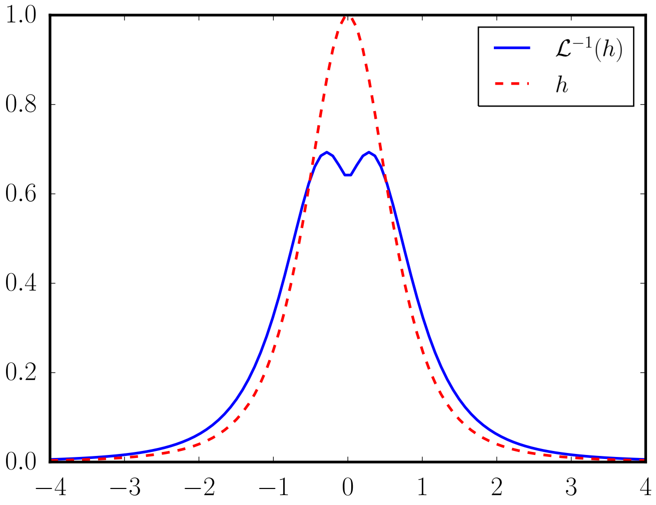


Figure 3.1: Functions h and $\mathcal{L}^{-1}(h)$.

To address these issues, we follow the previous works of [45, 213] and make use of an auto-regressive (AR) discretization of the sPDE (3.2). In contrast with these previous works, we use a second order AR(2) regression (in place of a first order AR(1) model). Using higher order recursions is crucial to make the output consistent with the continuous formulation (3.2). Indeed, numerical simulations show that AR(1) iterations lead to unacceptable temporal artifacts: in particular, the time correlation of AR(1) random fields typically decays too fast in time.

AR(2) synthesis without global translation, $v_0 = 0$. The discretization computes a (possibly infinite) discrete set of 2-D frames $(I_0^{(\ell)})_{\ell \geqslant \ell_0}$ separated by a time step Δ, and we approach at time $t = \ell\Delta$ the derivatives as

$$\frac{\partial I_0(\cdot,t)}{\partial t} \approx \Delta^{-1}(I_0^{(\ell)} - I_0^{(\ell-1)}) \quad \text{and} \quad \frac{\partial^2 I_0(\cdot,t)}{\partial t^2} \approx \Delta^{-2}(I_0^{(\ell+1)} + I_0^{(\ell-1)} - 2I_0^{(\ell)}),$$

and

$$\frac{\partial W(\cdot,t)}{\partial t} \approx \Delta^{-1}(W^{(\ell)} - W^{(\ell-1)})$$

which leads to the following explicit recursion $\forall\, \ell \geqslant \ell_0$,

$$I_0^{(\ell+1)} = (2\delta - \Delta\alpha - \Delta^2\beta) \star I_0^{(\ell)} + (-\delta + \Delta\alpha) \star I_0^{(\ell-1)} + \Delta(W^{(\ell)} - W^{(\ell-1)}), \quad (3.10)$$

where δ is the 2-D Dirac distribution and where $(W^{(\ell)} - W^{(\ell-1)})_\ell$ are i.i.d. 2-D Gaussian field with distribution $\mathcal{N}(0, \Delta\sigma_W)$, and $(I_0^{(\ell_0)}, I_0^{(\ell_0-1)})$ can be arbitrary initialized.

One can show that when $\ell_0 \to -\infty$ (to allow for a long enough "warmup" phase to reach approximate time-stationarity) and $\Delta \to 0$, then I_0^Δ defined by interpolating $I_0^\Delta(\cdot, \Delta\ell) = I^{(\ell)}$ converges (in the sense of finite dimensional distributions) toward a solution I_0 of the sPDE (3.2). Here we choose to use the standard finite difference however we refer to [191, 26] for more advanced discretization. We implemented the recursion (3.10) by computing the 2-D convolutions with FFT's on a GPU, which allows us to generate high resolution videos in real time, without the need to explicitly store the synthesized video.

AR(2) synthesis with global translation. The easiest way to approximate a sPDE cloud using an AR(2) recursion is to simply apply formula (3.1) to $(I_0^{(\ell)})_\ell$ as defined in (3.10), that is, to define

$$I^{(\ell)}(x) \overset{\text{def.}}{=} I_0^{(\ell)}(x - v_0\Delta\ell).$$

A second alternative approach would be to directly discretized the sPDE (3.4). We did not use this approach because it requires the discretization of spatial differential operators $\mathcal{G}$ and $\mathcal{H}$, and is hence less stable. A third, somehow hybrid, approach, is to apply the spatial translations to the AR(2) recursion, and define the following recursion

$$I^{(\ell+1)} = \mathcal{U}_{v_0} \star I^{(\ell)} + \mathcal{V}_{v_0} \star I^{(\ell-1)} + \Delta(W^{(\ell)} - W^{(\ell-1)}), \tag{3.11}$$

$$\text{where} \quad \begin{cases} \mathcal{U}_{v_0} \overset{\text{def.}}{=} (2\delta - \Delta\alpha - \Delta^2\beta) \star \delta_{-\Delta v_0}, \\ \mathcal{V}_{v_0} \overset{\text{def.}}{=} (-\delta + \Delta\alpha) \star \delta_{-2\Delta v_0}, \end{cases} \tag{3.12}$$

where δ_s indicates the Dirac at location s, so that $(\delta_s \star I)(x) = I(x - s)$ implements the translation by s. Numerically, it is possible to implement (3.11) over the Fourier domain,

$$\hat{I}^{(\ell+1)}(\xi) = \hat{\mathcal{U}}_{v_0}(\xi)\hat{I}^{(\ell)}(\xi) + \hat{\mathcal{V}}_{v_0}(\xi)\hat{I}^{(\ell-1)}(\xi) + \Delta\hat{\sigma}_W(\xi)(\hat{w}^{(\ell)}(\xi) - \hat{w}^{(\ell-1)}(\xi)),$$

$$\text{where} \quad \begin{cases} \hat{\mathcal{U}}_{v_0}(\xi) \overset{\text{def.}}{=} (2 - \Delta\hat{\alpha}(\xi) - \Delta^2\hat{\beta}(\xi))e^{-\mathrm{i}\Delta v_0\xi}, \\ \hat{\mathcal{Q}}_{v_0}(\xi) \overset{\text{def.}}{=} (-1 + \Delta\hat{\alpha}(\xi))e^{-2\mathrm{i}\Delta v_0\xi}, \end{cases}$$

and where $w^{(\ell)} - w^{(\ell-1)}$ is a 2-D white noise with distribution $\mathcal{N}(0, \Delta)$.

4 Synthesis from Examples

When developing a generative model of dynamic textures, it is important to quantify how well it is able to synthesize real dynamic textures. In this section, we present a way to perform dynamic textures synthesis from examples based on the sPDE model. Although the formulation is continuous, the approach is similar to the several AR methods presented in Section 1.1. However, as a continuous model the inference of sPDE coefficients can be improved, see [26].

4.1 sPDE with convolution coefficients

First let $T = \mathbb{R}^2/\mathbb{Z}^2$. Then, for all $(f, g) \in \mathcal{F}(T^2, \mathbb{C})^2$ where $\mathcal{F}(T^2, \mathbb{C})$ is the space of functions from T^2 to $\mathbb{C}$, we denote $P_{f,g}(X) = X^2 + fX + g$. We now define the stability set

$$\mathcal{S} = \{(f, g) | \forall x \in T, \forall z \in P^{-1}_{f(x),g(x)}(\{0\}), \Re(z) < 0\}. \tag{4.1}$$

The set $\mathcal{S}$ is the space of function that ensure the stability of solutions of a set of second order linear stochastic equations. One must think to the simple second order linear non-stochastic case where the stability is ensured when the eigenvalues of the linear operator have negative real part. We consider the second order linear sPDE with convolutive coefficients as in Equation (3.2):

$$\frac{\partial^2 F}{\partial t^2} + \alpha \star \frac{\partial F}{\partial t} + \beta \star F = \frac{\partial W}{\partial t} \tag{4.2}$$

where $\star$ is the spatial convolution and $(\hat{\alpha}, \hat{\beta}) \in \mathcal{S}$. The source term $\frac{\partial W}{\partial t}$ is a Gaussian process white in time and with spatial stationary covariance σ_W. As we work on the torus T. We do not use space Fourier transform but a space Karhunen-Loève representation of the Gaussian process. The goal is the same as using the Fourier transform, it allows to rewrite Equation (4.2) in the frequency domain.

Proposition 8. *Karhunen-Loève Transform of a Gaussian Process Let N be a Gaussian process white in time and with spatial stationary covariance σ_W. Then there exist $\hat{N}$,*

$$\forall t \in \mathbb{R}, \quad N(x, t) = \sum_{n \in \mathbb{Z}^2} \hat{N}(n, t) \exp\left(2i\pi\langle n,\, x\rangle\right)$$

where $N(n, t) \sim \mathcal{CN}(0, \hat{\sigma}_W(n))$ and $\mathcal{CN}(0, \hat{\sigma}_W(n))$ denotes the complex normal distribution of variance $\hat{\sigma}_W(n)$.

Proof. See [207] for details. □

Applying the Karhunen-Loève transform to Equation (4.2) is useful because it “diagonalizes” the convolution operators. The Karhunen-Loève transform of the solution of Equation (4.2) can therefore be obtained in the frequency domain by solving a set of second linear stochastic differential equations. We have the following proposition.

Proposition 9. *For all $n \in \mathbb{Z}^2$, $\hat{F}(n, \cdot)$ is solution of*

$$\forall\, t \in \mathbb{R}, \quad \frac{\partial^2 \hat{F}(n, t)}{\partial t^2} + \hat{\alpha}\frac{\partial \hat{F}(n, t)}{\partial t} + \hat{\beta}\hat{F}(n, t) = \frac{\partial \hat{W}(n, t)}{\partial t} \tag{4.3}$$

where $N(n, t) \sim \mathcal{CN}(0, \hat{\sigma}_W(n))$.

Proof. The proof is the direct application of the Karhunen-Loève transform and use the linearity of Equation (4.2). □

It useful to derive the probability distribution at a fixed time and frequency $(t, n) \in \mathbb{R} \times \mathbb{Z}$.

Proposition 10. *The solution $\hat{F}(n,t)$ has the following probability density function:*

$$\mathbb{P}(\hat{F}(n,t)|\hat{\alpha},\hat{\beta},\hat{C}) \propto \frac{1}{\sqrt{\hat{C}(n)}} \exp\left(-\frac{\left|\mathcal{L}_{\alpha,\beta}(\hat{F})(n,t)\right|^2}{2\hat{\sigma}_W(n)}\right)$$

where $\mathcal{L}_{\alpha,\beta}(\hat{F})(n,t) = \frac{\partial^2 \hat{F}(n,t)}{\partial t^2} + \hat{\alpha}(n)\frac{\partial \hat{F}(n,t)}{\partial t} + \hat{\beta}(n)\hat{F}(n,t)$.

Proof. As an invertible linear operator $\mathcal{L}_{\alpha,\beta}$ allows to express $\hat{F}(n,t)$ as a linear function of $\frac{\partial \hat{W}(n,t)}{\partial t}$ which gives the expected distribution. □

The probability density function expressed in Proposition 10 allows to adopt a maximum likelihood estimation strategy. The log-likelihood and the parameters $(\hat{\alpha}_m, \hat{\beta}_m, \hat{C}_m)$ that minimize it are summarized in the following proposition.

Proposition 11. *Assume that we have samples $(\hat{F}(n,t))_{(n,t)\in\hat{\Omega}_N\times\Omega_{N_T}}$ where $\hat{\Omega}_N = \{1,\dots,N\}^2$ and $\Omega_{N_T} = \{1,\dots,N_T\}$. The log-likelihood writes*

$$l(\hat{\alpha},\hat{\beta},\hat{C}) = \sum_{t\in\Omega_{N_T}} \sum_{n\in\hat{\Omega}_N} \frac{1}{2\hat{C}(n)} \left|\mathcal{L}_{\alpha,\beta}(\hat{F})(n,t)\right|^2 + \frac{1}{2}\log\left(\hat{C}(n)\right)$$

The triplet $(\hat{\alpha}_m, \hat{\beta}_m, \hat{C}_m)$ that minimizes l verify

$$\forall n \in \hat{\Omega}_N, \quad A(n)\begin{pmatrix} \hat{\alpha}_m(n) \\ \hat{\beta}_m(n) \end{pmatrix} = b(n)$$

where $\forall n \in \hat{\Omega}_N$,

$$A(n) = \begin{pmatrix} \sum\limits_{t\in\Omega_{N_T}} \left|\frac{\partial^2 \hat{F}(n,t)}{\partial t^2}\right| & \sum\limits_{t\in\Omega_{N_T}} \overline{\frac{\partial \hat{F}(n,t)}{\partial t}} F(n,t) \\ \sum\limits_{t\in\Omega_{N_T}} \overline{F(n,t)} \frac{\partial \hat{F}(n,t)}{\partial t} & \sum\limits_{t\in\Omega_{N_T}} \left|\hat{F}(n,t)\right| \end{pmatrix}$$

and

$$b(n) = \begin{pmatrix} -\sum_{t\in\Omega_{N_T}} \overline{\frac{\partial \hat{F}(n,t)}{\partial t}} \frac{\partial^2 \hat{F}(n,t)}{\partial t^2} \\ -\sum_{t\in\Omega_{N_T}} \overline{F(n,t)} \frac{\partial^2 \hat{F}(n,t)}{\partial t^2} \end{pmatrix}.$$

The parameter $\hat{C}_m$ thus writes

$$\forall n \in \Omega_N, \quad \hat{C}_m(n) = \frac{1}{N_t} \sum_{t\in\Omega_{N_T}} \left| \mathcal{L}_{\alpha_m,\beta_m}(\hat{F})(n,t) \right|^2.$$

Thus, we can implement a texture synthesis algorithm from examples based on Proposition 11.

4.2 Examples of Synthesis

We display here some examples of dynamic textures. Videos are available online[1]. We detail below the different steps of the algorithm. In particular, we follow the preprocessing used in [23]: we perform color synthesis by using a PCA on the RGB color channels and we handle edges by using the "Periodic+Smooth" decomposition [127].

Algorithm Assume that we have a dynamic texture sample $(F(x,t))_{(x,t)\in\Omega_N\times\Omega_{N_T}}$ projected on the first component of the PCA color space.

- For each time $t \in \Omega_{N_T}$, compute the 2D Fast Fourier Transform (fft) $\hat{F}(\cdot,t))$ of $F(\cdot,t)$,

- For each time $t \in \Omega_{N_T}$, compute the fft of the periodic component $\tilde{\hat{}}F(\cdot,t))$ of $\hat{F}(\cdot,t))$,

- Use Proposition 11 to infer $(\hat{\alpha}_m, \hat{\beta}_m, \hat{C}_m)$ (time derivative are approximated by finite differences),

- Use the algorithm describe in Section 3.3 to perform synthesis.

In Figure 4.1, we show frames extracted from natural texture examples *vs* frames extracted from synthesized textures. We observe that the model is not able to reproduce spatial edges and sharp contrasts unless they arise from approximate spatial periodicity. This is not surprising as our model suppose

[1] `https://jonathanvacher.github.io/mc_examples.html`

that textures are Gaussian and stationary which is generally not verified by natural textures. This remark is also valid for the temporal dynamic.

Our code is commented and available online[2].

5 Conclusion

We have proposed and detailed a generative model of dynamic textures based on a formalization of small perturbations from the observer's point of view during parameterized rotations, zooms and translations. We connected these transformations to descriptions of ecologically motivated movements of both observers and the dynamic world. The fast synthesis of naturalistic textures optimized to probe motion perception was then demonstrated, through fast GPU implementations applying auto-regression techniques with much potential for future experiments. Indeed, even if there exists some mathematical issues (delay,...) that we do not mention in details, the real-time synthesis algorithm allows to modify the model parameters over time (*ie* the covariance can be time dependent). We can imagine in a the future to control delay and to modify the parameters in real time to maximize the responses of neurons. This extends previous work from [169] by providing an axiomatic formulation. Finally, we detail a way to perform texture synthesis by maximum likelihood estimation of the sPDE coefficients. Such a synthesis algorithm can be useful for visual stimulation as it allows one to run experimental protocols that test natural dynamic textures *versus* their synthesis.

[2] `http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/motion_clouds`

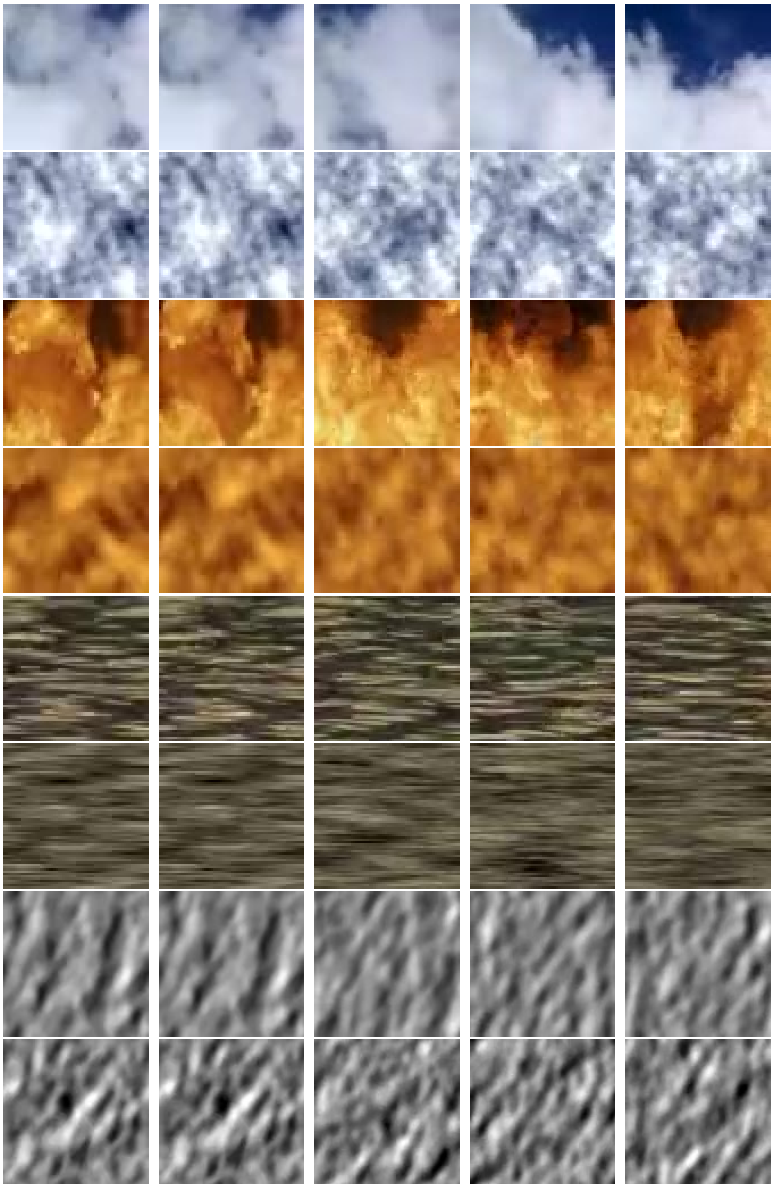


Figure 4.1: From top to bottom: alternating of five frames extracted from original textures and five frames of their synthesis (clouds, fire, goldenline, motion clouds).

⋆ II ⋆

Bayesian Inference Models for Psychophysics

In this chapter, we describe a mathematical formulation of an ideal Bayesian observer model. We develop a probabilistic formalism in order to properly define the concept of an observer's internal representations. In particular, we define in mathematical terms the notion of psychometric curve obtained in a two-alternate forced choice (2AFC) experiment. This general definition, combined with our ideal Bayesian observer model appears intractable in absence of specific assumptions. We thus exemplify the psychometric curve by making simplifying assumptions. Finally, we provide an algorithm which allows to solve the inverse Bayesian inference problem and we give numerical examples.

Contents

1 Introduction

This chapter aims at introducing a probabilistic framework for the analysis of psychophysical data that goes from stimulation parameters to their representation in the brain of an observer. In particular we focus on the modeling of discriminating tasks such as 2AFC or staircase procedure. While in this manuscript we focus on vision, we develop ideas that are not restrictive and can often be applied to other perception. This modeling is very much inspired by several recent work on the notion of inverse Bayesian inference (see bellow for the relevant litterature) and is in particular aiming for a rigorous formulation of the problem and a specific instantiation of a numerical inverse problem solver, that we use in chapter III to address speed discrimination.

1.1 Bayesian Brain

The "Bayesian brain" hypothesis is first a probabilistic theory. Among the probabilistic approaches to vision, it is important to mention the book of Desolneux, Moisan and Morel [41] who develop a set of statistical tools for image analysis based on Gestalt theory. The Bayesian interpretation of perception comes from the Helmholtz Machine principle [38]. The perceptual system is viewed as a statistical inference machine whose role is to infer the causes of sensory input. There is now a great interest for the Bayesian approaches as it is well suited to handle uncertainty, ambiguity and complexity [46, 34, 102]. However, when it comes to say that the brain performs Bayesian computations, there is still a lot of experimental data to confront [103, 106, 150]. The claim that the brain is Bayesian optimal is wisely criticized and discussed by Bowers [21]. Finally, we highlight some interesting works in physiology that formulates a way neural populations could perform Bayesian computations [94, 72, 151].

1.2 Inverse Bayesian Inference

The stochastic and dynamic generative models developed in Chapter I are closely related to the likelihood and prior models which serve to infer motion estimates from the dynamic visual stimulation [3] . In order to account for perceptual bias, a now well-accepted methodology in the field of psychophysics is to assume that observers are "ideal observers" and therefore make decisions using optimal statistical inference (typically a maximum-a-posteriori or MAP estimator) which combines this likelihood with some internal prior (see Introduction Equation (2.1)). Several experimental studies use this hypothesis as a justification for the observed perceptual biases by proposing some adjusted likelihood and prior models [46, 34], and more recent works pushes this ideas even further. Observing some perceptual bias, is it possible to "invert" this forward Bayesian decision making process, and infer the (unknown) internal prior that best fit a set of observed experimental choices made by observers? While a few previous works have raised similar questions (see below), its precise formulation and resolution is still an open problem both theoretically and numerically. Indeed, in traditional Bayesian inference approaches an estimator is computed from given likelihood and prior [22]. In contrast, here we have access to parameters and estimates from which we want to infer a likelihood and prior. Following [184], we coined this promising methodology "inverse Bayesian inference". This is of course an ill-posed inverse problem, in particular there is multiplicative ambiguity between the likelihood and prior. In addition, it is a highly non-linear. For all these reasons, it is clear that additional constraints on both the prior and the likelihood are needed to make it tractable. For instance [182, 184, 96] impose smoothness constraints in order to be able to locally fit the slope of the prior.

1.3 Contributions

We formalize the concept of observer's internal representations in a probabilistic model. In this model, we are able to formulate the "Bayesian brain" hypothesis: our brain estimates external parameters as if they have generated the sensory representation. Then, we are able to define in mathematical terms the notion of psychometric curve obtained in a two-alternate forced choice (2AFC) experiment. This general definition, combined with our ideal Bayesian observer model appears intractable in absence of specific assumptions. We thus exemplify the psychometric curve by making simplifying assumptions on the likelihood and prior and we give numerical examples. Finally, we provide an algorithm which allows to solve the inverse Bayesian inference problem and we also give a numerical example.

2 From Stimulation to Internal Representation

2.1 Model Description and Bayesian Assumptions

In a typical experimental context, the experimenter only knows the parameters of the stimulation $q \in \mathcal{Q}$, the stimuli $i \in \mathcal{I}$ and the *yes-no* answers of the different subject to some discrimination tasks. Chapter I deals with the connection between parameters and stimulation which we addressed by building a random generative model of stimulation. Figure 2.1 depicts the current situation. In order to understand what happens between the perception of an observer and the answer he produces, it is important to make some assumptions based on our current understanding of the brain architecture. Indeed,

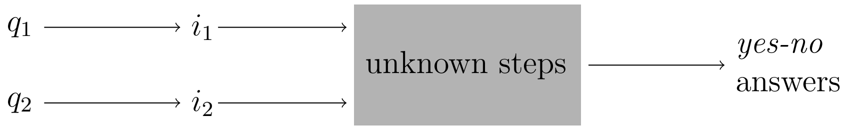


Figure 2.1: The partial knowledge provided by a psychophysics experiment. In the ideal case, parameters q_1 and q_2 allow to generate stimuli i_1 and i_2 from which the subject answers to an experimental *yes-no* question.

we know that some neurons respond or not to a particular stimulation. A typical example is the mostly known case of orientation selectivity in the primary visual cortex of many mammals [87]. It is thus reasonable to assume that the brain is making some measurements $m \in \mathcal{M}$ based on the perceived stimulus. Such measurements allow the observer to perform an estimation $\hat{q} \in \hat{\mathcal{Q}}$ of the parameters questioned by the experimenter. Finally, these estimations are used to provide a *yes-no* answer to a detection or a discrimination task. In a probabilistic setting, the variables q, i, m and $\hat{q}$ are realizations $Q(\omega), I(\omega), M(\omega)$ and $\hat{Q}(\omega)$ of the random variables Q, I, M and $\hat{Q}$. For simplicity, we assume that $\hat{Q}$ only depends on M which only depends on I, which only depends on Q and that these random variables have respectively the following densities $\mathbb{P}_{\hat{Q}|M}$, $\mathbb{P}_{M|I}$, $\mathbb{P}_{I|Q}$ and $\mathbb{P}_Q$. Hence, we complete the "unknown steps" box of Figure 2.1 by abstract measurement and estimation steps that we suppose to be performed by an observer, see Figure 2.2. This modeling pipeline is close to those presented in recent literature, see for instance [184]. However, it is not Bayesian yet.

$$\begin{tikzcd}
 & & \text{abstract completion} & & \\
q_1 = Q_1(\omega) \arrow[r] & i_1 = I_1(\omega) \arrow[r] & m_1 = M_1(\omega) \arrow[r] & \hat{q}_1 = \hat{Q}_1(\omega) \arrow[rd] & \\
Q_1 \sim \mathbb{P}_{Q_1} & I_1 \sim \mathbb{P}_{I_1|Q_1} & M_1 \sim \mathbb{P}_{M_1|I_1} & \hat{Q}_1 \sim \mathbb{P}_{\hat{Q}_1|M_1} & \textit{yes-no} \ \text{answer} \\
q_2 = Q_2(\omega) \arrow[r] & i_2 = I_2(\omega) \arrow[r] & m_2 = M_2(\omega) \arrow[r] & \hat{q}_2 = \hat{Q}_2(\omega) \arrow[ru] & \\
Q_2 \sim \mathbb{P}_{Q_2} & I_2 \sim \mathbb{P}_{I_2|Q_2} & M_2 \sim \mathbb{P}_{M_2|I_2} & \hat{Q}_2 \sim \mathbb{P}_{\hat{Q}_2|M_2} &
\end{tikzcd}$$

Figure 2.2: An abstract completion of the partial knowledge provided by a psychophysics experiment. The observer makes measures m_1 and m_2 of the stimulation which provides information to compute estimates $\hat{q}_1$ and $\hat{q}_2$ of the parameters questioned by the experiment.

First, let us focus on the underlying hypothesis of the model: the generative model of of stimulation (movies in the case of MC), the measurement and estimation steps. Obviously, the main underlying hypothesis in such Bayesian approaches is that the brain is able to encode probability distribution. Some works tackles this question experimentally in electrophysiology [94, 72, 111].

Generative Model First, we assume that stimuli came from a generative model conditioned by some relevant parameters that have their proper distribution. This perfectly fit the class of MC model designed in Chapter I. Such an assumption is strong because when confronted to natural images, their high complexity cannot be captured by a small number of parameters (see Introduction, Section 2.1). Therefore, generative models of natural movies are still out of reach.

Measurement Second, we assume that the measurement depends only on the stimulus. This hypothesis is based on the idea that neurons respond to certain stimulus features like spatio-temporal frequency, speed, orientation, motion direction, ...) [120, 153, 39, 13]. Although – a large part of neurons activity is still not understood, this assumption is reasonable.

Estimation Finally, we suppose an estimation step. Although it appears natural to perform an estimation from measurement, to the best of our knowledge, there is no clear experimental evidence that some neural circuits directly implements this function. This step can be understood as taking into account a direct comparison of the measurement made on different stimuli. We refer to the of Acuna *et al.* [2] that discuss the related question of whether the brain activity can be interpreted as rather sampling from some posterior distribution or only seeks for a maximum likelihood-type estimate.

Bayesian Assumptions Our model allows to formulate two different assumptions that are commonly made in Bayesian observer modeling: the Bayesian estimation and the natural prior hypotheses [150]. We formulate these hypothesis using our notations.

Assumption 1 (Bayesian Estimation)**.** *The random variable $\hat{Q}$ is estimating q as if it had directly generated the measurement m ie*

$$\mathbb{P}_{M|\hat{Q}}(m|q) = \mathbb{P}_{M|Q}(m|q).$$

Assumption 1 combined with Bayes theorem allows to compute the distribution of $\hat{Q}$ knowing M as:

$$\mathbb{P}_{\hat{Q}|M}(\hat{q}|m) = \frac{\mathbb{P}_{M|\hat{Q}}(m|\hat{q})\mathbb{P}_{\hat{Q}}(\hat{q})}{\mathbb{P}_M(m)} = \frac{\mathbb{P}_{M|Q}(m|\hat{q})\mathbb{P}_{\hat{Q}}(\hat{q})}{\mathbb{P}_M(m)}.$$

Before we impose this assumption, our model was only probabilistic because the probability $\mathbb{P}_{\hat{Q}|M}(\hat{q}|m)$ does not assume any estimation strategy but only a causal relation between M and $\hat{Q}$. Assumption 1 is the key that make our model a Bayesian one.

Assumption 2 (Natural Prior)**.** *The internal observer prior is supposed to reflect the natural environment ie*

$$\mathbb{P}_{\hat{Q}}(q) = \mathbb{P}_Q(q).$$

Assumption 2 means that the random variable $\hat{Q}$, which is internal to the observer, has the same distribution as Q which is external. Obviously here, $\mathbb{P}_Q$ is considered to represent the frequencies of different values taken by Q in a natural environment and not during an experiment. Therefore, when studying vision, the expected priors on image features is expected to ressemble those estimated from natural movies see [150, 67].

In order to analyze data using this model, one still needs to specify at least two distributions:

- the prior $\mathbb{P}_Q(q)$,
- and the likelihood $\mathbb{P}_{M|Q}(m|q)$.

We will exemplify these choices in Sections 3 and detail how they leads to several (closely related but different) data analysis methodologies in Section 4.

2.2 Psychometric Curve

A discrimination task experiment always involves at least two parameters (one for each of the two replications to discriminate between), so, for readability, we denote in bold any couple of variables $\boldsymbol{x} = (x_1, x_2)$. We call the estimated parameters $\hat{\boldsymbol{q}}$ "outputs" as opposed to the experimental parameters $\boldsymbol{q}$ called "inputs". Following the model described above (see also Figure 2.2), two stimuli $\boldsymbol{i}$ generated with inputs $\boldsymbol{q}$ are presented to a subject, the former makes some internal measurement $\boldsymbol{m}$ and estimates the inputs by $\hat{\boldsymbol{q}}$. He can therefore compare them to answer the *yes-no* question that actually represents one sample of a binary event $E \subset \mathcal{Q}^2$ specific to the subject. What is important to note is that the sample is obtained knowing that the stimulation have been generated independently using two "input" parameters $\boldsymbol{q}$. Then, we can define the abstract psychometric curve as a function of the input parameters.

Definition 3. *The psychometric function is the probability that $\hat{\boldsymbol{Q}}$ belongs to E knowing that $\boldsymbol{Q} = \boldsymbol{q}$*

$$\varphi_E(\boldsymbol{q}) \overset{\text{def.}}{=} \mathbb{P}_{\hat{\boldsymbol{Q}}|\boldsymbol{Q}}(E|\boldsymbol{q}) = \mathbb{E}_{\hat{\boldsymbol{Q}}|\boldsymbol{Q}}(1_E|\boldsymbol{q}).$$

where we denoted 1_E the indicator function of E

$$\mathbb{1}_E(\boldsymbol{q}) = \begin{cases} 1 & \text{if} \quad \boldsymbol{q} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Example 1. *A typical example, if q denotes a speed, is to test for speed discrimination by setting*

$$E \overset{\text{set}}{=} \{(\hat{q}_1, \hat{q}_2) \ ; \ \hat{q}_1 > \hat{q}_2\} \,,$$

i.e. whether the first stimulation appears faster than the second one to the user.

Now, we give a decomposition formula of the psychometric curve by taking into account the full model described in Figure 2.2 and the Bayesian Assumptions 1 and 2.

Proposition 12. *The psychometric curve verifies*

$$\varphi_E(\boldsymbol{q}) = \int_{\mathcal{Q}^2}\int_{\mathcal{M}^2} 1_E(\hat{\boldsymbol{q}}) \frac{\mathbb{P}_{\boldsymbol{M}|\boldsymbol{Q}}(\boldsymbol{m}|\hat{\boldsymbol{q}})\mathbb{P}_{\boldsymbol{M}|\boldsymbol{Q}}(\boldsymbol{m}|\boldsymbol{q})\mathbb{P}_{\boldsymbol{Q}}(\hat{\boldsymbol{q}})}{\mathbb{P}_{\boldsymbol{M}}(\boldsymbol{m})} \mathrm{d}\hat{\boldsymbol{q}}\mathrm{d}\boldsymbol{m}$$

with

$$\mathbb{P}_{\boldsymbol{M}|\boldsymbol{Q}}(\boldsymbol{m}|\boldsymbol{q}) = \int_{\mathcal{I}^2} \mathbb{P}_{\boldsymbol{M}|\boldsymbol{I}}(\boldsymbol{m}|\boldsymbol{i})\mathbb{P}_{\boldsymbol{I}|\boldsymbol{Q}}(\boldsymbol{i}|\boldsymbol{q})\mathrm{d}\boldsymbol{i}$$

and where we denoted 1_E the indicator function of E

$$\mathbb{1}_E(\boldsymbol{q}) = \begin{cases} 1 & \textit{if} \quad \boldsymbol{q} \in E, \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. First, we write

$$\varphi_E(\boldsymbol{q}) = \int_{\mathcal{Q}^2} 1_E(\hat{\boldsymbol{q}})\mathbb{P}_{\hat{\boldsymbol{Q}}|\boldsymbol{Q}}(\hat{\boldsymbol{q}}|\boldsymbol{q})\mathrm{d}\hat{\boldsymbol{q}}$$

Then, we plug successively the two following decompositions

$$\mathbb{P}_{\hat{\boldsymbol{Q}}|\boldsymbol{Q}}(\hat{\boldsymbol{q}}|\boldsymbol{q}) = \int_{\mathcal{M}^2} \mathbb{P}_{\hat{\boldsymbol{Q}}|\boldsymbol{M}}(\hat{\boldsymbol{q}}|\boldsymbol{m})\mathbb{P}_{\boldsymbol{M}|\boldsymbol{Q}}(\boldsymbol{m}|\boldsymbol{q})\mathrm{d}\boldsymbol{m}$$

and

$$\mathbb{P}_{\boldsymbol{M}|\boldsymbol{Q}}(\boldsymbol{m}|\boldsymbol{q}) = \int_{\mathcal{I}^2} \mathbb{P}_{\boldsymbol{M}|\boldsymbol{I}}(\boldsymbol{m}|\boldsymbol{i})\mathbb{P}_{\boldsymbol{I}|\boldsymbol{Q}}(\boldsymbol{i}|\boldsymbol{q})\mathrm{d}\boldsymbol{i}.$$

Finally, Assumptions 1 and 2 lead to the result. □

Proposition 12 is crucial to understand the Bayesian inverse inference problem. Indeed, we make the connection between the psychometric curve, which is a fundamental function usually sampled in a psychophysical experiment, and the likelihood $\mathbb{P}_{\boldsymbol{M}|\boldsymbol{Q}}$ and prior $\mathbb{P}_{\boldsymbol{Q}}$. The Bayesian inverse inference problem consists in determining the likelihood $\mathbb{P}_{\boldsymbol{M}|\boldsymbol{Q}}$ and prior $\mathbb{P}_{\boldsymbol{Q}}$ from the samples of psychometric curve $\varphi_E(\boldsymbol{q})$. In the form expressed in Proposition 12 and in absence of any assumption the inverse Bayesian inverse problem appear too difficult. In the following section, we make strong hypothesis that allow for closed form computations.

2.3 A Closed Form Example

In this section, we adopt the same assumption as Stocker [184] by assuming a Gaussian measure and a Laplacian prior. This allows for closed-form approximate of $\mathbb{P}_{\hat{Q}|Q}$ and thus of the psychometric curve. Although we do not assume any estimator, these assumptions allow for the computation of the bias $a\sigma^2$ that is equal to the one introduced in [184].

Proposition 13. *Suppose that*

- $\mathbb{P}_{M|Q}(m|q) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(m-q)}{2\sigma^2}\right)$ *with* $q \in \mathbb{R}$ *and* $\sigma > 0$ *,*
- *and* $\mathbb{P}_Q(q) = a \exp(-aq)$ *with* $a > 0$*.*

When $\sigma \to 0$*, one has*

$$\mathbb{P}_{\hat{Q}|Q}(\hat{q}|q) = \frac{1}{\sqrt{2\pi}(\sqrt{2}\sigma)} \exp\left(-\frac{1}{2(\sqrt{2}\sigma)^2}(\hat{q} - q + a\sigma^2)^2\right)(1 + o(1)).$$

This formula corresponds intuitively to the fact that the prior shifts the likelihood to give the posterior. Here, the posterior is approximately a Gaussian of standard deviation $\sqrt{2}\sigma$ and mean $q - a\sigma^2$. The shift comes from the combination of the prior parameter a and likelihood (v, σ). This is expected if one wants to explain perceptual bias.

Proof. First, we use the standard decomposition of probabilities and the Bayes formula combined with Assumption 1 and 2. Therefore,

$$\begin{aligned}\mathbb{P}_{\hat{Q}|Q}(\hat{q}|q) &= \int_{\mathcal{M}} \mathbb{P}_{\hat{Q}|M}(\hat{q}|m)\mathbb{P}_{M|Q}(m|q)\mathrm{d}m \\ &= \int_{\mathcal{M}} \frac{\mathbb{P}_{M|Q}(m|\hat{q})\mathbb{P}_Q(\hat{q})}{\mathbb{P}_M(m)}\mathbb{P}_{M|Q}(m|q)\mathrm{d}m \\ &= \mathbb{P}_Q(\hat{q}) \int_{\mathcal{M}} \frac{\mathbb{P}_{M|Q}(m|\hat{q})\mathbb{P}_{M|Q}(m|q)}{\mathbb{P}_M(m)}\mathrm{d}m. \qquad (2.1)\end{aligned}$$

In the expression above $\mathbb{P}_{M|Q}(.|q)$ is known and we need to compute $\mathbb{P}_M$ then,

$$\begin{aligned}\mathbb{P}_M(m) &= \int_{\mathcal{M}} \mathbb{P}_{M|Q}(m|q)\mathbb{P}_Q(q)\mathrm{d}q \\ &= \frac{a}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} \exp\left(-\frac{(m-q)^2}{2\sigma^2}\right) \exp(-aq)1_{\mathbb{R}_+}(q)\mathrm{d}q \\ &= \frac{a}{\sqrt{2\pi}\sigma} \exp\left(-am + \frac{a^2\sigma^2}{2}\right) \int_0^{+\infty} \exp\left(-\frac{(q-(m-a\sigma^2))^2}{2\sigma^2}\right)\mathrm{d}q \\ &= a\exp(-am)\exp(\frac{a^2\sigma^2}{2})\operatorname{erfc}\left(\frac{a\sigma}{\sqrt{2}} - \frac{m}{\sqrt{2}\sigma}\right)/2.\end{aligned}$$

Finally, when $\sigma \to 0$ we have

$$\mathbb{P}_M(m) \simeq \begin{cases} a\exp(-am)(1+o(\sigma^2)) & \text{si} \quad m > 0, \\ \frac{a}{2}\exp(-am)(1+o(\sigma)) & \text{si} \quad m = 0, \\ 0 & \text{si} \quad m < 0. \end{cases}$$

In order to avoid heavy technical details that are hidden in $o(\sigma^2)$, we choose to replace $\mathbb{P}_M(m)$ by

$$\tilde{\mathbb{P}}_M(m) = a\exp(-am)\mathbb{1}_{\mathbb{R}_+}(m).$$

By pluging the expression of $\tilde{\mathbb{P}}$ above into the integrand of Equation (2.1) and we obtain

$$2\pi a\sigma^2 \frac{\mathbb{P}_{M|Q}(m|\hat{q})\mathbb{P}_{M|Q}(m|q)}{\tilde{\mathbb{P}}(m)} = f_{q,\hat{q}}(m)$$

where

$$f_{q,\hat{q}}(m) = \exp\left(-\frac{(m-\hat{q})^2}{2\sigma^2} - \frac{(m-q)^2}{2\sigma^2} + am\right) 1_{\mathbb{R}_+^*}(m).$$

Consequently, we can write $\mathbb{P}_{\hat{Q}|Q}$ as

$$\mathbb{P}_{\hat{Q}|Q}(\hat{q}|q) \underset{\sigma\to 0}{\simeq} \frac{1}{2\pi a\sigma^2}\int_{\mathbb{R}} f_{q,\hat{q}}(m)\mathbb{P}_Q(\hat{q})\mathrm{d}m. \tag{2.2}$$

Using the equality

$$\begin{aligned} -\frac{(m-\hat{q})^2}{2\sigma^2} - \frac{(m-q)^2}{2\sigma^2} + am = &-\frac{1}{\sigma^2}\left(m - \frac{q+\hat{q}-a\sigma^2}{2}\right)^2 \\ &+ \frac{1}{4\sigma^2}(q+\hat{q}+a\sigma^2)^2 - \frac{1}{2\sigma^2}(q^2+\hat{q}^2), \end{aligned}$$

we can finally compute the integral in Equation (2.2) and when $\sigma \to 0$, we obtain

$$\mathbb{P}_{\hat{Q}|Q}(\hat{q}|q) \underset{\sigma\to 0}{\simeq} = \frac{1}{\sqrt{2\pi}(\sqrt{2}\sigma)} \exp\left(-\frac{1}{2(\sqrt{2}\sigma)^2}(\hat{q}-q+a\sigma^2)^2\right)(1+o(1)).$$

□

In order to illustrate this proof, we run a numerical simulation that approximate $\mathbb{P}_{\hat{Q}|Q}(\hat{q}|q)$. We use the following values $a = 1.0$, $\sigma = 1.2$ and $q = 10$ and compare the results obtained with $\tilde{\mathbb{P}}$ and $\mathbb{P}_M$. This shows that when σ is not too large the approximation holds as the proposition indicates. In addition, the numerical simulation highlights the fact that $\mathbb{P}_M$ (cyan in Figure 2.3) does

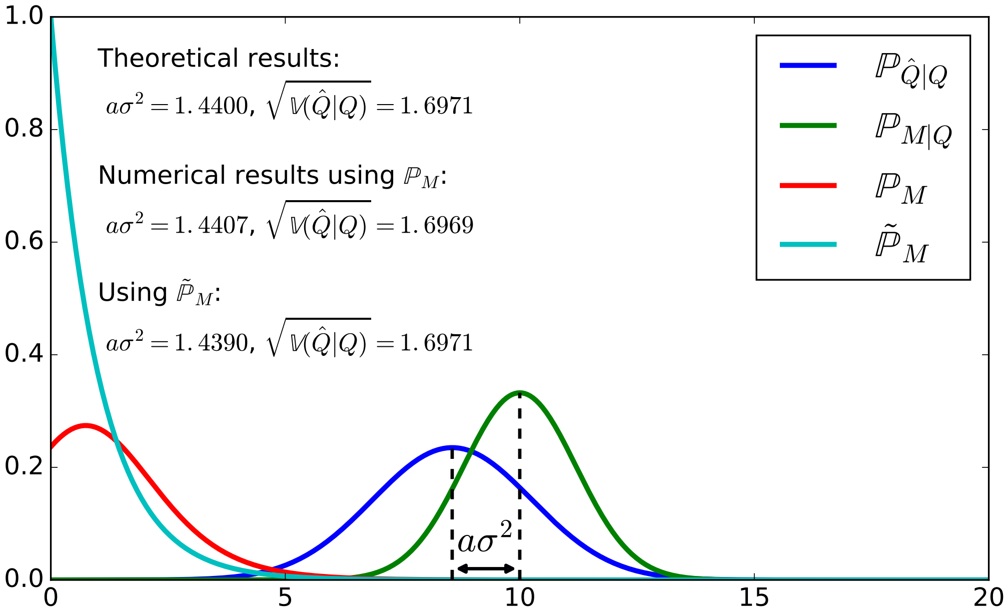


Figure 2.3: Numerical simulation of Proposition 13. The blue curve is obtain using $\mathbb{P}_M$, however the numerical results displayed for $a\sigma^2$ and $\sqrt{\mathbb{V}(\hat{q}|q)}$ indicates that the results for $\mathbb{P}_{\hat{Q}|Q}$ is very close when computed using $\tilde{\mathbb{P}_M}$.

not need to be very close to the approximation we use in the proof $\tilde{\mathbb{P}}$ (red in Figure 2.3). This indicates that this step is probably not necessary but we do not expand more on this.

By knowing an estimate of the posterior $\mathbb{P}_{\hat{Q}|Q}$, we can estimate the psychometric curve, see the following example.

Example 2. *Let us define the set E as given in Example 1. Assume different likelihoods respectively parametrized by $(q_1, \sigma_1) \in \mathbb{R} \times \mathbb{R}_+^*$ and $(q_2, \sigma_2) \in \mathbb{R} \times \mathbb{R}_+^*$ and different priors respectively parametrized by $a_1 > 0$ and $a_2 > 0$. We have*

$$\varphi_E(q_1, q_2) = \psi\left(\frac{q_1 - q_2 - a_1\sigma_1^2 + a_2\sigma_2^2}{\sqrt{2(\sigma_1^2 + \sigma_2^2)}}\right)$$

where $\psi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-s^2/2}\mathrm{d}s$ is the cumulative normal function of sigmoid shape. See Proposition 16 in the following chapter for a demonstration. An example of psychometric curve is shown in Figure 2.4.

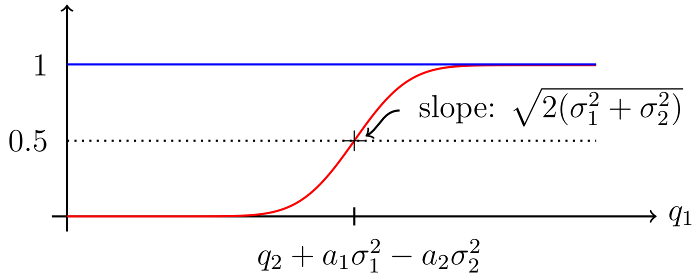


Figure 2.4: An example of the psychometric curve given in Example 2. It is shown as function of q_1 while q_2 is fixed.

3 Simplified Model: Deterministic Measures and Estimation

In this section, we assume that the estimation $\hat{q}$ is deterministic *ie* $\hat{q}$ is in fact computed from an estimation mapping $\Psi : m \mapsto \Psi(m) = \hat{q} \in \mathcal{Q}$ and the probability density is therefore $\mathbb{P}_{\hat{Q}|M}(|m) = \delta_{\Psi(m)}$. Such an asumption is the most frequent in the literature, see for instance [184], most likely because it leads to the simpler computations and numerical schemes. In the same way, we assume that the measurement is computed from an image by a mapping $\Phi : i \mapsto \Phi(i) = m \in \mathcal{M}$ and that the probability density is therefore $\mathbb{P}_{M|I}(|i) = \delta_{\Phi(i)}$. These assumptions do not rule out the Bayesian estimation hypothesis as one can still use a Bayesian estimator to design $\Psi(m)$. In the following, we give few standard examples for the mapping Ψ and Φ.

3.1 Examples of Mapping

3.1.1 Measurement

In order to design the mapping Φ that computes the measurement m the most natural way to proceed is to use the concept of neuron's receptive field (see Introduction, Section 2.3 or Section V1.1.3 for further details). The measurement performed on an image i by a neural population of size $n \in \mathbb{N}$ is typically modeled as a succession of a linear transform and a non-linear rectification, for instance $m = (\max(\langle \varphi_1, i\rangle, 0), \ldots, \max(\langle \varphi_n, i\rangle, 0))$ where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product on $\mathcal{I} = \mathbb{R}^{N^2}$ for an image of size $N \in \mathbb{N}$. The linear impulse responses of neurons $(\varphi_k)_{k\in\{1,\ldots,n\}}$ are typically wavelet-like oriented multiscale filters. We refer to the book of Mallat [113] for mathematical details about wavelet transform.

3.1.2 Estimation

A class of estimation mapping is defined in variational form as

$$\Psi(m) \overset{\text{def.}}{=} \underset{\tilde{q}\in\mathcal{Q}}{\operatorname{argmin}} \int L(q,\tilde{q})\mathbb{P}_{Q|M}(q|m)\mathrm{d}q.$$

where L is some loss function. In our framework $\mathbb{P}_{Q|M}$ is not given but computed by the Bayes rule

$$\mathbb{P}_{Q|M}(q|m) = \frac{\mathbb{P}_{M|Q}(m|q)\mathbb{P}_{Q}(q)}{\mathbb{P}_{M}(m)}$$

where the normalization probability $\mathbb{P}_M$ can be computed from $\mathbb{P}_Q$ and $\mathbb{P}_{M|Q}$ alone through the integration

$$\mathbb{P}_M(m) = \int_{\mathcal{Q}} \mathbb{P}_{M|Q}(m|\tilde{q})\mathbb{P}_Q(\tilde{q})\mathrm{d}\tilde{q}.$$

Maximum A Posteriori For instance, choosing $L(q,\hat{q}) = 1 - \delta_{\hat{q}}(q)$, where $\delta_{\hat{q}}$ is the Dirac located at $\hat{q}$, one obtains the Maximum A Posteriori (MAP) estimator

$$\Psi_{\text{MAP}}(m) = \underset{\hat{q}\in\mathcal{Q}}{\operatorname{argmax}}\ \mathbb{P}_{Q|M}(\hat{q}|m) = \underset{\hat{q}\in\mathcal{Q}}{\operatorname{argmin}}\ -\log\mathbb{P}_{M|Q}(m|\hat{q}) - \log\mathbb{P}_Q(\hat{q}). \quad (3.1)$$

Mean Square Error Choosing $L(q,\hat{q}) = \|q-\hat{q}\|^2$ one obtains the Mean Square Error (MSE) estimator (the conditional expectation)

$$\Psi_{\text{MSE}}(m) \overset{\text{def.}}{=} \int_{\mathcal{Q}} \tilde{q}\,\mathbb{P}_{Q|M}(\tilde{q}|m)\mathrm{d}\tilde{q} = \frac{\int_{\mathcal{Q}} \tilde{q}\,\mathbb{P}_{M|Q}(m|\tilde{q})\mathbb{P}_Q(\tilde{q})\mathrm{d}\tilde{q}}{\int_{\mathcal{Q}} \mathbb{P}_{M|Q}(m|\tilde{q})\mathbb{P}_Q(\tilde{q})\mathrm{d}\tilde{q}}$$

It is more intricate to compute than the MAP, mainly because of the integration, and of the normalizing constant, that itself depends on $\mathbb{P}_Q$.

3.2 Psychometric curve

As before, to ease of notations, we denote the estimation mapping on pairs as

$$\forall\, \boldsymbol{m} = (m_1, m_2) \in \mathcal{M}\times\mathcal{M}, \quad \boldsymbol{\Psi}(\boldsymbol{m}) = (\Psi(m_1), \Psi(m_2)) \in \mathcal{Q}^2.$$

The following proposition give a simplified expression for the psychometric curve associated with an event E.

Proposition 14. *Assume that there exists mappings $\Psi : m \mapsto \Psi(m) = \hat{q} \in \mathcal{Q}$ and $\Phi : i \mapsto \Psi(i) = m \in \mathcal{M}$. Then,*

$$\varphi_E(\boldsymbol{q}) = \int_{\mathcal{I}^2} 1_E(\boldsymbol{\Psi} \circ \boldsymbol{\Phi}(i))\mathbb{P}_{\boldsymbol{I}|\boldsymbol{Q}}(\boldsymbol{i}|\boldsymbol{q})\mathrm{d}\boldsymbol{i}$$

Proof.

$$\begin{aligned}\forall\, \boldsymbol{q} \in \mathcal{Q}^2, \quad \varphi_E(\boldsymbol{q}) &= \int_{\mathcal{Q}^2}\int_{\mathcal{I}^2} 1_E(\hat{\boldsymbol{q}})\mathbb{P}_{\hat{\boldsymbol{Q}}|\boldsymbol{M}}(\hat{\boldsymbol{q}}|\boldsymbol{\Phi}(\boldsymbol{i}))\mathbb{P}_{\boldsymbol{I}|\boldsymbol{Q}}(\boldsymbol{i}|\boldsymbol{q})\mathrm{d}\boldsymbol{i}\mathrm{d}\hat{\boldsymbol{q}} \\ &= \int_{\mathcal{Q}^2}\int_{\mathcal{I}^2} 1_E(\hat{\boldsymbol{q}})\delta_{\boldsymbol{\Psi}\circ\boldsymbol{\Phi}(i)}(\hat{\boldsymbol{q}})\mathbb{P}_{\boldsymbol{I}|\boldsymbol{Q}}(\boldsymbol{i}|\boldsymbol{q})\mathrm{d}\boldsymbol{i}\mathrm{d}\hat{\boldsymbol{q}} \\ &= \int_{\mathcal{I}^2} 1_E(\boldsymbol{\Psi} \circ \boldsymbol{\Phi}(i))\mathbb{P}_{\boldsymbol{I}|\boldsymbol{Q}}(\boldsymbol{i}|\boldsymbol{q})\mathrm{d}\boldsymbol{i}\end{aligned}$$

where $\delta_{\boldsymbol{\Psi}\circ\boldsymbol{\Phi}(i)}(\hat{\boldsymbol{q}}) = \delta_{\Psi\circ\Phi(i_1)}(\hat{q}_1)\delta_{\Psi\circ\Phi(i_2)}(\hat{q}_1)$. □

Proposition 14 is useful as it expresses the psychometric curve as a function of $(\boldsymbol{\Psi}, \boldsymbol{\Phi}, \mathbb{P}_{\boldsymbol{I}|\boldsymbol{Q}})$. We have given some examples of mappings in the section above. Moreover the generative model $\mathbb{P}_{\boldsymbol{I}|\boldsymbol{Q}}$ can be for instance the Motion Cloud model developed in Chapter I.

4 Inference Bayesian Inference Algorithms

After having detailing the structure of the psychophysical function φ_E under several Bayesian model, we are now in the position to formally define an inverse Bayesian inference procedure to estimate the prior from the output of psychophysical experiments.

4.1 Prior fitting

Let us first put in statistical terms the outcome of n independent replications of 2AFC tests $\{(\varepsilon_1, \boldsymbol{i}_1, \boldsymbol{q}_1), \ldots, (\varepsilon_n, \boldsymbol{i}_n, \boldsymbol{q}_n)\} \in (\{0,1\} \times \mathcal{I}^2 \times \mathcal{Q}^2)^n$ where for all $k \in \{0, \ldots, n\}$, $\boldsymbol{i}_k$ is a pair of stimulations generated with density $\mathbb{P}_{\boldsymbol{I}|\boldsymbol{Q}}(.|\boldsymbol{q}_k)$ and ε_k is a Bernoulli random variable equals to 1 with some probability p if $q_k \in E$. Following Chapter 3, for each pair of stimulation $\boldsymbol{i}_k$, the brain measures $\boldsymbol{m}_k = \boldsymbol{\Phi}(\boldsymbol{i}_k)$ from which it computes a pair of estimations $\hat{\boldsymbol{q}}_k = \boldsymbol{\Psi}(\boldsymbol{m}_k)$. For simplicity of the exposition, we assume that measures are directly the image's pixels *ie* $\Phi = \mathrm{Id}$. Moreover, we restrict our attention to a MAP estimator, but similar derivations can be carried out in the general setting. Using a MAP estimator leads to $\hat{q}_k = \Psi_{MAP}(m_k) = \Psi_{MAP} \circ \Phi(i_k) = \Psi_{MAP}(i_k)$ which can be written as

$$\hat{q}_k = \underset{q\in\mathcal{Q}}{\operatorname{argmin}}\; -\log \mathbb{P}_{I|Q}(i_k|q) - \log \mathbb{P}_Q(q).$$

However, the output of an 2AFC experiment is not $\hat{q}_k$ but only the Bernouilli random variable ε_k. For the moment, we can only expect to simulate an experiment *ie* generating two stimuli, estimating their parameters and decide whether these estimates belong to E or not.

Example 3. *As a typical example of instantiation of this formula, consider MC stimulation (see Proposition 2). In this case, the stimuli are Gaussian and have the power spectrum*

$$\forall\,(\xi,\tau)\in\mathbb{R}^2\times\mathbb{R},\;\hat{\gamma}_q(\xi,\tau)=\frac{\mathbb{P}_Z\left(\|\xi\|\right)}{\|\xi\|^2}\mathbb{P}_\Theta\left(\angle\xi\right)\mathcal{L}(\mathbb{P}_{\|V-v_0\|})\left(-\frac{\tau+\langle v_0,\,\xi\rangle}{\|\xi\|}\right),$$

where the linear transform $\mathcal{L}$ is such that

$$\forall\,u\in\mathbb{R},\quad\mathcal{L}(f)(u)\overset{\text{def.}}{=}\int_{-\pi}^{\pi}f(-u/\cos(\varphi))\mathrm{d}\varphi.$$

and $q=(v_0,\sigma_V,\theta_0,\sigma_\Theta,z_0,B_Z)$ is the vector of the MC parameters. One can therefore compute $\log\mathbb{P}_{I|Q}(i_k|q)$.

4.2 Prior fitting when samples from $\hat{q}(m)$ is accessible.

Let us here rephrase in our language the approach detailed in [144]. It shed some lights on the (convex!) class of constraints that a prior should typically, so it is quite informative. Unfortunately, it cannot be used for psychophysical studies because one never have direct access to the internal estimate hat q made by the brain. The goal is to estimate the prior function $g(q)$ from psychophysical experiments. As remarked by [144], if one directly has access to values $\hat{q}(m)$ for some set $m\in\mathcal{M}$ of stimulations, then finding g can be obtained by solving a convex program, since g is only constrained to satisfy

$$\left\{g\;;\;g\geqslant 0,\;\int g=1,\;\forall\,(m,w)\in\tilde{\mathcal{M}}\times\mathcal{Q},\;\langle g,\,h_{w,m}\rangle\leqslant 0\right\}$$

$$\text{where}\quad h_{w,m}=(L(\cdot,\hat{q}(m))-L(\cdot,w))\mathbb{P}(m|\cdot).$$

An even simpler set-up is obtained when using the MAP estimator. In this case, the first order optimality conditions of (3.1) reads

$$\nabla G(\hat{q}(m))=-\nabla F(\hat{q}(m),m)\tag{4.1}$$

where ∇ is the gradient with respect to the $\hat{q}$ variables, and where we denoted $G(\hat{q})=\log(g(\hat{q}))$ and $F(\hat{q},m)=\log(\mathbb{P}(m|\hat{q}))$. Interestingly, the optimality conditions (4.1) can be integrated which constitute an natural improvement of [144] for the special case of a MAP estimator.

Proposition 15. *Assume a series of stimulations/measures* $(m_t)_{t=0}^1$ *between* m_0 *and* m_1*, one has*

$$G(\hat{q}_1) = G(\hat{q}_0) - \int_0^1 \langle \nabla F(\hat{q}_t, m_t),\, \hat{q}_t' \rangle \mathrm{d}t$$

where $\hat{q}_t = \hat{q}(m_t)$ *and thus also* $\hat{q}_t'$ *(time derivative of* $\hat{q}_t$*) is supposed to be known.*

4.3 Prior fitting when samples from φ_E are accessible.

We now detail the much more complicate (but realistic for psycophysics) setting where one only has access to the output of a 2AFC experiment, so when one has at its disposal an approximation $\hat{\varphi}_E$ of the true psychophysical function. In order to make the process computationally tractable, we assume that the prior belongs to a parametric family $\mathbb{P}(q) = g_\alpha(q)$ parametrized by some α. According to Proposition 12, the psychometric curve thus also depends on α, which we denote as $\varphi_E(\boldsymbol{q}) = \varphi_E(\boldsymbol{q}, \alpha)$.

We denote $\varepsilon_i \in \{0, 1\}$ for $i \in \mathcal{I}$ the output of the psychophysical experiment for the i^{th} trial, which is obtained by a stimulation with some parameters $\boldsymbol{q}_i \in \mathcal{Q}$. We set $\varepsilon_i = 1$ if he estimated for this trial that $(\hat{q}_1, \hat{q}_2) \in E$, and $\varepsilon_i = 0$ otherwise. Then, according to our model, the ε_i are samples from independent Bernoulli distributions of parameter $\varphi_E(\boldsymbol{q}_i, \alpha)$. The parameter α can thus be estimated using a maximum likelihood estimate

$$\min_\alpha \sum_{i \in \mathcal{I}} \ell(\varepsilon_i, \varphi_E(\boldsymbol{q}_i, \alpha)) \tag{4.2}$$

where we denoted $\ell(\cdot, p)$ the anti-log likelihood of the Bernoulli variable of parameter p

$$\ell(\varepsilon, p) = \begin{cases} -\log(p) & \text{if} \quad \varepsilon = 0, \\ -\log(1-p) & \text{if} \quad \varepsilon = 1. \end{cases}$$

Assuming for simplicity that for each tested $\boldsymbol{q} \in \mathbb{V} = \{q_i\}_{i \in \mathcal{I}}$, the cardinal of trials $|\{i \; ; \; \boldsymbol{q}_i = \boldsymbol{q}\}|$ is the same, the optimization (4.2) can be elegantly rewritten as a Kullback-Leibler minimization

$$\min_\alpha \sum_{\boldsymbol{q} \in \mathbb{V}} \mathrm{KL}(\hat{\varphi}_E(\boldsymbol{q}) | \varphi_E(\boldsymbol{q}, \alpha)) \tag{4.3}$$

$$\text{where} \quad \mathrm{KL}(\hat{p}|p) = \hat{p} \log\left(\frac{\hat{p}}{p}\right) + (1-\hat{p}) \log\left(\frac{1-\hat{p}}{1-p}\right),$$

where we denoted $\hat{\varphi}_E(\boldsymbol{q})$ the empirical psychometric curve

$$\hat{\varphi}_E(\boldsymbol{q}) = \frac{|\{i \in \mathcal{I}\ ;\ \varepsilon_i = 1, \boldsymbol{q}_i = \boldsymbol{q}\}|}{|\{i \in \mathcal{I}\ ;\ \boldsymbol{q}_i = \boldsymbol{q}\}|} \in [0,1].$$

The problem (4.3) is a non-convex, but typically smooth and low dimensional optimization problem. We thus advocate the use of standard quasi-newton technics (LBFGS) in order to capture a local optimal α.

⋆ III ⋆

Speed Discrimination in Psychophysics

To exploit biologically-inspired parameterization of the MC model and provide a proof of concept of its usefulness based on motion perception, we consider here the problem of discriminating the relative speed of moving dynamical textures. The overall aim is to characterize the impact of both average spatial frequency and average duration of temporal correlations on perceptual speed estimation based on the empirical evidences. By simplifying the general Bayesian framework developed in Chapter II, we assume a Gaussian likelihood and we estimate a Laplacian prior to account for the bias observed in the psychophsysical experiment. We focus here our attention on the use of a maximum a posteriori (MAP) estimator, which leads to a numerically tractable fitting procedure.

Contents

1 Introduction

1.1 Previous Works

Previous works of Brooks [27] and Smith [179] have revealed that spatial frequency positively bias our speed perception. In the following, we reproduce these experiments and embed into the Bayesian framework exposed in the previous chapter. In particular, our approach is mainly based on the works [184, 182, 96]. Stocker *et al.* [184] develop a Bayesian inverse methodology to infer both parametric likelihood and prior that are able to explain the well known negative effect of contrast on speed perception. By testing a large range of speed, they are able to integrate the prior that appears to favor slow speed. Their work has been reproduced by Sotiropoulos [182] that gives further practical details. More recently, Jogan and Stocker [96] successfully adapt this methodology to account for multiple spatial frequency channels. Finally, our experiment is conducted using MCs developed in Chapter I and it is important to highlight the work of Schrater *et al.* [171, 170] that use MC-like stimulation. In [171], they use stimuli with decreasing spatial frequencies to study how humans estimate expansion rates from scale-changes information. In [170], they use homogeneously oriented and heterogeneously oriented stimuli to probe the mechanisms of energy summation over orientation during the estimation of motion.

1.2 Contributions

We run psychophysical experiments to probe speed perception in humans using zoom-like changes in MCs spatial frequency content. We simplify the general Bayesian model developed in Chapter II by assuming a Gaussian likelihood and a Laplacian prior. As the MC model allows for the derivation of a local motion-energy model, we use it to estimate speed in the experimental stimuli. By comparing the estimated variances of observers' likelihood to the distribution of the motion-energy model estimates of speed we show that they are not compatible. We validate the fitting process of the model using synthesized data. The human data replicates previous findings [27, 179] that relative perceived speed is positively biased by spatial frequency increments. The effect cannot be fully accounted for by previous models, but the current prior acting on the spatio-temporal likelihoods has proved necessary in accounting for the perceptual bias. We provide an online[1] example of data synthesis and analysis.

2 Experimental Settings and Model

2.1 Methods

The task is to discriminate the speed $v \in \mathbb{R}$ of a MC stimuli moving with a horizontal central speed $\mathbf{v} = (v, 0)$. We refer to Section I2.3 for the parameter notations. We assign as independent experimental variable the most represented spatial frequency z_0, that we denote in the following z for easier reading. The other parameters are set to the following values

$$\sigma_V = \frac{1}{t^\star z}, \quad \theta_0 = \frac{\pi}{2}, \quad \sigma_\Theta = \frac{\pi}{12}.$$

[1] `http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/bayesian_observer/`

Note that σ_V is thus dependent of the value of z to ensure that $t^\star = \frac{1}{\sigma_V z}$ stays constant. This parameter $t^\star$ controls the temporal frequency bandwidth, as illustrated on the middle of Figure I2.2. We used a two alternative forced choice (2AFC) paradigm (see Section II)2). In each trial, a gray fixation screen with a small dark fixation spot was followed by two stimulus intervals of 250 ms each, separated by an uniformly gray 250 ms inter-stimulus interval. The first stimulus had parameters (v_1, z_1) and the second had parameters (v_2, z_2). At the end of the trial, a gray screen appears asking the participant to report which one of the two intervals was perceived as moving faster by pressing one of two buttons, that is whether $v_1 > v_2$ or $v_2 > v_1$.

Given reference values $(v^\star, z^\star)$, for each trial, (v_1, z_1) and (v_2, z_2) are selected such that

$$\begin{cases} v_i = v^\star, \; z_i \in z^\star + \Delta_Z \\ v_j \in v^\star + \Delta_V, \; z_j = z^\star \end{cases} \quad \text{where} \quad \Delta_V = \{-2, -1, 0, 1, 2\},$$

where $(i, j) = (1, 2)$ or $(i, j) = (2, 1)$ (i.e. the ordering is randomized across trials), and where z values are expressed in cycles per degree (c/°) and v values in °/s. The range Δ_Z is defined below. Ten repetitions of each of the 25 possible combinations of these parameters are made per block of 250 trials and at least four such blocks were collected per condition tested. The outcome of these experiments are summarized by sampled psychometric curves $\hat{\varphi}_{v^\star, z^\star}$ (see Definition 3), where for all $(v - v^\star, z - z^\star) \in \Delta_V \times \Delta_Z$, the value $\hat{\varphi}_{v^\star, z^\star}(v, z)$ is the empirical probability (each averaged over the typically 40 trials) that a stimulus generated with parameters $(v^\star, z)$ is moving faster than a stimulus with parameters $(v, z^\star)$.

To assess the validity of our model, we tested different scenarios summarized in Table 2.1. Each row corresponds to 35 minutes of testing per participant and was always performed by at least two of the participants. Stimuli were generated on a Mac running OS 10.6.8 and displayed on a 20" Viewsonic p227f monitor with resolution 1024 × 768 at 100 Hz. Routines were written using Matlab 7.10.0 and Psychtoolbox 3.0.9 controlled the stimulus display. Observers sat 57 cm from the screen in a dark room. Four observers, three male and one female, with normal or corrected to normal vision took part in these experiments. They gave their informed consent and the experiments received ethical approval from the Aix-Marseille Ethics Committee in accordance with the declaration of Helsinki.

To increase the statistical power of the data set during analysis, psychometric functions were generated following the observed effect in the data and a sampling was carried out to obtain a synthetic data set for the validation of the Bayesian fitting procedure (see Chapter II4). The steps involved are detailed in section 4.2.

Case	$t^\star$	σ_Z	B_Z	$v^\star$	$z^\star$	Δ_Z
A1	200 ms	1.0 c/°	×	5 °/s	0.8 c/°	$\{-0.27, -0.16, 0, 0.27, 0.48\}$
A2	200 ms	1.0 c/°	×	5 °/s	1.28 c/°	$\{-0.48, -0.21, 0, 0.32, 0.85\}$
A3	200 ms	1.0 c/°	×	10 °/s	0.8 c/°	$\{-0.27, -0.16, 0, 0.27, 0.48\}$
A4	200 ms	1.0 c/°	×	10 °/s	1.28 c/°	$\{-0.48, -0.21, 0, 0.32, 0.85\}$
B1	100 ms	1.0 c/°	×	10 °/s	0.8 c/°	$\{-0.27, -0.16, 0, 0.27, 0.48\}$
B2	100 ms	1.0 c/°	×	10 °/s	1.28 c/°	$\{-0.48, -0.21, 0, 0.32, 0.85\}$
C1	100 ms	×	1.28	5 °/s	1.28 c/°	$\{-0.48, -0.21, 0, 0.32, 0.85\}$
C2	100 ms	×	1.28	10 °/s	1.28 c/°	$\{-0.48, -0.21, 0, 0.32, 0.85\}$
C3	200 ms	×	1.28	5 °/s	1.28 c/°	$\{-0.48, -0.21, 0, 0.32, 0.85\}$
C4	200 ms	×	1.28	10 °/s	1.28 c/°	$\{-0.48, -0.21, 0, 0.32, 0.85\}$

Table 2.1: A and B are both bandwidth controlled in °/s with high and low $t^\star$ respectively, C is bandwidth controlled in octaves.

2.2 Bayesian modeling

To make full use of our MC paradigm in analyzing the obtained results, we follow the methodology of the Bayesian observer used for instance in [184, 182, 96] that we have formalized and refined in Chapter II. We assume the observer makes its decision using a Maximum A Posteriori (MAP) estimator

$$\hat{v}_z(m) = \underset{v}{\operatorname{argmin}} \, [-\log(\mathbb{P}_{M|V,Z}(m|v,z)) - \log(\mathbb{P}_{V|Z}(v|z))] \tag{2.1}$$

computed from some internal representation $m \in \mathbb{R}$ of the observed stimulus (see Section II3.1). For simplicity, we assume that the observer estimates z from m without bias. To simplify the numerical analysis, we assume that the likelihood is Gaussian, with a variance independent of v. Furthermore, we assume that the prior is Laplacian as this gives a good description of the a priori statistics of speeds in natural images [43]:

$$\mathbb{P}_{M|V,Z}(m|v,z) = \frac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{|m-v|^2}{2\sigma_z^2}} \quad \text{and} \quad \mathbb{P}_{V|Z}(v|z) \propto e^{a_z v} 1_{[0,v_{\max}]}(v). \tag{2.2}$$

where $v_{\max} > 0$ is a cutoff speed ensuring that $\mathbb{P}_{V|Z}$ is a well defined density even if $a_z > 0$.

Both a_z and σ_z are unknown parameters of the model, and are obtained from the outcome of the experiments by a fitting process we now explain.

3 Experimental Likelihood vs. the MC Model

The approach we propose in this chapter is to use the model (2.2), which thus corresponds to directly fitting the likelihood $\mathbb{P}_{M|V,Z}(m|v,z)$ from the ex-

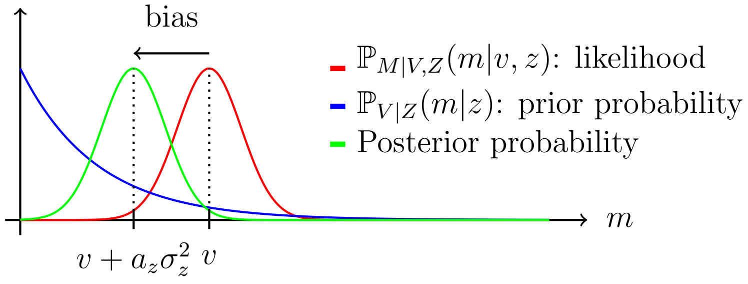


Figure 2.1: Multiplying the likelihood by such a prior gives a posterior that looks like a shifted version of the likelihood. Such an idea shows that the prior is responsible for bias when a Bayesian inference is performed.

perimental psychometric curve. While this makes sense from a data-analysis point of view, this required strong modeling hypothesis, in particular, that the likelihood is Gaussian with a variance σ_z^2 independent of the parameter v to be estimated by the observer.

Before actually analyzing the output of the experiments in Section 4 and 5, we first propose in this section to derive a likelihood model directly from the stimuli. We assume the hypothesis that the observer uses a standard motion estimation process, based on the motion energy concept [3], an idea we incorporate here into the MC distribution. In this setting, this corresponds to using a MLE estimator, and making use of the sPDE formulation of MC.

3.1 MLE Speed Estimator.

We first show how to compute this MLE estimator. To be able to achieve this, we use the sPDE formulation provided by Proposition 5. Equation (3.4) is useful from a Bayesian modeling perspective, because, informally, it can be interpreted as the fact that the Gaussian distribution of MC has the following appealing form, for any video $\mathcal{I} : \Omega \times T \to \mathbb{R}$ observed on a bounded space-time domain $\Omega \times [0,T]$,

$$\begin{aligned} -\log(\mathbb{P}_I(\mathcal{I}|v_0)) = Z_I + \int_\Omega \int_0^T & |\mathcal{D}(K_W \star \mathcal{I})(x,t) \\ & + \langle \mathcal{G}(K_W \star \mathcal{I})(x,t),\, v_0\rangle + \langle \mathcal{H}(K_W \star \mathcal{I})(x,t)v_0,\, v_0\rangle|^2 \mathrm{d}t\mathrm{d}x \end{aligned} \tag{3.1}$$

where K_W is the spatial filter corresponding to the square-root inverse of the covariance Σ_W, i.e. which satisfies $\hat{K}_W(\xi) \overset{\text{def.}}{=} \hat{\sigma}_W(\xi)^{-1}$, where $\mathcal{D}$ is defined in (3.2), $\mathcal{G}$ and $\mathcal{H}$ are defined in (3.5), where Z_I is a normalization constant which is independent of v_0 where $\hat{\sigma}_W$ is defined in (3.7). Equation (3.1) can be seen as a direct generalization of the initial energy model (2.3), when the first order luminance conservation sPDE (2.2) is replaced by the second order MC sPDE model (3.4).

It is however important to realize that the expression (3.1) is only formal, since the rigorous definition of the likelihood of infinite dimensional Gaussian distribution is more involved [70]. It is possible to give a simple rigorous expression for the case of discretized clouds satisfying the AR(2) recursion (3.11). In this case, for some input video $\mathcal{I} = (\mathcal{I}^{(\ell)})_{\ell=1}^{L}$, the log-likelihood reads

$$-\log(\mathbb{P}_I(\mathcal{I})) = \tilde{Z}_I + K_{v_0}(\mathcal{I}) \quad \text{where}$$

$$K_{v_0}(\mathcal{I}) \overset{\text{def.}}{=} \frac{1}{\Delta^4}\sum_{\ell=1}^{L}\int_\Omega |K_W \star \mathcal{I}^{(\ell+1)}(x) - \mathcal{U}_{v_0} \star K_W \star \mathcal{I}^{(\ell)}(x) - \mathcal{V}_{v_0} \star K_W \star \mathcal{I}^{(\ell-1)}(x)|^2 \mathrm{d}x$$

where $\mathcal{U}_{v_0}$ and $\mathcal{V}_{v_0}$ are defined in (3.12). This convenient formulation can be used to re-write the MLE estimator of the horizontal speed v parameter of a MC as

$$\hat{v}^{\text{MLE}}(\mathcal{I}) \overset{\text{def.}}{=} \underset{v}{\operatorname{argmax}}\ \mathbb{P}_I(\mathcal{I}) = \underset{v}{\operatorname{argmin}}\ K_{v_0}(\mathcal{I}) \quad \text{where} \quad v_0 = (v, 0) \in \mathbb{R}^2 \tag{3.2}$$

where we used the fact that $\tilde{Z}_I$ is independent of v_0. The solution to this optimization problem with respect to v is then computed using the Newton-CG optimization method implemented in the python library `scipy`.

3.2 MLE Modeling of the Likelihood.

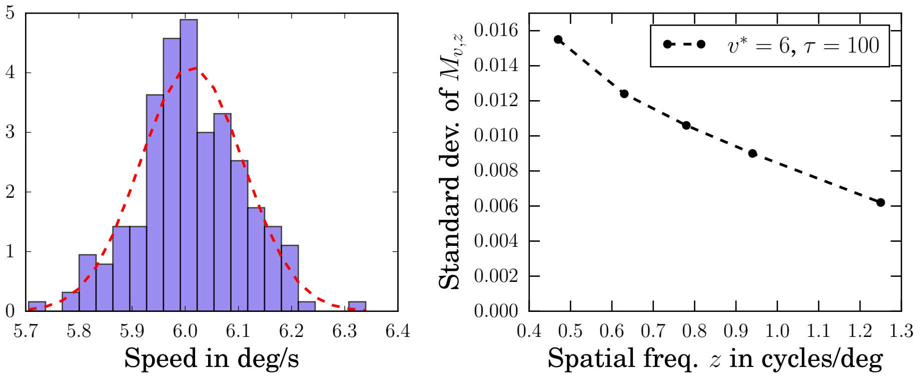
Following several previous works such as [184, 182], we assumed the existence of an internal representation parameter m, which was assumed to be a scalar, with a Gaussian distribution conditioned on (v, z). We explore here the possibility that this internal representation could be directly obtained from the stimuli by the observer using an "optimal" speed detector (an MLE estimate).

Denoting $I_{v,z}$ a MC, which is a random Gaussian field of power spectrum (2.5), with central speeds $v_0 = (v, 0)$ and central spatial frequency z (the other parameters being fixed as explained in the experimental section of the paper), this means that we consider the internal representation as being the following scalar random variable

$$M_{v,z} \overset{\text{def.}}{=} \hat{v}_z^{\text{MLE}}(I_{v,z}) \quad \text{where} \quad \hat{v}_z^{\text{MLE}}(\mathcal{I}) \overset{\text{def.}}{=} \underset{v}{\operatorname{argmax}}\ \mathbb{P}_{M|V,Z}(\mathcal{I}|v, z), \tag{3.3}$$

which corresponds to the optimization (3.2) and can be solved efficiently numerically.

As shown in Figure 3.1(a), we observed that $M_{v,z}$ is well approximated by a Gaussian random variable. Its mean is very close to v, and Figure 3.1(b) shows the evolution of its variance for different spatial frequencies z. An important point to note here is that this optimal estimation model (using an MLE) is not consistent with the experimental finding because the estimated standard deviations of observers do not show a decreasing behavior as in Figure 3.1(b).



(a) Histogram of $M_{v,z}$

(b) Standard deviation $M_{v,z}$

Figure 3.1: Estimates of $M_{v,z}$ for $z = 0.8$ c/° defined by (3.3) and its standard deviation as a function of z.

4 Model Fitting Evaluation

4.1 Likelihood and Prior Estimation

Adopting an approach from previous literature [184, 182, 96] and developed in Section II2, the theoretical psychometric curve obtained by a Bayesian decision model is

$$\varphi_{v^\star,z^\star}(v,z) \overset{\text{def.}}{=} \mathbb{E}(\hat{v}_{z^\star}(M_{v,z^\star}) > \hat{v}_z(M_{v^\star,z}))$$

where $M_{v,z} \sim \mathcal{N}(v, \sigma_z^2)$ is a Gaussian variable having the distribution $\mathbb{P}_{M|V,Z}(\cdot|v,z)$. The definition corresponds to the one introduced in Definition 3, however we adapt the notations to the experimental context.

The following proposition shows that in our special case of Gaussian prior and Laplacian likelihood, it can be computed in closed form. Its proof follows closely the derivation of [182, Appendix A]. This proposition must be related to Proposition 13. The difference is that here we assume a MAP estimator

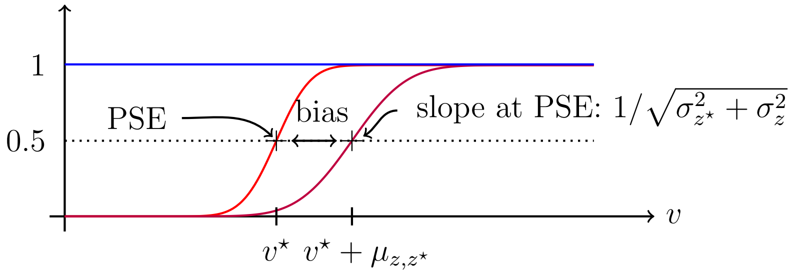


Figure 4.1: The shape of the psychometric function follows the estimation of the two speeds by Bayesian inference 2.1. This figure illustrates Proposition 16. The bias ensues from the difference between the bias on the two estimated speeds.

whereas in Proposition 13 the posterior is obtained by integration. Importantly, in both cases the bias is the same while the standard deviation of the posterior is not scaled by $\sqrt{2}$ in the following Proposition.

Proposition 16. *In the special case of the estimator* (2.1) *with a parameterization* (2.2)*, one has*

$$\varphi_{v^\star,z^\star}(v,z)=\psi\left(\frac{v-v^\star-a_{z^\star}\sigma_{z^\star}^2+a_z\sigma_z^2}{\sqrt{\sigma_{z^\star}^2+\sigma_z^2}}\right) \tag{4.1}$$

where $\psi(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t}e^{-s^2/2}\mathrm{d}s$ *is the cumulative normal function of sigmoid shape.*

Proof. One has the closed form expression for the MAP estimator

$$\hat{v}_z(m)=m-a_z\sigma_z^2,$$

and hence, denoting $\mathcal{N}(\mu,\sigma^2)$ the Gaussian distribution of mean μ and variance σ^2,

$$\hat{v}_z(M_{v,z})\sim\mathcal{N}(v-a_z\sigma_z^2,\sigma_z^2)$$

where $\sim$ means equality of distributions. One thus has

$$\hat{v}_{z^\star}(M_{v,z^\star})-\hat{v}_z(M_{v^\star,z})\sim\mathcal{N}(v-v^\star-a_{z^\star}\sigma_{z^\star}^2+a_z\sigma_z^2,\sigma_{z^\star}^2+\sigma_z^2),$$

which leads to the results by taking expectation. □

Fitting procedure In order to fit this model to our data we use a two-step method each consisting in minimizing the Kullback-Leibler divergence between the model and its samples (see Section II4.3). Numerically, the Nelder-Mead simplex method implemented in the python library `scipy` has been used. Before going further let us introduce

$$\varphi^{a,\sigma}_{v^\star,z^\star}(v,z) = \psi\left(\frac{v - v^\star - a_{z^\star}\sigma^2_{z^\star} + a_z\sigma^2_z}{\sqrt{\sigma^2_{z^\star} + \sigma^2_z}}\right),$$

$$\varphi^{\mu,\Sigma}_{v^\star,z^\star}(v,z) = \psi\left(\frac{v - v^\star + \mu_{z^\star,z}}{\Sigma_{z^\star,z}}\right)$$

$$\text{and}\quad \mathrm{KL}(\hat{p}|p) = \hat{p}\log\left(\frac{\hat{p}}{p}\right) + (1-\hat{p})\log\left(\frac{1-\hat{p}}{1-p}\right)$$

where $\mu_{z^\star,z} = a_z\sigma^2_z - a_{z^\star}\sigma^2_{z^\star}$, $\Sigma^2_{z^\star,z} = \sigma^2_{z^\star} + \sigma^2_z$ and KL is the Kullback-Leibler divergence between samples $\hat{p}$ and model p.

- Step 1: for all $z, z^\star$, initialize at a random point, compute

$$(\hat{\mu}, \hat{\Sigma}) = \underset{\mu,\Sigma}{\operatorname{argmin}} \sum_v \mathrm{KL}(\hat{\varphi}_{v^\star,z^\star}|\varphi^{\mu,\Sigma}_{v^\star,z^\star})$$

- Step 2: solve the linear relation shown above between $(\hat{\mu}, \hat{\Sigma})$ and $(\hat{a}, \hat{\sigma})$

- Step 3: initialize at $(\hat{a}, \hat{\sigma})$, compute

$$(\hat{\hat{a}}, \hat{\hat{\sigma}}) = \underset{a,\sigma}{\operatorname{argmin}} \sum_{z,z^\star}\sum_v \mathrm{KL}(\hat{\varphi}_{v^\star,z^\star}|\varphi^{a,\sigma}_{v^\star,z^\star})$$

Remark 3. *This method is coupled with a repeated stochastic initialization for the first step in order to overcome the number of local minima encountered during the fitting process. The approach was found to exhibit better results than a direct and global fit (third point). The potential problem of KL fits producing misleading results after convergence to local minima made it necessary to extend the empirical data by generating synthetic analogous data from the psychometric fits. Through this process detailed in Section 4.2 a more robust test of the validity of the analysis can be carried out.*

Remark 4. *Note that in practice we perform a fit in a log-speed domain ie we consider $\varphi_{\tilde{v}^\star,z^\star}(\tilde{v},z)$ where $\tilde{v} = \ln(1 + v/v_0)$ with $v_0 = 0.3°/\mathrm{s}$ following [184].*

4.2 Results on Synthetic Data

To avoid the dangerous aspect of undefined local minima convergence during KL fitting to empirical data, the quality of fitting can be assessed more objectively on derived synthetic data. The parameters a_z and σ_z were chosen so that they reproduce the increasing behavior of $\mu_{z^\star,z} = a_z\sigma_z^2 - a_{z^\star}\sigma_{z^\star}^2$. Then, the values of the psychometric functions $\varphi_{v^\star,z^\star}^{a;\sigma}(v,z)$ at the experimental points (v_1,z_1) and (v_2,z_2) described in Section 2.1 and rows $(A1)$ and $(A2)$ of Table 2.1 were used as the parameters of a binomial distribution from which we can generate any number n_b of blocks of 10 repetitions. The ten corresponding psychometric curves are shown in Figure 4.2 along with their fitted version. Following the fitting procedure described above in 4.1, we show in Figures 4.3 and 4.4 our results for $(\hat{a},\hat{\sigma})$ and $(\hat{\hat{a}},\hat{\hat{\sigma}})$. The quality of fitting naturally increases with the number of blocks, this effect is most striking for the likelihood width. The fitted log-prior slope shows a significant offset that is due to the under determination of the linear relations between $(\hat{\mu},\hat{\Sigma})$ and $(\hat{a},\hat{\sigma})$. Indeed solutions of the associated linear system lies in one dimensional affine space. However, even though the true values of a_z remain intractable the decreasing behavior of a is well captured within the trends generated by the synthetic data sets and by implication the same trends are valid in the empirical data.

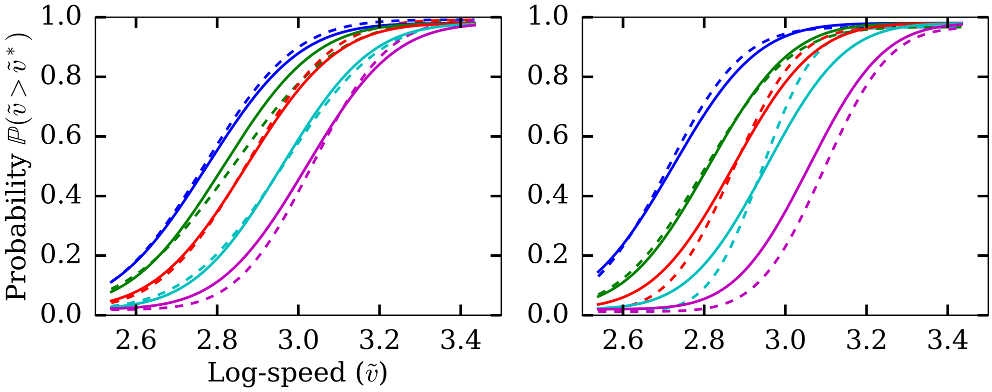


Figure 4.2: On the left the psychometric curves that simulate case $A1$, on the right the psychometric curves that simulate $A2$. Simulated psychometric curves resulting from the synthetic data are represented by the plain lines and the empirically fitted psychometric curves are represented by the dotted lines.

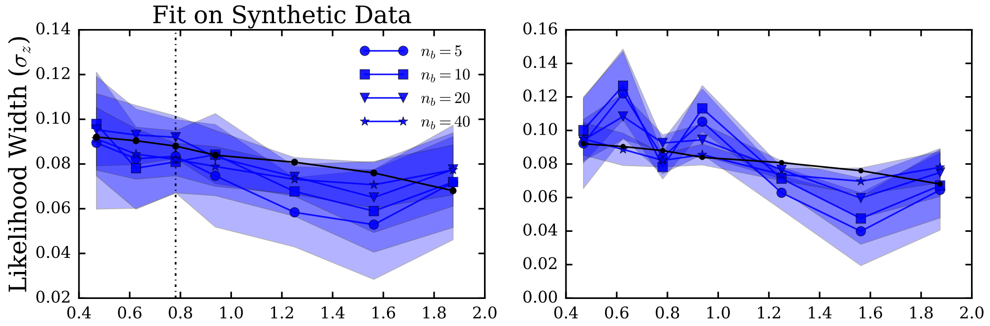


Figure 4.3: On the left the likelihood width $\hat{\sigma}$ obtained after the first optimization step 4.1, on the right the likelihood width $\hat{\hat{\sigma}}$ obtained after the third optimization step 4.1. These estimations are represented for different numbers of block with one standard deviation error. The black line represents the ground truth values of the likelihood widths used to generate the synthetic data.

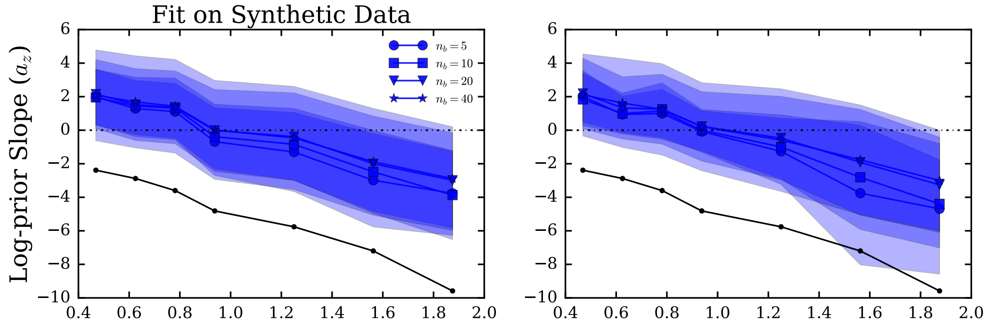


Figure 4.4: On the left the prior slope $\hat{a}$ obtained after the first optimization step 4.1, on the right the prior slope $\hat{\hat{a}}$ obtained after the third optimization step 4.1. These estimations are represented for different numbers of block with one standard deviation error. The black line represents the ground truth values of the prior slopes used to generate the synthetic data.

5 Experimental Data and Discussion

5.1 Results on Experimental Data

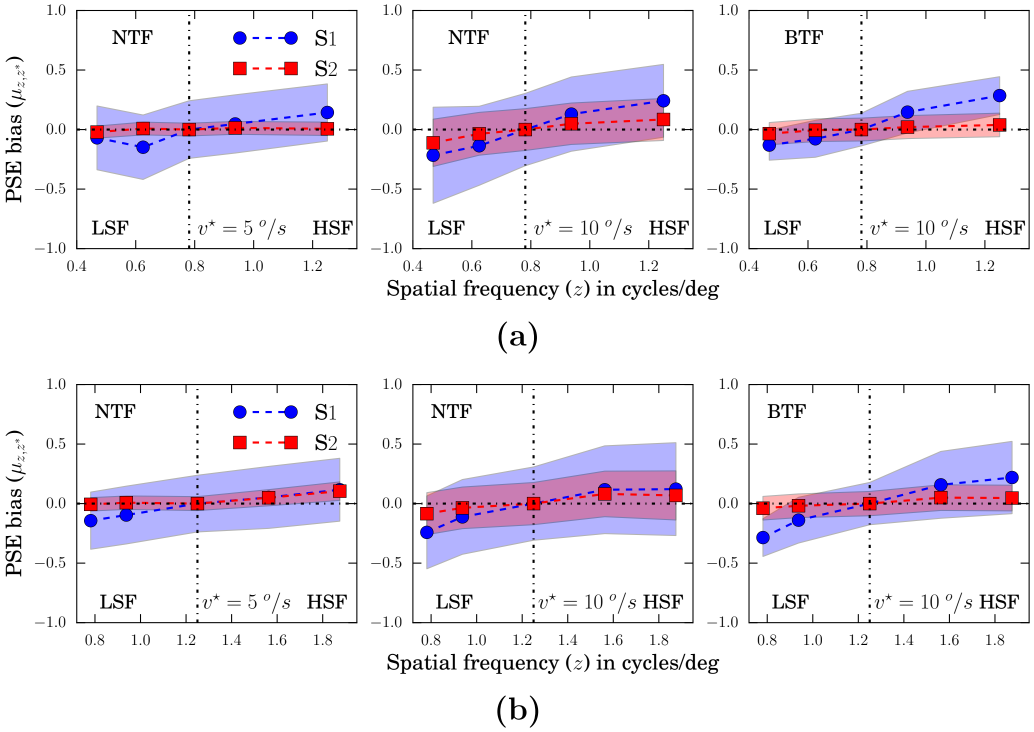
Estimating speed in dynamic visual scenes is undoubtedly a crucial skill for the successful interaction of any animal with its environment. Human judgements of perceived speed have therefore generated much interest, and been studied with a range psychophysics paradigms. The different results obtained in these studies suggest that rather than computing a veridical estimate, the visual system generates speed judgements influenced by contrast [189], speed range [190], luminance [79], spatial frequency [27, 178, 180] and retinal eccentricity [80]. There are currently no theoretical models of the underlying mechanisms serving speed estimation which capture this dependence on such a broad range of image characteristics. One of the reasons might be that the simplified grating stimuli used in most of the previous studies do not shed light on the possible elaborations in neural processing that arise when more complex stimulation. Such elaborations, such as nonlinearities in spatio-temporal frequency space can be seen in their simplest form even with a superposition of a pair gratings [152]. In the current work, we used our formulation of motion cloud stimuli which allowed the separate parametric manipulation of peak spatial frequency (z), spatial frequency bandwidth (B_z, σ_z) and stimulus lifetime ($t^\star$) which is inversely related to the temporal variability. The stimuli are all broadband, closer resembling visual inputs under natural stimulation. In the plotted data, we avoid cluttering by restricting traces to a subset of data, S1/S2, from the pair of participants who completed the full set of parametric conditions. Our approach was to test fewer participants (4) but under several parametric conditions using a large number trials analyzed alongside the synthetic data. The data that is not plotted here shows trends that lie within the range of patterns seen from S1/S2.

Before going into the details of analysis let us introduce convenient abbreviations.

- NTF/BTF: Narrow/Broad band Temporal Frequency;
- LSF/HSF: Low/High Spatial Frequency.

Cycle-controlled bandwidth conditions The main manipulation in each case was the direct comparison of the speed of a range of five stimuli in which the central spatial frequency was varied between five values, but all other parameters were equated under the different conditions. In a first manipulation in which bandwidth was controlled by fixing it at a value of 1 c/° for all stimuli (conditions A* and B* in Table 2.1), we found that lower frequencies were consistently perceived to be moving slower than higher frequencies (see Figure 5.1). The bias was generally smaller at 5 °/s than at 10 °/s (compare first column on the left with remaining two columns). This trend was the same for both the lower and the higher spatial frequency ranges used in the tasks (see Table 2.1 for details) when we compare the top row, Figure 5.1**(a)** with the bottom row, Figure 5.1**(b)**. This means the effect generalizes across the two scales used. The temporal variability of the stimulus manipulated via $t^\star$

was found to increase the variability of the bias estimates, though this did not significantly increase the biases (compare the shaded errors in the pair of plots in both the second and the third columns of Figure 5.1).



(a)

(b)

Figure 5.1: *Relative perceived speeds from the Point of Subjective Equality (PSE).* **(a)** From left to right A1, A3, B1. **(b)** From left to right A2, A4, B2. Task generates psychometric functions which show shifts in the point of subjective equality for the range of test z. Stimuli of lower frequency with respect to the reference (intersection of dotted horizontal and vertical lines gives the reference stimulus) are perceived as going slower, those with greater mean frequency are perceived as going relatively faster. This effect is observed under all conditions but is stronger for subject 1. Error bars are computed from those obtained for $(\hat{a}_z, \hat{\sigma}_z)$ which explains their amplitude. In case of a direct fitting of $\mu_{z,z^\star}$ they are significantly smaller (not shown).

Octave-controlled bandwidth conditions The octave-bandwidth controlled stimuli of conditions C* (see Table 2.1), allowed us to vary the spatial frequency manipulations (z) in a way that generated scale invariant bandwidths exactly as would be expected from zooming movements towards or away from scene objects (see Figure 2.1). Thus if trends seen in Figure 5.1 were the result of ecologically invalid fixing of bandwidths at 1 c/° in the manipulations, this would be corrected in the current manipulation. Only the higher frequency comparison range was used. We found that the trend was the same as that seen in Figure 5.1, indeed higher spatial frequencies were consistently perceived as faster than lower ones, shown in Figure 5.2. Interestingly, for the bandwidth controlled stimuli, the biases do not change across speed conditions (compare left column with right hand side columns of Figure 5.2). A small systematic change in the bias is seen with the manipulation of $t^\star$, reducing temporal variability going from the upper to the lower row reduces the measured biases. The bias at the highest frequency averaged for S1/S2 is equal to 0.13 for $t^\star = 100$ ms (BTF) and equal to 0.08 for $t^\star = 200$ ms (NTF).

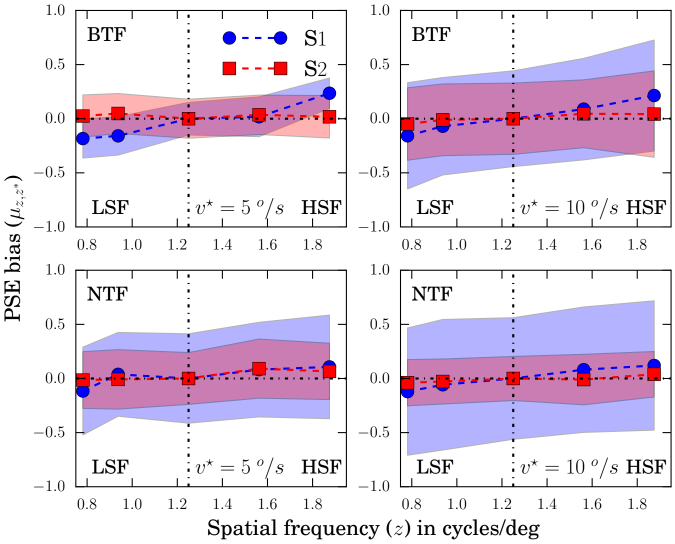
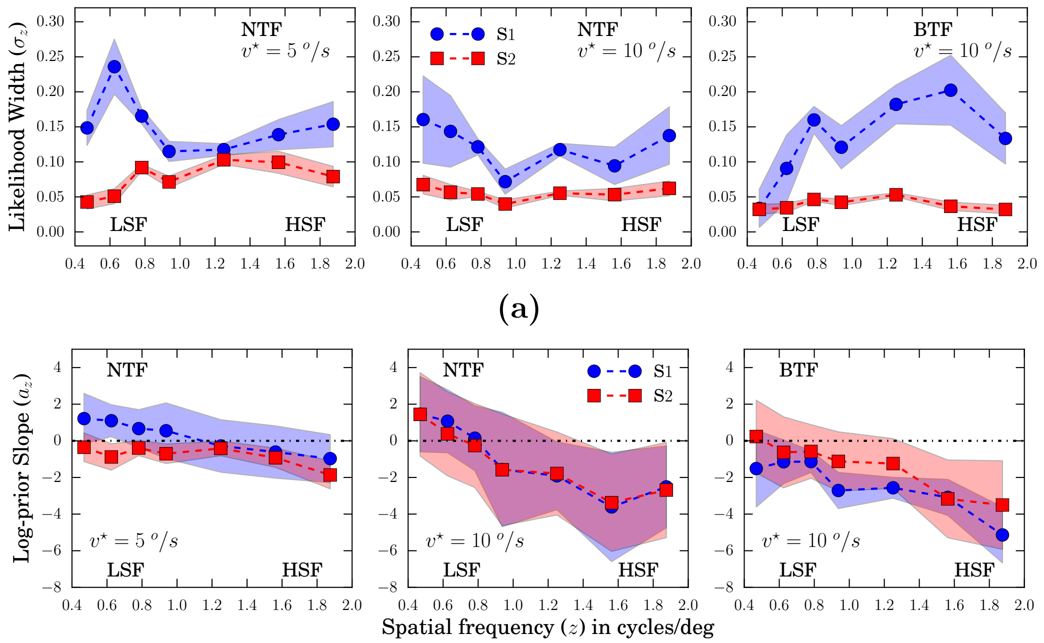


Figure 5.2: *Relative perceived speeds from the Point of Subjective Equality (PSE).* Top: C1, C2. Bottom: C3, C4. Same comment as Figure 5.1. The effect does not appear for subject 2 in case C2 and C3. Error bars are computed from those obtained for $(\hat{a}_z, \hat{\sigma}_z)$ which explains their amplitude. In case of a direct fitting of $\mu_{z,z^\star}$ they are significantly smaller (not shown).

Measured biases and corresponding sensory likelihoods and priors We used the Bayesian formulation detailed in Section 4.1 to estimate the likelihood widths and the corresponding prior slopes under the tested experimental conditions. There is no systematic trend within the likelihoods in the cycle-bandwidth controlled condition fits in Figure 5.3**(a)** and there is also individual variability in the trends. We conclude that the sensory variability of the speed estimates obtained from the Bayesian modeling cannot explain the spatial frequency driven bias in perceived speed that is measured. The log prior slopes show a systematic reduction as spatial frequency is increased, see in Figure 5.3**(b)**. Under all conditions, the data is best explained by a decreasing log prior as spatial frequencies are increasing. Under the octave-bandwidth controlled stimulus condition, the trends in changes in the best fitted likelihoods as the spatial frequency is increased are again not systematic (Figure 5.4**(a)**). The log prior slopes do however show a small systematic reduction as spatial frequencies are increased, in Figure 5.4**(b)**. The slopes are less steep than under the cycle-bandwidth manipulations (linear regression gives an average of -2.08 for the log-prior slopes in Figure 5.3**(b)** and -1.31 for the log-prior slopes in Figure 5.4**(b)**). Under both bandwidth configurations, we conclude that the prior slope explains at least part of the systematic effect of spatial frequency on perceived speed.

5.2 Insights into Human Speed Perception

We exploited the principled and ecologically motivated parameterization of MC to study biases in human speed judgements under a range of parametric conditions. Primarily, we considered the effect of scene scaling on perceived speed, manipulated via central spatial frequencies in a similar way to previous experiments which had shown spatial frequency induced perceived speed biases [27, 179]. In general, our experimental result confirmed that higher spatial frequencies were consistently perceived to be moving faster than compared lower frequencies; the same result reported in a previous study using both simple gratings and compounds of paired gratings, the second of which can be considered as a relatively broadband bandwidth stimulus [27]. In that work, they noted that biases were present, but slightly reduced in the compound (broadband) stimuli. That conclusion was consistent with a more recent psychophysics manipulation in which up to four distinct composite gratings were used in relative speed judgements. Estimates were found to be more veridical as bandwidth increased by adding additional components from the set of four, but increasing spatial frequencies generally biased towards faster perceived speed even if individual participants showed different trends [96].



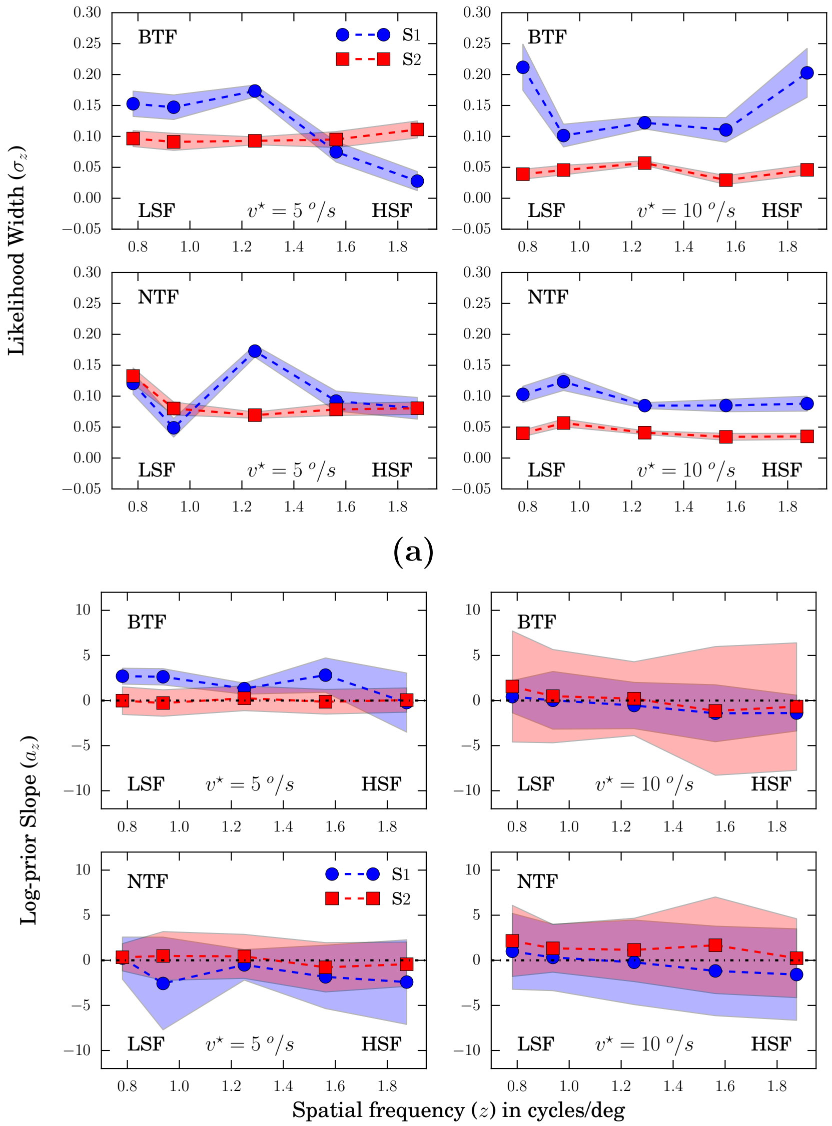
(a)

(b)

Figure 5.3: *Likelihood widths and log-prior slopes.* **(a)** Likelihood widths for A1-A2, A3-A4 and B1-B2. Likelihood widths do not show any common behavior, different behavior are observed for subject 1 whereas it is almost constant for subject 2. **(b)** Log prior slopes for A1-A2, A3-A4 and B1-B2. Despite the amplitude of error bars the log prior slopes have a common decreasing behavior in all subjects and in all cases.

Indeed, findings from primate neurophysiology studies have also noted that while responses are biased by spatial frequency, the tendency towards true speed sensitivity (measured as the proportion of individual neurons showing speed sensitivity) increases when broadband stimulation is used [152, 147].

It is increasingly being recognized that linear systems approaches to interrogating visual processing with single sinusoidal luminance grating inputs represents a powerful, but limited, approach to studying speed perception as they fail to capture the fact that naturalistic broadband frequency distributions may support speed estimation [27, 124, 125]. A linear consideration for example would not account for the fact that estimation in the presence or multiple sinusoidal components results in non-linear optimal combination [96]. The current work sought to extend the body of previous work by looking at spatial frequency induced biases using a parametric configuration in the form of the motion clouds which allowed a manipulation across a continuous scale



(a)

(b)

Figure 5.4: *Likelihood widths and log-prior slopes.* **(a)** Likelihood widths. Top: C1, C2. Bottom: C3, C4. Same as Figure 5.3(a). **(b)**Log prior slopes. Top: C1, C2. Bottom: C3, C4. Same as Figure 5.3(b) except for subject 2 in case C3.

of frequency and bandwidth parameters. The effect of frequency interactions across the broadband stimulus defined along the two dimensional orthogonal spatio-temporal luminance plane to allowed us to measure the perceptual effect of the projection of different areas (e.g. see Figure 2.2) onto the same speed line. The measurement should rely on proposed inhibitory interactions which occur during spatio-temporal frequency integration for speed perception [178] which cannot be seen with component stimuli separated by octaves [96].

We used a faster and slower speed because previous work using sinusoidal grating stimuli had shown that below the slower range (< 8 °/s), uncertainty manipulated through lower contrasts caused an under estimation of speeds while at faster speeds (> 16 °/s) it caused an overestimation [190, 79]. Our findings show that under the cycle-controlled bandwidth conditions, biases were larger at the faster speed than the slower ones while under the octave controlled bandwidths, the biases were almost identical for both speeds. The projections made from the frequency plane onto the speed line at these two speeds, once corrected with a scale invariance assumption, was therefore the same at these two speeds which typically show differences in contrast manipulations. Indeed the Bayesian fitting did not identify a systematic shift of either likelihood or prior slope parameters that could explain the biases observed particularly for the bandwidth controlled condition. While the current work does not resolve the ongoing gaps in our understanding of speed perception mechanisms particularly as it did not tackle contrast related biases, it showed that known frequency biases in speed perception also arise from orthogonal spatial and temporal uncertainties when RMS contrast is controlled. Bayesian models such as the one we applied, which effectively project distributions in the spatiotemporal plane onto a given speed line in which a linear low speed prior applies [184] may be insufficient to capture the actual spatiotemporal priors. Indeed the Bayesian models which successfully predict speed perception with more complex or composite stimuli often require various elaborations away from simplistic low speed priors [96, 182]. Indeed even imaging studies considering the underlying mechanisms fail to find definitive evidence for the encoding of a slow speed prior [202].

5.3 Conclusions

We used the MC stimuli in a psychophysical task and showed that these textures allow one to further understand the processes underlying speed estimation. We used broadband stimulation to study frequency induced biases in visual perception, using various stimulus configuration including octave bandwidth and RMS contrast controlled manipulations which allowed us to manipulate central frequencies as scale invariant stimulus zooms. We showed that measured biases under these controlled conditions were the same at both a faster and a slower tested speed. By linking the stimulation directly to the standard Bayesian formalism, we demonstrated that the sensory representation of the stimulus (the likelihoods) in such models can be described directly from the generative MC model. The widely accepted Bayesian model which assumes a slow speed prior showed that the frequency interactions could not be fully captured by the current formulation. We conclude that an extension to that formulation is needed and perhaps a two dimensional prior acting on the frequency space and mediated by underlying neural sensitivity has a role to play in computational modeling of complex spatiotemporal integration behind speed perception. We propose that more experiments with naturalistic stimuli such as MCs and a consideration of more generally applicable priors will be needed in future.

⋆ IV ⋆

A Review of Supervised Classification Techniques

This Chapter presents techniques from supervised machine learning classification to further analyze two types of brain recordings in Chapters V and VI. First, we introduce classification as a statistical inference problem and describe three different approaches: a deterministic approach, and two probabilistic approaches. The difference between the probabilistic approaches lies in the fact that one of them uses a discriminative probability model while the other uses a generative model combined with Bayes formula. We then details four different algorithm based on the generative approach – Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (GNB), Nearest Centroid (NC) – and one based on the discriminative approach – Logistic Classification (LC). We use these approaches in the following chapters. Finally, we define useful tools for the analysis conducted in the following chapters, as well as an original error classification measure.

Contents

1 Introduction

1.1 Generalities on Classification

In statistics and machine learning, supervised learning tackles the problem of modeling the relationship between features x in $\mathcal{X}$ to labels y in $\mathcal{Y}$. To put it briefly, it aims at inferring a function that maps any feature x to its label y based on the knowledge of $(x_i, y_i)_{i \in I}$. In absence of any assumption over the set $\mathcal{Y}$, the problem is known as supervised regression or simply regression. When we assume that $\mathcal{Y}$ is finite the problem becomes a classification issue, also known as supervised classification. Contrary to unsupervised classification (also called clustering) the class of features is known *a priori*. There is a great diversity of approaches and algorithms that are explored both theoretically and practically; we refer to the following handbooks for detailed descriptions [198, 199, 197, 93, 81]. There are approaches that search for empirical rules based on the data to build a decision tree [166]. Others aim at separating the data by a frontier; this is the principle of Support Vector Machine (SVM). Finally, some methods rely on simple estimations of statistics of each class like Nearest Centroid (NC) or k-Nearest Neighbors (kNN), when others are built on a parametric probabilistic model like Logistic Classification (LC) or Quadratic and Linear Discriminant Analysis (QDA/LDA). The performances of these methods might vary significantly and depend largely on the type of data and the evaluation criteria that are used. We refer to [30, 29, 107] and the references therein for an exemple of empirical evaluations. As a statistical tool, supervised learning is used in a variety of fields like social sciences [123], geology [138], finance [186], medicine [62], biology [85], *etc.*.

1.2 Contributions

From a mathematical point of view this Chapter provides very few contributions. We give some useful and sometimes original examples to the different supervised learning approaches. At the end, we give precise definitions of the different tools we use in the following chapters. In particular, we define a notion of distance between labels $(y_i)_{i\in I}$ and predicted labels $(\hat{y}_i)_{i\in I}$ based on optimal transport that takes into account the structure (when it exists) of labels. In summary, this Chapter is closer to a graduate course in machine learning than to a contribution to research. However, as an interdisciplinary work, this manuscript is not only addressed to mathematician and we find it necessary to set up the general problem of supervised classification and to introduce the different algorithms as particular cases of a common framework before we apply them in Chapters V and VI. The goal is to introduce these tools to experimental neuroscientists and psychophysicists so they can imagine relevant data analysis based on supervised learning. We provide the source code[1] of Examples 6 and 7 that illustrate Section 2.

2 Classification as a Statistical Inference Problem

For simplicity purposes in this section, we assume that $\mathcal{X}$ is a subset of $\mathbb{R}^n$ and $\mathcal{Y}$ is a finite subset of $\mathbb{N}$. We denote $(x_i, y_i)_{i\in I} \subset \mathcal{X} \times \mathcal{Y}$ the features extracted from the data and their associated labels. One can use the observed data as features but it is often necessary to process the raw data to remove at least outliers and obvious recorder noise. We detail the PCA feature selection technique in Section 5. In the following subsections, we introduce three approaches that tackle the problem of classification; the first is deterministic,

[1] `http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/examples_classif/`

the two others are probabilistic and based on either a discriminating model or a generative model.

2.1 Deterministic Approach

The goal is to find a deterministic function $f : \mathcal{X} \to \mathcal{Y}$ that maps any features $(x_i)_{i\in I}$ to their label $(y_i)_{i\in I}$ as well as possible in order to be able to predict the class $y = f(x)$ for any unknown observation $x \in \mathcal{X}$. Such a mapping f can be determined by minimizing a loss function $V : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that penalizes the distance between the two labels

$$\hat{f} = \operatorname*{argmin}_{f} \sum_{i\in I} V(y_i, f(x_i)). \tag{2.1}$$

In such a form, this problem is ill-posed and unstable due to potential noise in the data. For that reason, it is common to assume that functions f lie in a parametric space $F_{\mathcal{P}} = \{f_\theta | \theta \in \mathcal{P}\}$ where $\mathcal{P}$ is a space of parameters. Under such a hypothesis we obtain

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\mathcal{P}} \sum_{i\in I} V(y_i, f_\theta(x_i)). \tag{2.2}$$

Example 4. *In the binary classification case $\mathcal{Y} = \{-1, 1\}$, the desired mapping can be chosen as a linear function even if it does not take its value in $\mathcal{Y}$. In this case the parameters is $\theta = w$ and it lies in $\mathcal{P} = \mathcal{X}$, therefore we have $f_w(x) = \langle w, x\rangle$. By taking the loss to be the squared difference $V(y, f_w(x)) = (y - f_w(x))^2$, the problem comes down to a linear regression and it is therefore simple to estimate $\hat{w}$. Finally, we can make predictions using* $\operatorname{sign}(f_{\hat{w}})$ *as the line defined by $\langle \hat{w}, x\rangle = 0$ aims at separating the two classes, see Figure 2.1.*

2.2 Discriminative Approach

The discriminative approach relies on a conditional probabilistic model for a feature to belong to a particular class. We denote the associated density $\mathbb{P}_{Y|X}$. Therefore, the function mapping f emerges in a maximum likelihood framework, for all $x \in \mathcal{X}$,

$$f(x) = \operatorname*{argmax}_{y\in\mathcal{Y}} \mathbb{P}_{Y|X}(y|x). \tag{2.3}$$

Although the function mapping – being defined as a maximum over a finite set – is easy to compute, the difficulty lies in designing the probabilistic model.

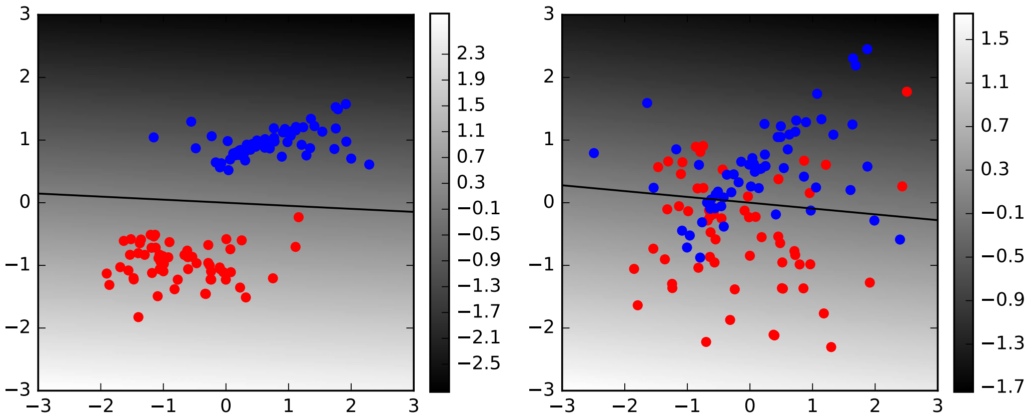


Figure 2.1: Two illustrations of the method introduced in Example 4. The gray levels are the values of the linear function $x \mapsto \langle w, x \rangle$ and the black line represents its zeros set. Left: the two classes are linearly separable. Right: the two classes are not linearly separable.

Again, it is appropriate to assume that the desired densities are parametrized $\left\{\mathbb{P}_{Y|X,\theta} | \theta \in \mathcal{P}\right\}$ and that the data $(x_i, y_i)_{i \in I}$ are i.i.d.. Hence, the problem becomes tractable and the parameter $\hat{\theta}$ that best represents the density that generates the data can be estimated using a Maximum Likelihood Estimation (MLE).

Definition 4. *Assume that observations $(x_i, y_i)_{i \in I}$ are i.i.d. and have the conditional density $\mathbb{P}_{Y|X,\theta}$. Their associated likelihood is*

$$\mathcal{L}(\theta) \overset{\text{def.}}{=} \prod_{i \in I} \mathbb{P}_{Y|X,\theta}(y_i | x_i). \tag{2.4}$$

Denoting $\tilde{V}(y, x, \theta) = -\log\left(\mathbb{P}_{Y|X,\theta}(y|x)\right)$, the negative log-likelihood is

$$\ell(\theta) \overset{\text{def.}}{=} \sum_{i \in I} \tilde{V}(y_i, x_i, \theta). \tag{2.5}$$

Given these definitions, the maximum of the likelihood or equivalently the minimum of the negative log-likelihood are written

$$\hat{\theta} = \underset{\theta \in \mathcal{P}}{\operatorname{argmax}}\ \mathcal{L}(\theta) = \underset{\theta \in \mathcal{P}}{\operatorname{argmin}}\ \ell(\theta). \tag{2.6}$$

In the form of a minimum, the optimization is consistent with Equation (2.2). As a consequence, $\tilde{V}$ can be interpreted as a modified loss function. In particular when there exists a loss function V, a parameter θ, a function f_θ and two constants (c_1, c_2) such that

$$\tilde{V}(y, x, \theta) = c_1 V(y, f_\theta(x)) + c_2 \tag{2.7}$$

the deterministic approach and the discriminative approaches are equivalent, see Example 5. However, these are strong hypotheses that are generally not verified, see Example 6 below.

Example 5. *In the binary classification case, although this does not respect the binary assumption, we can assume that the label y is the realization of a Gaussian random variable with unknown mean $\langle w,\, x\rangle$ and known variance σ^2. The parameter $\theta = w$ lies in the space $\mathcal{P} = \mathcal{X}$. In short, the density is written*

$$\mathbb{P}_{Y|X,w}(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \langle w,\, x\rangle)^2}{2\sigma^2}\right)$$

which yields to the following loss

$$\tilde{V}(y, x, w) = \log(\sqrt{2\pi}\sigma) + \frac{1}{2\sigma^2}(y - \langle w,\, x\rangle)^2.$$

Thus, $\tilde{V}(y, x, w) = c_1 V(y, f_\theta(x)) + c_2$ with $V(y, y') = (y - y')^2$, $\theta = w$, $f_\theta = \langle \theta,\, w\rangle$, $c_1 = 1/(2\sigma^2)$ and $c_2 = \log(\sqrt{2\pi}\sigma)$. This probabilistic approach is therefore equivalent to Example 4.

Example 6. *Again, consider a binary classification problem. We assume a paremetrization $\theta = (w_{-1}, w_1) = (-w, w) \in \mathcal{P} = \mathcal{X}^2$ and that the discriminative probability is given by*

$$\mathbb{P}_{Y|X,\theta}(y|x) = \frac{1}{\pi} \arctan\left(\langle w_y,\, x\rangle\right) + \frac{1}{2}.$$

This assumption is more realistic for a binary variable than in Example 5 since it is discrete. In this case the loss is

$$\tilde{V}(y, x, w) = -\log\left(\frac{1}{\pi} \arctan\left(\langle w_y,\, x\rangle\right) + \frac{1}{2}\right)$$

which cannot be set in the form of Equation (2.7)*. However the line $\langle w_1,\, x\rangle = 0$ can still be used to separate the two classes, see Figure 2.2.*

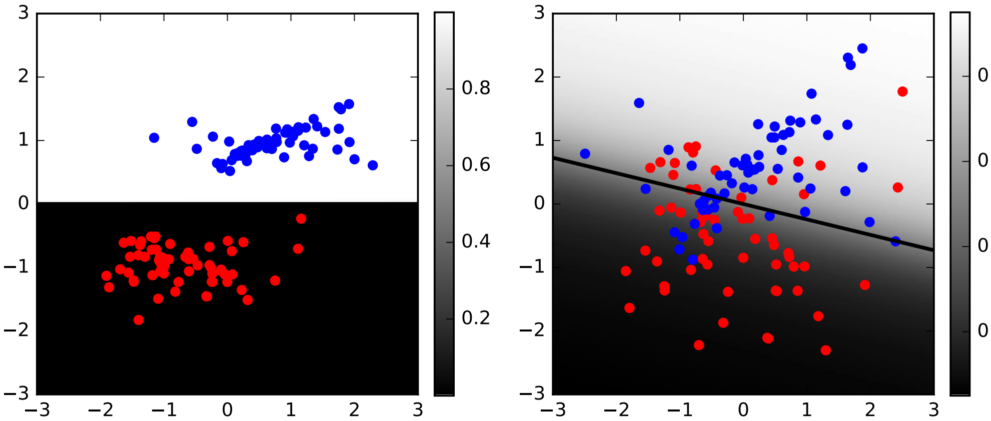


Figure 2.2: Two illustrations of the method introduced in Example 6. The gray levels are the values of the probability density function $x \mapsto \mathbb{P}_{Y|X,\theta}(1|x)$ and the black line represents the set where this probability is equal to 0.5. Left: the two classes are linearly separable. Right: the two classes are not linearly separable.

2.3 Generative Approach

Instead of searching for a conditional probability $\mathbb{P}_{Y|X}$, we can apply Bayes Theorem introduced in Chapter III. For all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\mathbb{P}_{Y|X}(y|x) = \frac{\mathbb{P}_{X|Y}(x|y)\mathbb{P}_Y(y)}{\mathbb{P}_X(x)} \tag{2.8}$$

Such a Bayesian approach provides a different expression for the decision rule (2.3),

$$f(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_{X|Y}(x|y)\mathbb{P}_Y(y). \tag{2.9}$$

Therefore the probabilistic model we are searching for is now twofold: on the one hand a generative model of the data knowing their class, on the other hand the occurrence probability of classes. Contrary to the direct design of a discriminant model $\mathbb{P}_{Y|X}$, the Bayesian Theorem allows to introduce some knowledge about the data conditioned by their class in $\mathbb{P}_{X|Y}$ and it takes into account the differences in proportion of each class in the discrete probability $\mathbb{P}_Y$. For the sake of simplicity, it is also convenient to assume that the generative probabilities lie in a parametric space $\left\{\mathbb{P}_{X|Y,\theta_1} | \theta_1 \in \mathcal{P}_1\right\}$. Such an assumption is implicit for discrete probabilities and we write $\{\mathbb{P}_{Y,\theta_2} | \theta_2 \in \mathcal{P}_2\}$ its parametric space. Thereby, denoting $\theta = (\theta_1, \theta_2) \in \mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2$, we can use the negative log-likelihood to estimate the parameter θ. Moreover, using the fact that following Equation (2.8)

$$\tilde{V}(y,x,\theta) = -\log\left(\mathbb{P}_{X|Y,\theta_1}(x|y)\right) - \log\left(\mathbb{P}_{Y,\theta_2}(y)\right) + \log\left(\mathbb{P}_X(x)\right), \tag{2.10}$$

the minimization of the negative log-likelihood comes down to

$$\hat{\theta} = \underset{\theta=(\theta_1,\theta_2)\in\mathcal{P}}{\operatorname{argmin}} \; -\sum_{i\in I} \log\left(\mathbb{P}_{X|Y,\theta_1}(x_i|y_i)\right) - \log\left(\mathbb{P}_{Y,\theta_2}(y_i)\right). \tag{2.11}$$

Example 7. *Once more, let us assume a binary classification problem. We use the parametrization $\theta_1 = (\mu_{-1}, \mu_1, \sigma_{-1}, \sigma_1) \in \mathcal{P}_1 = \mathbb{R}^2 \times \mathbb{R}^{\star 2}$ and consider the following generative probability*

$$\mathbb{P}_{X|Y,\theta_1}(x|y) = \begin{cases} \frac{\sigma_y^2}{2\pi Z} \exp\left(-\frac{1}{1-\|\sigma_y x - \mu_y\|^2}\right) & \text{if} \quad \|\sigma_y x - \mu_y\| \leqslant 1, \\ 0 \quad \text{otherwise,} \end{cases}$$

where $Z = \int_0^1 s \exp\left(-\frac{1}{1-s^2}\right) \mathrm{d}s$ is a normalizing constant. In the case where each class has the same probability (ie that $\forall y \in \mathcal{Y}, \mathcal{P}_Y = 1/2$) the minimization of the negative log-likelihood writes as follows

$$\hat{\theta} = \underset{\theta_1\in\mathcal{P}_1}{\operatorname{argmin}} \sum_{i\in I} \frac{1}{1 - \|\sigma_{y_i} x_i - \mu_{y_i}\|^2} - 2\log(\sigma_{y_i}).$$

Since the negative log-likelihood is convex and sufficiently smooth, $\hat{\theta}$ can be computed using conjugate gradient descent. An illustration is given in Figure 2.3.

3 Gaussian Generative Analysis

The goal of this section is to introduce the particular framework of Gaussian generative classification that results in simpler classifiers when particular assumptions are made on the covariance. We assume that $\mathcal{X} = \mathbb{R}^n$ and we also set $\mathcal{Y} = \{1, \ldots, K\}$ *ie* there are K classes. The parameters $\theta = (\theta_1, \theta_2)$ of the distributions are $\theta_1 = (\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$ and $\theta_2 = (p_1, \ldots, p_K)$, hence we write the densities as

$$\mathbb{P}_{X|Y,\theta_1}(x|y) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_y)}} \exp\left(-\frac{1}{2}(x-\mu_y)^T \Sigma_y^{-1} (x-\mu_y)\right)$$
$$\text{and} \quad \mathbb{P}_{Y,\theta_2}(y) = p_y \quad \text{with} \quad \sum_{y\in\mathcal{Y}} p_y = 1.$$

The negative log-likelihood ℓ follows from Definition 4 and Equation (2.11)

$$\ell\left(\theta\right) = \frac{1}{2} \sum_{i\in I} n\log(2\pi) + \log(\det(\Sigma_{y_i})) + (x_i - \mu_{y_i})^T \Sigma_{y_i}^{-1} (x_i - \mu_{y_i}) + \log\left(p_{y_i}\right)$$

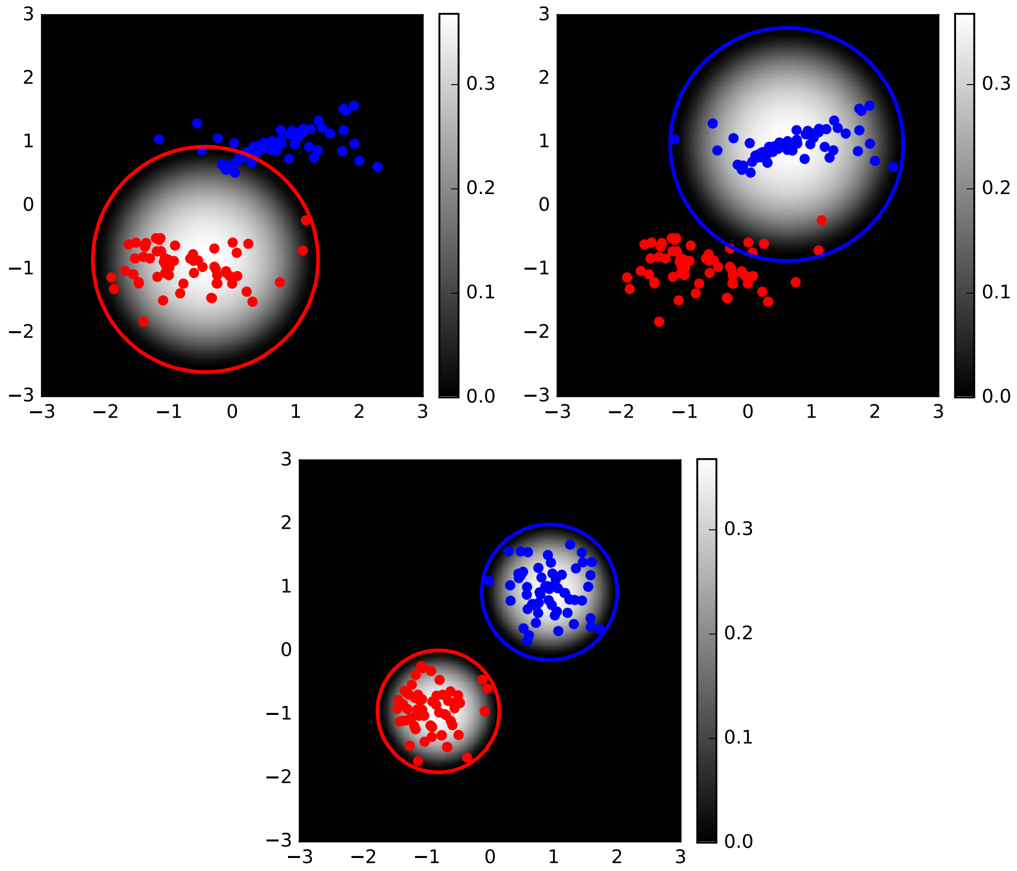


Figure 2.3: Illustrations of the method introduced in Example 7. The gray levels are the values of the probability density functions $x \mapsto \mathbb{P}_{Y|X,\theta}(x|1)$ (right) and $x \mapsto \mathbb{P}_{Y|X,\theta}(x|-1)$ (left). The red and blue circles represents the set where these probabilities are equal to 0.0001. Down: circularly separable data. Top: non-circularly separable data.

Proposition 17. *The negative log-likelihood ℓ reaches its minimum at $(\hat{\mu}, \hat{\Sigma}, \hat{p})$*

$$
\begin{aligned}
\forall y \in \mathcal{Y}, \quad \hat{\mu}_y &= \frac{1}{N_y} \sum_{i \in I} x_i \delta_{y_i}^y, \\
\hat{\Sigma}_y &= \frac{1}{N_y} \sum_{i \in I} (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^T \delta_{y_i}^y \\
\textit{and} \quad \hat{p}_y &= \frac{N_y}{|I|}
\end{aligned}
$$

where $\delta_k^l = \begin{cases} 1 & \textit{if} \quad k = l \\ 0 & \textit{if} \quad k \neq l \end{cases}$ *and* $N_y = \sum_{i \in I} \delta_{y_i}^y$.

Proof. The proof is standard and appear in under-graduate machine learning course. □

We can now use the decision rule (2.9) to classify any data. Note that the prior probability $\mathbb{P}_Y$ is estimated by the relative proportion of each class among the dataset which shifts the decision toward the most represented classes. This approach is also known as Quadratic Discriminant Analysis (QDA) [93] although it is based on a generative model $\mathbb{P}_{X|Y}$ and not on a discriminant model $\mathbb{P}_{Y|X}$. The QDA is an attractive option and appears to be more efficient than a linear classifier, as the decision boundaries are quadratic curves. However, the potential high dimension of features compared to the number of samples has a harmful effect on its performance as it becomes difficult to estimate the covariance matrix. Indeed, when the dimension n of the features is higher than the number of samples $|I|$, the estimated covariance matrix $\hat{\Sigma}$ is not full-rank and then is singular. In other cases it can still be ill-conditioned. Note that in the method described above, the dimension of θ_1 is $Kn(n+1)$. This phenomenon is also known as the "curse of dimensionality" and we refer to [105] for details. The dimensionality issue is generally mitigated by dimension reduction methods, sparsity or relevent subspace selection [19, 20, 173, 154]. In the following sections, we make three different assumptions that aim at reducing the number of parameters in the covariance matrix.

3.1 Linear Discriminant Analysis

In the Linear Discriminant Analysis (LDA) framework, all the classes share the same covariance matrix, in other terms

$$\exists \Sigma \in \mathcal{S}_n(\mathbb{R}), \quad \forall y \in \mathcal{Y}, \quad \Sigma_y = \Sigma. \tag{3.1}$$

Such a hypothesis can make the classification more tractable in high dimension as there is only one covariance Σ to estimate from all samples whatever their class. However, the dimension of the parameter $\theta_1 = (\mu_1, \dots, \mu_K, \Sigma)$ can still be far higher than the size of the dataset. For instance, this dimension is equal to $n(n+K)$, *ie* the covariance matrix dimension plus the mean of each of the K different classes.

3.2 Gaussian Naive Bayes

The Gaussian Naive Bayes (GNB) approach takes its name from the assumption, called "naive", that each component of a feature vector is independent from the others. Formally,

$$\forall y \in \mathcal{Y}, \quad \exists(\sigma_{1,y}, \ldots, \sigma_{n,y}) \in \mathbb{R}_+^n, \quad \Sigma_y = \operatorname{diag}(\sigma_{1,y}, \ldots, \sigma_{n,y}) \tag{3.2}$$

where $\operatorname{diag}(\sigma_{1,y}, \ldots, \sigma_{n,y})$ is the diagonal matrix with coefficients $(\sigma_{1,y}, \ldots, \sigma_{n,y})$. Under such an assumption the dimension of the covariance matrices is no more quadratic in n and the dimension of θ_1 is reduced to $2Kn$.

3.3 Nearest Centroid

The Nearest Centroid (NC) method does not require such a heavy framework. In fact, it can be described as a deterministic approach in which the function $f : \mathcal{X} \mapsto \mathcal{Y}$ is prescribed by the user. As its name indicates, it consists in computing the estimated centroid of each class $\hat{\mu}_y$ following the formula in Proposition 17 and classifying a sample x according to nearest centroid rule *ie*

$$f(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \|x - \hat{\mu}_y\|. \tag{3.3}$$

However, it is interesting to relate this rule to the discriminant analysis framework to see it as a simple particular case. Indeed, if we assume a common covariance Σ (3.1) that is diagonal (3.2) and an equal occurrence of each class (*ie* that $\forall y \in \mathcal{Y}, p_y = 1/K$) then

$$f(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} \ \exp\left(-\frac{1}{2}(x - \hat{\mu}_y)^T \hat{\Sigma}^{-1}(x - \hat{\mu}_y)\right)$$

which is equivalent to a weighted nearest centroid rule. In the particular case where we do not optimize for the weights and set them to 1 *ie* $\Sigma = \mathrm{Id}_n$, formula (3.3) is equivalent to

$$f(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} \ \exp\left(-\frac{1}{2}\|x - \hat{\mu}_y\|^2\right)$$

which is exactly the nearest centroid rule. In this framework, the dimension of θ_1 is $n(K+1)$ for weighted nearest centroid and nK for nearest centroid.

4 Multinomial Logistic Classification

The formulation of multinomial Logistic Classification (LC) resembles Example 6 and belongs to the discriminative approach presented in Section 2.2. A vector x belongs to class $y \in \mathcal{Y}$ with the following probability:

$$\mathbb{P}_{Y|X,\theta}(y|x) = \frac{e^{\langle x, \omega_y \rangle}}{\sum_{y' \in \mathcal{Y}} e^{\langle x, \omega_{y'} \rangle}} \tag{4.1}$$

where $\theta = (\omega_1, \dots, \omega_K)$ are called weight vectors and represent each class in $\mathcal{Y}$. Following Definition 4, the modified loss is

$$\tilde{V}(y, x, \theta) = -\langle x, \omega_y \rangle - \log \left(\sum_{y' \in \mathcal{Y}} e^{\langle x, \omega_{y'} \rangle} \right)$$

which yields, using the knowledge of samples and their labels $(x_i, y_i)_{i \in I}$, to the negative log-likelihood

$$\ell(\theta) = -\sum_{i \in I} \langle x_i, \omega_{y_i} \rangle + \log \left(\sum_{y' \in \mathcal{Y}} e^{\langle x_i, \omega_{y'} \rangle} \right).$$

As the function ℓ is C^2, it is possible to use fast optimization algorithms that use gradients and Hessian to compute estimates of the weight vectors. Usually, conjugate gradient and quasi-Newton with limited memory (L-BFGS) are used to perform this optimization. Moreover this is a convex problem which ensures the existence of a solution despite it cannot be written in closed form. We use Equation (2.3) to estimate the class of a feature x. Intuitively, when a vector x tends to be colinear with $\hat{\omega}_y$, the estimated probability $\mathbb{P}_{Y|X,\hat{\theta}}(y|x)$ is high and x most likely belongs to class y. The LC is always compared to Support Vector Machines (SVM, see [199], Chapter 5) and they offer similar results. However for our concern, we prefer LC because it provides a probabilistic model (although the probabilities obtained with LC are not reliable, especially in high-dimensional settings), the extension to multilabel classification appears natural and the optimization is a smooth convex problem.

5 Tools for Analysis

Before going into the details of analysis, let us introduce some useful concepts to have a full understanding of the results. As we saw in Section 2, each algorithm is able to predict the class $f(x) \in \mathcal{Y}$ of any feature $x \in \mathcal{X}$. To assess their classification performances we split a dataset into two parts the learning set $(x_i, y_i)_{i \in I_{\text{train}}} \in (\mathcal{X} \times \mathcal{Y})^{I_{\text{train}}}$ that is used to estimate the parameters of each algorithm and the test set $(x_i, y_i)_{i \in I_{\text{test}}} \in (\mathcal{X} \times \mathcal{Y})^{I_{\text{test}}}$ that is used to check the quality of predictions. This is measured by the score.

Definition 5 (Score). *The score ι_I over a particular test set I_{test} is defined as follows*

$$\iota_{I_{test}} \overset{\text{def.}}{=} \frac{1}{|I_{test}|} \sum_{i \in I_{test}} \delta_{y_i}^{f(x_i)},$$

where $|I|$ denotes the cardinal of I and δ_k^l is the Kronecker symbol defined in Proposition 17.

Another tool, named confusion matrix, allows to visualize the proportion of correctly predicted labels with respect to the ground truth labels.

Definition 6 (Confusion Matrix)**.** *The confusion matrix over a particular test set I_{test} is the matrix $M_{I_{test}} = \left(m_{y,y'}^{I_{test}}\right)_{(y,y')\in\mathcal{Y}^2}$ defined by*

$$m_{y,y'}^{I_{test}} \overset{\text{def.}}{=} \frac{\sum_{i\in I_{test}} \delta_y^{y_i} \delta_{y'}^{f(x_i)}}{\sum_{i\in I_{test}} \delta_y^{y_i}},$$

where δ_k^l is the Kronecker symbol.

If the confusion matrix verifies $M_{\mathrm{Id}_{\mathrm{test}}} = \mathrm{Id}_{|\mathcal{Y}|}$ then the classifier has perfectly predicted the classes of the features in the test set. The score and confusion matrix are subject to bias due to the choice of I_{train} and I_{test}. In order to reduce this bias, we perform a n_{folds} cross-validation procedure. Note that in our setting, the classes have the same number of samples, ensuring that $|I|/|\mathcal{Y}|$ is an integer.

Definition 7 (n_{folds} Cross-Validation)**.** *Let n_{folds} divide $|I|/|\mathcal{Y}|$. Assume that I is a partition of n_{folds} non overlaying test sets of equal cardinal i.e.*

$$I = \cup_{i=1}^{n_{folds}} I_{test}^{(i)} \quad \textit{with} \quad \forall i \neq j, \quad |I_{test}^{(i)}| = |I_{test}^{(j)}| \quad \textit{and} \quad I_{test}^{(i)} \cap I_{test}^{(j)} = \emptyset$$

The n_{folds} cross-validation consists in learning the algorithm parameters n_{folds} times using the training sets $I_{train}^{(i)} = I \backslash I_{test}^{(i)}$ for $i \in \{1, \dots, n_{folds}\}$.

The cross-validation allows to compute the average and standard deviation of the different scores over the folds. It also permits to compute an averaged confusion matrix. Finally, the averaged of the fitted parameter of each fold $\hat{\theta}_{I_{\mathrm{test}}^{(i)}} \in \mathcal{P}$ are computed both with its standard deviation.

Definition 8. *The average of the scores μ_ι and the standard deviation of the scores σ_ι are*

$$\mu_\iota = \frac{1}{n_{folds}} \sum_{i=1}^{n_{folds}} \iota_{I_{test}^{(i)}} \quad \textit{and} \quad \sigma_\iota = \left(\frac{1}{n_{folds}-1} \sum_{i=1}^{n_{folds}} (\iota_{I_{test}^{(i)}} - m_{\iota_I})^2 \right)^{\frac{1}{2}}. \tag{5.1}$$

The averaged confusion matrix is

$$\Lambda = \frac{1}{n_{folds}} \sum_{i=1}^{n_{folds}} M_{I_{test}^{(i)}}. \tag{5.2}$$

The average of the parameter and its standard deviation are

$$\hat{\theta}^{(av.)} = \frac{1}{n_{folds}} \sum_{i=1}^{n_{folds}} \hat{\theta}_{I_{test}^{(i)}} \quad \textit{and} \quad \hat{\theta}^{(st.d.)} = \left(\sum_{i=1}^{n_{folds}} \frac{(\hat{\theta}_{I_{test}^{(i)}} - \hat{\theta}^{(av.)})^2}{n_{folds}-1} \right)^{\frac{1}{2}}. \tag{5.3}$$

As we mention in Section 3, dimension can limit the performances of classifiers. In order to reduce the dimension of our data we use the Principal Component Analysis (PCA). This method aims at projecting the data onto a space of lower dimension n_{pca} spanned by the eigenvectors associated to the n_{pca} highest eigenvalues of the covariance matrix of the data.

Proposition 18. *Assume that the vectors $(x_i)_{i\in I}$ have a zero sample mean. Let $\mathtt{X}$ be the matrix whose columns are the vectors $(x_i)_{i\in I}$. Let $\mathtt{C} = \mathtt{X}^T\mathtt{X}/(\dim \mathcal{X} - 1)$ be the sample covariance matrix and $n_{pca} \leqslant |I|$.*

$$\exists \mathtt{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_{|I|}) \in \mathcal{M}_{|I|}(\mathbb{R}) \quad \textit{and} \quad \exists \mathtt{P} \in \mathcal{O}_{|I|}(\mathbb{R}), \quad \mathtt{C} = \mathtt{PDP}^T$$

with $\lambda_1 \geqslant \dots \geqslant \lambda_{|I|}$. Denote $\tilde{\mathtt{P}}$ the matrix whose columns are the n_{pca} first columns of $\mathtt{P}$. Then, the data of lower dimension $(\tilde{x}_i)_{i\in I}$ are the columns of $\tilde{\mathtt{X}} = \tilde{\mathtt{P}}^T\mathtt{X}$.

Finally, we define an original error measure of classification. In particular, this error measure is useful when the classes have structure. For instance, in an visual perception experiment, we associate a recorded signal to a common class when it is obtained under the same stimulation. When the parameters of two stimulations are close these stimulations are close. It can be then harder for supervised learning algorithm to discriminate between close stimulations than between distant stimulations. We face this in the two following chapters.

Definition 9. *Let $\Lambda = (\Lambda_{y,y'})_{(y,y')\in\mathcal{Y}^2}$ be the averaged confusion matrix and $(\theta_y)_{y\in\mathcal{Y}}$ the parameters associated to the different classes. Moreover, assume that we have a appropriate distance d between the parameters $(\theta_y)_{y\in\mathcal{Y}}$. Then, we define the error measure as*

$$d_p = \sum_{(y,y')\in\mathcal{Y}^2} d(\theta_y, \theta_{y'})\Lambda_{y,y'}.$$

Figure 5.1 displays examples of error measure for different confusion matrices in order to get an idea of the values it takes in simple cases. In the following chapter, we consider parameter $\theta_y = (\theta_y^{(1)}, \theta_y^{(2)}) \in \mathbb{R} \times \mathbb{R}/l\mathbb{Z}$. Therefore, we use the following distance between parameters:

$$d(\theta_y, \theta_{y'}) = \frac{1}{2M_1}|\theta_y^{(1)} - \theta_{y'}^{(1)}| + \frac{1}{2M_2}\max\left(|\theta_y^{(2)} - \theta_{y'}^{(2)}|, l - |\theta_y^{(2)} - \theta_{y'}^{(2)}|\right)$$

where

$$M_1 = \max_{y^{(1)},y'^{(1)}} (|\theta_y^{(1)} - \theta_{y'}^{(1)}|)$$

and

$$M_2 = \max_{y^{(2)},y'^{(2)}} \max\left(|\theta_y^{(2)} - \theta_{y'}^{(2)}|, l - |\theta_y^{(2)} - \theta_{y'}^{(2)}|\right).$$

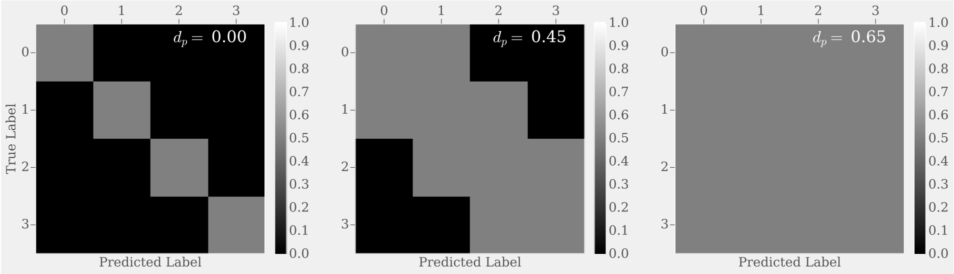


Figure 5.1: Some error measure d_p computed for different confusion matrices. Here $\theta_y = y \in \{0, 1, 2, 3\}$ and $d(\theta_y, \theta_{y'}) = |\theta_y - \theta_{y'}|/3$.

⋆ V ⋆

Supervised Classification For VSD Imaging

In this chapter, we first describe the human visual pathway, which is similar for several mammal species. We then describe intra-cortical processing and the properties of neurons and their organization. We present the principle of Voltage Sensitive Dye optical imaging (VSDi) and standard processing, and go on to reviewing the interactions between machine-learning and experimental neurosciences. We apply the algorithms described in Chapter IV to different datasets. First, we validate these methods by comparing their results to those of standard approaches. We provide an automatic method to determine the appropriate number of PCA components. Then, we introduce spatially and temporally localized analysis methods in order to identify relevant spatio-temporal features. Finally, we build a simple model of activation maps obtained under oriented stimulations.

Contents

1 Introduction

In this introduction, we first recall some basic knowledge about the visual system by describing its biological components at different scales and their interactions. Then, we focus on the VSDi recording technique. After a brief historical reminder, we give details about their principles and classical processing. We broadly review different machine learning approach that deal with data acquired in neurosciences. Finally, we detail our working environment and put our work in a more precise context before detailing our contributions.

1.1 Visual System and Intracortical Processing

This Chapter aims at analyzing brain recordings in order to probe vision. We therefore first outline the functional organization of the visual system from large to fine scales. This review will help to situate the origins of acquired data using VSDi and ER. Our description is valid for many mammals, however our data are collected in cats.

1.1.1 The Visual Pathway

First, visual information reaches the eyes in the form of a light beam, which is projected onto the retina. The retina is covered in photoreceptor cells containing protein molecules called opsins (rods and cones), which are able to absorb the photons that compose light and to transmit a signal to bipolar and ganglion cells. Then, the signal, in the form of action potentials, is conducted through the optical nerves to the Lateral Geniculate Nucleus (LGN), which is itself connected to the visual cortex through axons. The visual cortex is composed of different areas called V1, V2,..., V5/MT, etc. They are hierarchically connected to each other and basically correspond to higher and higher levels of processing. The connectivity differs from one species to another, however we generally distinguish feedforward and feedback connectivity. Feedforward connections link areas from low to high processing levels (LGN $\rightarrow$ V1 $\rightarrow$ V2 $\rightarrow$ ...) whereas feedback connections link areas from high to low processing levels (... $\rightarrow$ V2 $\rightarrow$ V1 $\rightarrow$ LGN). In fact, connectivity is a more complex process; for instance, the V1 area has direct feedforward connections to MT, which itself has direct feedback connections to V1, V2, etc. We refer to the review [28] for further details. Figure 1.1 shows a drawing of the visual pathway from the retina to the primary visual cortex (V1). The recordings analyzed in this chapter are performed in the cat's primary visual cortex (V1), which is known to process some basic information about perceived images like position, orientation, spatial and temporal frequencies. In the next paragraph we explain how the primary visual cortex is organized.

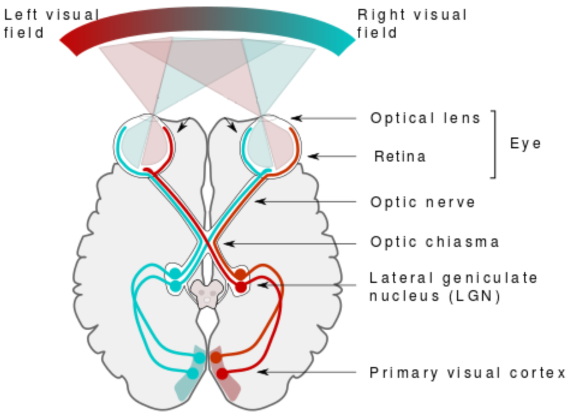


Figure 1.1: A simplified schema of the human visual pathway. Source: Miquel Perello Nieto (Creative Commons license), url: `https://commons.wikimedia.org/wiki/File%3AHuman_visual_pathway.svg`.

1.1.2 A Layered Cortex

The cortex is organized into six layers, mainly identified by the characteristics of the neurons they contain (size, distribution). For example, layers II-III contain small- to medium-sized pyramidal neurons whereas layer V contains large pyramidal neurons. The six layers form a neural network in which we distinguish horizontal connections from vertical connections. Horizontal connections link neurons of the same layer together whereas vertical connections link neurons from different layers. The distinction is important because, in the primary visual cortex (V1), layers IV-VI are known to receive most of the output connections from the LGN. The neural signal is then conducted to layers II-III through verticals connections. When the signal is carried by horizontal connections we talk about lateral propagation or diffusion waves, whereas when the signal is carried by vertical connections, we talk about standing waves. Figure 1.2 gives realistic and cartoon views of the cortex layers.

1.1.3 Receptive Fields and Columnar Organization

Before starting to establish the functional organization of V1, one has to understand in which way V1 neurons are encoding visual information. In 1959, Hubel and Wiesel published their first paper about receptive fields observed in the cat's primary visual cortex [87]. They are not the first authors to use the concept of receptive fields; however, they provide strong experimental re-

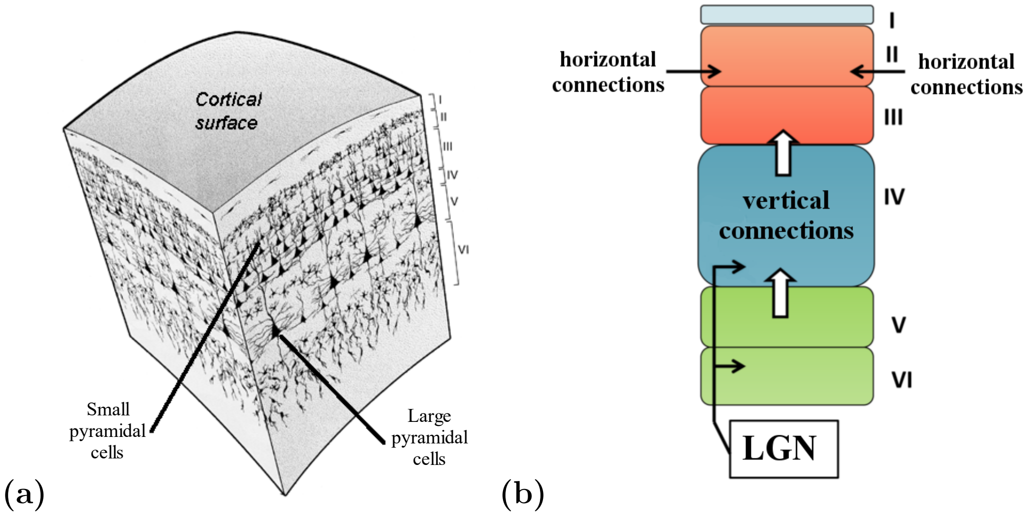


Figure 1.2: Cortex layers. **(a)** Distribution of neurons across layers: small pyramidal neurons in layer II and large pyramidal neurons in layer V. Source: online material of Professor Mike Claffey (UC San Diego), url: `http://mikeclaffey.com/psyc2/images/organization-cortical-layers.gif` **(b)** Horizontal connections in layers II-III. The LGN outputs arrive in layers IV-VI and reach other layers through vertical connections. Source: own work of Yannick Passarelli (Phd student at UNIC, CNRS).

sults and give many examples of their variety (see Figure 1.3). A receptive field characterizes the responses of a single neuron to visual stimulations. The activity of a neuron is measured by its spiking rate: when it increases, the neuron's response is excited, while when it decreases the neuron's response is inhibited. A receptive field is a local area of the visual field that produces variations in the spiking rate of the considered neuron. For example, the center of an area of cortex can be excitatory and its surround inhibitory. Depending on neurons, there are many possible arrangements of the local excitatory and inhibitory areas in the visual fields. Figure 1.3 shows different examples of receptive field arrangements observed by Hubel and Wiesel. Not all neurons show a receptive field with two or more clearly identified excitatory and inhibitory regions, this is why they separated neurons in two categories: simple cells, that have a simple receptive field with two distinct excitatory and inhibitory regions and complex cells. In V1, a remarkable property of receptive fields is their orientation. Following the idea of cortical columns introduced by Mountcastle *et al.* [128], Hubel and Wiesel found that receptive fields sharing a same orientation are distributed in a slab shape perpendicularly to the cortex surface [88]. Two different columnar organizations have now been thoroughly mapped using optical imaging techniques: orientation columns and ocular-dominance columns. As explained in Section 1.1.2 there exists horizontal connections in layers II-III. These connections are organized in a very specific way, as shown by Bosking *et al.* [18]. Neurons in the same orientation columns are mostly connected to neurons in adjacent orientation columns and to neurons in co-oriented, co-axially aligned columns.

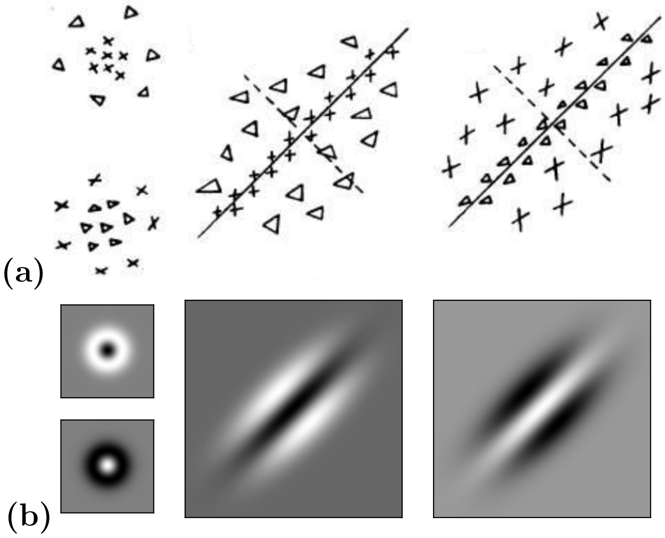


Figure 1.3: **(a)** Examples of receptive fields extracted from [89]. **(b)** Their corresponding examples using a mathematical formulation (Gabor functions, Gaussian functions and their derivatives). Triangles and white: excitatory region. Crosses and black: inhibitory region.

1.1.4 Toward a Model of the Primary Visual Cortex

Strong of the receptive field concept, models of the visual cortex are always built on the first *a priori* linear transform made by neurons: the linear dot product between their receptive field and the stimulus. The result of such a linear transform is always passed trough a non-linearity to be converted to a firing rate. Such a model is known as the Linear Non-Linear Poisson spiking model (LNP) [165]. Other models are explained and compared by Vintch *et al.* [203], they show that a cascade Linear Non-Linear Linear Non-Linear (LNLN) model outperforms three alternative models: the LNP model, the

Spike Trigger Covariance based LNP (see Section 1.4.1 and [172]) and the Energy model (see [4]). Using a similar cascade (a LNLN model called Nonlinear Input Model (NIM)), McFarland *et al.* [121] successfully explain variety of neuronal responses and perform better than the STA and STC based LNP models.

All these models underlie the receptive field concept. However, Fournier *et al.* [55] show that receptive fields are stimulus dependent: they can be more simple-like or more complex-like depending on the stimulus statistics. Moreover, these approaches are purely bottom-up *ie* they model the parallel feedforward computations performed by the visual cortex neglecting the local recurrent computations. As the review of Fregnac *et al.* [59] suggests, it is promising to develop a mathematical framework that goes beyond the receptive field concept and that takes local feedback into account. However, these developments are beyond the scope of this chapter which aims at showing how useful machine learning techniques can be for the analysis of neurophysiological and neuroimaging data.

1.2 Principles of Voltage Sensitive Dye Optical Imaging (VSDi)

In this chapter, we focus on data obtained with Voltage Sensitive Dye Optical Imaging (VSDi). Before going into the details of how informative they are about brain functions, we briefly recall some background about these methods.

The development of voltage sensitive molecules provides neuroscientists with a new way to observe neurons activity [205]. Embedded in a dye, these molecules called “fluorophores” are able to bind to neuron membranes over a small area of cortex from which one could monitor its activity. The local electrical field variations on the dentritic membranes in superficial layers cause small re-emited fluorescence that is captured by a highly sensitive optical camera, see [33]. A recording consists in a video of a few squared millimeters of cortex for a duration of several hundreds of milliseconds at a spatial resolution down to dozens of micrometers and a temporal resolution of a few milliseconds. Such a spatial extent corresponds to thousands of neurons, which provides a way to investigate their spatial organization at a population level *ie* at the mesoscale. Moreover, when studying vision, the spatial organization is directly related to visual field since the layer plane receives a homeomorphic projection of visual space (contralateral hemifield). Since the cortex consists of six layers, the fluorophores cannot penetrate deep and their major part binds in layers II-III (see Section 1.1). If intrinsic optical imaging –which detects luminance variation due to hemoglobin oxygenation changes (see [16] for a complete description)– can reach these spatial resolutions, its temporal resolution remains limited by the few seconds duration of the biological phenomenon on which it relies on. Therefore, the use of VSD in optical imaging offers a much refined temporal resolution to study the population dynamic. The first use of voltage sensitive dye (VSD) dates back to 1973 with the work of Davila, Salzberg, Cohen *et al.* that monitored the action potentials of single giant axons [37, 167]. Such a method now holds a place of choice for the study of cortical activity between large and small scale recordings (fMRI/EEG *vs* extra/intra-cellular recordings) [75]. Figure 1.4 outlines the experimental setup of VSD imaging used to study the visual cortex at UNIC laboratory. We refer to [76, 61] for practical methodology.

1.3 Processing of VSDi Data

In neuroscience, there are many recording modalities in addition to VSDi, each of them being related in a certain way to brain activity. For instance, electroencephalography (EEG) records the electric field generated by a large amount of neurons [187] whereas functional magnetic resonance imaging (fMRI) measures blood flow variations due to neurons activity [82]. The different signals correspond to different physical observables, they have their own drawbacks and are not directly comparable. Therefore, it is necessary to model the observed signals in order to discriminate the relevant information from the specific noise of the technique. This becomes more important when the amount of collected data is limited because the human and financial toll of the experimental protocols is significant. To this purpose, signal separation and dimensional reduction methods like Principal Component Analysis (PCA) and Indenpendant Component Analysis (ICA) or non-linear sparse methods have shown their usefulness. We review the recent approaches developed for VSDi. In particular, it is necessary to perform some pre-processing to get relevant features before applying more advanced machine learning techniques, especially for VSDi.

Despite the promising perspectives of the VSDi technique, the signal is perturbed by many biological and physical artifacts comprising animal breathing, heartbeat noticeable in blood vessels, decreasing sensitivity of fluorophores to light (the so-called “bleaching”) and optical noise of the camera. Experimenters overcome these drawbacks by including blank acquisitions (*ie* with no stimulation) in their protocols in order to compare other acquisitions to this blank reference signals. Although this method has shown its efficiency for further statistical analysis, it does not allow for a trial to trial analysis.

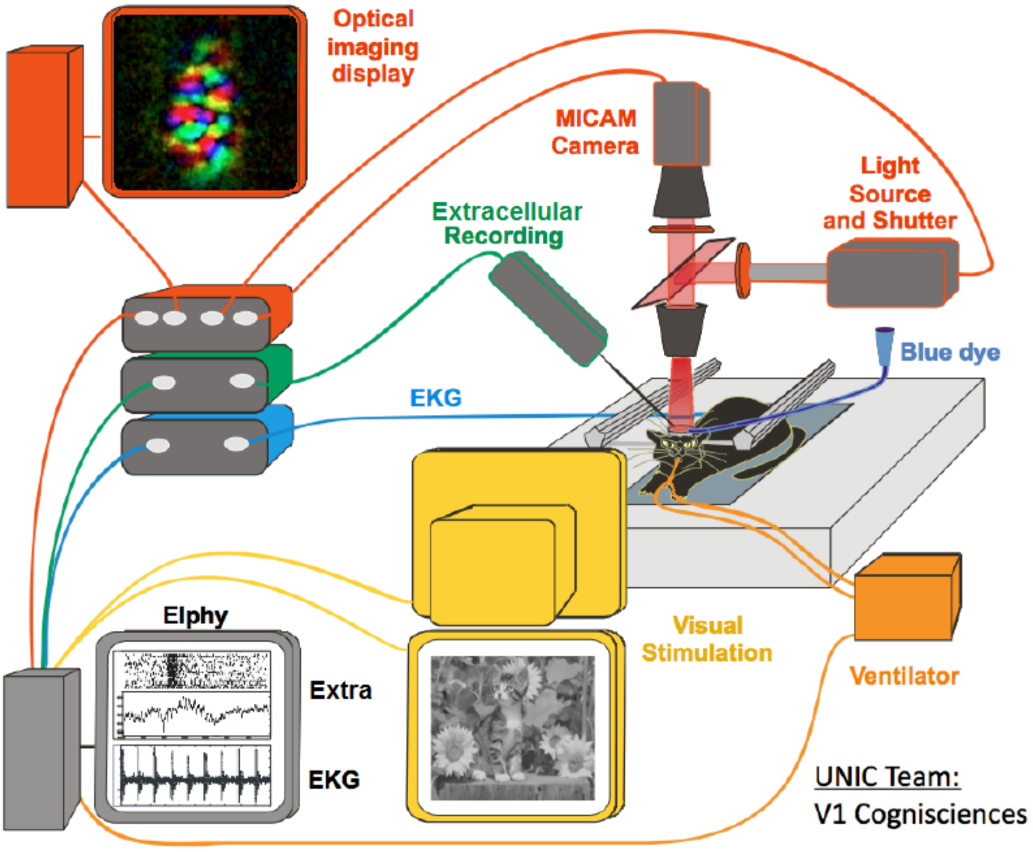


Figure 1.4: Typical layout of a VSD of ER experiment at UNIC lab. The visual stimulation equipment is represented in yellow. The optical imaging setup is represented in red. The extracellular recording setup is depicted in green. Finally, the biological monitoring materiel is depicted in orange and cyan. Modified figure from UNIC lab.

Some solutions coming from signal processing and in particular source separation analysis [159, 216] show good results using generalized linear models. In particular, the work of Raguet [155, 156] at UNIC tackles this problem using sparse regularization signal processing techniques. Focusing on spatial biological artifacts Fekete *et al.* [51] use the resemblance between artifact to identify and substract them. We also refer to the review [33] and references therein for details about the importance of modeling the VSDi signal.

Using VSDi, neuroscientists confirmed the existence of functional maps inside the cortex of many mammals, such maps are what we call columns in

Section 1.1.3. A functional or cortical map is the spatial representation of different areas of cortex activated when processing different features of the same operation, we use the term activation map when we refer to areas activated by the same feature. For instance, when the operation is to detect orientations, the different features are the different possible values of orientations. Although spatial resolution has increased compared to intrinsic signal imaging, the study of cortical maps does not benefit from the high temporal resolution of VSD imaging. Indeed, these maps are always computed by averaging the recording over time and repetition of the same experimental condition (*eg* orientation). Yet, it is crucial to investigate the dynamic aspect of such a population of neurons to better understand how a stimulus is encoded in a neural network and how information is propagated in the network. Let us know give a few examples of use of VSDi to analyze the dynamical neuronal activity. The paper [174] highligths the increasing-decreasing dynamic of the difference between preferred and orthogonal orientation responses. A two-layers neural field model [116] (excitatory and inhibitory neurons) can explain the spatio-temporal activity recorded during the presentation of different types of simple stimuli. An appropriate use of Singular Value Decomposition (SVD) shows that some components of the VSDi dynamic match the drifting-grating (DG) temporal frequency [142]. A comparison of VDSi activity evoked by DG and natural images reveals that natural stimuli continuously modulate the responses indicating more complex excitation/inhibition mechanisms [141]. In [32], VSDi is used to understand the lateral spread of orientation selectivity by comparing responses evoked by local/full-field and center/surround stimuli. The problem of the transient dynamic due to changes in motion direction is tackled in [212] where it is shown that cortical dynamic combined with population coding is well suited to encode these changes. This work supports the theory developed by Hansel [14]. In our work, we limit the preprocesing to debleaching by fitting exponential functions that are further removed from the signal. Sometimes it appears useful to perform a spatial Gaussian smoothing in order to remove spatial noise. We also apply some dimension reduction methods such as Principal Component Analysis (PCA) and make an extensive use of supervised learning to analyze the dynamic of VSDi recordings.

1.4 Previous Works: Machine Learning for fMRI and VSDi

Recently, there is a growing interest to apply machine learning data analysis tools to neurophysiology and neuroimaging data. These approaches appear fruitful for neuroscience, however they are not always known from biologists and they still require close collaboration with mathematicians, computer scientists, *etc.* In experimental contexts in which one collects data under different particular conditions, supervised classification appears to be well suited to evaluate the relevance of these conditions (viewed as labels). However, as we now detail, it is important to note that the Machine Learning (ML) techniques used in neurosciences applications vary considerably from one experimental setup to another.

1.4.1 fMRI

In fMRI data analysis, a lot of effort is dedicated to using ML techniques, typically in "brain reading"-like scenario. Thirion *et al.* [188] build an encoding/decoding model that allows to predict and even to reconstruct the stimuli presented to subject from its fMRI acquisition. Michel *et al.* [126] perform a feature selection method based on a mutual information criteria that increases the classification performances of the SVM classifiers. This allows for a better understanding of where the information is encoded in the brain. In a similar way, Gramfort and Thirion [73] make use of a TV-ℓ_1 penalization as a feature selector. Again, the work of Varoquaux *et al.* [200] makes feature selection using a randomized lasso based method which shows better results than standard ℓ_1/ℓ_2 penalization methods. The works of Jenatton [95] and Murphy [129] base their analysis on logistic classification to identify activation maps. The review [130] by Naselaris *et al.* lists the previous studies that use supervised learning, and exposes the limit of the approach, which is only a decoding tool and does not explain how signal is encoded. This highlights the importance of developping encoding models. The team of Gallant developped several approaches that rely on an encoding model to increase the predictive power of the decoding model. In the papers [101, 204], they predict the natural images perceived by subjects using linear and sparse models. Going further, Neselaris [131] and Nishimoto [136] combine motion-energy models (see [4] and Section III3) with a Bayesian decoder to reconstruct perceived natural images from brain activity.

1.4.2 VSDi

In optical imaging, the computation of activation maps [15], although usually not expressed this way, can be cast as a way to perform supervised classification. Indeed, it makes use of the label of each recording by computing the class centroids. However, such a computation does not provide any error quantification nor is used to make predictions. In intrinsic optical imaging, SVM are used to compute functional maps [214] and it shows significant improvement compared to class centroids computations that correspond to the method usually used. In VSDi, the work of Macke *et al.* [112] makes use of 2D Gaussian processes to estimate activation maps and directly provides relevant error bars on these estimates. Ayzenshtat *et al.* [7, 8] make use of SVM and kNN to perform classification and prediction in order to quantify how much information conveyed by neural activity is related to the stimulus. Using these classifiers, they also localize both in space and time the most predictive features. Finally, Briggman *et al.* [24] perform a LDA to identify neuronal populations that can discriminate before any single neurons. To the best of our knowledge this is the only work that makes use of a classification method to probe the dynamic of population activation. It is the purpose of our contributions to develop a principled ML framework to analyze VSDi data and in particular compute functional maps.

This review of the state of the art reveals that the usage of ML techniques in the fields of VSDi is very limited. It is one of the purposes of this chapter to explicitly and systematically propose ML-based solutions of the exploration of such VSDi datasets.

1.5 Contributions

1.5.1 Recording at UNIC

This chapter is the outcome of a strong interdisciplinary collaboration with the experimental and theoretical neuroscience team led by Yves Frégnac and Cyril Monier at UNIC laboratory. As a graduate student in applied mathematics, I did not have any specific training to work among neuroscientist experimenters. During the course of my PhD, as a team member, I have been confronted to the ordeals that are usually faced by experimenters when collecting data. Being able to conduct a full experiment requires various skills, from surgery and biological monitoring to electronic and signal processing. Although, many of these experimental aspects remained out of reach (especially surgery and fine biology) I have been able to punctually assist experimenters, for instance during anesthetic injections, fundus examination or contact lenses positioning. During the other steps, I was always invited to watch the intermediate states of the animal preparation, for instance scalp removing, craniotomy, electrodes descent or cortex staining. My role was to design and run the protocols involving MC developed in Chapter I by the use of the software Elphy used at UNIC. The discussions with experimenters were fundamental to understand the new parameters provided by MC and to decide on the optimal ordering of the stimulation protocols to run.

1.5.2 Main Contributions

The first major contribution of this chapter is an automatic method to select the number of components of the Principal Component Analysis (PCA) based on the classification performances (see Section 4.1). The second main contribution is a methodology of local space-time analysis of classification performances which enables to identify the most predictive pixels and to precisely quantify the temporal dynamic (see Section 4.2). The third major contribution is the definition of a simple and efficient model of the VSDi signal obtained when using oriented stimuli (see Section 5). In addition, we have several minor biological contributions related to the experimental protocols that we analyze. In particular, we find that activation of neural populations is faster when stimulated after a blank than when stimulated after a first oriented stimulus (see Section V4.3.2). Moreover, the simple proposed model supports the role of lateral connections for a neural population to handle an abrupt change of stimulus orientation (see 5.3). We provide an online[1] example of data synthesis using the proposed model. Moreover, additional Figures are also available online[2].

1.5.3 Related Works

In order to accurately define the context in which spots this chapter we go back to the references that are the most relevant for our work.

To our knowledge Ayzenshtat *et al.* [7, 8] are the only ones to make use of supervised classification as a complementary tool to analyze VSDi data. They assess the question of how much stimulus related information is conveyed in the VSD signal. They also assess the question of where this information is located in space and time by looking for the features that offer the best prediction performances. Indeed, these are natural questions that emerge when using a supervised classification approach and that are tackle in fMRI by Gallant and Thirion's teams, see Section 1.4.1. Briggman *et al.* [24] use supervised learning combined with PCA to make prediction over time and detect a discriminatory threshold. Their experiments are different from us because they are performed on the leech in which they record the activity of many dozens of neurons that are discernible. However, supervised learning help them to draw strong conclusions about population *versus* single neuron coding.

Benucci *et al.* [12] established that activity in the orientation domain emerges as a standing wave whereas in the spatial domain activity spreading indicates a traveling wave. Finally, Wu *et al.* [212] perform some experiment that make

[1] http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/model_vsd/

[2] https://jonathanvacher.github.io/chapV-supp.html

use of a similar protocol to the second one (random dot moving in a single direction abruptly rotated). They draw the conclusion that cortical dynamic and population coding combine to handle abrupt changes in motion direction.

2 Material and Methods

The analysis is performed on 6 different datasets recorded from cat primary visual cortex, area 17. Datasets 1 and 2 were obtained in 2000 at the Weizmann Institute of Science and were used in the publication by Chavane *et al.* [32] that we refer to for details. Datasets 3 to 6 are obtained at UNIC laboratory by Luc Foubert during my PhD. We describe the experimental details below for these datasets 3 to 6.

2.1 Animal Preparation for Dataset 3 to 6

Experiments are performed on adult cats, male or female, aged of 12 to 24 months, weighting 2 to 4 kg. Animals are initially anesthetized with intramuscular alfaxolone (1 mL kg^{-1}) followed by a cannulation of the femoral vein in order to anesthetized intravenously. After tracheotomy, animals are artificially respirated, continuously anesthetized with 1.5 % (0.5 % during recording) isoflurane added to the 1:2-1:3 mixture of O_2/N_2O. Minimum alveolar concentration is kept above 1 %. Animals are head fixed on an anti-vibration table. Craniotomies of about 1.5 cm diameter are performed above area 17 and 18 and the dura is resected. Paralysis is maintained by intravenous injection of rocuronium bromide (4 mg $\mathrm{kg}^{-1}\,\mathrm{h}^{-1}$ + glucose + Na/Cl) administered starting less than three hours before running protocols in order to suppress eye movements. Accommodation and pupil contraction is blocked by atropine and neosynephrine. Appropriate corrective optical lenses are added depending on the animal. Area centralis positions are measured before and after imaging. Stainless steel chambers are mounted and fixed using dental cement and the cortices are stained for 23 h with voltage-sensitive dye (RH-1691), unbound dye is washed out after staining. The chambers is then filled with CSF-saline or silicone oil and closed. Electrocardiogram (ECG), expired CO2, body temperature, and EEG are continuously monitored during the entire experiment.

2.2 Setup

Acquisition and visual stimulation is controlled by the Elphy software (Gérard Sadoc, CNRS), communicating with the acquisition program provided by the imaging system for VSD recordings. A CMOS MiCam camera is used, providing 100 × 100 pixel resolution and up to 10 kHz temporal resolution.

The recording is performed at 200 Hz temporal resolution *i.e.* a temporal sampling of $\Delta_t = 5$ ms (104.17 Hz for datasets 1 and 2 *i.e.* $\Delta_t = 9.6$ ms). One pixel in the recording corresponds approximately to 60 × 60 µm^2 of cortical sheet. Image acquisitions are synchronized with ECG and respiration signals. For detection of changes in fluorescence the cortex is illuminated with light of 630 nm.

2.3 Visual Stimulation

Data are collected under 3 different protocols described below and summarized in Figure 2.1. A protocol consists in presenting a number $C \in \mathbb{N}$ of stimuli variously parametrized. This operation is then repeated a certain number $R \in \mathbb{N}$ of times. Except for datasets 1 and 2, a LCD screen (ASUS) with a resolution of 1920 × 1080 pixels and a refreshing rate of 120 Hz was placed at 57 cm of the animal so that 1 cm on the screen is equal to one visual degree. All visual stimuli were generated with Elphy, maximum and background luminance were set at 40 $\mathrm{cd\,cm^{-2}}$ and 12 $\mathrm{cd\,cm^{-2}}$ respectively.

Dataset 1 and 2 Stimuli were contrast sinusoidal luminance gratings with a spatial frequency of 0.6 c/° and drifting in a single direction for 576 ms at speed $v_0 = 3.33\ ^\circ\,\mathrm{s}^{-1}$ starting 174 ms after recording onset. Four orientations were presented (0 °, 45 °, 90 ° and 135 °) both full field and locally, totaling $C = 8$ stimulation conditions. Local stimuli had a diameter of 2 ° and were presented at an eccentricity of 1-15 °, depending on the cortical area location exposed by the craniotomy. The stimuli presentation were pseudo-randomly interleaved and were displayed binocularly using VSG series three stimulator with 38 × 29 cm^2, 640 × 480 pixel^2 monitor, at a distance of 57 cm from cat's eyes at a refresh rate of 144 Hz. For each repetition of the height tested stimulation conditions a trial under blank stimuli was recorded.

Dataset 3 to 5 Stimuli are high contrast sinusoidal luminance gratings with a spatial frequency of 0.6 c/° and drifting in a single direction for 800 ms at speed $v_0 = 3.33\ ^\circ\,\mathrm{s}^{-1}$ starting 140 ms after recording onset. Four orientations (and two directions) are presented (0 °, 45 °, 90 ° and 135 °) and instantaneously rotated of +135 (resp. +90) ° 400 ms after stimulus onset, totaling $C = 8$ stimulation conditions. The stimuli presentation are pseudo-randomly interleaved and are displayed binocularly. For each repetition of the four tested orientations a trial under blank stimuli is recorded.

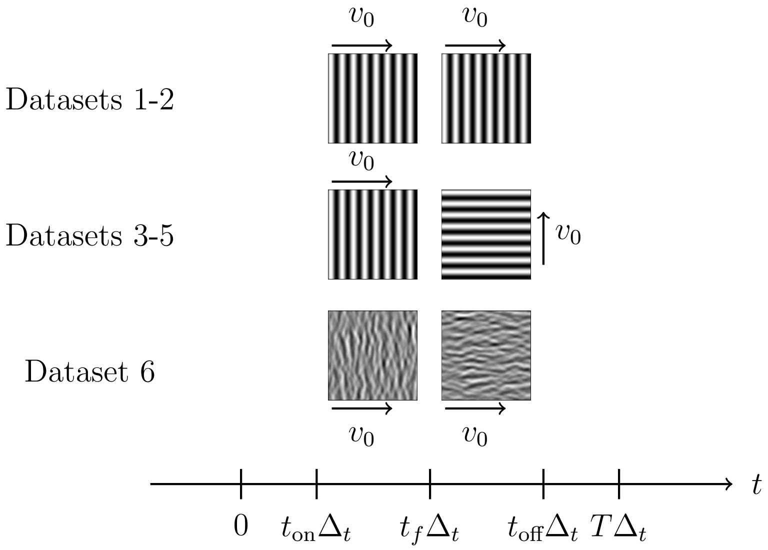


Figure 2.1: The 3 different protocols used to collect the data. The sampling time is Δ_t whereas $t_{\text{on}}, t_f, t_{\text{off}}$ and T are integers corresponding respectively to the frames numbers of stimulus onset, stimulus rotation, stimulus offset and recording offset.

Dataset 6 Stimuli are Motion Clouds with parameters $z_0 = 0.6$ c/°, $B_Z = 1.35, \sigma_V = \frac{1}{t^\star z_0}$ with $t^\star = 0.666$ ms and drifting in a single direction for 800 ms at speed $v_0 = 2.5$ c/s starting 140 ms after recording onset. Four orientations are presented ($\theta_0 = 0$ °, 45 °, 90 ° and 135 °) with four orientation bandwidths ($\sigma_\Theta = 0.5, 1.4, 2.0, 2.8$) and instantaneously rotated of +90 ° 400 ms after stimulus onset, totaling $C = 16$ stimulation conditions. The stimuli presentation are pseudo-randomly interleaved and are displayed binocularly. For each repetition of the sixteen tested stimulation conditions a trial under blank stimuli is recorded.

2.4 Preprocessing Using an Exponential Fitting

As we saw in Section 1.3, VSDi signal is corrupted by different artifacts (bleaching, heartbeat, breathing), we handle only the bleaching using a exponential fit to the data on every pixel temporal trace. Artifacts were already removed on dataset 1 and 2, see [32]. For other dataset we apply the following method. Due to the spatial heterogeneity of illumination during a recording, before any processing, we delimit a region of interest. To this purpose, we compute the mean frame by averaging a whole dataset over time, stimulation conditions and repetitions. Then, we keep the pixels that have values above

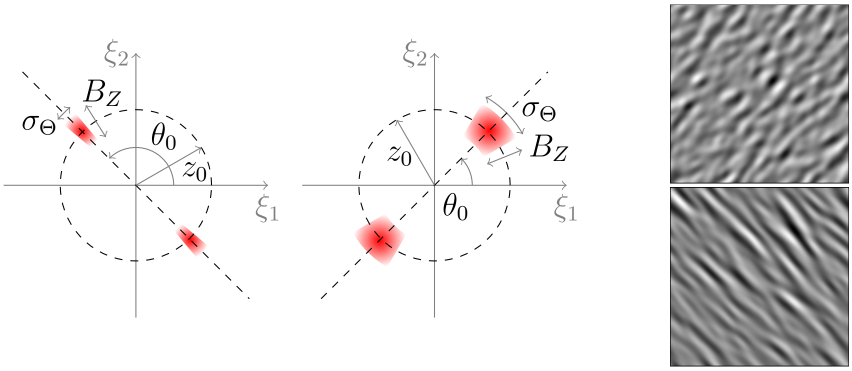


Figure 2.2: An example of 2 different stimuli tested in the recordings of dataset 6. Left: the spatial power spectrum with $(\theta_0, \sigma_\Theta) = (45\ °, 5\ °)$ and $(\theta_0, \sigma_\Theta) = (135\ °, 15\ °)$. Right: the two corresponding frames.

a certain percentage threshold r_{thresh} of the maximum. Then, let $A = (a_{q,t})_{q,t}$ denotes the pixel values of any recording and

$$h_{\kappa_q,\tau_q,\eta_q}(t) = \kappa_q \exp\left(-\frac{t}{\tau_q}\right) + \eta_q$$

denotes the bleaching model with $(\kappa_q, \tau_q, \eta_q)_q$ being the parameters to be fitted (t denotes the frame number and q denotes the pixel's position). In order to avoid the bias due to pixels' activation during stimulation, we only use the set $\hat{\mathcal{T}} = \{0, \ldots, t_{\text{on}}, t_{\text{off}} + 20, \ldots, T\}$ as samples. Therefore, we compute

$$(\tilde{\kappa}_q, \tilde{\tau}_q, \tilde{\eta}_q) = \underset{\kappa_q,\tau_q,\eta_q}{\operatorname{argmin}} \sum_{t \in \hat{\mathcal{T}}} \left(a_{q,t} - h_{\kappa_q,\tau_q,\eta_q}(t)\right)^2$$

using a gradient descent method. Finally, we remove these fitted functions to obtain the neuronal response signal as $S \overset{\text{def.}}{=} (a_{q,t} - h_{\tilde{\kappa}_q,\tilde{\tau}_q,\tilde{\eta}_q}(t))_{q,t}$. Figure 2.3 shows an example of such a preprocessing.

2.5 Recorded Datasets Organization

Such a signal S is computed for every stimulation condition $c \in \mathcal{C} = \{1, \ldots, C\}$ and repetition $r \in \mathcal{R} = \{1, \ldots, R\}$. In the following, we denote by $\mathcal{T} = \{0, \ldots, T\}$ the set of time samples and by $\mathcal{Q} = \{1, \ldots, Q_1\} \times \{1, \ldots, Q_2\}$ the set of pixels where $Q = Q_1 Q_2$. The dataset number $k \in \mathcal{K} = \{1, \ldots, 6\}$ (corresponding to the numbers of Section 2.3) is therefore denoted

$$S^{(k)} = (s^{(k)}_{q,t,c,r})_{(q,t,c,r) \in \mathcal{Q} \times \mathcal{T} \times \mathcal{C} \times \mathcal{R}}. \tag{2.1}$$

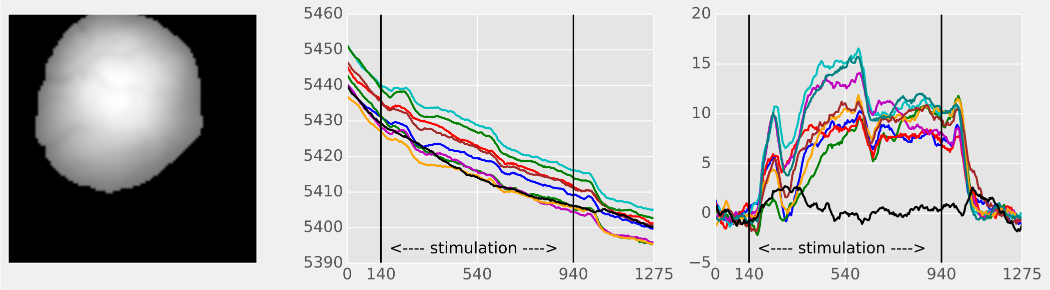


Figure 2.3: Left: the region of interest selected using r_{tresh}. Middle: the temporal traces of each condition averaged over all pixels and repetitions. Right: the temporal traces after preprocessing.

Formulating a Machine Learning Classification Problem For the sake of clarity and to make a proper connection with Section IV, we further describe the feature, label and sample spaces, respectively $\mathcal{X}$, $\mathcal{Y}$ and I previously introduced. These choices are critical to examine the data as they restrain the reading of the fitted model parameters to chosen feature and label sets. A dataset $S^{(k)}$ has 4 parameters and there is plenty of ways to define the features and their label $(x_i, y_i)_{i \in I} \in (\mathcal{X} \times \mathcal{Y})^I$. For instance, one can decide to use just a subset of the stimulation conditions $\mathcal{C}$ or to consider that time samples can be concatenated with the repetition set ignoring the time correlation, *etc.* In such a case, the drawn conclusions can only concern the spatial distribution of information. The simplest choice for the label set is to use $\mathcal{Y} = \mathcal{C}$ *i.e.* the stimulation conditions, however one could decide to group them in an appropriate way in order to avoid large number of classes. For example, dataset $S^{(6)}$ has $C = 16$ different conditions and thus as much classes, but one can decide to group the shared orientations in the different pairs of parameter $(\theta_0, \sigma_\Theta)$ reducing the number of classes to $|\mathcal{Y}| = 4$. Finally, it is essential to balance the dimension of the feature space $\dim \mathcal{X}$ and the cardinal of samples $|I|$ as there are only $R \in \{10, \dots, 30\}$ repetitions of the C different conditions in each dataset.

Cross-Validation We have defined the cross-validation using a partition of I. As in our context, there are various possibilities to decide for the feature and sample spaces, it is important to avoid breaking the independence assumption between the train and test sets to avoid overfitting. Indeed, for instance one can decide not to take into account the temporal dependencies and consider a set of samples $I = \mathcal{T} \times \mathcal{Y} \times \mathcal{R}$ (with $\mathcal{Y} \subset \mathcal{C}$) supposed to be independent when training an algorithm. In this case, taking an arbitrary partition of I is unfortunate because samples that come from different instants of the same recording can be both in the train and test sets, which will cause an artificial increase of scores. To avoid that artifact, cross-validation must always be performed on non-overlapping partitions of $\mathcal{Y} \times \mathcal{R}$.

Technical Details All the analysis that follow were performed using Scikit-Learn [146]. The numerical computations were run in parallel on the cluster of CEREMADE - Paris Dauphine University - PSL.

3 Comparison of the Different Algorithms

We compare in this section the classification performances, on different VSDi datasets, of the methods introduced in Chapter IV:

- Quadratic Discriminant Analysis (QDA, see Section 3),
- Linear Discriminant Analysis (LDA, see Section 3.1),
- Gaussian Naive Bayes (GNB, see Section 3.2),
- Nearest Centroid (NC, see Section 3.3),
- Logistic Classification (LC, see Section 4).

This comparison is performed by choosing the stimuli orientations as labels. As explained in Section 1, there exists an orientation maps in cat's visual cortex, therefore it should be possible to classify the frames of the recordings according to stimulation of different orientations.

3.1 Pixels as a Feature Space

Design of $\mathcal{X}$ and $\mathcal{Y}$ To this purpose, we choose the values of pixels (the frame) as feature space $\mathcal{X} = \mathbb{R}^Q$ *i.e.* we implicitly suppose that the frames are independent from each other, discarding any temporal correlation. We consider 4 classes corresponding to the 4 orientations tested in each dataset $\mathcal{Y} = \{0, 1, 2, 3\}$, in fact $\mathcal{Y}$ can be considered as a subset of $\mathcal{C}$, however we renumber these conditions from 0 to 3 for convenience. In datasets $S^{(1)}$ and $S^{(2)}$, we consider the full field stimulations only; in dataset $S^{(6)}$, we consider the orientations tested with parameter $\sigma_\theta = 0.5$ only. Moreover, we restrict our attention to the set of time samples between $t_{\text{on}} + t_{\text{wait}}$ and partial offset t_{f}. Hence we have $I = \{t_{\text{on}} + t_{\text{wait}}, \dots, t_f\} \times \mathcal{Y} \times \mathcal{R}$ and therefore

$$\forall i = (t, c, r) \in I, \quad x_i = s^{(k)}_{\cdot,t,c,r} \in \mathcal{X},$$

where $s^{(k)}_{\cdot,t,c,r}$ is defined in (2.1). We set $t_{\text{wait}} = 5$ for datasets $S^{(1)}$ and $S^{(2)}$ and $t_{\text{wait}} = 10$ for all other datasets. Table 3.1 summarizes the relevant experimental and pre-processing parameters of every dataset comprising the threshold r_{thresh} defined in Section 2.4, the temporal sampling of the camera Δ_t defined in Section 2.2 and the other parameters defined above.

	r_{thresh} (%)	Q	Δ_t (ms)	$t_{\text{on}}\Delta_t$	$t_{\text{f}}\Delta_t$	$t_{\text{off}}\Delta_t$	$T\Delta_t$	C	R
$S^{(1)}$	×	9398	9.6	173	749	$t_{\text{f}}\Delta_t$	902	8	32
$S^{(2)}$	×	9398	9.6	173	749	$t_{\text{f}}\Delta_t$	902	8	28
$S^{(3)}$	35		5	200	600	1000	1280	8	10
$S^{(4)}$	35		5	200	600	1000	1280	8	10
$S^{(5)}$	35		5	200	600	1000	1280	8	10
$S^{(6)}$	45		5	200	600	1000	1280	16	10

Table 3.1: Experimental and pre-processing parameters of the different datasets.

Results Figure 3.1 summarizes the results showing the average score over folds μ_ι (defined in Equation (5.1)) and their standard deviation σ_ι (defined in Equation (5.1)). For each dataset, LC shows the best scores with reasonable standard deviations. It is followed by LDA with 0 to 30 percentage points of score below also with reasonable standard deviations. The NC and GNB methods perform similarly but far worst than LC and LDA (scores around 30-40 %). Finally, the QDA performs at chance level. The bad performances of QDA is certainly due to the dimension issue mentioned in Section 3. Indeed, there are only 400 to 1200 samples per class to estimate the covariance of each class compared to the high dimension of the feature space, a few thousands. In contrast, the covariance in LDA is better estimated as the samples number and the features dimension are of the same order of magnitude, which explains the correct results. Finally, the poor performances of GNB can be explained by the strong assumptions of feature components (*i.e.* pixels) independence which is not assumed in LDA and LC. Indeed, the pixels, from which we compute an orientation map (see next Section 4.1), show a strong specific correlation which gives to the orientation map its geometrical organization, see [6] for a detailed study of their mathematical properties. In order to check if the performance are equivalent for each class, Figure 3.2 shows the averaged confusion matrices Λ (defined in Equation (5.2)) obtained using the five methods for datasets $S^{(2)}$ and $S^{(5)}$. When overall performances are good, the prediction for the different classes are similar. On the contrary, when the performances are bad, strong biases appear. In Figure 3.2**(a)**, the GNB and NC methods tend to predict any

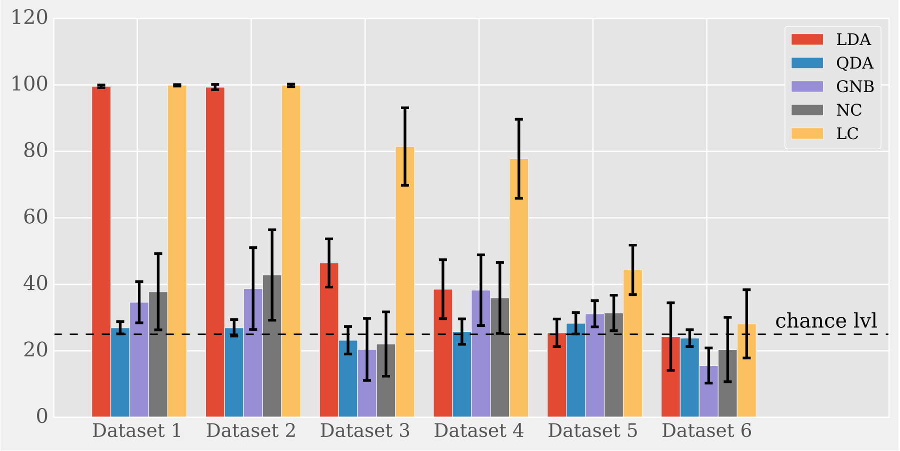


Figure 3.1: Classification performances of the different algorithms on the 6 datasets. QDA: Quadratic Discriminant Analysis. LDA: Linear Discriminant Analysis. GNB: Gaussian Naive Bayes. NC: Nearest Centroid. LC: Logistic Classification.

class to belong to class 0 or 2, as the high values of columns 0 and 2 indicate. In Figure 3.2**(b)**, these two methods tend to predict any class to belong to class 1 or 3, as the high value of the corresponding columns indicate. Finally, the distance d_p (defined in 9) between stimulation parameters associated to predicted labels and parameters associated to true labels is close to 0 when the performances are good.

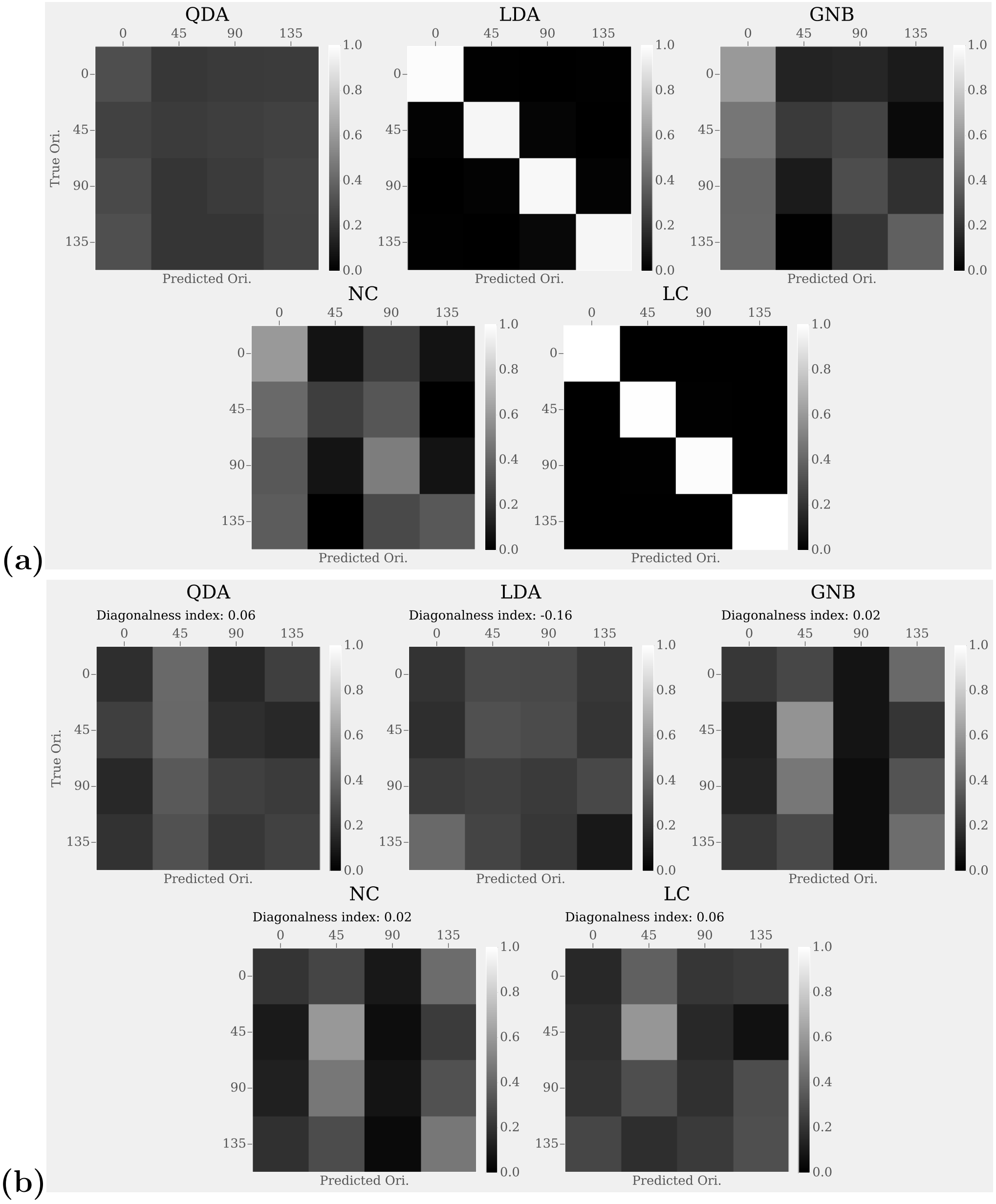


Figure 3.2: Confusion matrices Λ (defined in (5.2)) and the distance d_p (defined in 9) obtained for each algorithm for dataset $S^{(2)}$ **(a)** and $S^{(5)}$ **(b)**. QDA: Quadratic Discriminant Analysis. LDA: Linear Discriminant Analysis. GNB: Gaussian Naive Bayes. NC: Nearest Centroid. LC: Logistic Classification.

3.2 Dimension Reduction Using PCA

In the previous section, we choose to use the pixels' value as feature space $\mathcal{X} = \mathbb{R}^Q$ which limits the performances of QDA and GNB respectively for two reasons: the high number of features compared to samples and the assumption of pixels independence. In particular, the performances of LDA and QDA can be improved by regularizing the covariance matrix (shrinkage to identity). However, we choose to overcome these drawbacks by using Principal Component Analysis (PCA, see Section 5) which allows to compute new features in a low dimensional space: $(\tilde{x}_i)_{i \in I} \in \tilde{\mathcal{X}}$ with $n_{\text{pca}} = \dim \tilde{\mathcal{X}} < Q$. In addition, the component $(x_i^1, \dots, x_i^{n_{\text{pca}}})$ of the vector x_i are uncorrelated. In order to select an appropriate number of PCA features we test different values $n_{\text{pca}} \in \mathcal{E}_{\text{pca}} = \{5, 10, 20, 40, 80, 160\}$ and compute the associated average score over the folds $\mu_\iota(n_{\text{pca}})$ (defined in Equation (5.1)). The number of PCA features that provides the highest score is selected, see Figure 3.3. Then, using this number of PCA features, we compare the scores of the five methods, see Figure 3.4.

Results In Figure 3.3, the left graph shows the average score as a function of the number of PCA features $\mu_\iota(n_{\text{pca}})$ with $n_{\text{pca}} \in \mathcal{E}_{\text{pca}}$ for the five methods (QDA, LDA, GNB, NC, LC) applied to dataset $S^{(4)}$. In every case, the score increases until $n_{\text{pca}} = 10$ features are selected and then it decreases. The right graph shows a normalized version of the score

$$\underline{\mu_\iota}(n_{\text{pca}}) \overset{\text{def.}}{=} \frac{\exp\left(\mu_\iota(n_{\text{pca}})\right)}{\max\limits_{n \in \mathcal{E}_{\text{pca}}} \left(\exp\left(\mu_\iota(n)\right)\right)}$$

as function of n_{pca} highlighting that the highest score is obtained for $n_{\text{pca}} = 10$. Such an normalization is sometime necessary to clearly identify the appropriate number of features. Significant enough to be noted, over the six datasets the number of selected PCA features is often the same for every algorithms and is often equal to 10, 20 or 40. Such a behavior means that the most relevant features about stimuli orientation is among the first dozens of PCA features. This result is in accordance with the previous work of Yavuz [216] who empirically selects the PCA features that are able to discriminate the different orientations among the first PCA features.

Using the appropriate number of features ($n_{\text{pca}} = 10$), Figure 3.4 shows the average scores μ_ι and their standard deviation σ_ι (defined in Equation (5.1)). Compared to the previous section, the scores of QDA and GNB slightly improve as expected, showing scores similar to LC. The score of LDA also reaches the level of LC and even outperforms LC for the datasets $S^{(3)}$ and $S^{(4)}$. The

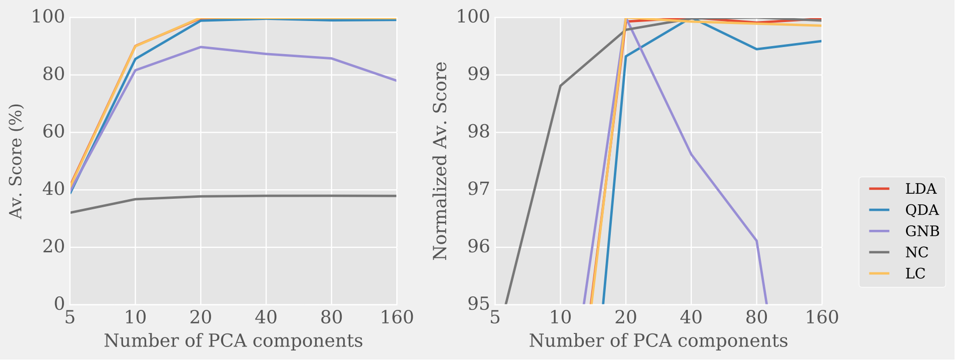


Figure 3.3: Dataset $S^{(4)}$. Left: the average score $\mu_\iota(n_{\text{pca}})$. Right: the normalized average score $\underline{\mu_\iota}(n_{\text{pca}})$.

PCA dimension reduction has almost no effect only on NC that performs equally as with no PCA, see Figure 3.1. The standard deviations appear to be stable or reduced (check LC for datasets $S^{(3)}$ and $S^{(4)}$). Furthermore and although the duration are not shown here, it is important to note that the computation time reduces significantly because algebraic computation are faster in space of dimension at least hundred times smaller than the original pixels' space.

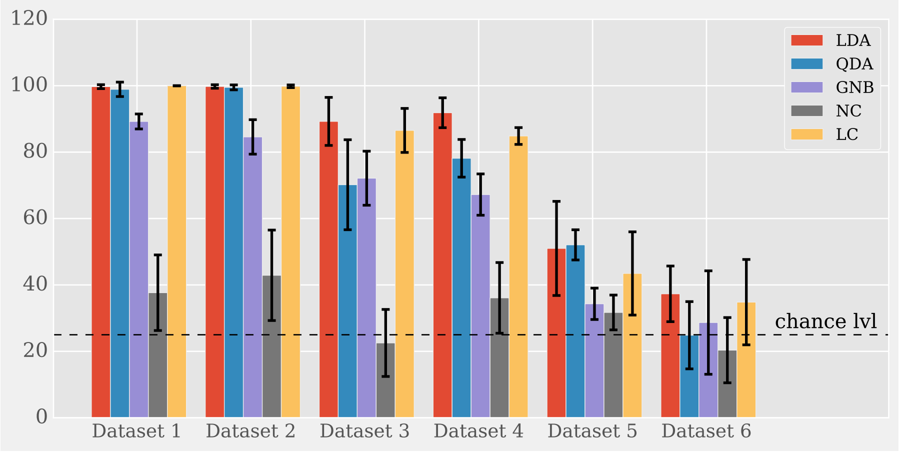


Figure 3.4: Classification performances of the different algorithms on the six datasets and using $n_{\text{pca}} = 10$ PCA features. QDA: Quadratic Discriminant Analysis. LDA: Linear Discriminant Analysis. GNB: Gaussian Naive Bayes. NC: Nearest Centroid. LC: Logistic Classification.

Main Conclusions

- LC and LDA perform best both on pixels and PCA features.
- The discriminant information about stimuli orientation is contained in the first dozens of PCA features.
- The dimension reduction using PCA slightly improves the performances of GNB and QDA which shows the existence of significant spatial correlations.

4 Data Analysis Using Logistic Classification

The LC shows the best classification performances and is therefore selected for further analysis. Although this performance quantify how much information the signal carries about its class, the scores do not provide any clue about how this information is encoded. Answers are to be found in the way LC is classifying the data *i.e.* in the weight vectors. We first compare in Section 4.1 the average of the fitted weight vectors $(\hat{\omega}_1^{\text{ave.}}, \ldots, \hat{\omega}_K^{\text{ave.}})$ with the estimated class centroids $(\hat{\mu}_1^{\text{ave.}}, \ldots, \hat{\mu}_K^{\text{ave.}})$ that can be used to compute orientation maps ($\omega_y^{\text{ave.}}$ and $\hat{\mu}_y^{\text{ave.}}$ for $y \in \mathcal{Y}$ are defined in Equation (5.3)). This first step ensures the consistency with previous works. Then in Section 4.3, we use LC to make prediction locally in time in order to explore the dynamic of classification performances and of the orientation maps emergence. Finally, we use a feature selection method to find the most informative pixels for classification. The examination of important spatio-temporal features enables a precise analysis of the structure of the information representation in cortex (when observed through the VSDOI prism).

Recall that in our analysis of orientation $\mathcal{Y} = \{0, 1, 2, 3\}$, *i.e.* there are 4 classes corresponding to the 4 tested orientations denoted $(\theta_y)_{y \in \mathcal{Y}} = \left(0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\right)$ in radian.

4.1 Comparison of Orientation Maps

We assume that each class $y \in \mathcal{Y}$ corresponds to a particular stimulus orientation θ_y. An orientation map $O = (o_q)_{q \in \mathcal{Q}}$ is computed from every single orientation activation map $M^{(y)} = (m_q^{(y)})_{q \in \mathcal{Q}}$. We compare the maps computed using the class centroids as activation maps *i.e.* $M^{(y)} = \hat{\mu}_y^{\text{ave.}}$ to the ones computed using the weight vectors of LC as activation maps *i.e.* $M^{(y)} = \hat{\omega}_y^{\text{ave.}}$.

Definition 10 (Orientation Map)**.** *Assume that every activation map* $(M^{(y)})_{y \in \mathcal{Y}}$

have a zero mean, then

$$\forall q \in \mathcal{Q}, \quad o_q = \frac{1}{2} \operatorname{Arg} \left(\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} m_q^{(y)} \exp\left(2\mathrm{i}\pi\theta_y\right) \right),$$

where Arg *is the function that give the argument of a complex number in* $[0, 2\pi)$.

Results with no PCA Figures 4.1 and 4.2 show activation maps $(\hat{\omega}_y^{\text{ave.}}, \hat{\mu}_y^{\text{ave.}})$ and their corresponding orientation maps respectively for datasets $S^{(2)}$ and $S^{(3)}$. These figures also show the standard deviation of the class centroids and weight vectors $(\mu_y^{\text{st.d.}}, \omega_y^{\text{st.d.}})$. In Figure 4.1, the weight vectors $\hat{\omega}_y^{\text{ave.}}$ are much noisier than class centroids $\hat{\mu}_y^{\text{ave.}}$ which leads to a map that is noisier for LC. However, the orientation columns appear to be the same in each map. The standard deviation $\mu_y^{\text{st.d.}}$ of NC is slightly higher than $\omega_y^{\text{st.d.}}$ relatively to their mean. The high level of $\mu_y^{\text{st.d.}}$ is mainly due to the illumination conditions. In Figure 4.2, again class centroids appear smoother than weight vectors and so the resulting maps are. The orientation columns are approximately the same in each maps, however the map obtained with class centroids shows a prevalence of green (60 °) and purple (150 °) domains whereas the map obtained with weight vectors shows more balanced columns. Such a difference is probably due to illumination conditions. Finally, the standard deviations are quantitatively higher for the class centroids than for the weight vectors. Moreover, the standard deviations of weight vectors show that the blood vessels are a source of noise whereas the standard deviations of class centroids are mainly related to illumination conditions. Figure 4.3 shows the absolute difference between maps obtained using class centroids and weight vectors. As orientation maps take their value in the torus $\mathbb{Z}/\pi\mathbb{Z}$, we choose an appropriate distance on the torus, for all orientation maps O and O' and for all pixels $q \in \mathcal{Q}$

$$\mathrm{d}\left(o_q, o'_q\right) = \min\left(|o_q - o'_q|, 180 - |o_q - o'_q|\right). \tag{4.1}$$

The Figure 4.3 show that the difference between maps is mainly due to the noise of maps obtained using LC.

Results using PCA Figures 4.4 and 4.5 are similar to Figure 4.1 and 4.2 except that a PCA has been performed to reduce the dimension before the classification, see Section 3.2. The effect of PCA is particularly visible in the weight vectors of LC, in which a large amount of noise is removed. The map obtained with LC displays a new orientation column in the top right corner of the map when compared to the map obtained with NC. The absolute differences between maps (Figure 4.6) are less contaminated by such a noise and show that the differences are noticeable at pinwheels and where the color gradient is high, this is particularly true for Figure 4.6**(a)**. For dataset $S^{(2)}$, the PCA has a limited effect on the errors $(\mu_y^{\text{st.d.}}, \omega_y^{\text{st.d.}})$, which are quantitatively the same as when no PCA is used. For dataset $S^{(3)}$, the PCA reduces a little the errors $\omega_y^{\text{st.d.}}$ but not $\mu_y^{\text{st.d.}}$ relatively to the weight vectors and the class centroids. The error $\mu_y^{\text{st.d.}}$ shows more structure than when the PCA is not used. Finally, the difference between maps shown in Figure 4.6 appears less corrupted by noise when PCA is used. However, this is only visible in Figure 4.6**(a)**.

Remarkable Properties of Activation Maps As this is the first time that we show some activity maps in this manuscript, we find important to precise two remarkable mathematical properties that are verified by activation maps:

- Two activation maps evoked by stimuli of orientations that differ from 45 ° are orthogonal.

- Two activation maps evoked by stimuli of orientation that differ from 90 ° are in phase opposition.

The phase opposition can be checked on the different mentioned figures. However, we check our claim by computing a cosine similarity index defined by

$$\forall (y_0, y_1) \in \mathcal{Y}^2, \quad \texttt{csi}(y_0, y_1) = \frac{\langle \hat{\omega}_{y_0}^{\text{ave.}}, \hat{\omega}_{y_1}^{\text{ave.}} \rangle}{\|\hat{\omega}_{y_0}^{\text{ave.}}\| \|\hat{\omega}_{y_1}^{\text{ave.}}\|}.$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean scalar product and $\| \cdot \|$ is the Euclidean norm. When the cosine similarity index is equal to 1, the vectors are colinear. When it is equal to 0, the vectors are orthogonal. Finally, when it is equal to -1, the vectors are colinear but in opposite directions. Table 4.1 shows values that support our claim.

y_0/y_1	0/1	1/2	2/3	3/0	0/2	1/3
`csi`	-0.040	0.170	-0.120	-0.014	-0.990	-0.997

Table 4.1: Cosine similarity index for relevant activation maps among $(\hat{\omega}_y^{\text{ave.}})_{y \in \mathcal{Y}}$ obtained with LC on dataset $S^{(2)}$.

Main Conclusions

- Activation maps and orientation maps obtained with LC are consistent with those obtained with NC.
- Cross-validation provides a way to quantify errors in the estimation of activation maps.
- PCA reduces noise in activation maps and orientation maps.

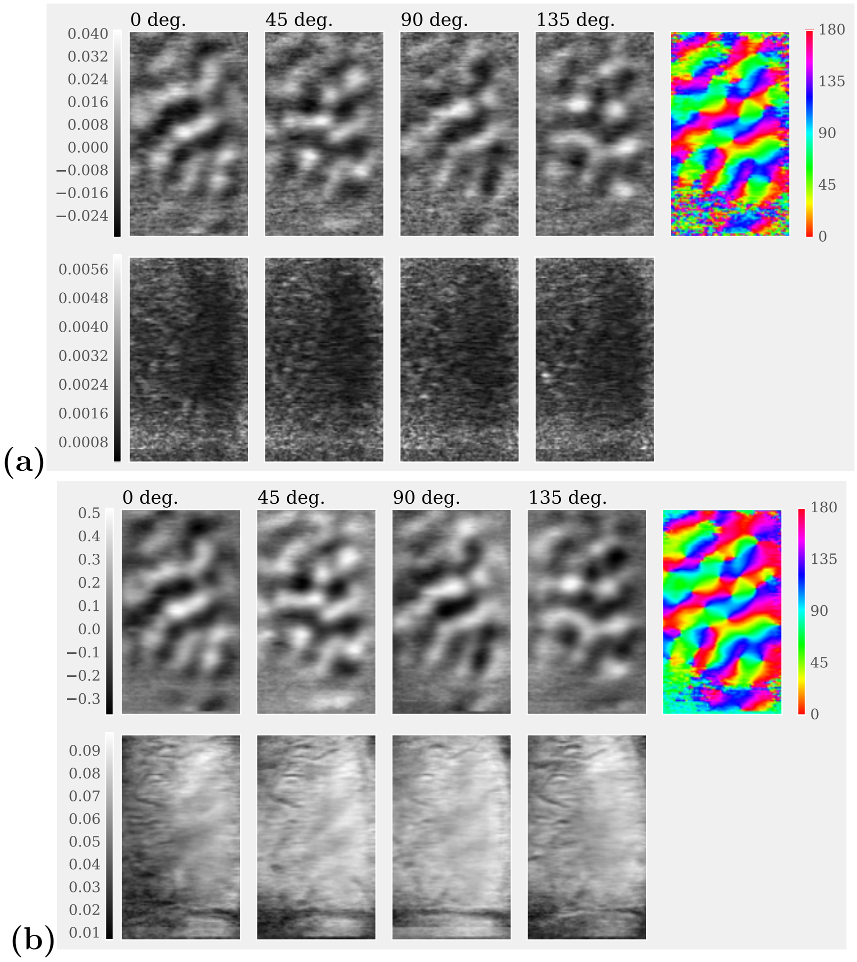


Figure 4.1: Dataset $S^{(2)}$. **(a)** Top: weight vectors $\hat{\omega}_y^{\text{ave.}}$ and their corresponding orientation map. Bottom: errors on the weight vectors $\omega_y^{\text{st.d.}}$. **(b)** Top: class centroids $\hat{\mu}_y^{\text{ave.}}$ and their corresponding orientation map. Bottom: errors on the class centroids $\mu_y^{\text{st.d.}}$.

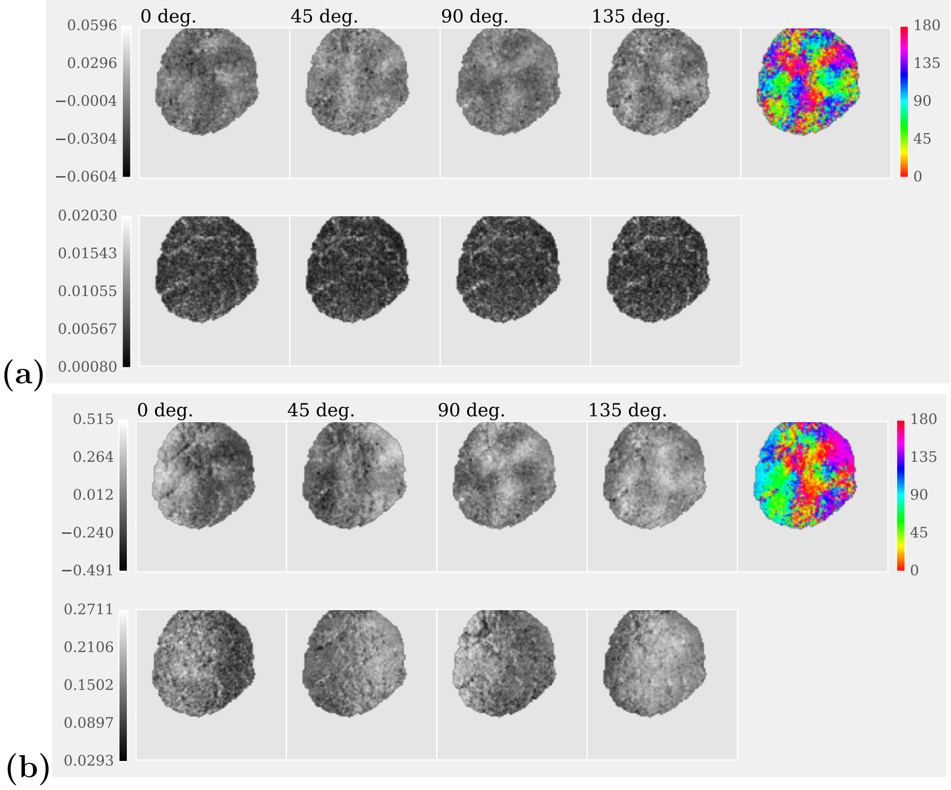


Figure 4.2: Dataset $S^{(3)}$. **(a)** Top: weight vectors $\hat{\omega}_y^{\text{ave.}}$ and their corresponding orientation map. Bottom: errors on the weight vectors $\omega_y^{\text{st.d.}}$. **(b)** Top: class centroids $\hat{\mu}_y^{\text{ave.}}$ and their corresponding orientation map. Bottom: errors on the class centroids $\mu_y^{\text{st.d.}}$.

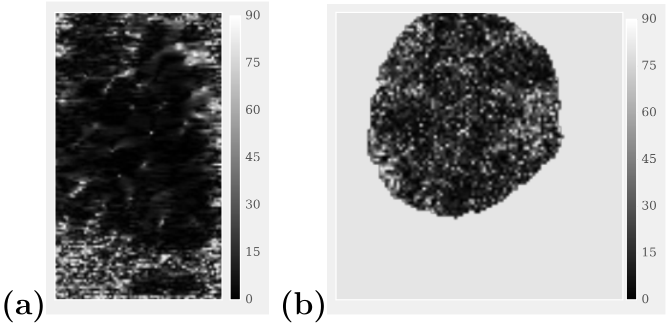


Figure 4.3: Distance d (see Equation (4.1)) between orientation maps obtained using weight vectors and class centroids as activation map, see 4.1. **(a)** Dataset $S^{(2)}$. **(b)** Dataset $S^{(3)}$.

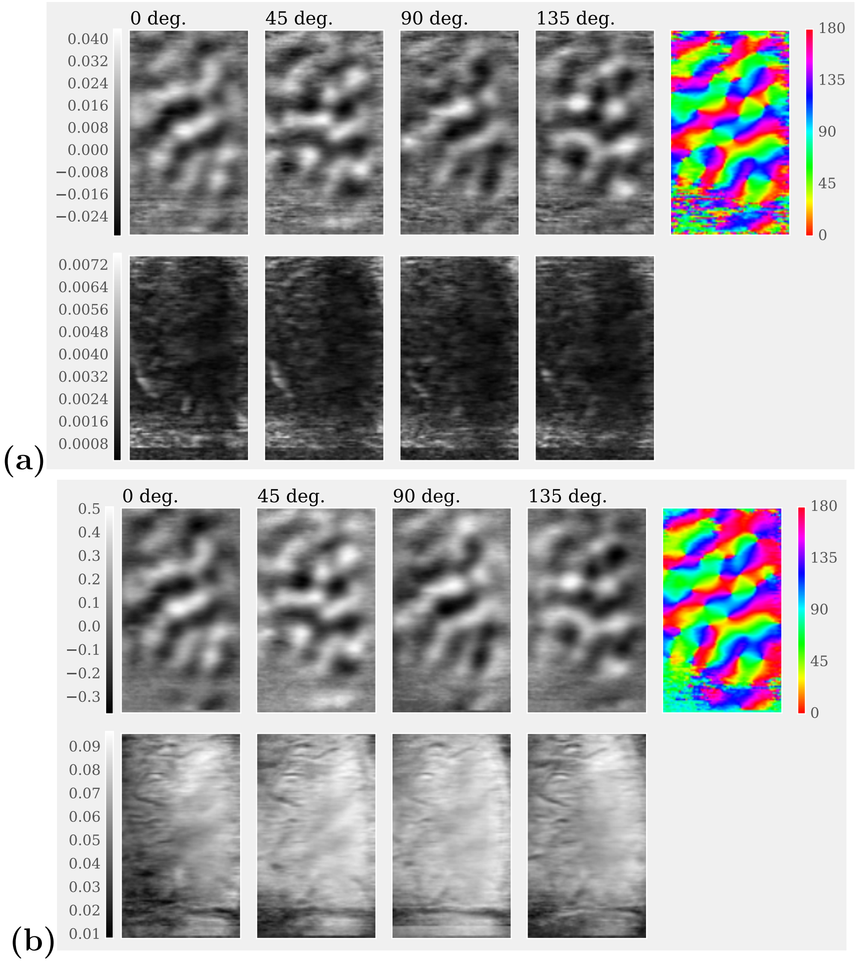


Figure 4.4: Dataset $S^{(2)}$ after dimension reduction to $n_{\text{pca}} = 20$. **(a)** Top: weight vectors $\hat{\omega}_y^{\text{ave.}}$ and their corresponding orientation map. Bottom: errors on the weight vectors $\omega_y^{\text{st.d.}}$. **(b)** Top: class centroids $\hat{\mu}_y^{\text{ave.}}$ and their corresponding orientation map. Bottom: errors on the class centroids $\mu_y^{\text{st.d.}}$.

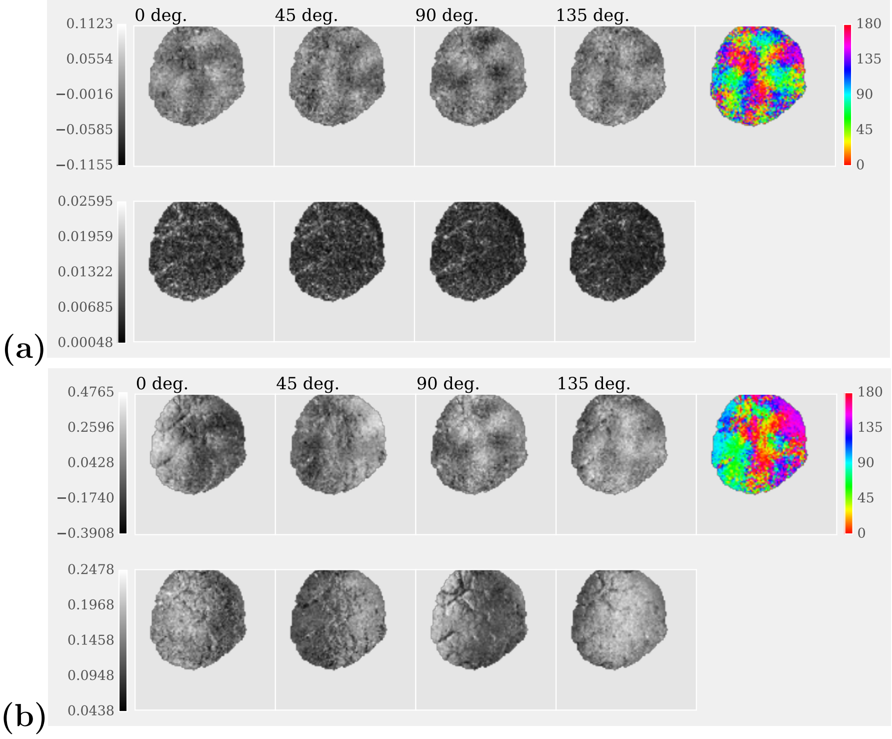


Figure 4.5: Dataset $S^{(3)}$ after dimension reduction to $n_{\mathrm{pca}} = 20$. **(a)** Top: weight vectors $\hat{\omega}_y^{\text{ave.}}$ and their corresponding orientation map. Bottom: errors on the weight vectors $\omega_y^{\text{st.d.}}$. **(b)** Top: class centroids $\hat{\mu}_y^{\text{ave.}}$ and their corresponding orientation map. Bottom: errors on the class centroids $\mu_y^{\text{st.d.}}$.

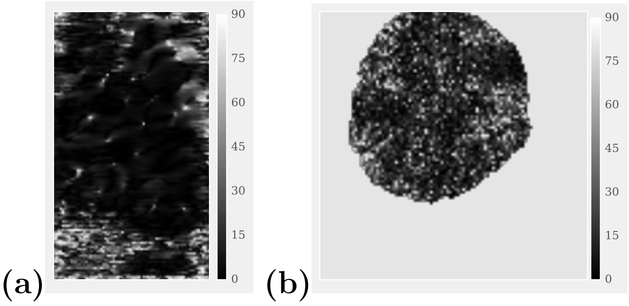


Figure 4.6: Distance d (see Equation (4.1)) between orientation maps obtained using weight vectors and class centroids as activation map, see 4.1. **(a)** Dataset $S^{(2)}$. **(b)** Dataset $S^{(3)}$. The dimension is reduced to $n_{\mathrm{pca}} = 20$ using PCA.

4.2 Spatially Localized Predictions

The LC method has shown good classification performances using all pixels (projected or not onto the low dimensional PCA features space). One interesting question is to determine which pixels are most useful for prediction. To this purpose we evaluate the prediction performances obtained using a local subset of the weight vector's pixels. We use a two-dimension Gaussian sliding window

$$\forall q' \in \mathcal{Q}, \quad g_q(q') = \exp\left(-\frac{\|q'-q\|^2}{2\sigma_g^2}\right)$$

where σ_g is the window size. The window weights the pixels q' of the weight vectors around each pixel q. We use these windowed weight vectors to make predictions. The windowing is performed with null conditions at the border of the weight vectors *i.e.* $\forall q' \in \mathcal{Q}, g_q(q') = 0$ if and only if $q' - q \notin \mathcal{Q}$. The probabilities of logistic classification defined in Equation (4.1) is therefore modified in order to obtained the following localized predictor centered at pixel $q \in \mathcal{Q}$,

$$\mathbb{P}_{Y|X,\theta,q}(y|x) = \frac{e^{\langle x,\, g_q \omega_y \rangle}}{\sum_{y' \in \mathcal{Y}} e^{\langle x,\, g_q \omega_{y'} \rangle}}.$$

This probability is then plugged in Equation (2.3) to make prediction. We can then compute an average score $\mu_{\iota,q}$ for each pixel q.

Results For the datasets $S^{(3)}$ to $S^{(6)}$ we use a Region of Interest (ROI) that is composed of pixels which luminance is above a percentage threshold r_{tresh} of the maximum. In order to assess the relevance of this ROI, we show in Figure 4.7 the average score $\mu_{\iota,q}$ for each pixel $q \in \mathcal{Q}$ obtained on dataset $S^{(3)}$ with no ROI selected beforehand and $\sigma_g = 15$. First, the region of highly predictive pixels (*ie* with high local score $\mu_{\iota,q}$) correspond approximately to the selected ROI using the thresholding, see Figure 4.7**(b)**. This is not surprising because VSD signal is highly dependent on illumination. Second, this region corresponds actually to the region where we identify the orientation map, the remaining pixels of the image consist of noise and are not useful for prediction. The parameter σ_g indicates the level of locality of the prediction. When it becomes too small, the score reaches the chance level, when it becomes too high, for all pixels the prediction score converges to the score obtained using the entire weight vectors. Figure 4.8 shows the local scores $\mu_{\iota,q}$ for each pixel $q \in \mathcal{Q}$ computed using $\sigma_g = 2$, $\sigma_g = 5$ and $\sigma_g = 15$. Above 30, σ_g is not informative because each pixel tends to have the same prediction score. However, very small value (1 to 5) highlights a kind of map skeleton. This skeleton follows iso-oriented lines that separate areas of small color gradient. Moreover, the lines of

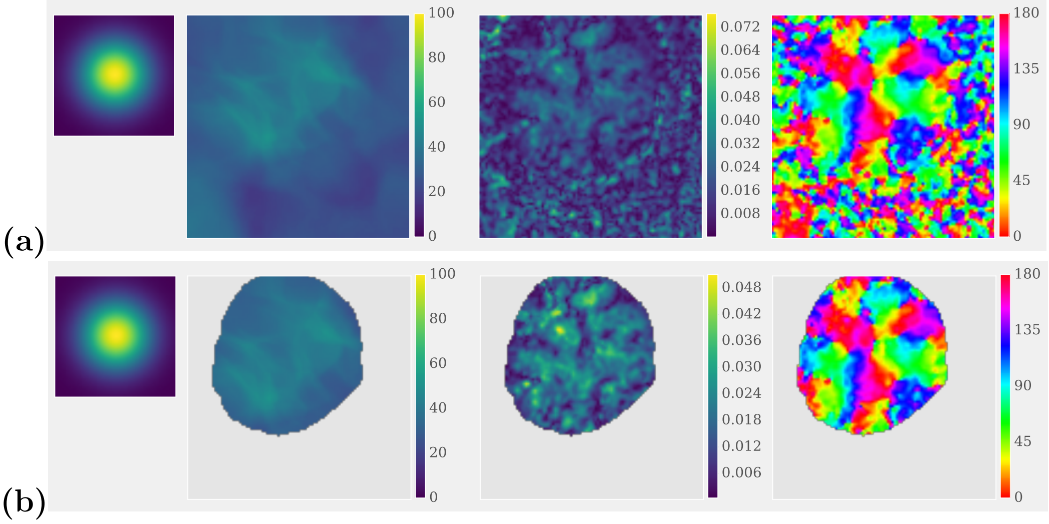


Figure 4.7: Results for dataset $S^{(3)}$. From left to right: Gaussian window g_q then for each pixel $q \in \mathcal{Q}$, local scores $\mu_{\iota,q}$, orientation selectivity index oi_q (defined in (11)) and orientation preference o_q (defined in (10)). **(a)** No ROI selected during preprocessing. **(b)** ROI selected using $r_{\text{tresh}} = 35$.

this skeleton cross at pinwheels which are singularities of the map where all the orientations are represented in their neighborhood. The reason why prediction is maintained at a high level above chance around pinwheels and between areas of small color gradient is probably that the value of certain pixels highly increase when a particular orientation is presented whereas others increase moderately. In datasets $S^{(1)}$ and $S^{(2)}$, the protocols involve global versus local

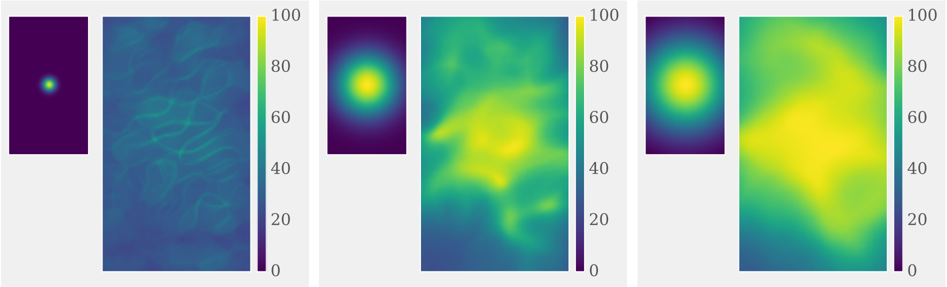


Figure 4.8: Results for dataset $S^{(2)}$. Gaussian window g_q and local scores $\mu_{\iota,q}$ for each pixel $q \in \mathcal{Q}$. From left to right: $\sigma_g = 2$, $\sigma_g = 10$ and $\sigma_g = 15$.

stimulation. Since the visual field is mapped to the visual cortex [114], a local stimulation is supposed to activate only a subset of the recorded cortical area. Pixels' activation is measured by their mean luminance, activated pixels are distinct from orientation selective pixels measured by an orientation selective index.

Definition 11 (Orientation Selectivity Index)**.** *Assume that every activation map* $(M^{(y)})_{y\in\mathcal{Y}}$ *have a zero mean, then*

$$\forall q \in \mathcal{Q}, \quad oi_q = \left| \frac{1}{|\mathcal{Y}|} \sum_{y\in\mathcal{Y}} m_q^{(y)} \exp\left(2\mathrm{i}\pi\theta_y\right) \right|.$$

In Figure 4.9, we compare orientation selectivity index to local scores. The first quantifies how much a pixel is selective whereas the second quantifies how much the neighborhood of a pixel is able to discriminate between orientations. Since a large enough neighborhood of a highly selective pixel must be able to discriminate between orientation such a comparison make sense. As expected the discriminant areas correspond approximately to highly selective areas. However, the pinwheels are not highly selective whereas their neighborhood is discriminant. For the local stimulation, the discriminant area keep located at the corresponding retinotopic position of the visual stimulation, as for the selective area.

Main Conclusions

- Local prediction performances is consistent with the ROI selection based on luminance levels.
- The size σ_g of the local neighborhood allows to evaluate local scores prediction at various scales.
- Small local neighborhoods provide a skeleton that is not yet fully understood.
- Local prediction performance is a new measure that quantifies the discriminative power of a pixel and its neighbors.

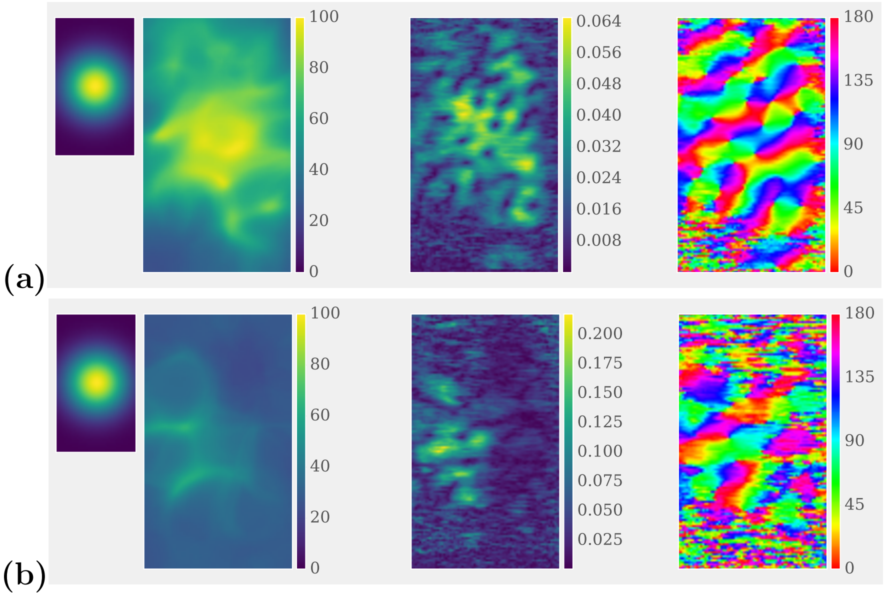


Figure 4.9: Results for dataset $S^{(2)}$. From left to right: Gaussian window g_q then for each pixel $q \in \mathcal{Q}$, local scores $\mu_{\iota,q}$, orientation selectivity index oi_q (defined in (11)) and orientation preference o_q (defined in (10)). **(a)** Full field stimulation. **(b)** Local stimulation.

4.3 The Temporal Dynamic of Orientation Maps and Prediction

The previous section shows that LC and NC give similar results for the computation of orientation maps. It enables the computation of a localized prediction score that we can use to quantify how informative an area around a pixel is. As we mention in Section 1, the VSDi has a good temporal resolution which allows to probe the dynamic of neural activation. Therefore, it is of particular interest to estimate weight vectors, orientation maps and scores over time.

Design of $\mathcal{X}$ and $\mathcal{Y}$ For this purpose we consider a sample set indexed by the time $I_t = \{t\} \times \mathcal{Y} \times \mathcal{R}$ for all $t \in \mathcal{T}$ with $\mathcal{Y} = \{0, 1, 2, 3\}$ *i.e.* we consider only 4 classes consisting of the 4 orientations tested. We keep the same feature space $\mathcal{X} = \mathbb{R}^Q$. Therefore

$$\forall t \in \mathcal{T}, \quad \forall i = (t, c, r) \in I_t, \quad x_{t,i} = s_{.,t,c,r} \in \mathcal{X}.$$

In such a setting, we can compute, at each time t, the average of the scores $\mu_{\iota,t}$, their standard deviation $\sigma_{\iota,t}$, the average of the estimated weight vectors $\hat{\omega}_{y,t}^{\text{ave.}}$, the corresponding orientation maps O_t, the local score $\mu_{\iota,q',t}$ for each pixel $q' \in \mathcal{Q}$ and the confusion matrix Λ_t.

Such a design of $\mathcal{X}$ and $\mathcal{Y}$ is necessary to compute the weight vectors $\hat{\omega}_{y,t}^{\text{ave.}}$ and its corresponding orientation maps O_t at each time $t \in \mathcal{T}$. However the scores $\mu_{\iota,t}$, their standard deviation $\sigma_{\iota,t}$ and the local scores $\mu_{\iota,q',t}$ can be computed using the weight vectors $\hat{\omega}_{y}^{\text{ave.}}$ of Section 4.1. When computed this way, we denote respectively $\tilde{\mu}_{\iota,t}$, $\tilde{\sigma}_{\iota,t}$, $\tilde{\mu}_{\iota,q',t}$, $\tilde{\Lambda}_t$ the scores, their standard deviation, the local scores and the confusion matrix.

4.3.1 Full Field *vs* Local Stimulation

First, we compare the dynamic of global scores $\tilde{\mu}_{\iota,t}$ for all $t \in \mathcal{T}$ for the full field and local stimulation tested in datasets $S^{(1)}$ and $S^{(2)}$ (see Figure 4.10). For the global stimulation, the score starts to increase significantly above chance level about 20 ms after stimulus onset, it reaches more than 95% of correct prediction about 50 ms after stimulus onset. For the local stimulation, the dynamic is a little slower than the dynamic of the full field stimulation. In Figure 4.10, the score starts to increase about 30 ms after stimulus onset and reaches almost 80% of correct prediction about 90 ms after stimulus on set. For the local stimulation, the score reaches a plateau level between 60% and 80% whereas it stays near 100% for the full field stimulation. The global scores over time measure the dynamic of orientation discriminability. The more selective the pixels are, the more they are able to discriminate the orientations. Thus, it is interesting to compare the dynamic of selectivity index measured by Sharon *et al.* [174] and our measure of discriminability. Not surprisingly, the times at which selectivity index starts to increase and peaks are similar to the ones we report. However, our scores are computed over the whole cortex whereas Sharon limits the analysis of selectivity index to pixels in a highly reproducible region. The score $\tilde{\mu}_{\iota,t}$ is a global measure of discriminability. Now, we compare the dynamic of local scores $\tilde{\mu}_{\iota,q,t}$ to the dynamic of orientation maps O_t for both local and full field stimulation. Figure 4.11 and 4.12 displays these local scores (with $\sigma_g = 10$) and maps for the dataset $S^{(2)}$. For both full field and local stimulation, the different orientation columns are visible in less than 20 ms after stimulus onset. It is striking that the map shows its geometrical structure in regions where the local score is low. Around time 211.2 ms in Figure 4.11(**b**) and 4.12(**b**), the map is geometrically structured outside the highly discriminatory pixels inside the black line. After 40 ms for full field stimulus (resp. 60 ms for local stimulus), the map is stable until the end of the stimulation. For full field stimulation, the local score $\mu_{\iota,q,t}$ first

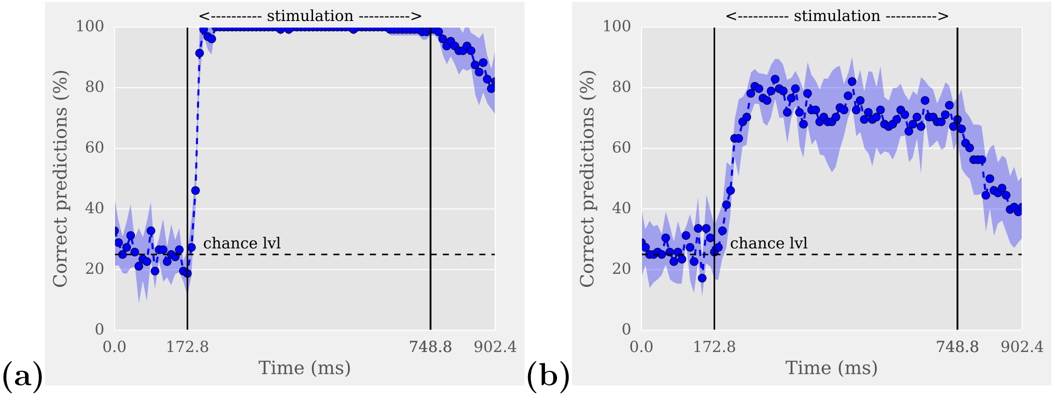


Figure 4.10: Score $\tilde{\mu}_{\iota,t}$ for all $t \in \mathcal{T}$ for dataset $S^{(2)}$. **(a)** Full field stimulation. **(b)** Local stimulation.

starts to increase locally (in a skeleton form, probably due to the value of σ_g) then it builds up until half of pixels is highly discriminant ($\tilde{\mu}_{\iota,q',t} > 80$ %, 70 ms after stimulus onset). For the local stimulation, the local score starts to increase locally and then builds up but stays limited to the retinotopic region corresponding to the visual field where the stimulus is presented ($\tilde{\mu}_{\iota,q',t} > 40$ %, 70 ms after stimulus onset).

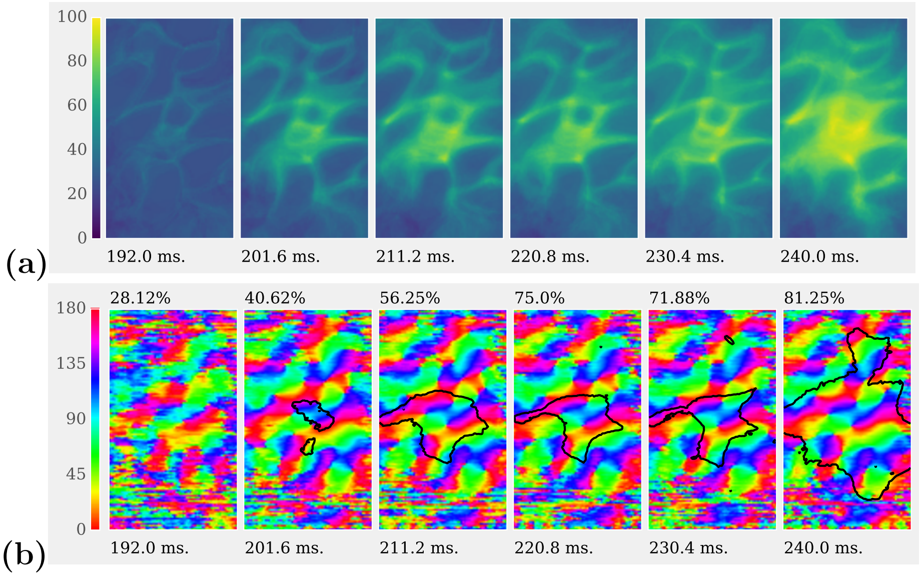


Figure 4.11: Full field stimulation for dataset $S^{(1)}$. **(a)** Local score $\mu_{\iota,q,t}$ for each pixel $q \in \mathcal{Q}$ using $\sigma_g = 10$). **(b)** Maps O_t at time t indicated below each frame. The black line delimits the pixels with a local score above 80%. Supplementary figure online.

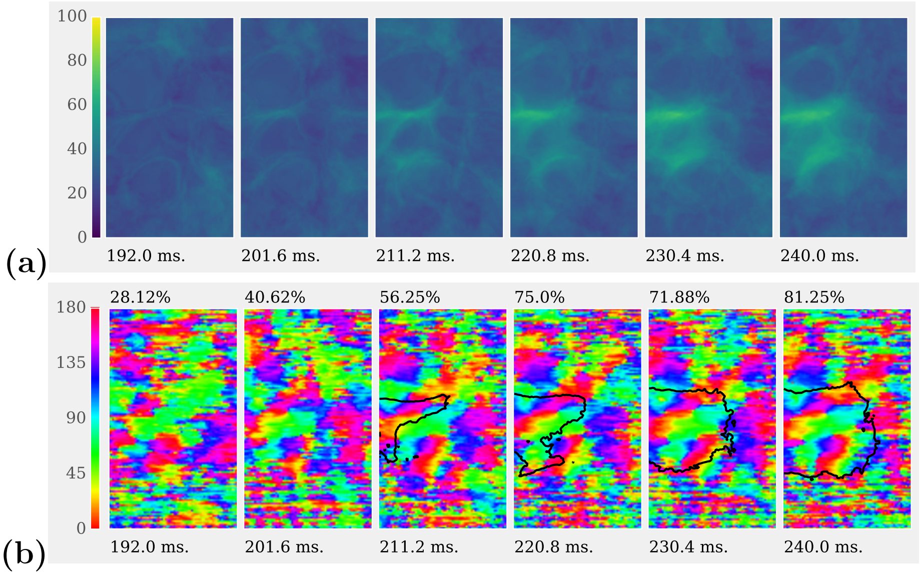


Figure 4.12: Local field stimulation for dataset $S^{(1)}$. **(a)** Local score $\mu_{\iota,q,t}$ for each pixel $q \in \mathcal{Q}$ using $\sigma_g = 10$). **(b)** Maps O_t at time t indicated below each frame. The black line delimits the pixels with a local score above 40%. Supplementary figure online.

Main Conclusions

- Classification score grows slower for the local stimulation than for the full field stimulations.
- The map becomes visible in areas before pixels reach a significant score.
- For the local stimulation, highest local scores are limited to the retinotopic limit of the stimulus.

4.3.2 Sharp Rotation of the Stimulation

Analyzing the dynamic of datasets $S^{(3)}$ and $S^{(4)}$ is important to understand how the neural population switches between activation maps. Indeed, the protocols described in Section 2.3 make use of stimuli with a temporal discontinuity (instantaneous rotation). The consequences of such a discontinuity over the VSD signal is likely informative to tackle the question of propagating and standing waves. First, we focus on the dynamic of global scores $\tilde{\mu}_{\iota,t}$ for the two datasets $S^{(3)}$ and $S^{(4)}$. At mid stimulation time 400 ms after stimulus onset it rotates of 135 ° for dataset $S^{(3)}$ and of 90 ° for dataset $S^{(4)}$. Such a rotation corresponds to a circular permutation of the labels (three step right for the 135 ° rotation and two step right for the 90 ° rotation), it is therefore useful to consider $\tilde{\mathcal{Y}} = \mathbb{Z}/|\mathcal{Y}|\mathbb{Z}$ as label set instead of $\mathcal{Y}$. Figure 4.13 displays two scores: the blue score $\tilde{\mu}^b_{\iota,t}$ is computed using the original labels of the features in the test set and the red score $\tilde{\mu}^r_{\iota,t}$ is computed using the permuted labels of the features in the test set,

$$\tilde{\mu}^b_{\iota,t} = \frac{\sum_{y\in\tilde{\mathcal{Y}}} \tilde{\Lambda}_{y,y,t}}{\sum_{(y,y')\in\tilde{\mathcal{Y}}^2} \tilde{\Lambda}_{y,y',t}} \quad \text{and} \quad \tilde{\mu}^r_{\iota,t} = \frac{\sum_{y\in\tilde{\mathcal{Y}}} \tilde{\Lambda}_{y+k,y,t}}{\sum_{(y,y')\in\tilde{\mathcal{Y}}^2} \tilde{\Lambda}_{y,y',t}}.$$

where $k = 2$ for the 90 ° rotation and $k = 3$ for the 135 ° rotation. These scores allow to compare the dynamic of activation and the dynamic of activation after the rotation. In both datasets, the activation score $\tilde{\mu}^b_{\iota,t}$ (blue) reaches its maximum level about 55 ms after stimulus onset and it starts to decrease 50 ms after stimulus rotation and reaches chance level 100 ms after stimulus rotation. Again in both dataset, the activation score after rotation $\tilde{\mu}^r_{\iota,t}$ (red) increase progressively 50 ms after stimulus rotation. For dataset $S^{(3)}$ the score after rotation $\tilde{\mu}^r_{\iota,t}$ reaches 60 % about 135 ms after stimulus rotation, for dataset $S^{(4)}$ the score after rotation $\tilde{\mu}^r_{\iota,t}$ reaches 60 % about 160 ms after stimulus rotation. Finally, the score after rotation $\tilde{\mu}^r_{\iota,t}$ starts to decrease about

90 ms after stimulus offset. These results suggest that the neural population is activated faster when stimulated after blank than when stimulated after a first stimulation. Another interesting computation is the mean distance between

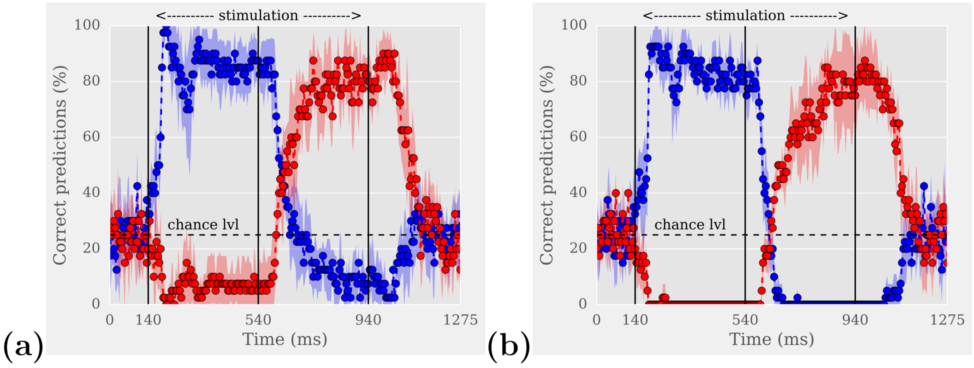


Figure 4.13: Score $\tilde{\mu}^b_{\iota,t}$ (blue) and $\tilde{\mu}^r_{\iota,t}$ (red) for all $t \in \mathcal{T}$. **(a)** Dataset $S^{(3)}$. **(b)** Dataset $S^{(4)}$.

the maps O_t and the map O (shown in Figure 4.2), at each time $t \in \mathcal{T}$

$$\mathtt{md}_t = \frac{1}{V} \sum_{q \in \mathcal{Q}} d(o_{t,q}, o_q).$$

The mean md is displayed in Figure 4.14 for dataset $S^{(3)}$ and $S^{(4)}$ showing the dynamic of the map after the stimulus rotation. For both datasets, the map is established 50 ms after stimulus onset. Then, still for both dataset, the map differentiate 50 ms after stimulus rotation. The map reaches its stationary state about 100 ms after stimulus rotation.

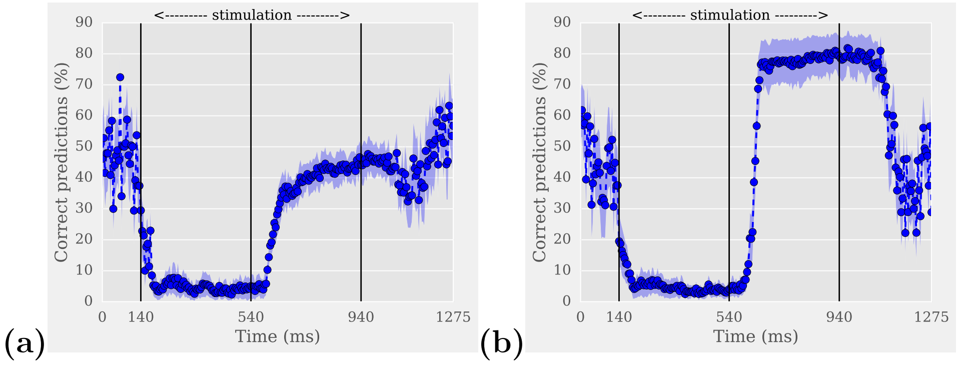


Figure 4.14: The mean distance between maps $\mathtt{md}_t$ for $t \in \{t_{\text{on}}\Delta_t, \dots, t_{\text{off}}\Delta_t\}$. **(a)** Dataset $S^{(3)}$. **(b)** Dataset $S^{(4)}$.

In order to check if logistic classification is correctly switching to the correct orientation after the rotation, we check the class predictions. This information is contained in the confusion matrix Λ_t and is illustrated in Figure 4.15 corresponding to datasets $S^{(3)}$ and $S^{(4)}$.

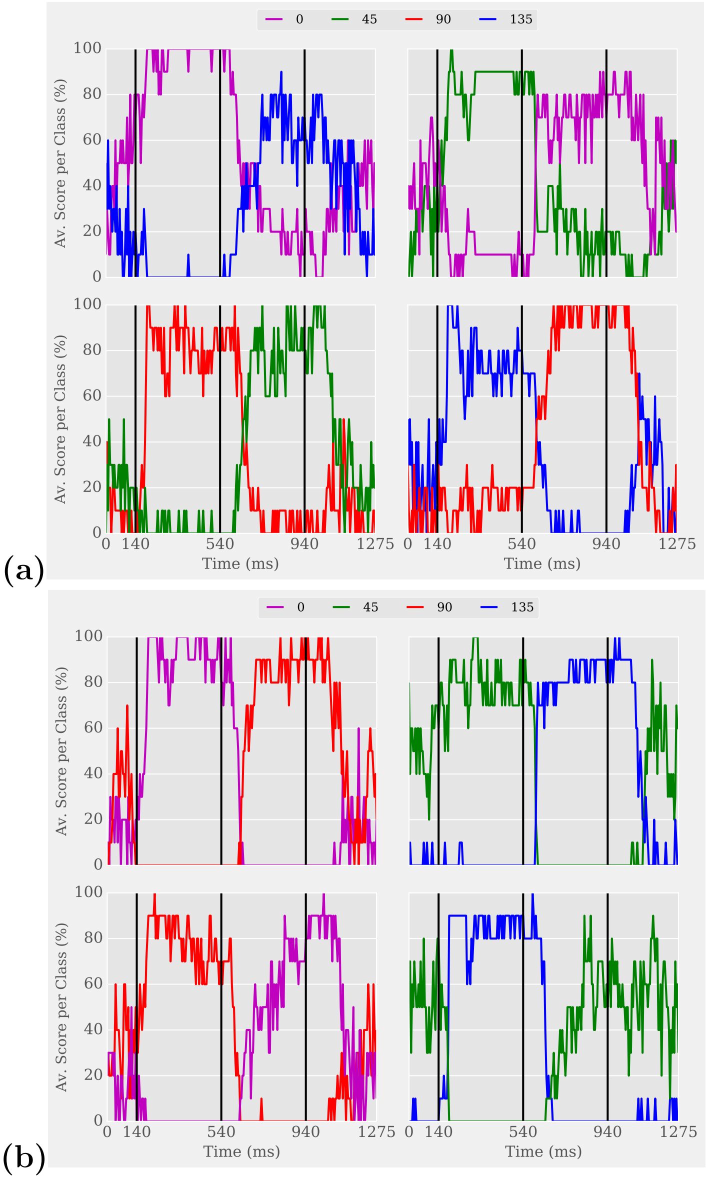


Figure 4.15: Relevant coefficients of the average confusion matrix $\tilde{\Lambda}_t$ computed at each time $t \in \mathcal{T}$. **(a)** Dataset $S^{(3)}$. Coefficients $\tilde{\Lambda}_{y,y,t}$ and $\tilde{\Lambda}_{y+3,y,t}$ for $y \in \tilde{\mathcal{Y}}$. **(b)** Dataset $S^{(4)}$. Coefficients $\tilde{\Lambda}_{y,y,t}$ and $\tilde{\Lambda}_{y+2,y,t}$ for $y \in \tilde{\mathcal{Y}}$. Supplementary figure online.

For example in dataset $S^{(4)}$, a recording under horizontal stimulation (label $y = 0$) is supposed to be correctly predicted by the activity map $\hat{\omega}_0^{\text{ave.}}$. After the orthogonal rotation, the signal is recorded under vertical stimulation (label $y = 2$) and it is, therefore, supposed to be correctly predicted by the activity map $\hat{\omega}_2^{\text{ave.}}$. In both cases, this is what happened and the transitions are as fast as for the global score $\tilde{\mu}_{\iota,t}$. However for the two bottom graphs of Figure 4.15**(b)**, after the rotation, the prediction does not reach a high probability. These two classes slow down the global score switching dynamic of dataset $S^{(4)}$ compared to dataset $S^{(3)}$, explaining the differences in the reported times 135 ms *vs* 160 ms, see above.

Finally, to understand the dynamic of the neural population we show in Figure 4.16 and 4.17 the weight vectors $\hat{\omega}_{y,t}^{\text{ave.}}$ for each class $y \in \mathcal{Y}$ and the scores $\tilde{\mu}_{\iota,t}$ (see Figure 4.13) at transition times t indicated under each frame. Figure 4.16 shows that for the 135 ° rotation, the area of activity of the initial orientation is moving continuously towards the area of activity of the new orientation. However, this transition is much confused for the 90 ° rotation, see Figure 4.17. For dataset $S^{(3)}$, the key times that support our claim are $t = 620$ ms and $t = 630$ ms where activity has begun to switch but has not completely switched yet. For dataset $S^{(4)}$, the key time is $t = 630$ ms where high activity level are spread in a noisy fashion (**(a)(b)**), absent (**(c)**) or prevalent (**(d)**).

Main Conclusions

- Activation of the neural population after blank stimulation is faster than activation after a first stimulation.
- Activation maps smoothly switch for the 135 ° rotation.
- Activation maps abruptly switch for the 90 ° rotation.

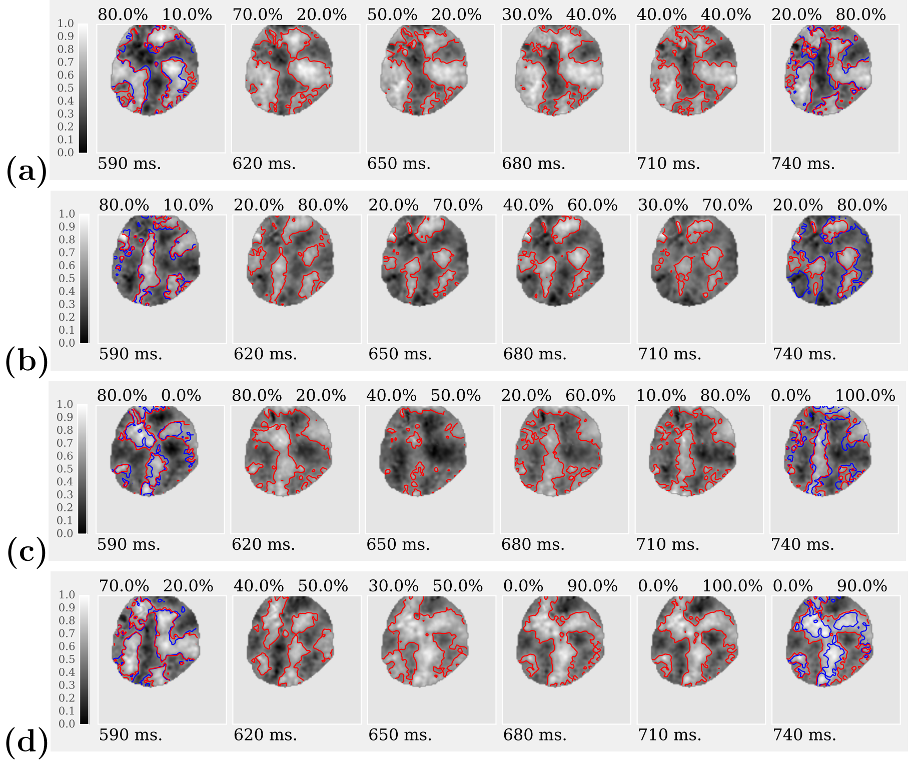


Figure 4.16: Dataset $S^{(3)}$ (orientation shift +135 °). Normalized weight vectors $\hat{\omega}_{y,t}^{\text{ave.}}$ and scores $\tilde{\mu}_{\iota,t}$ (Left: blue scores. Right: red scores. See Figure 4.13) for time indicated under each frame. The red contours are the level set of the current normalized weight vectors at 0.5. The blue contours are the level set of the normalized weight vectors $\hat{\omega}_{y}^{\text{ave.}}$ computed in Section 4.1 at 0.5. **(a)** Label $y = 0$ corresponding to a 0 ° orientation. **(b)** Label $y = 1$ corresponding to a 45 ° orientation. **(c)** Label $y = 2$ corresponding to a 90 ° orientation. **(d)** Label $y = 3$ corresponding to a 135 ° orientation. Supplementary figure online.

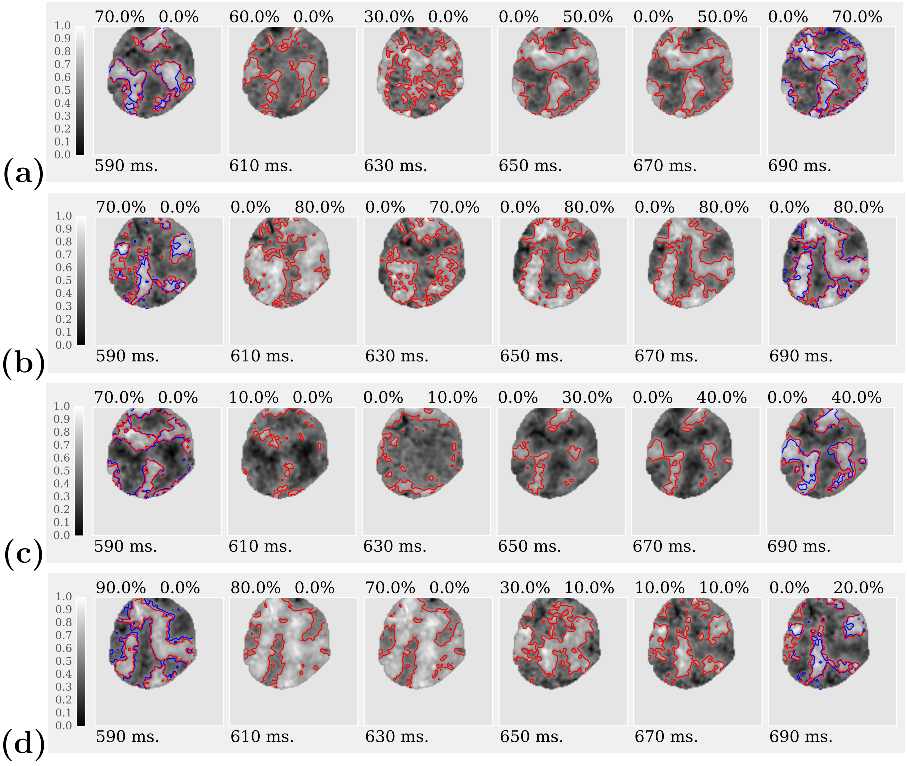


Figure 4.17: Dataset $S^{(4)}$ (orientation shift +90 °). Normalized weight vectors $\hat{\omega}^{\text{ave.}}_{y,t}$ and scores $\tilde{\mu}_{\iota,t}$ (Left: blue scores. Right: red scores. See Figure 4.13) for time indicated under each frame. The red contours are the level set of the current normalized weight vectors at 0.5. The blue contours are the level set of the normalized weight vectors $\hat{\omega}^{\text{ave.}}_{y}$ computed in Section 4.1 at 0.5. **(a)** Label $y = 0$ corresponding to a 0 ° orientation. **(b)** Label $y = 1$ corresponding to a 45 ° orientation. **(c)** Label $y = 2$ corresponding to a 90 ° orientation. **(d)** Label $y = 3$ corresponding to a 135 ° orientation. Supplementary figure online.

5 A Steerable Model of Activation Maps

In this section, we develop a simple model of activation maps. Such a model provides a way to interpolate between activation maps, it is then possible to simulate the activation maps learned after the abrupt rotation of the stimulus in protocols 3 and 4, see Figures 4.16 and 4.17. We also detail some consequences of the model on the computation of orientation maps defined in Equation (10). Our model is related to the seminal paper of Swindale [185] about tuning curves (Proposition 19). The main advantage of our method is its simplicity compared to other approaches [100, 9].

5.1 The Steerable Model

The model is similar to the concept of steerable filters [113]. Knowing two filters f_1 and f_2, one can compute an oriented version of these filters by linear combination. For instance, let $f_1 = \partial/\partial x$ and $f_2 = \partial/\partial y$ be the horizontal and vertical derivatives on the plane. Then, for any $\theta \in [0, 2\pi], g_\theta = \cos(\theta) f_1 + \sin(\theta) f_2$ is the directional derivative of angle θ. By the use of complex numbers, we have the following Definition.

Definition 12. *Let $\theta_0 \in \mathbb{R}/\pi\mathbb{Z}$ and $\theta_1 = \theta_0 + \frac{\pi}{4}$. The complex maps $(Z_\theta)_{\theta\in\mathbb{R}/\pi\mathbb{Z}}$ associated to the activation maps $(M^{(\theta_0)}, M^{(\theta_1)})$ is*

$$\forall\theta \in \mathbb{R}/\pi\mathbb{Z}, \quad Z_\theta = (M^{(\theta_0)} + \mathrm{i}M^{(\theta_1)}) \exp\left(-2\mathrm{i}(\theta - \theta_0)\right).$$

The activation map evoked by a stimulus with orientation $\theta \in \mathbb{R}/\pi\mathbb{Z}$ is

$$M^{(\theta)} = \mathcal{R}\mathrm{e}(Z_\theta).$$

The interpolation between maps is straightforward because it follows the linear interpolation between orientations. For any $(\theta_0, \theta_1) \in (\mathbb{R}/\pi\mathbb{Z})^2$, for all $t \in [0, 1]$, the activation maps $M^{(\theta_0(1-t)+\theta_1 t)}$ interpolate between activation maps $M^{(\theta_0)}$ and $M^{(\theta_1)}$. The following Proposition gives an expression of activation maps as linear combination of trigonometric functions.

Proposition 19. *The activation maps $M^{(\theta)}$ verifies*

$$\forall\theta \in \mathbb{R}/\pi\mathbb{Z}, \quad M^{(\theta)} = M^{(\theta_0)} \cos\left(2(\theta - \theta_0)\right) + M^{(\theta_1)} \sin\left(2(\theta - \theta_0)\right).$$

Proof. The proof follows simple arithmetic by computing $\mathcal{R}\mathrm{e}(Z_\theta)$. □

An interesting aspect of this model is that it provides a way to compute orientation maps using a continuous sum over all orientations. First, we recast the Definition 10 of orientation maps.

Definition 13 (Orientation Map)**.** *An orientation map O is defined by*

$$O = \int_0^\pi M^{(\theta)} \exp\left(2\mathrm{i}\theta\right) \mathrm{d}\theta.$$

Such Definition combined with Proposition 19 provides a way to compute an orientation maps as a function of only two activation maps $M^{(\theta_0)}$ and $M^{(\theta_0)}$.

Proposition 20. *An orientation map O verifies*

$$O = \frac{1}{2}\operatorname{Arg}\Bigg(\frac{\pi}{2}\cos(2\theta_1)M^{(\theta_1)} - \pi\cos(\theta_1)\sin(\theta_1)M^{(\theta_2)} \\ + \mathrm{i}\bigg(\pi\cos(\theta_1)\sin(\theta_1)M^{(\theta_1)} + \frac{\pi}{2}\cos(2\theta_1)M^{(\theta_2)}\bigg)\Bigg)$$

where Arg *is the function that give the argument of a complex number in* $[0, 2\pi)$*.*

Proof. Plug the expression of $M^{(\theta)}$ established in Proposition 19 into the Definition 13 of an orientation map. Then, we use fact that

$$\int_0^{\pi}\cos(2(\theta-\theta_0))\cos(2\theta)\mathrm{d}\theta = \int_0^{\pi}\sin(2(\theta-\theta_0))\sin(2\theta)\mathrm{d}\theta = \frac{\pi}{2}\cos(2\theta_0) \quad \text{and}$$
$$\int_0^{\pi}\cos(2(\theta-\theta_0))\sin(2\theta)\mathrm{d}\theta = -\int_0^{\pi}\sin(2(\theta-\theta_0))\cos(2\theta)\mathrm{d}\theta = \pi\cos(\theta_0)\sin(\theta_0).$$

□

5.2 Validation on Dataset 2

First, we validate this model on Dataset $D^{(2)}$. To this purpose, we use the activation maps $(\omega_y^{\text{ave.}})_{y\in\mathcal{Y}}$ obtained using logistic regression (see Section 4.1. Therefore, we use $(M^{(\theta_0)}, M^{(\theta_1)}) = (\omega_0^{\text{ave.}}, \omega_1^{\text{ave.}})$ where $\theta_0 = 0°$ and $\theta_1 = 45°$. Then, we compute M^{θ} for $\theta \in \{90°, 135°\}$ that we compare to $(\omega_2^{\text{ave.}}, \omega_3^{\text{ave.}})$. This comparison is shown in Figure 5.1. In the top row, activation maps and orientation maps are consistent with the ones obtained in Figure 4.1. In the bottom row, the differences with activation maps $(\omega_y^{\text{ave.}})_{y\in\mathcal{Y}}$ are obviously equal to zeros for $y \in \{0, 1\}$ and stay low with respect to the range of the activation maps for $y \in \{2, 3\}$. Finally, the differences between maps are low and reach up to 10 ° in some areas. The high differences are saturated to 10 ° but they appear only in some pixels around pinwheels or in noisy areas. Second, we validate Proposition 20 in Figure 5.2 by comparing the maps computed using Definition 10 and Proposition 20. Again the differences are low. In fact, these differences corresponds very closely (difference is below 1 °) to the ones obtained at bottom right of Figure 5.1.

5.3 Interpolation on Dataset 3

The initial goal of this model is to provide a way to interpolate between activation maps. Therefore, we test such an interpolation on Dataset 3. Recall

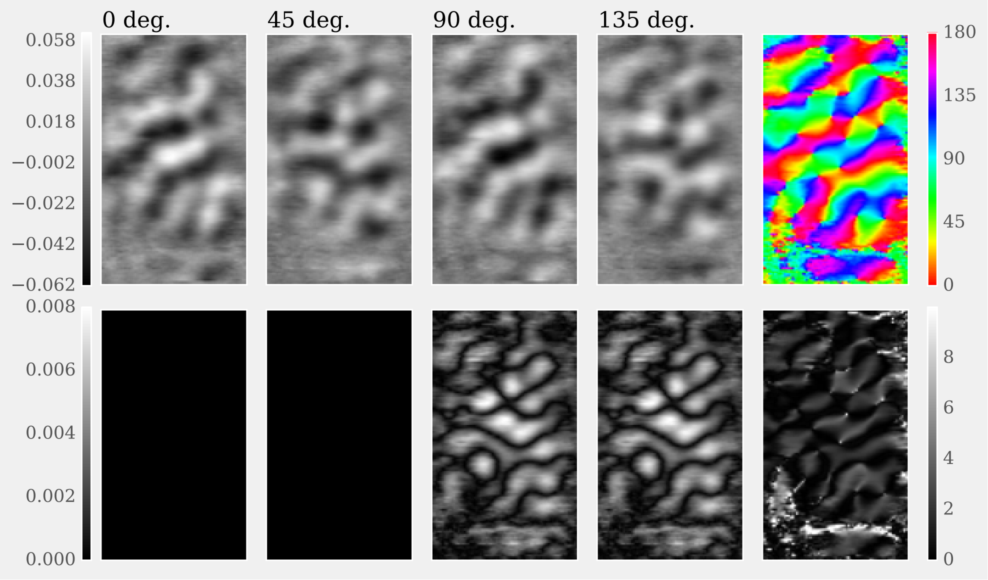


Figure 5.1: Top: activation maps obtained on dataset $S^{(2)}$ with Definition 12 and the corresponding orientation map computed using Definition 10. We use $M^{(\theta_0)} = \hat{\omega}_0^{\text{ave.}}$ and $M^{(\theta_1)} = \hat{\omega}_1^{\text{ave.}}$. Bottom: the differences with experimental results $|\hat{\omega}_y^{\text{ave.}} - M^{(\theta_y)}|$ and the differences between their corresponding orientation maps computed using Equation 4.1. The difference is saturated at 10 degree in order to highlight the small differences. Supplementary figure online.

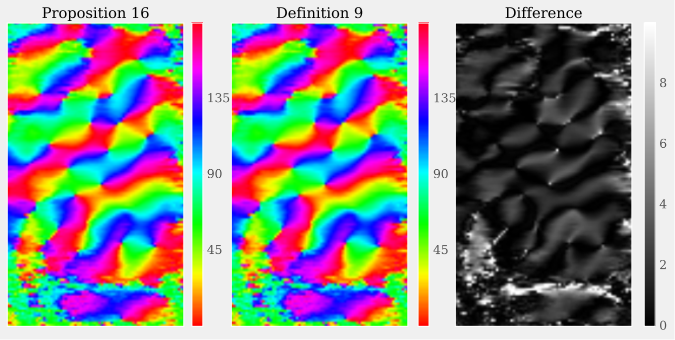


Figure 5.2: Orientation maps obtained on dataset $S^{(2)}$ with Proposition 20 (left), Definition 10 (center) and their difference (right). The difference is saturated at 10 degree in order to highlight the small differences.

that in this protocol the stimulus is rotated from 135° at middle stimulation time. First, we compute the weight vectors $\hat{\omega}_{y,t}^{\text{ave.}}$ and we use them to obtain the diagonal coefficients $(\Lambda_{y,y,t})_{y\in\mathcal{Y}}$ of the average confusion matrices Λ_t at each time $t \in \mathcal{T}$. Second, we model the activation maps using Proposition 19 by $M^{(y,\theta_t)}$ were for all all $y \in \tilde{\mathcal{Y}}$ (defined in Section 4.3.2) and $t \in \mathcal{T}$,

$$\theta_t = \theta_y \mathbb{1}_{[0,575]}(t\Delta_t) + \theta_{y+3}\mathbb{1}_{[700,1280]}(t\Delta_t) + (\theta_y + \theta_{y+3}\frac{t\Delta_t - 575}{125})\mathbb{1}_{[575,700]}(t\Delta_t).$$

Figure 5.3 shows the comparison of the diagonal coefficients of the confusion matrices. The modeled activation maps are able to predict the classes as good as the learned weight vectors. Moreover, in Figure 5.4, we show the normalized weight vectors $\hat{\omega}_{y,t}^{\text{ave.}}$ both with their contours (red) at 0.5 and the contours of the modeled activation maps $M^{(y,\theta_t)}$ (blue). The red contours are generally inside the blue contours, see Figure 5.4**(b)(c)**. The contours coincide for Figure 5.4**(a)** whereas they differ slightly in Figure 5.4**(d)** (640 and 665 ms).

Main Conclusions

- We build a simple model of activation maps.
- The model successfully accounts for the data.
- The model supports the role of lateral connections in case of a non-orthogonal rotation of the stimulus, see [14, 212]

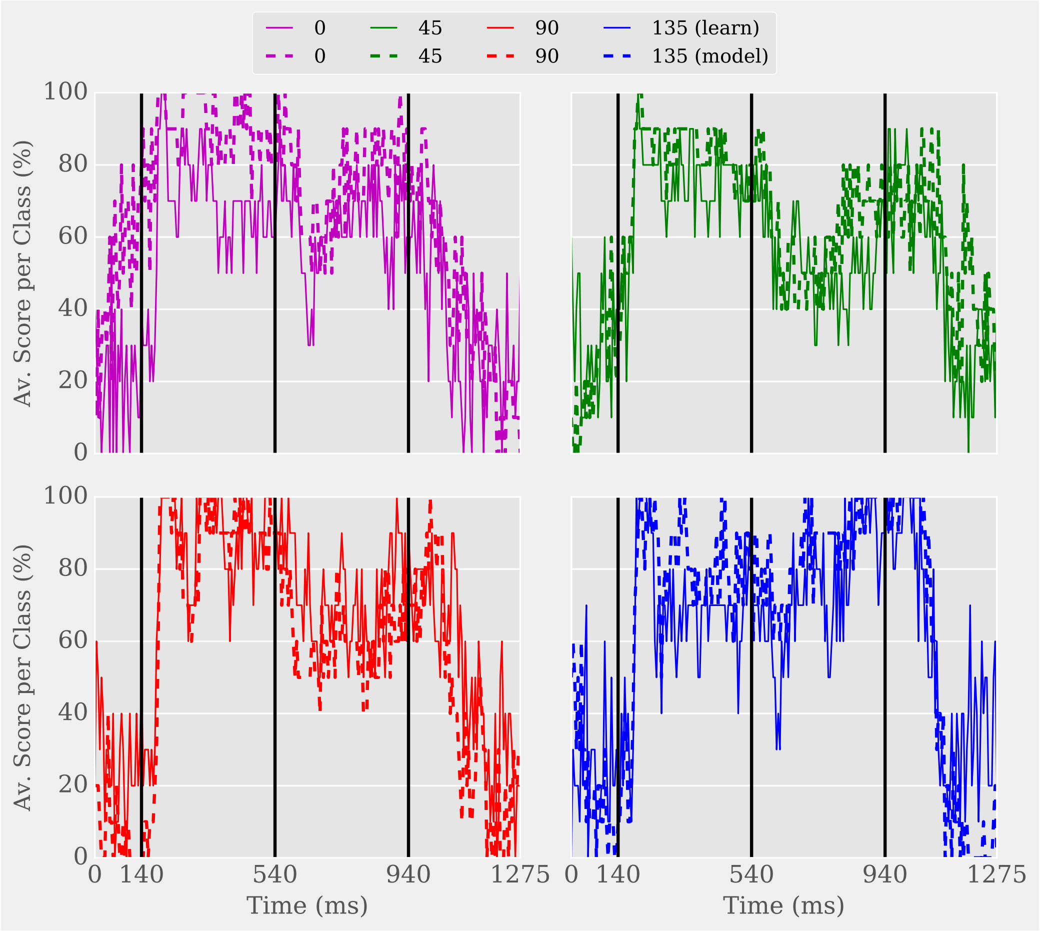


Figure 5.3: Plain lines: coefficients $\Lambda_{y,y,t}$ of the average confusion matrix Λ_t computed at each time $t \in \mathcal{T}$ on dataset $S^{(3)}$. Dotted lines: the same coefficients except that we use the interpolated maps $M^{(\theta_t)}$ to compute the confusion matrices (see Equations 6 and 5.2).

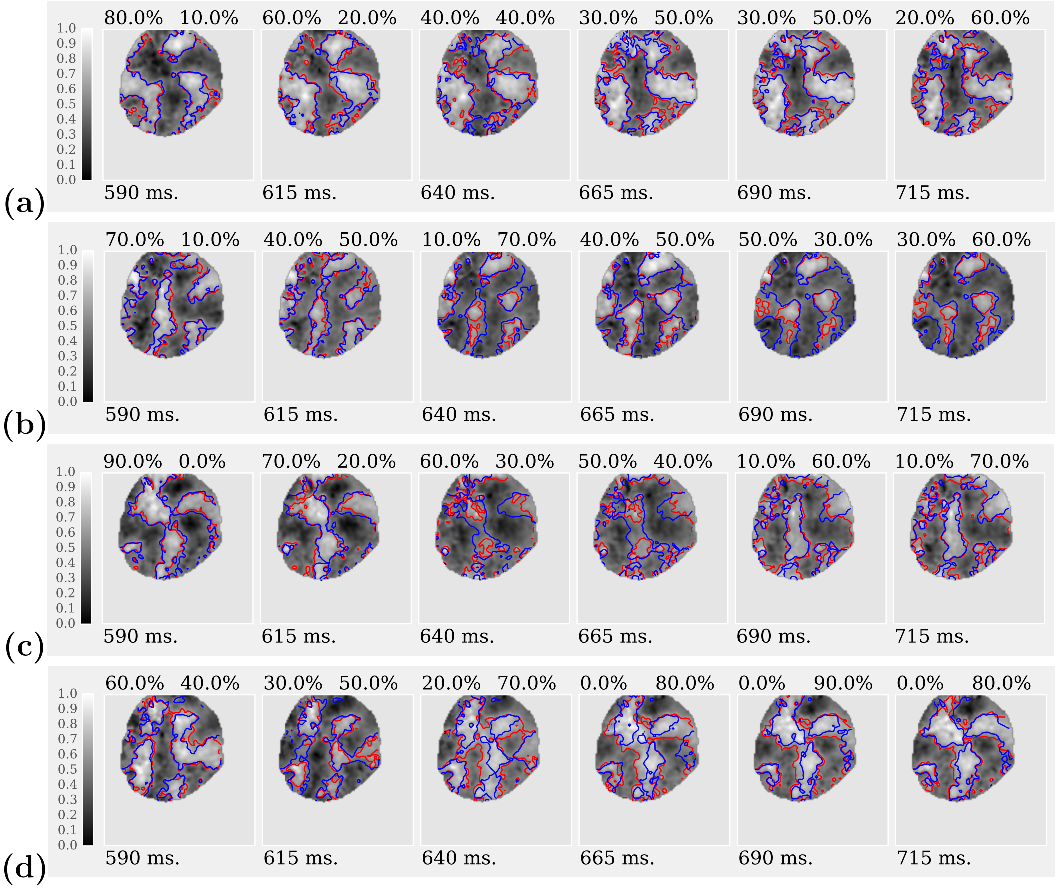


Figure 5.4: Dataset $S^{(3)}$ (orientation shift +135 °). Normalized weight vectors $\hat{\omega}^{\text{ave.}}_{y,t}$ and scores $\tilde{\mu}_{\iota,t}$ (Left: blue scores. Right: red scores. See Figure 4.13) for time indicated under each frame. The red contours are the level set of the current normalized weight vectors at 0.5. The blue contours are the level set of the interpolating activation maps $M^{(y,\theta_t)}$ at 0.5. **(a)** Label $y = 0$ corresponding to a $\theta_1 = 0$ ° orientation. **(b)** Label $y = 1$ corresponding to a $\theta_1 = 45$ ° orientation. **(c)** Label $y = 2$ corresponding to a $\theta_1 = 90$ ° orientation. **(d)** Label $y = 3$ corresponding to a $\theta_1 = 135$ ° orientation.

⋆ VI ⋆

Supervised Classification For Extracellular Recording

In this chapter, we first review the principles and processing of Extracellular Recording (ER). Then, we compare the results obtained using the different algorithms introduced in Chapter IV on several datasets, including datasets that were recorded under MC simulations. In particular, we show that the signals recorded under oriented stimulations with a variable orientation contents or a variable spatial frequency contents can be discriminated. Then, we present spatially and temporally localized analysis methods. These analyses reveal differences between the information contained in the spiking activity of single neurons and that of a neural population. Finally, we build a simple Linear/Non-linear Poisson model which enables us to reproduce the behaviour of the recorded datasets.

Contents

1 Introduction

In this chapter, we focus on data obtained with Extracellular Recordings (ER). Before going into the details of how informative it is about brain functions, we briefly recall some background about this method.

1.1 Principles of ER

The extracellular recording technique is the oldest way to measure neural activity; it consists in inserting an electrode in brain tissue to measure current variations. The path to electrophysiology started in the second half of the 18th century, with Luigi Galvani's experiment on frogs which lead him to hypothesize an intrinsic "animal electricity" [148]. A century later, Santiago Ramón y Cajal laid the foundation of neuron theory by describing the nervous system as a network of polarized nerve cells in contact with each other at the synapse level [77]. Another forty years went by before Edgar Adrian was able to record the first electrical signal from nerve fibers using a Lippmann electrometer [5]. The year 1940 marked the beginning of miniaturization when Renshaw, Forbes and Morrison used microelectrodes in the cat's hippocampus [158]. Miniaturization led to major results in neuroscience thanks to intracellular recording [117]. It allowed Hodgkin and Huxley to formalize the electrical behavior of neurons by describing the famous equation named after them [84]. It also permitted Hubel and Wiesel to identify receptive fields (see Section 1.1) of single neurons in the cat's primary visual cortex [87]. In the last fifty years, microelectrodes diversified in terms of sizes, numbers and shapes, enabling researchers to record multiple neurons at the same time over larger and deeper volumes of cortex [36, 47]. Figure 1.1 displays a sketch of the laminar electrode used to collect the data analyzed in this work.

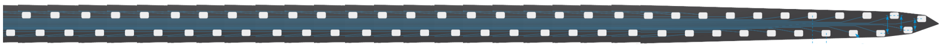


Figure 1.1: The laminar electrode used during ER experiment at UNIC lab. It has 64 recording points staggered along a silicon array. This Figure is extracted from technical documentation of Neuronexus and edited by Yannick Passarelli.

1.2 Processing of ER Data

Laminar electrodes are used to record up to a few dozens of neurons distributed in the different layers of the cortex or across the laminar plane (for tangential penetrations). In extracellular recordings the signal is twofold: the high-frequency component (400Hz to few thousands) which corresponds to spikes of single neurons and is known as multiunit activity (MUA), and the low-frequency component (cut off at about 300Hz which represents an averaged activity of multiple neurons and is known as local field potential (LFP). These raw signals can be treated directly, however, for MUA it is of a particular interest to associate the spikes to their corresponding neuron, especially in order to study them in particular. This corresponds to the spikes sorting problem, which is a multi-canal blind-deconvolution inverse problem, and can be tackled using a variety of inverse problem sparse regularization methods, see for instance [48, 118, 161] for a few relevant previous works. After performing such a signal separation, the signal is known as Single Unit Activity (SUA). At UNIC, they now use the toolbox developed by Rossant *et al.* [161] called Klusta.

Understanding how information is encoded along the vertical and horizontal connections is fundamental to explain the relation between different areas of the cortex like, for instance, cortical columns. Indeed, many studies highlight the importance of feedforward and feedback mechanisms for information processing. For instance, Martinez *et al.* [119] compare orientation selectivity of neurons across the different layers and show that orientation tuning curves have common properties in the same layer but different ones across layers. Later, Hirsch and Martinez [83] examine the micro-circuitry of the visual cortical column in order to outline the connectivity between layers and offer a better understanding of their role. Lamme *et al.* [109] review the role of the feedforward, horizontal and feedback connections. They conclude that feedforward connections are central to the receptive field concept whereas horizontal and feedback connections involve higher level tasks like perceptual organization and attention. Another interesting paper of Lamme [108] suggests that the blindsight phenomenon (the ability to respond to visual stimuli without consciously perceiving them) occurring in some patients with lesions in V1 can be caused by the absence of feedback to V1. Ferster and Miller [52] compare the feedforward and feedback models and what they imply in the visual processing. A feedback model can only encode a finite number of stimuli depending on its number of connection whereas a feedforward model is more flexible and encode different stimuli differently. Our study focuses on how information is distributed in time and among a neural population. We mainly make use of supervised learning to probe if neural responses contain information related to the Motion Cloud parameters (mainly bandwidths).

1.3 Previous Works: Machine Learning for ER

Electrophysiological data are often used to compute neurons' receptive field using reverse correlation techniques [98, 40]. Also known as Spike Triggered Average (STA) and Spike Triggered Covariance (STC), these methods are well documented [92] and extended beyond the hypothesis of Gaussian stimuli [168]. Moreover, they have worthy consistency and efficiency properties [143]. The STA and STC are useful to identify the linear component in standard Linear-Nonlinear Poisson (LNP) models used to mimic the spiking behavior of neurons [172]. Although these methods are not considered as standard supervised classification algorithms, they consist in averaging the values of the stimulus presented at times before spikes. So, spikes are implicitly associated with their stimulus (belonging to a similar stimulation class, *e.g.* with constant orientation). Important to mention, the work of Park [145] uses a Bayesian estimation of receptive field that provides smooth results. In relation with Chapters II and III, these works of Neri *et al.* [132, 134] makes use of reverse correlation techniques in psychophysics to study detection and stereoscopic vision. Neri and Levi [133] further compare neurons' receptive field to humans' perceptual field and highlight their common features.

However, the receptive field is not a fixed characteristic of the neuron because it is to be stimulus dependent [55]. Variation of the estimated receptive field as a function of the stimulation class is a crucial information. This approaches thus has a strong machine learning flavor. Even so, there exists few papers that make use of standard supervised learning techniques. In particular, Hung *et al.* [90] use a kernel linear regression to classify the responses of IT neurons to different stimuli and make stimulus predictions knowing the neurons' response. More recently Yamins *et al.* [215] use a SVM classifier to classify the responses of IT neurons and make predictions about the stimulus. They compare this performance to the ones obtained using the same classifier over features that are computed from the stimuli using different models of the visual cortex (V1, V2) and other computational models. They show that the model based on Convolution Neural Networks (CNNs) is the only one that closely follow the performances (across all experimental conditions) obtained using the IT neurons' response as features.

This review of the state of the art reveals that the usage of ML techniques in the fields of ER is very limited. It is the purpose of this chapter to explicitly and systematically propose ML-based solutions of the exploration of such ER datasets.

1.4 Contributions

1.4.1 Main Contributions

The first major contribution results from the use of MCs as stimuli. We find that small neural populations (few dozens of neurons) contain enough information to discriminate between homogeneously and heterogeneously oriented stimuli. Such populations contain also enough information to discriminate between stimuli with narrow and broad spatial frequency bands (see Section 4). The second major contribution is a methodology of temporal analysis of prediction performances. We find that neural populations have, systematically, better classification performances than any single neurons when stimulated by MCs (see Section 5). However, when stimulated with natural movies, there exists neurons that provide classification performances that are similar to the entire population. The third major contribution is a simple Linear/Non-linear Poisson (LNP) spiking neurons model that generate synthetic data (see Section 6). When generated with MCs, the synthetic data provide results that are similar to the one obtained on the experimental recordings. We provide an online[1] example of data synthesis using the proposed model and MCs.

1.4.2 Related Works

In order to accurately define the context in which spot this chapter we go back to the most important references.

The papers of Hung [90] and Yamins [215] are the most related papers to our work. They both make use of supervised classification in different ways that we use together in this chapter. In Hung's paper, they make an extensive use of their classifier: they compare the performances obtained using different combinations of the recorded signal (single unit activity, multi unit activity, local field potential), different time bins, different times and different stimulation conditions. Such an approach is similar to feature selection, this helps understand where the information is concentrated in a population of neurons. On the contrary in Yamins paper, they make a single use of the classifier. However they build many models that are able to simulate data, by then using the classifier on these synthetic data they obtain different classification performances that are compared to the ones obtained with the real data. Such an approach allows to discriminate between good an bad models and is therefore another way to quantify their quality.

[1] http://nbviewer.jupyter.org/github/JonathanVacher/projects/tree/master/lnp_spiking_neurons/

Goris *et al.* [71] make use of MC-like stimuli with different orientation bandwidth to tackle the question of tuning diversity in neurons. Once the diversity is established they asked whether the observed diversity is an advantage to represent local patches of natural images which appears to be true. Although we are using the same kind of stimulations, our analysis is different as we are checking if the recorded neural population contains information that allows for the discrimination of the stimuli. By using another physiologically-based texture synthesis algorithm (see, [149]), Freeman *et al.* [57, 58] precise the role of the V2 area. Their main finding is that V2 neurons are sensitive to the high-order statistical correlations present in natural images. The works of Goris and Freeman employ the most promising strategy to understand the visual processing: they both use complexified artificial stimuli (mixture of drifting gratings) and mimicked natural images (also called naturalistic textures) to answer a precise scientific question. They subsequently test their finding over natural images.

2 Material and Methods

2.1 Animal Preparation

The animal preparation follows closely the description given in Section 2.1 except for the craniotomy. It is smaller, a few millimeters squared and located at anteroposterior coordinate $(-2, -2)$ to allow for the electrode implantation in A17.

2.2 Setup

One type of Neuronexus probe was used : 1×64, on shank 64 channels (A1x64-Poly2-6mm-23s-160) The electrodes were lowered through cortex using a micromanipulator (Luigs & Neumann). Silicon probes being quite large, in order to avoid as much damage as possible of the tissue, advancement through the brain was made very slowly and 1m at a time. Acquisition was made with a Blackrock Cerebus system. Signal from the probes was amplified, filtered and digitized by an amplifier then transmitted to a Neural Signal Processor (NSP) via a optic-fiber. The amplifier filters the signals with a first order highpass filter at 0.3 Hz and a third-order low-pass filter at 7.5 kHz. The filtered neural signals from each electrode are digitized with 16-bit resolution at 1 µV per bit with a sampling rate of 30000 kHz. The analog filtering of the electrode signals allow both low frequency field potentials extracellular spike signals to pass through. The neural signals are later separated into low frequency (filtering between $1 - 250$ Hz) and and spike signals (highpass at

250 Hz) by digital filtering in the Neural Signal Processor (NSP). The NSP does an online analysis and then transmits the processed data to a host PC system via an Ethernet cable. On the host PC a homemade software, Elphy (Gérard Sadoc, CNRS), was in communication with the Blackrock system in order to save the acquired data.

2.3 Visual Stimulation

Data are collected under three different protocols described below. A protocol consists in presenting a number $C \in \mathbb{N}$ of stimuli variously parametrized. This operation is then repeated a certain number $R \in \mathbb{N}$ of times. Figure 2.1 is illustrating the protocols used for datasets 1, 2 and 4 to 6. A LCD screen (ASUS) with a resolution of 1920×1080 pixels and a refreshing rate of 120 Hz was placed at 57 cm of the animal so that 1 cm on the screen is equal to one visual degree. All visual stimuli were generated with Elphy, maximum and background luminance were set at 40 $\mathrm{cd\,cm^{-2}}$ and 12 respectively.

Dataset 1 Stimuli were Motion Clouds with parameters $z_0 = 0.6$ c/°, $B_Z = 1.35, \sigma_V = \frac{1}{t^\star z_0}$ with $t^\star = 0.666$ ms and drifting in a single direction for 1000 ms at speed $v_0 = 2.5$ c/s starting 500 ms after recording onset. Six orientations were presented ($\theta_0 = 0$ °, 60 °, 120 °, 180 °, 240 ° and 300 °) with five orientation bandwidths ($\sigma_\Theta = 0.35$, 0.79, 1.12, 1.58 and 2.24), totaling $C = 30$ stimulation conditions. The stimuli presentation were pseudo-randomly interleaved and were displayed monocularly.

Dataset 2 Stimuli were Motion Clouds with parameters $z_0 = 0.6$ c/°, $B_Z = 1.35, \sigma_V = \frac{1}{t^\star z_0}$ with $t^\star = 0.666$ ms and drifting in a single direction for 1000 ms at speed $v_0 = 2.5$ c/s starting 500 ms after recording onset. Six orientations were presented ($\theta_0 = 0$ °, 60 °, 120 °, 180 °, 240 °, 300 °) with five orientation bandwidths ($\sigma_\Theta = 0.79$, 1.12, 2.74, 3.87 and 5.48), totaling $C = 30$ stimulation conditions. The stimuli presentation were pseudo-randomly interleaved and were displayed monocularly.

Dataset 3 Stimuli were Motion Clouds with parameters $z_0 = 0.6$ c/°, $\sigma_\Theta = 0.5, \sigma_V = \frac{1}{t^\star z_0}$ with $t^\star = 0.666$ ms and drifting in a single direction for 1000 ms at speed $v_0 = 2.5$ c/s starting 500 ms after recording onset. Six orientations were presented ($\theta_0 = 0$ °, 60 °, 120 °, 180 °, 240 ° and 300 °) with six spatial frequency bandwidths ($B_Z = 1.76$, 2.38, 2.55, 2.77, 2.84, 2.9), totaling $C = 36$ stimulation conditions. The stimuli presentation were pseudo-randomly interleaved and were displayed monocularly.

Dataset 4 Stimuli were Motion Clouds with parameters $z_0 = 0.6$ c/°, $\sigma_\Theta = 0.5, \sigma_V = \frac{1}{t^\star z_0}$ with $t^\star = 0.666$ ms and drifting in a single direction for 1000 ms at speed $v_0 = 2.5$ c/s starting 500 ms after recording onset. Six orientations were presented ($\theta_0 = 0$ °, 60 °, 120 °, 180 °, 240 °, 300 °) with six spatial frequency bandwidths ($B_Z = 1.2$, 1.84, 2.11, 2.31, 2.46 and 2.61), totaling $C = 30$ stimulation conditions. The stimuli presentation were pseudo-randomly interleaved and were displayed monocularly.

Dataset 5 Stimuli were 10 different natural movies. The stimuli presentation were pseudo-randomly interleaved and were displayed monocularly centered on the approximate receptive fields center.

Dataset 6 Stimuli were 15 different natural movies with scrambled frame (*ie* with removed temporal correlations). The stimuli presentation were pseudo-randomly interleaved and were displayed monocularly centered on the approximate receptive fields center.

2.4 Preprocessing

A signal $(s_t)_{t\in\mathcal{T}}$ only consists of spiking activity *i.e* it is binary and discrete. It is, therefore, not appropriate for the algorithms we use. Instead we compute an approximate spike density $(d_t)_{t\in\mathcal{T}}$ by convolving the spike train against a Gaussian kernel, for all time $t \in \mathcal{T}$

$$d_t = \frac{\sum_{t'\in\{-w_f,\dots,w_f\}} \exp\left(-\frac{t'^2}{2\sigma_f^2}\right) s_{t+t'}}{\sum_{t'\in\{-w_f,\dots,w_f\}} \exp\left(-\frac{t'^2}{2\sigma_f^2}\right)},$$

where σ_f is the width of the averaging window. An example of spike density approximation is shown in Figure 2.1. When, the width of the averaging window increases the density flatten around the average number of spike over the entire window. We usually choose $\sigma_f = 10$ to be consistent with standard averaging methods (see peristimulus time histogram for more information).

2.5 Recorded Datasets Organization

In this section, we consider both MUA and SUA. The MUA consists of spiking activity recorded on 64 channels whereas the SUA consists of spiking activity of individual neurons. In both cases we denote by $\mathcal{Q} = \{1, \dots, Q\}$ the number of channels or neurons. Such a signal is recorded for every stimulation

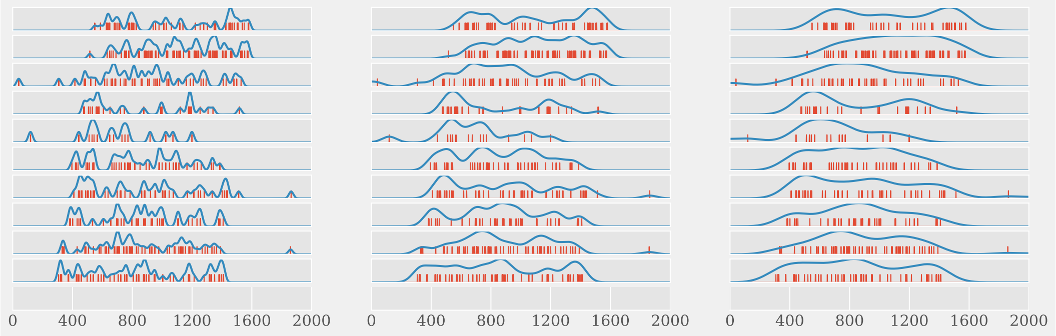


Figure 2.1: An example of the approximate spike density computation for $\sigma_f = 15, 50$ and 100 (from left to right). Abscissa indicates time in ms. The lines represents different repetitions of a same condition recorded in dataset $D^{(1)}$ (see below).

conditions $c \in \mathcal{C} = \{1, \ldots, C\}$ and repetition $r \in \mathcal{R} = \{1, \ldots, R\}$. Moreover, we denote by $\mathcal{T} = \{0, \ldots, T\}$ the set of time samples. Therefore, dataset number $k \in \mathcal{K} = \{1, \ldots, 6\}$ is therefore denoted

$$D^{(k)} = (d^{(k)}_{q,t,c,r})_{(q,t,c,r) \in \mathcal{Q} \times \mathcal{T} \times \mathcal{C} \times \mathcal{R}}.$$

The same remarks hold for the formulation of a machine learning classification problem, cross-validation and technical details, see Section 2.5.

3 Comparison of the Different Algorithms

In order to compare the different algorithms introduce in Chapter IV we evaluate their prediction performances on each entire dataset, *ie* we consider every stimulation parameter as a different label. For the natural movies, we consider that each movie corresponds to a label. The contribution is twofold because we provide a comparison different algorithms and in addition this comparison is performed on innovative protocols. Indeed, to our knowledge similar protocols to the ones used to obtain datasets $D_{(1)}$ and $D_{(2)}$ are used only in [71]. We do not know any study that tests various spatial frequency bandwidths corresponding to protocols used to collect datasets $D_{(3)}$ and $D_{(4)}$.

3.1 Channels and Time Samples as Feature Space

Design of $\mathcal{X}$ and $\mathcal{Y}$ We choose to concatenate channel and time spaces as a feature space $\mathcal{X} = \mathbb{R}^{\mathcal{Q} \times \mathcal{T}}$ *ie* the entire recorded signal space. The number of classes varies from one dataset to another. But, in each case we set $\mathcal{Y} = \mathcal{C} \times \mathcal{R}$. Datasets $D_{(1)}$ and $D_{(2)}$ have 30 classes. Datasets $D_{(3)}$ and $D_{(4)}$ have 36 classes.

Finally dataset $D_{(5)}$ has 10 classes and dataset $D_{(6)}$ has 15. Therefore,

$$\forall i = (r, c) \in, \quad x_i = d^k_{\cdot,\cdot,c,r} \in \mathcal{X},$$

where $d^k_{\cdot,\cdot,c,r}$ is defined in (2.5). Table 3.1 summarizes the relevant experimental parameters.

	Δ_t (ms)	$t_{\text{on}}\Delta_t$	$t_{\text{off}}\Delta_t$	$T\Delta_t$	C	R
$D^{(1)}$	1	500	1500	2000	30	20
$D^{(2)}$	1	500	1500	2000	30	15
$D^{(3)}$	1	500	1500	2000	36	10
$D^{(4)}$	1	500	1500	2000	36	20
$D^{(5)}$	1	1000	11000	12000	10	15
$D^{(6)}$	1	1000	11000	12000	15	10

Table 3.1: Experimental parameters of the different datasets.

Results Figure 3.1 summarizes the results by showing the average score over folds μ_ι and their standard deviation σ_ι (both defined in Equation (5.1)). For each dataset, LC and NC shows the best scores with an small advantage for LC. The low standard deviations ensure that scores are at least $3\sigma_\iota$ above chance level (*ie* there are more than 99.7% chance that the score is above chance). The score of LDA varies from one dataset to another.

- Datasets $D^{(1)}$ and $D^{(4)}$: the score is at a reasonable level similarly to LC and NC with an smaller standard deviation.
- Datasets $D^{(3)}$, $D^{(5)}$ and $D^{(6)}$: the score is around chance level.
- Dataset $D^{(2)}$: although the score is medium the standard deviation is extremely large.

As for the VSDi datasets with no dimensionality reduction, GNB and QDA fail to classify the data except, surprisingly, for GNB on dataset $D^{(5)}$. The reasons that explain such a poor performance are the same as for the VSDi signal *ie* features independence assumption and high dimension of the feature space. These are detailed in Section 4.1. In order to check the structure of predictions, we show in Figure 3.2 the confusion matrices Λ (defined in Equation (5.2)) and the distance d_p (defined in Equation (9)) obtained for each algorithm. We associate to each couple of stimulation parameters $(\theta_0, \sigma_\Theta)$ (or (θ_0, B_Z)) the label y corresponding to its rank in the increasing lexicographical

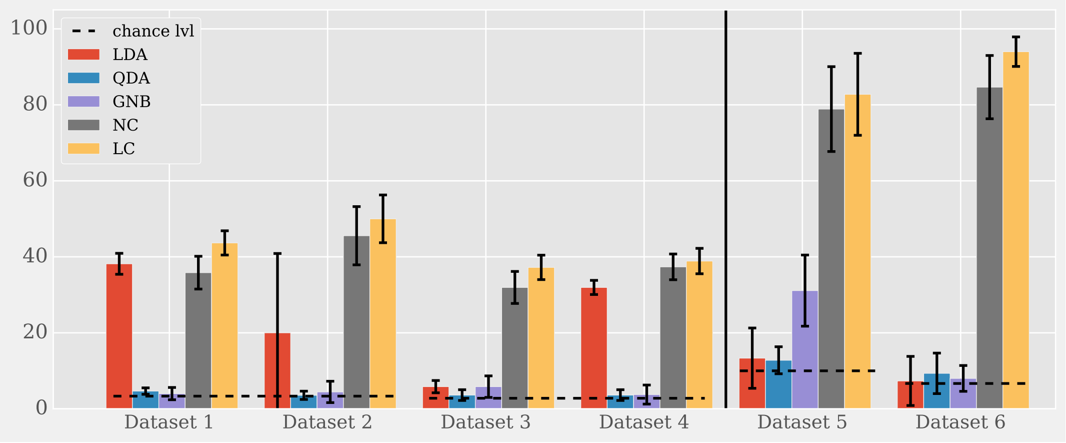


Figure 3.1: Classification performances of the different algorithms on the 6 datasets. QDA: Quadratic Discriminant Analysis. LDA: Linear Discriminant Analysis. GNB: Gaussian Naive Bayes. NC: Nearest Centroid. LC: Logistic Classification.

order. For dataset $D^{(1)}$ (Figure 3.2**(a)**), there is no apparent bias using the QDA, the classification is bad. The GNB tends to predict any label to be between 10 and 14 corresponding to a stimulation with orientation $\theta_0 = 120$ °. The LDA, NC and LC methods show good results with the expected diagonal block structure reflecting the correct predictions of orientations. Inside each block, the predictions are concentrated around the diagonal. Therefore, if bandwidths are not well predicted they are in fact mixed up with the neighboring bandwidths.

- QDA and GNB: the distances d_p are around 0.5 meaning that stimulation parameters are on average at half the maximum possible distance between labels (uniform matrix corresponds to $d_p = 0.52$).
- LDA, NC and LC: the distances d_p are between 0.17 and 0.20 which reinforces the fact that bandwidths are mixed up with their neighboring bandwidths (uniform blocks along the diagonal corresponds to $d_p = 0.29$).

For dataset $D^{(3)}$ (Figure 3.2**(b)**), there is no apparent bias using QDA and LDA. These two methods performs badly with a distance $d_p = 0.47$ (close to the distance of uniform matrix). The GNB method shows strong bias in the prediction as the visible columns indicates. Its performances are similar to LDA, however the distance $d_p = 0.43$ is a little smaller showing that the errors made are closest to real labels than the ones made by LDA and QDA. Finally, LC and NC show good prediction scores with the expected diagonal block structure indicating the correct predictions of orientation. Again, the distances d_p around 0.15 show that predicted spatial frequency bandwidths are close to the true spatial frequency bandwidths.

Partial Conclusions of Section 3.1

- LC and NC perform best.
- LDA results are variable.
- GNB and QDA perform badly showing the existence of significant spatial and temporal correlations.
- GNB shows significant prediction bias.

3.2 Dimension Reduction Using PCA

In order to increase the prediction performances of QDA, LDA and GNB, we perform a dimension reduction method using PCA. Following Section 3.2, we test $n_{\text{pca}} \in \mathcal{E}_{\text{pca}} = \{5, 10, 20, 40, 80, 150\}$ and compute the associated average score over the folds $\mu_\iota(n_{\text{pca}})$ (defined in Equation (5.1)). Then, we choose the number of PCA components n_{pca} that provide the highest score.

Results Contrary to VSDi data, the chosen number of PCA components n_{pca} depends highly on the algorithm we have used. As the right hand side of Figure 3.3 shows, QDA performs best with $n_{\text{pca}} = 10$. The GNB and LDA methods perform best using $n_{\text{pca}} = 40$. The NC method performs best with $n_{\text{pca}} = 80$. Finally, the LC methods performs best using $n_{\text{pca}} = 150$. Generally, the dimension reduction method improves the prediction performances of all algorithms. As expected, the QDA and GNB methods significantly perform above the chance level. For each dataset excepted for dataset $D^{(3)}$, the hierarchy of prediction performances is preserved. From the highest to the worst performances we have: LDA $>$ LC $>$ NC $>$ GNB $>$ QDA. Again, for each dataset, the error bar are small ensuring that each performances is at least $2\sigma_\iota$ above chance. The structure of prediction is shown in Figure 3.5. For dataset $D^{(1)}$ (Figure 3.5(**a**)), the diagonal block structure is visible for each algorithm. Small bias is observed in GNB in which we observe residual columns (column 10 for example). The QDA method correctly predicts orientations but sometimes mixes up directions as indicate the blocks above and under the diagonal. Not surprisingly, the QDA performs best with a small number of PCA components because a small feature dimension increases the class covariance estimates. For dataset $D^{(3)}$ (Figure 3.5(**b**)), the diagonal

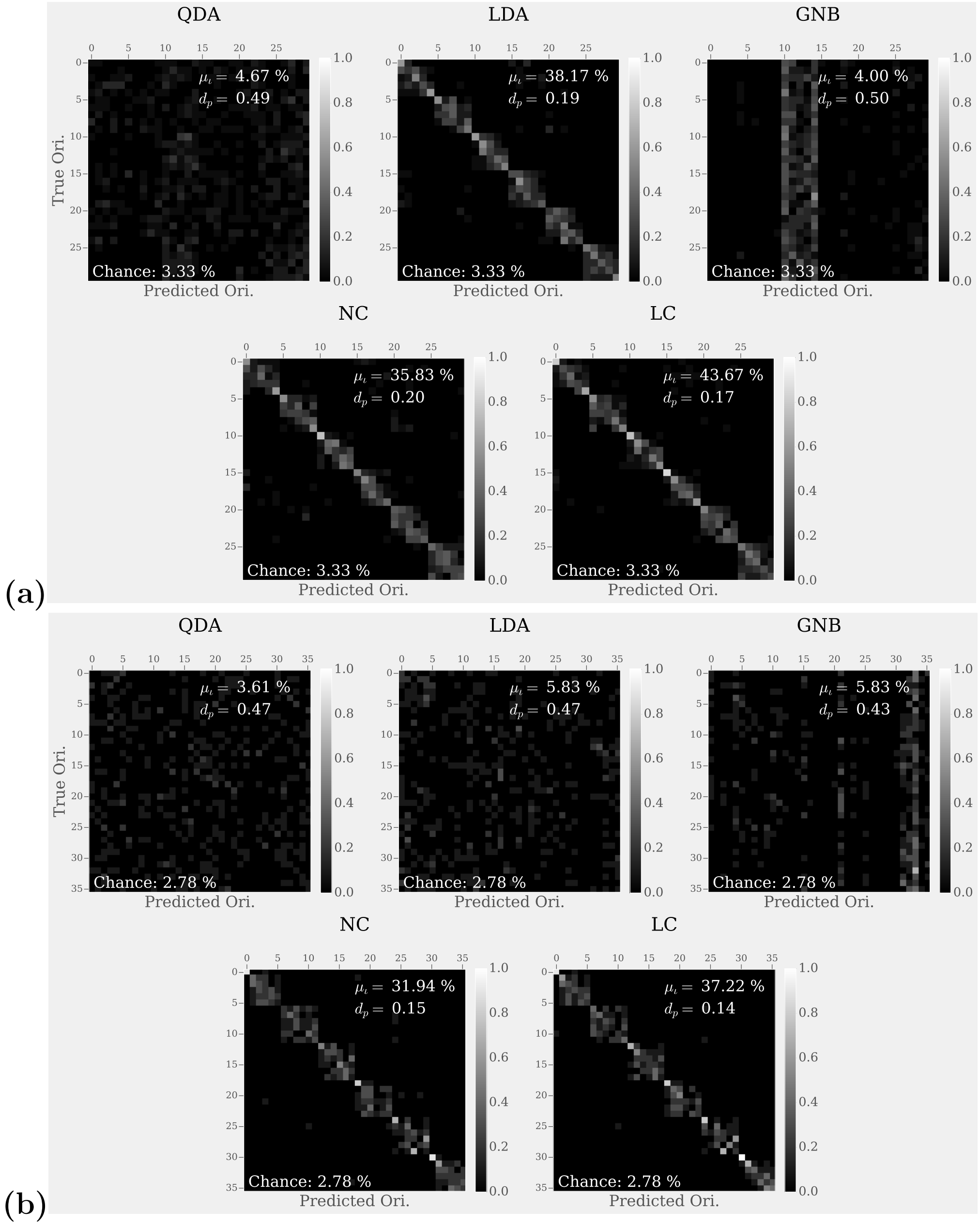


Figure 3.2: Confusion matrices Λ and distance d_p obtained for each algorithm for dataset $D^{(1)}$ **(a)** and $D^{(3)}$ **(b)**. QDA: Quadratic Discriminant Analysis. LDA: Linear Discriminant Analysis. GNB: Gaussian Naive Bayes. NC: Nearest Centroid. LC: Logistic Classification.

block structure is again visible for each algorithm. No bias is observed in GNB

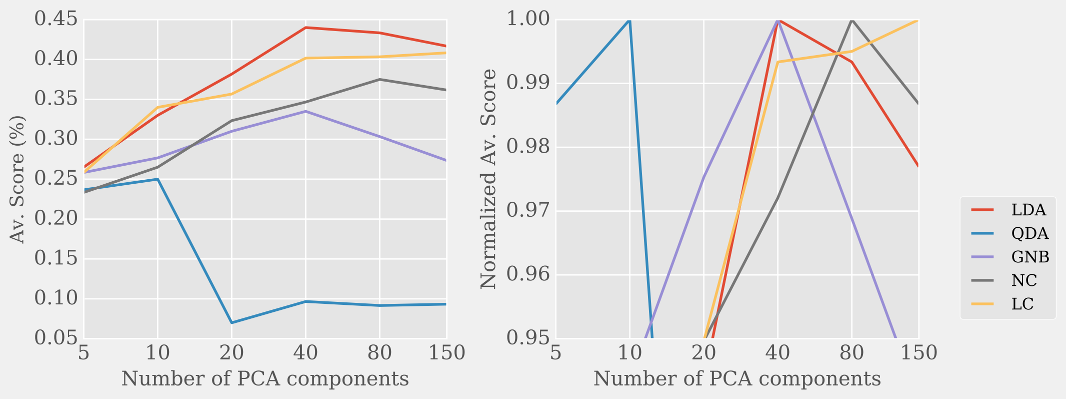


Figure 3.3: Dataset $D^{(1)}$. Left: the average score $\mu_\iota(n_{\text{pca}})$. Right: the normalized average score $\underline{\mu_\iota}(n_{\text{pca}})$.

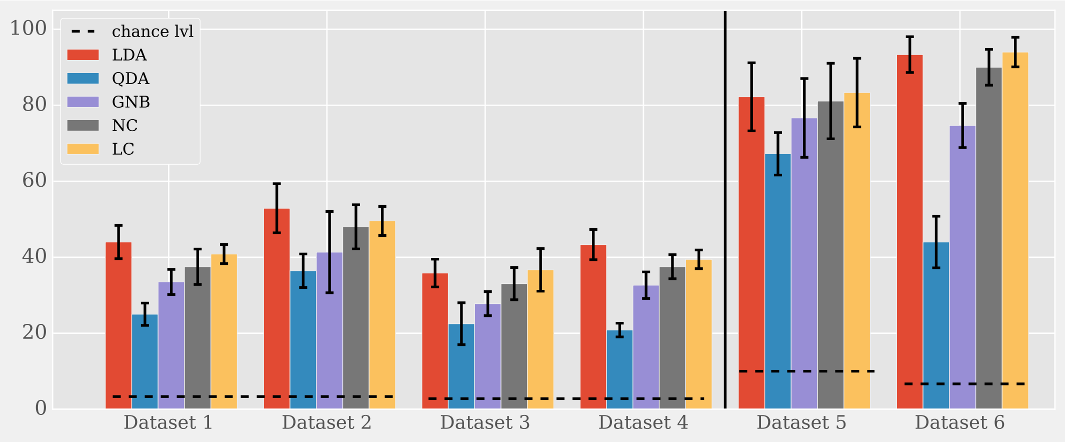


Figure 3.4: Classification performances of the different algorithms on the 6 datasets. QDA: Quadratic Discriminant Analysis. LDA: Linear Discriminant Analysis. GNB: Gaussian Naive Bayes. NC: Nearest Centroid. LC: Logistic Classification.

and the QDA method still mixes up directions as indicate the blocks above and under the diagonal. Again, the QDA performs best with a small number of PCA components. Finally, in both datasets the distances d_p are reduced showing a improvement of predictions.

Motion Clouds *vs* Natural Movies Remarkably, the prediction performances for the datasets involving natural movies stimulation ($D^{(5)}$, $D^{(6)}$) are higher than for other datasets involving motion clouds. Such a difference occurs for at least two reasons. First, the motion clouds are random stimuli.

In each repetition, although the stimuli have the same parameters they are generated with different seeds *ie* the values of pixels are different. Second, the tested motion clouds are sampled along lines or meshes in the space of parameters which make them being close to each other in quantifiable way (for example one can define the distance between two motion clouds as being the distance between their parameters). Considering natural movies is slightly different, there is no small dimension parametric model of natural movies and their diversity is high. Understanding the good prediction levels obtained on natural images would require to properly evaluate the differences between the natural movies in a way that reflects the processing of V1 neurons. As a perspective for future works, the importance of the stochastic stimulation (MC) in the prediction results could be evaluated by running a control experiment in which the motion clouds are generated with same seed.

Main Conclusions of Section 3

- The LC and NC methods perform best without dimension reduction.
- The LDA performs better than LC and NC after dimension reduction.
- The GNB and QDA perform badly with no dimension reduction showing the existence of significant spatial and temporal correlations.
- There are important differences in the prediction performances of motion clouds and natural movies.

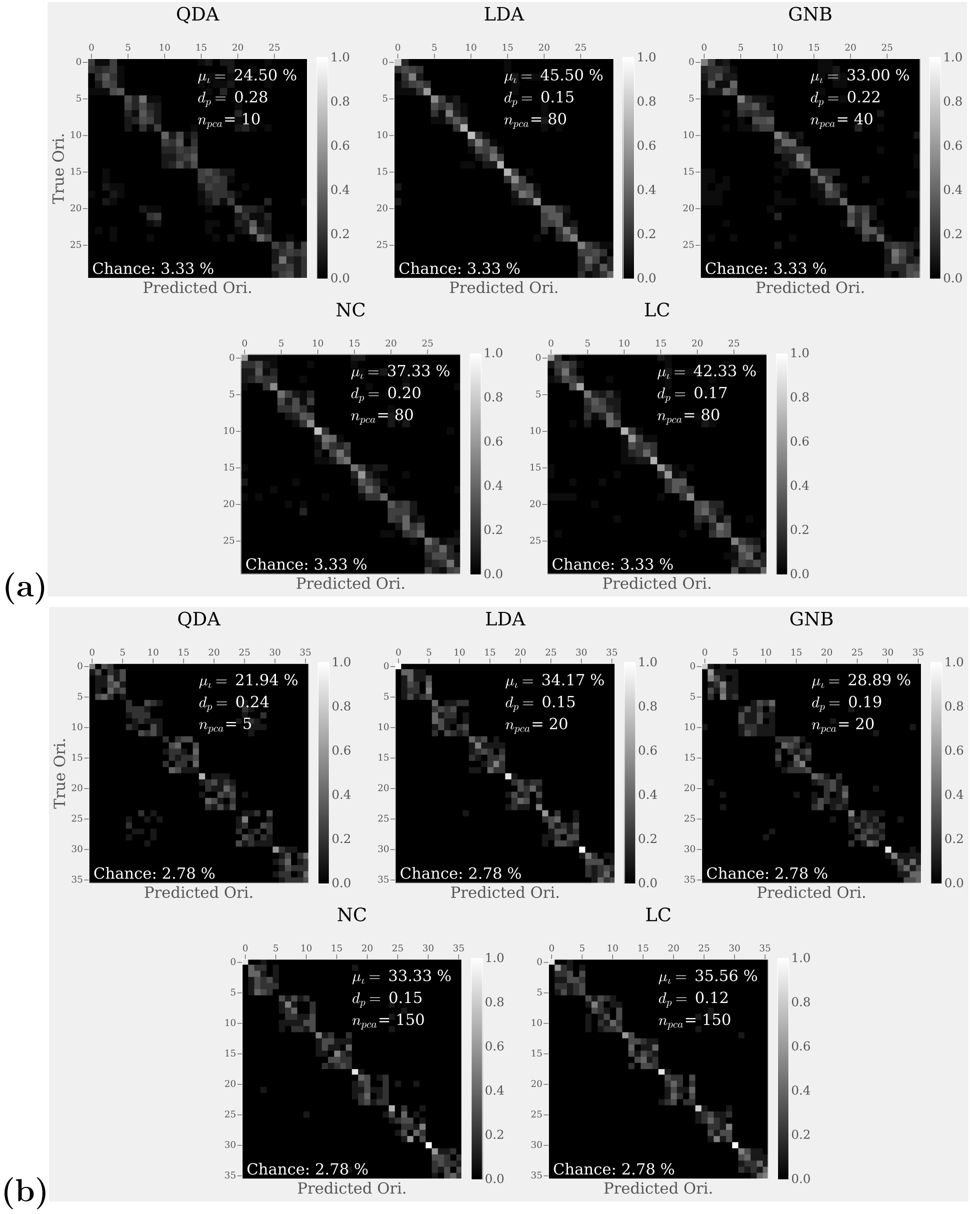


Figure 3.5: Confusion matrices Λ and distance d_p (defined in 9) obtained for each algorithm for dataset $D^{(1)}$ **(a)** and $D^{(3)}$ **(b)**. QDA: Quadratic Discriminant Analysis. LDA: Linear Discriminant Analysis. GNB: Gaussian Naive Bayes. NC: Nearest Centroid. LC: Logistic Classification.

4 Bandwidths Encoded in Neurons

The classification tasks studied in the previous section does not allow to clearly evaluate the prediction performances over the bandwidths. In order to circumvent this issue, we reduce the confusion matrices by summing the coefficients that correspond to labels associated with a common orientation or a common bandwidth. Thus, we can check if one of the two tested parameters is better predicted than the other. The most accurate way to perform this evaluation is to rerun the algorithm on datasets in which we only label one of two tested parameters. However, collapsing the confusion matrix is much simple and does not call into question the conclusions.

The reduced confusion matrices presented here are performed on SUA after dimension reduction *ie* we are considering the spiking activity of single neurons instead of the spiking activity recorded by electrodes. In each case we choose the algorithm that shows the best prediction performances.

4.1 Collapsing a Confusion Matrix

First, let us define precisely how we collapse a confusion matrix. Here, we assume that $\mathcal{Y} = \mathcal{C}_0 \times \mathcal{C}_1$. For instance, in dataset $D^{(2)}$, we have $\mathcal{C}_0 = \{0, 60, 120, 180, 240, 300\}$ and $\mathcal{C}_1 = \{0.79, 1.12, 2.74, 3.87, 5.48\}$, see the definiton of protocols in Section 2.3. The collapsed version of the average confusion matrix Λ is defined as it follows.

Definition 14 (Collapsed Confusion Matrix)**.** *Let* $\mathcal{Y} = \mathcal{C}_0 \times \mathcal{C}_1$ *and* $\Lambda = (\lambda_{y,y'})_{(y,y')\in\mathcal{Y}^2}$ *an average confusion matrix. For* $i \in \mathbb{Z}/2\mathbb{Z}$*, the collapsed confusion matrix over* $\mathcal{C}_i$ *is* $\Lambda^{(i)} = (\lambda^{(i)}_{c,c'})_{(c,c')\in\mathcal{C}_i^2}$ *where*

$$\forall (c, c') \in \mathcal{C}_i^2, \quad \lambda^{(i)}_{c,c'} = \frac{1}{|\mathcal{C}_{i+1}|^2} \sum_{(u,v)\in\mathcal{C}_{i+1}^2} \lambda_{(c,u),(c',v)}.$$

In the following sections, we collapse the confusions matrices on $\mathcal{C}_0$ (tested orientations) and $\mathcal{C}_1$ (tested bandwidths) for datasets $D^{(2)}$ and $D^{(3)}$.

4.2 Orientation Bandwidths

The confusion matrix shown in Figure 4.1 reflects the good prediction performances obtained with SUA (see a definition of SUA in Section 1.2). Despite the fact that we sometimes observe a decrease in the performances compared to MUA, this indicates that the SUA contains approximately as much information as MUA. The observed differences can be of different origins. The

MUA can quantitatively dominate the SUA or the spike sorting step can be hard to perform. The reduced confusion matrices in the center ($\Lambda^{(0)}$) and on the right ($\Lambda^{(1)}$) demonstrate that orientations are much better predicted than orientation bandwidths. One interesting aspect is the two-blocks structure of the reduced confusion matrix corresponding to bandwidths (right) obtained on dataset $D^{(2)}$. It reflects the gap presents in the bandwidth parameter ($\sigma_\Theta = 0.79, 1.12 \mid 2.74, 3.87$ and 5.48). Therefore, it is likely that the performances must increase by testing a coarser set of bandwidths.

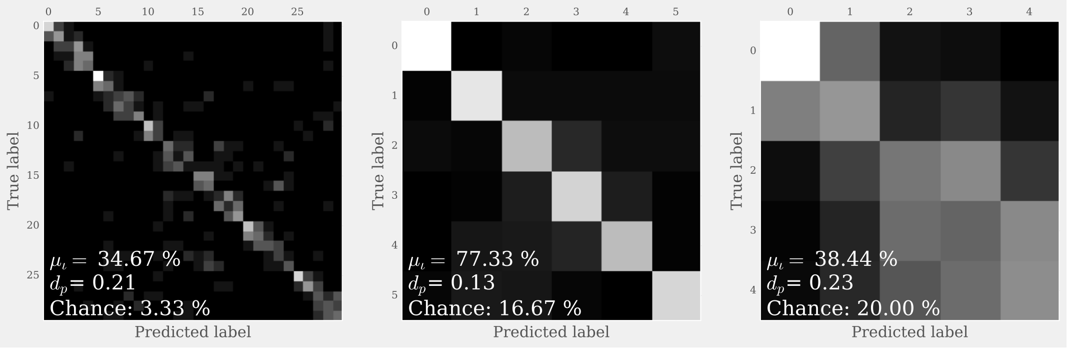


Figure 4.1: Dataset $D^{(2)}$. Confusion matrices Λ (left) and its reduction to orientation $\Lambda^{(0)}$ (center) and to orientation bandwidths $\Lambda^{(1)}$ (right).

4.3 Spatial Frequency Bandwidths

Again, the Figure 4.2 demonstrates the good prediction performances obtained with SUA. The reduced confusion matrices in the center ($\Lambda^{(0)}$) and on the right ($\Lambda^{(1)}$) demonstrate that orientation are much better predicted than spatial frequency bandwidths. Moreover, the reduced confusion matrix corresponding to spatial bandwidths has three-blocks structure that reflects the double gap in the tested parameters ($B_Z = 1.76 \mid 2.38, 2.55 \mid 2.77, 2.84$ and 2.9). Thus, a coarser set of spatial frequency bandwidths must provide better prediction performances.

Main Conclusions of Section 4

- Neurons of V1 are sensitive to both spatial frequency and orientation bandwidths (reinforcing the claims of Goris *et al* [71]).
- These preliminary experiments must be refined by wisely choosing the bandwidths to be tested.

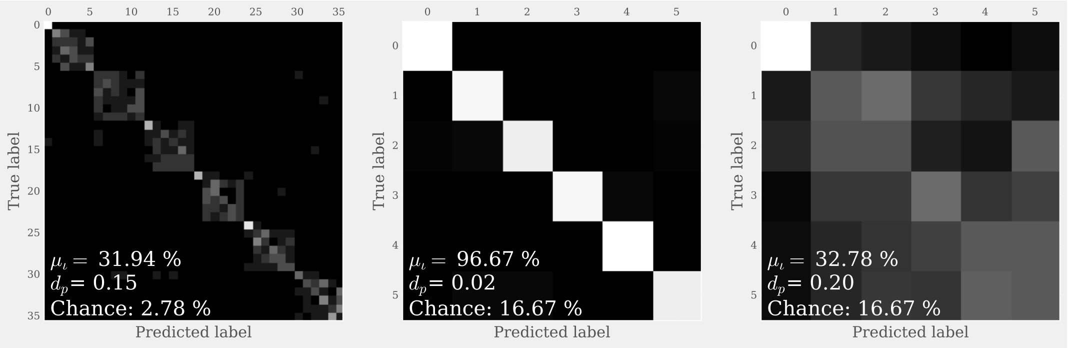


Figure 4.2: Dataset $D^{(3)}$. Confusion matrices Λ (left) and its reduction to orientation $\Lambda^{(0)}$ (center) and to spatial frequency bandwidths $\Lambda^{(1)}$ (right).

5 The Temporal Dynamic of Predictions

The previous Section establishes that the recorded signals contain stimulus-related information. Now, we are going to inspect where and when this information is located. By “where” we mean in which neurons the information is located.

5.1 Temporal Localization Method

In order to evaluate the prediction performances of individual neurons we simply set to zeros the values of other neurons and then we make the predictions. The temporal analysis is more involved and we restrict the analysis to the use of LC.

Sliding Window First, we use a Gaussian sliding window

$$\forall t' \in \mathcal{T}, \quad h_t(t') = \exp\left(-\frac{\|t'-t\|^2}{2\sigma_h^2}\right)$$

where σ_h is the window size. The window “localizes” the influence of the weight vector around each time t. We use these windowed weight vectors to make predictions. The windowing is performed with null conditions at the border of the weight vectors *i.e.* $\forall t' \in \mathcal{T}, g_t(t') = 0$ if and only if $t' - t \notin \mathcal{T}$. The probabilities of logistic classification defined in Equation (4.1) is therefore modified in order to obtained the following localized predictor centered at pixel $t \in \mathcal{T}$,

$$\mathbb{P}_{Y|X,\theta,t}(y|x) = \frac{e^{\langle x,\, h_t\omega_y\rangle}}{\sum_{y'\in\mathcal{Y}} e^{\langle x,\, h_t\omega_{y'}\rangle}}.$$

This probability is then plugged in Equation (2.3) to make the prediction. Then, we compute an average score $\mu_{\iota,t}$ for each time sample t. This first approach allows to localize information in time.

Growing Window Second, we use a window that linearly grows from $t = t_0$ to $t = T'$. In this setting, the probabilities of logistic classification become

$$\mathbb{P}_{Y|X,\theta,<t}(y|x) = \frac{e^{\langle x,\, \mathbf{1}_t \omega_y \rangle}}{\sum_{y' \in \mathcal{Y}} e^{\langle x,\, \mathbf{1}_t \omega_{y'} \rangle}}.$$

where $\forall t' \leqslant t, \mathbf{1}_t(t') = 1$ and $\forall t' > t, \mathbf{1}_t(t') = 0$. Then, it is possible to compute an average score $\mu_{\iota,<t}$ for each time window $[0, t]$. This second approach allows to evaluate how the information is building up along time.

5.2 Results on Motion Clouds

The results of the local analysis are shown in Figure 5.1 with $\sigma_h = 10$, $t_0 = 0$ ms and $T' = 1500$ ms. We do not observe any particular differences between datasets $D^{(2)}$ and $D^{(3)}$. First, on the left, we remark that despite the small value of σ_h, the prediction performances of the population often reach more than 10% whereas they stay below 7% for single neurons. This indicates that population coding is efficient at a small time scale. Another interesting point is the stationary behavior of predictions during stimulation that must be related to the stationary stimulus we are using. On the right, we observe that prediction performances increase faster and higher in the population than in any single neurons. Moreover, there are important differences between neurons: in some of them prediction performances are growing up whereas in some others they stay close to the chance level.

5.3 Results on Natural Movies

We perform the same analysis on datasets $D^{(5)}$ and $D^{(6)}$ with $\sigma_h = 10$, $t_0 = 500$ ms and $T' = 1500$ ms, see Figure 5.2. First, on the left, we observe in Figures **(a)** and **(b)** that the population provides better prediction performances locally in time. Population coding is efficient at a small time scale. The dynamic of prediction is not stationary as for the natural stimulation (even if the frames are randomly permuted **(b)**). Some time local events in the movie can sometime provide more information. We do not draw any conclusion about the differences in the stimulation (randomized frames *vs* standard movies) which requires more experiments. Second, on the right, we

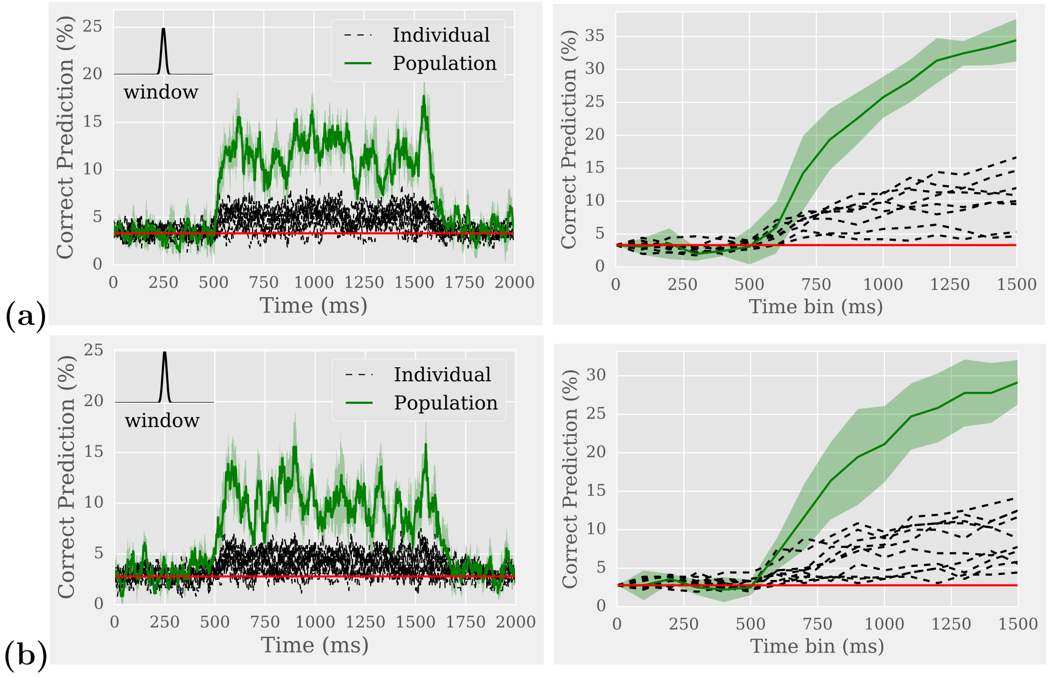


Figure 5.1: Left: the scores $\mu_{\iota,t}$ for all $t \in \mathcal{T}$. Right: the scores $\mu_{\iota,<t}$ for all $t \in \{0, \dots, t_{\text{off}}\Delta_t\}$. The red line represents the chance level. **(a)** Dataset $D^{(2)}$. **(b)** Dataset $D^{(3)}$.

observe that, in both cases, one neuron encodes as much information as the entire population which reveals that the population signal is redundant. This result must be related to the concepts of sparsity and redundancy reduction, however we do not expand on this topic as it requires further analysis. We refer to the following papers for more information on these topics [139, 201, 78, 10, 11]. Finally, we observe, in Figure **(a)** right, that the population prediction performances do not increase linearly after the stimulus onset. The increase is slow until 1000 ms and then it is faster. The fast increase occurs while the prediction performances obtained with one single neuron is increasing. This effect could be explained by the similarity of the ten movies during their first 500 ms at their center. However, this is unlikely after looking briefly at the movies. This effect requires a precise analysis of the presented movies to be understood.

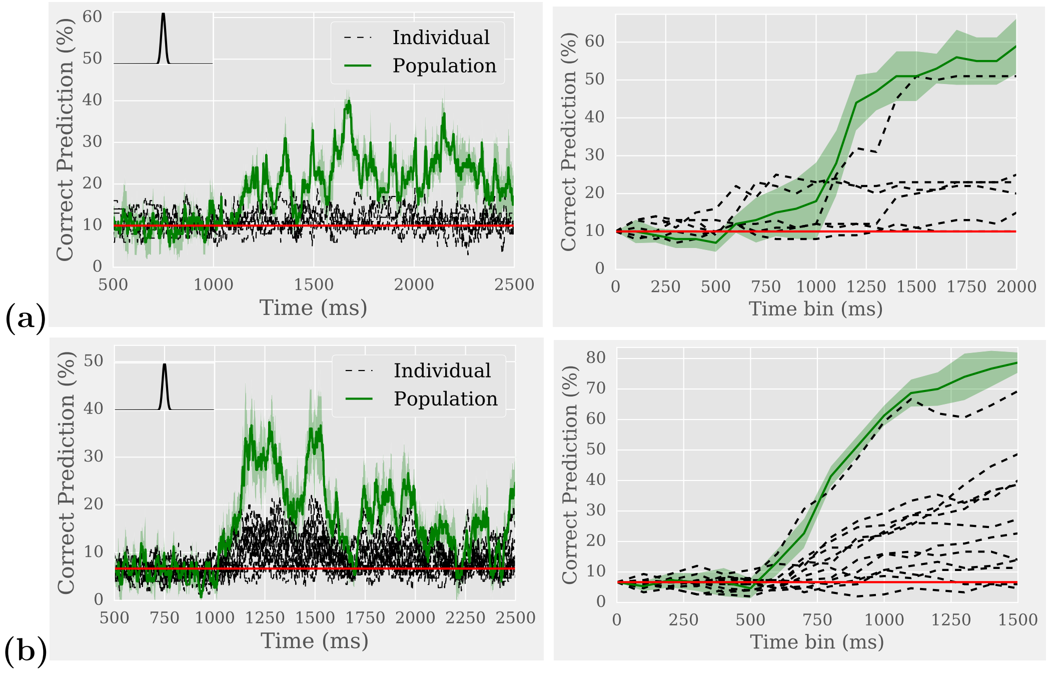


Figure 5.2: Left: the scores $\mu_{\iota,t}$ for all $t \in \mathcal{T}$. Right: the scores $\mu_{\iota,<t}$ for all $t \in \{0, \dots, t_{\text{off}}\Delta_t\}$. The red line represents the chance level. **(a)** Dataset $D^{(5)}$ **(b)** Dataset $D^{(6)}$.

Main Conclusions of Section 5

- For MC stimulations, prediction performances are stationary over time and population predicts faster and better than any single neurons.
- For natural movie stimulations, prediction performances are not stationary and one single neuron shows prediction performances similar to the entire population.

6 Comparison with a Simple V1 model

Although the previous analysis are informative, they do not identify which cortical mechanisms are involved. To this purpose, it is important to build generative models of data and to conduct similar analysis. We refer to Section 1.4.1 for a discussion about such forward models. In the following, we use a simple Linear/Non-linear Poisson model [172] (LNP) to simulate a small population of independent neurons. The simple assumption made on the population allow to partly reproduce the result of our previous analysis.

6.1 Linear/Non-linear Model

The LNP model [172] is the most simple spiking neuron model. It is able to reproduce the behavior of simple cells of V1 with oriented receptive fields of different scales. We assimilate a neuron i to its receptive field f_{θ_i,D_i,a_i} which, following [99], we model using a Gabor oriented filter

$$\forall x \in \Omega_N, \quad f_{g_i,\theta_i,D_i,a_i}(x) = g_i \exp\left(-\frac{1}{2}(R_{\theta_i}x)^T D_i (R_{\theta_i}x)\right) \cos\left(2\pi a_i \langle R_{\theta_i}x,\, e\rangle\right)$$

where $e = (1,0)$ and R_θ is the rotation of angle θ. The angle θ_i represents the orientation of the receptive field, the matrix D_i controls its spatial aspect, the real number a_i is the spatial frequency and g_i is a gain. The set $\Omega_N = \{-\frac{N}{2}, \dots, \frac{N}{2} - 1\}^2$ corresponds to the pixel positions. When stimulated with an image f_t at time $t \in \mathcal{T}$ the firing activity R_i of neuron i follows a Poisson law of parameter $\lambda_{i,t} = r + \max\left(0, \langle f_{g_i,\theta_i,D_i,a_i}(\Omega_N),\, f_t\rangle\right)$ where r is the residual spiking activity. Therefore the probability that r_i spikes occur between t and $t + \Delta t$ is

$$\forall r_i \in \mathbb{N}, \quad \mathbb{P}_{R_i|F_t}(r_i|f_t) = \lambda_{i,t}^{r_i} \Delta t \frac{\exp\left(-\lambda_{i,t}\Delta t\right)}{r_i!}. \tag{6.1}$$

In order to simulate data, we choose to reproduce the protocols presented in Section 2.3 that involve motion clouds.

Stimulation 1 First, we generate motion clouds with parameters $z_0 = 0.6$ c/°$, B_Z = 1.0, \sigma_V = \frac{1}{t^\star z_0}$ with $t^\star = 0.3$ ms and drifting in a single direction for 1000 ms at speed $v_0 = 2.5$ c/s. Six orientations were used ($\theta_0 =$ 0 °, 60 °, 120 °, 180 °, 240 ° and 300 °) with five orientation bandwidths ($\sigma_\Theta =$ 0.35, 0.79, 1.12, 1.58 and 2.24), totaling $C = 30$ stimulation conditions.

Stimulation 2 Second, we therefore generate motion clouds with parameters $z_0 = 0.6$ c/°$, \sigma_\Theta = 1.0, \sigma_V = \frac{1}{t^\star z_0}$ with $t^\star = 0.3$ ms and drifting in a single direction for 1000 ms at speed $v_0 = 2.5$ c/s. Six orientations were used ($\theta_0 =$ 0 °, 60 °, 120 °, 180 °, 240 ° and 300 °) with six spatial frequency bandwidths ($B_Z =$ 1.2, 1.84, 2.11, 2.31, 2.46 and 2.61), totaling $C = 30$ stimulation conditions.

In both cases, we use a set of 8 neurons with receptive field f_{g_i,θ_i,D_i,a_i} where for all $i \in \{1, \dots, 8\}$,

$$g_i = 0.005, \quad a_i = z_0, \quad \theta_i \in \left\{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\right\}$$

$$\text{and} \quad D_i \in \left\{ \begin{pmatrix} 0.1N & 0 \\ 0 & 0.1N \end{pmatrix}, \begin{pmatrix} 0.2N & 0 \\ 0 & 0.2N \end{pmatrix} \right\} \quad \text{with} \quad N = 128.$$

The receptive fields of the 8 neurons are shown in Figure 6.1. We generate spikes according to Equation (6.1). For $t \in \{0, \delta_t, \ldots, T'\delta_t\}$, we consider that at each time $t' \in \{t, t+\Delta_t, \ldots, t+\delta_t\}$ spikes occur with probability $\mathbb{P}_{R_i|F_t}(r_i|f_t)$. We use $T' = 200$ *ie* a frame refresh rate of 100 Hz and $\delta_t = 10\Delta_t = 10$ ms *ie* a recording at 1000 Hz. We set the residual spiking activity $r = 0.015$. An example of three simulated neurons is shown in Figure 6.2.

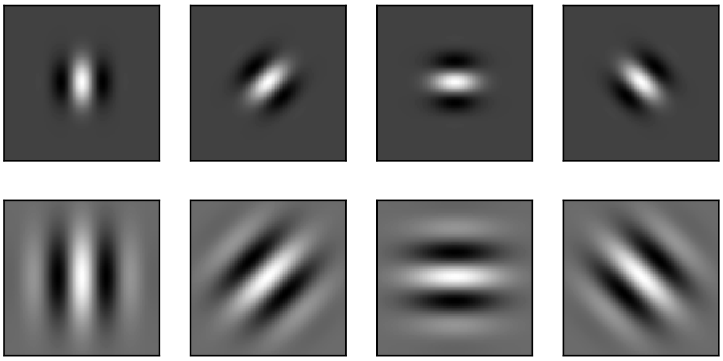


Figure 6.1: The receptive fields of the 8 neurons used for data simulation.

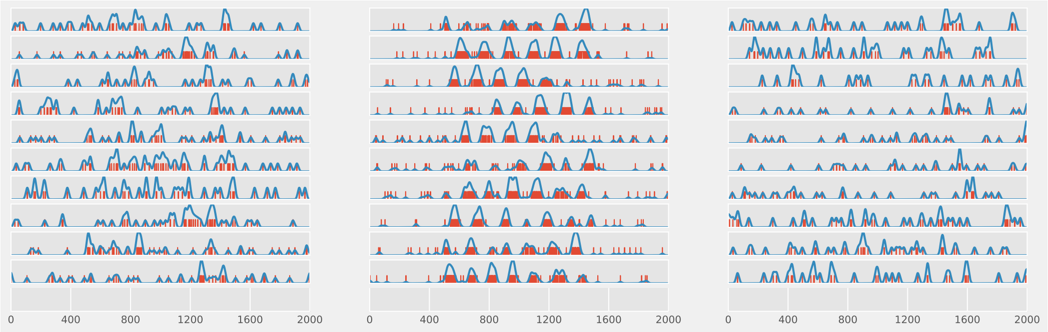


Figure 6.2: Example of 3 simulated neurons. Each column corresponds to a different neuron and each line corresponds the 10 different repetitions of a single condition. The red bars are the individual spikes. The blue line represents the spike density computation defined in Section 2.4.

6.2 Results of Supervised Classification

We conduct the same analysis as in Section 4 and 5.

6.2.1 Orientation Bandwidths Manipulation

The first simulated dataset reproduces the protocols used for datasets $D^{(1)}$ and $D^{(2)}$. The classification performances obtained are similar to the ones obtained on the real datasets. In Figure 6.3, the confusion matrix has the diagonal block structure. Moreover, the prediction performances are higher on the orientations than on the orientation bandwidths. The main difference comes from the miss prediction of directions. However, this behavior is expected because our model only considers spatial receptive fields and discards its temporal component which makes possible the discrimination of directions. Concerning the temporal aspects of prediction performances (see Figure 6.4), the results are again very similar to the ones obtained on real data. We observe that population shows better prediction performances than any single neuron. The prediction performances are also stable over time. We conclude that such a simple V1 model is able to explain the results we observe in the recorded neurons.

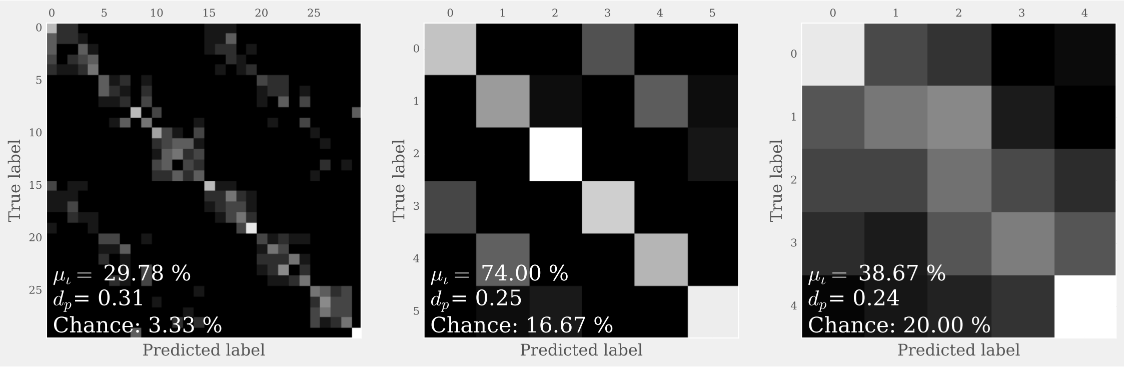


Figure 6.3: Simulated dataset 1. Confusion matrices Λ (left) and its reduction to orientation (center) and to orientation bandwidths (right).

6.2.2 Spatial Frequency Bandwidths Manipulation

The second simulated dataset reproduces the protocoles used for datasets $D^{(3)}$ and $D^{(4)}$. Again, we observe similar performances as obtained on the real

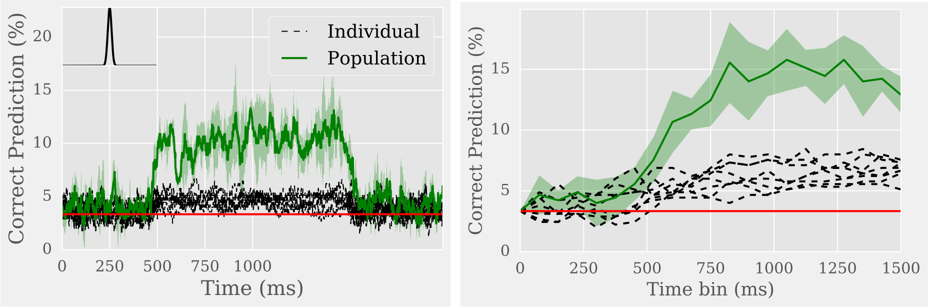


Figure 6.4: Simulated dataset 1. Left: the scores $\mu_{\iota,t}$ for all $t \in \mathcal{T}$. Right: the scores $\mu_{\iota,<t}$ for all $t \in \{0,\ldots,t_{\mathrm{off}}\Delta_t\}$. The red line represents the chance level.

datasets. In Figure 6.3, the confusion matrix has the diagonal block structure and the prediction performances are higher on the orientations than on the spatial frequency bandwidths. Obviously, we also observe the miss prediction of direction. The reduced confusion matrix to spatial frequency bandwidths has a two blocks structure: the first bandwidth is very well discriminated from the others. This is probably due to the choice of the neuron spatial aspect parameters D_i. Concerning the temporal aspects of prediction performances (see Figure 6.6), the results are again very similar. The population shows better prediction performances than any single neuron. Finally, we also conclude that the simple model we have build is able to explain the observed results on these protocols.

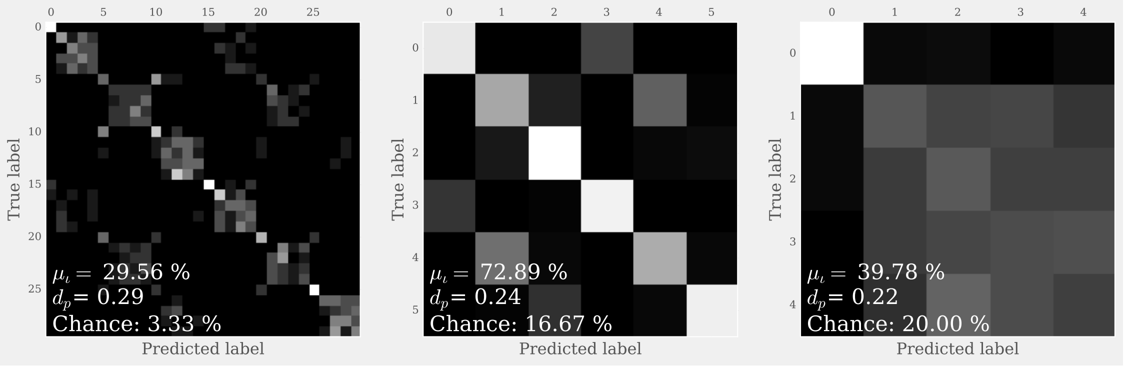


Figure 6.5: Simulated dataset 2. Confusion matrices Λ (left) and its reduction to orientation (center) and to spatial frequency bandwidths (right).

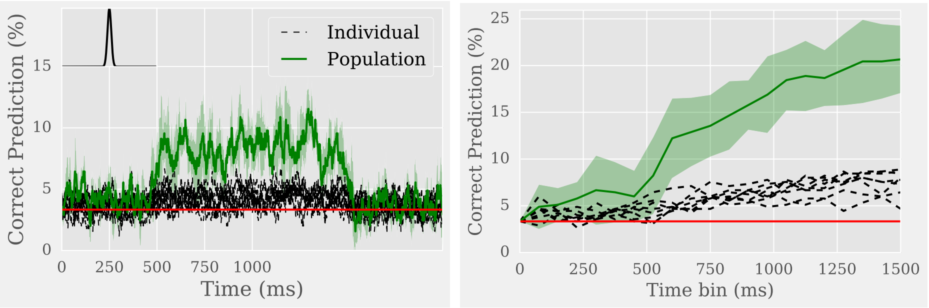


Figure 6.6: Simulated dataset 2. Left: the scores $\mu_{\iota,t}$ for all $t \in \mathcal{T}$. Right: the scores $\mu_{\iota,<t}$ for all $t \in \{0, \ldots, t_{\text{off}}\Delta_t\}$. The red line represents the chance level.

Conclusion and Perspectives

1 Neurosciences: Mathematics and Experiments

After my mathematical curriculum, starting a PhD on such a multidisciplinary topic was a real challenge, both scientifically and personally. Along this manuscript, we propose several contributions to texture models, neurosciences and psychophysics that we will not recap here. Instead, we prefer to insist on three main aspects:

- dynamic, stochastic, yet parametric, natural movie models;
- the “Bayesian brain” hypothesis;
- the relevancy of supervised learning for physiological data analysis.

Developing natural movie models for experimental neurosciences is fundamental, as they must be realistic, stochastic and well controlled. However, we are far from having a good and efficient model of natural movies that verifies these three assumptions (see Section 2 of the introduction). At the moment, we are able to capture some relevant properties of natural movies like texture, geometry and motion. We have focused here our attention on a dynamic texture model by taking into account only a small number of parameters. Indeed, parameters need to be biologically relevant to our current understanding of brain functions and in small number to be easily manipulable. We subsequently use these stimuli both in psychophysics and electrophysiology where we get interesting results. While the obtained results are encouraging, it is thus clearly important to use more involved dynamic texture models. In particular, non-Gaussian models should allow to design generative models of geometry and of complex motion.

Our approach to psychophysics is purely Bayesian. This framework offers a nice way to interpret psychophysical bias in a mathematically and statistically sound way. Our contribution in this direction is mostly methodological, and we believe that the notion of ”inverse Bayesian inference” offers a nice framework for further mathematical and statistical exploitation. It is however clear that the Bayesian brain hypothesis is questionable, and most importantly, it does not shed light on the actual neural implementation of the observed phenomena. In particular, it requires further physiological investigations.

We choose to use supervised learning to analyze our different physiological datasets. The first reason is that it fits the collected data labeled by their stimulation conditions perfectly. The second reason is that it is able to answer the standard question that goes with this kind of protocols: does the recorded signal contains stimulus related information ? Finally, supervised learning approaches usually perform better at detecting relevant features in noisy data than standard averaging methods. Therefore, we strongly encourage neuroscientists to use these machine learning based algorithm to analyze data by each time formalizing a proper classification task associated to an experimental acquisition campaign.

2 Perspectives: Toward a Unified Bayesian Model of Vision

Bayesian probability is a powerful tool to model information processing (Chapters II and III). In perceptual neuroscience, the brain is often viewed as an information processing machine [38]. In this context, the "Bayesian brain" hypothesis (Chapter II and reviews [103, 60]) emerges naturally: sensory information builds up in the brain as a likelihood that is combined to an internal prior in order to detect, discriminate, take decisions, *etc.* Confronting such an hypothesis to experimental data is still a challenge, especially in neurophysiology. The Bayesian inference theory is rich and applies in many fields that involve data analysis. However, when applied to perceptual neuroscience, the approaches are disparate because they highly depend on the experimental techniques used (electrophysiology, psychophysics, imaging, Chapters III, V and VI). To extend the methology developed in my dissertation, we must tackle the question of unifying the Bayesian brain models of vision to address both electrophysiological and psychophysical data. Such a project will help us gain a better understanding of what neural computations are. In addition, it will provide a complete Bayesian model that implies systematic characterization of neural responses through generative models of textures and images.

2.1 Related works

Ideal Bayesian Observer Model Recent advances in the field of Bayesian modeling ([194, 184, 183, 96] and Chapter III) have justified the outcome and bias observed in psychophysical studies of vision. In order to explain how the brain processes sensory information, the Bayesian model assumes that the brain performs some abstract measurements interpreted as a likelihood. These measurements are subsequently biased by an internal prior that is able to explain observed sensory bias, see Figure 2.1. However, these psychophysics experiments are rarely combined to their neurophysiological counterparts, as explained in [103]. In particular, our experimental works in psychophysics

(Chapter III) and electrophysiology (Chapter V and VI) have been conducted separately. In this context, such a model can be perceived by experimentalists as a mathematical and abstract black box.

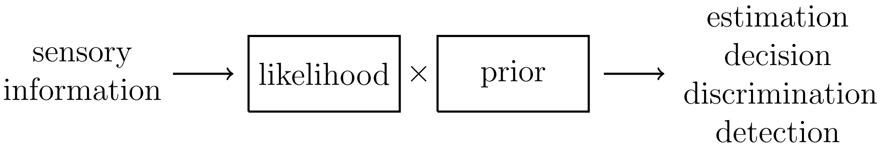


Figure 2.1: The basic principle of a Bayesian observer.

Likelihood Functions Implemented by Neural Populations In electrophysiology, work from the Movshon laboratory at NYU has focused on the computation made by neural populations (see Section III.3.1) the main claim is that neurons implement likelihood functions through the combinations of their tuning curve and stimulus responses [94, 72]. The work of Ma *et al.* [111] has the same approach. In short, a stimulus s elicits a number of spikes m_i in neuron i which has a characteristic response function f_i called a tuning curve. In the case of Poisson spiking neurons, the number of spikes m_i is generated from the Poisson law $P_{M_i|S}$ of parameter $f_i(s)$. The combination of the different neurons results in the computation of a likelihood function, see Figure 2.2. Future works must focus on two essential components: the question of the prior and that of the stimulus complexity (high dimensional movie stimulations) in order to improve the limited connections between Chapter II and Chapters III/VI.

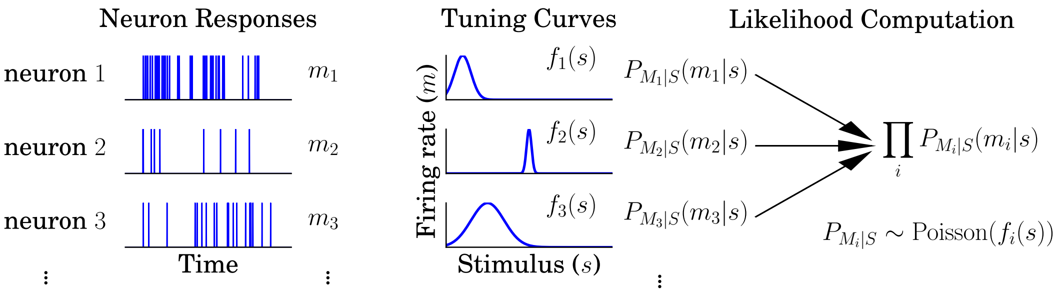


Figure 2.2: Neural implementation of likelihood computations.

Priors Encoded in the Heterogeneity of Neural Populations The recent work of Ganguli and Simoncelli [66] provides a framework that tackles in a mostly theoretical way the problem of how a prior is encoded in neural populations. In this work, the neurons implement a likelihood function. In addition, they assume that the tuning curves of neurons are distributed according to a certain law of density d, see Figure 2.3. By maximizing a lower bound of the mutual information between stimulus s and the measurements of neurons $\mathbf{m} = (m_1, \dots, m_n)$, they show that the prior (*ie* the density of s) is equal to the "tuning curve density" d (see [66] for details). Yet, this result does not take into account the complexity of the stimulus and reduces neurons to their tuning curves. However, it offers a way to tackle the physiological relevance of the bi-variate prior used in Chapter III.

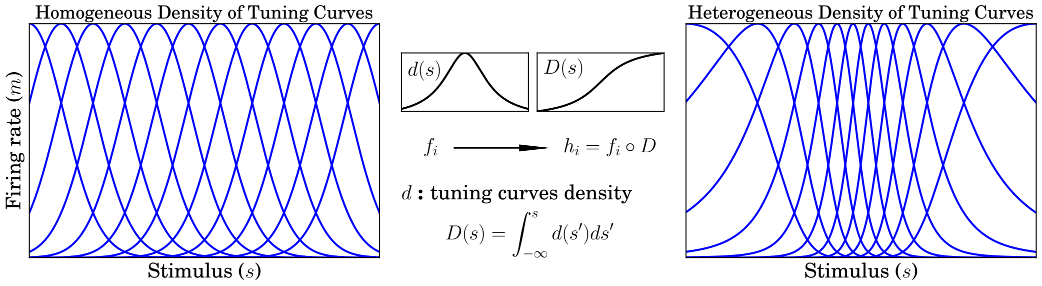

Homogeneous Density of Tuning Curves
Firing rate (m)
Stimulus (s)
$d(s)$
$D(s)$
f_i
$h_i = f_i \circ D$
d : tuning curves density
$D(s) = \int_{-\infty}^{s} d(s')ds'$
Heterogeneous Density of Tuning Curves
Stimulus (s)

Figure 2.3: Density of tuning curves.

2.2 Future Works

In my thesis contributions and in this related work, I identify three clear avenues for improvements. First, the lack of neurophysiological experiments that test the "Bayesian brain" hypothesis. Second, the simplifying assumption on the stimulus that discards its complexity *ie* the lack of connection between the Bayesian model (Chapter II) and the Motion Cloud model (Chapter I).. Third, following the simple stimulus, neurons are reduced to uni-variate tuning curves. Extend the neuron representation should allow to question the physiological relevance of our bi-variate prior III. Let us start by a natural improvement of the MC model that we mention in Section I.5.

Real-Time Stimulation The sPDE formulation of MC allows for real-time stimulation (Chapter I.3). However the spatial stationary covariance σ_W does not depend on time. Such an extension appears naturally and one can imagine designing stimuli parametrized by trajectories instead of constant values. Yet, this extension has some drawbacks. In particular the parameters of the solution are not equal to the input parameters but follow them with some delay. Some theoretical work must conducted to properly handle this delay. Then, it will possible to control the parameters with respect to neural responses using Bayesian prediction models (for instance, by seeking to maximize the responses).

Take Back the Complexity of Visual Stimuli The class of models to be developed must assume an underlying generative model of images, see Figure 2.4. An image i is generated with probability distribution $P_{I|S}$ parametrized by s (for instance speed, spatial frequency, orientation). When presented to an ideal Bayesian observer, it elicits measurements of neurons $\mathbf{m} = (m_1, \dots, m_n)$. Typically m_k is the spike counts of neuron k. Finally, the estimation $\hat{s}$ is computed from the combination of the measurements' distribution $P_{M_k|S}$ and of an internal prior $P_{\hat{S}}$ (see Section II.2). By requiring a generative model of measurements knowing the stimulus, such models are able to take back into account part of its complexity. One consequence is that a Poisson spiking neuron is modeled by the probability distribution $P_{M_k|I}$ (instead of $P_{M_k|S}$, see Figure 2.2) which depends on the image i. Instead of being parametrized by the neuron's tuning curve, the Poisson law is parametrized by the neuron's receptive field[1]. The goal is to refine the results of Ganguli [66] in a more general context where the concept of "tuning curve density" is replaced by an equivalent concept of "receptive field density". For instance, the Bayesian estimation perfomred by Park and Pillow [145] can be adopted: they use densities for the parameters of receptive fields. In parallel, such models allow to run simulations that are essential to evaluate their strengths and weaknesses.

$$\begin{tikzcd}
s \sim P_S \arrow[r] & i \sim P_{I|S} \arrow[r] & \overset{\text{likelihood}}{m_k \sim P_{M_k|I}} \arrow[rr] & & \hat{s} \sim P_{\hat{S}|M_k} \\
& P_{M_k|S} = \int_{\mathcal{I}} P_{M_k|I}(m_k|i) P_{I|S}(i|s) \mathrm{d}i & & \underset{\text{prior}}{P_{\hat{S}} = d} \arrow[u] &
\end{tikzcd}$$

Figure 2.4: The ideal Bayesian observer model that takes into account the stimulus complexity.

[1] For a neuron: the region of the visual field in which a stimulus modifies its firing rate.

A Bridge Between Electrophysiology and Psychophysics Using this framework to explain effects observed in psychophysics (*eg* the effect of spatial frequency/contrast over speed perception (see Chapter III)) provides hypothesis about the "receptive fields density" of the neural population observable using electrophysiology. In the same way, electrophysiology enables a characterization of the neural population by its "receptive fields density". Therefore, it provides hypothesis about potential effects that could be observed in psychophysics. Finally, electrophysiology and psychophysics become unified in a common framework which is not done in this dissertation. The goal is to run experiments both in psychophysics and electrophysiology using similar stimuli generated from a common generative model. Such stimuli has shown useful both in electrophysiology and psychophysics [71, 194]. The collected data will enable the comparison between a theoretical "receptive fields density" that explains the psychophysical effect and the empirical "receptive field density" measured using electrophysiology. We can therefore answer the question of the relevance of our bi-variate prior (see Chapter III).

Can We Fool the Brain Through Adaptation Mechanisms ? In the model developed above and implicitely assumed along this manuscript (see Section II.3.1 and VI.6), neurons are rigid units that combine each other to compute a likelihood that is biased by a "receptive field density". However, neurons are known to adapt their receptive field to the statistics of sensory inputs [55]. In this context, the "receptive field density" is not fixed anymore and thus is the Bayesian prior. Therefore, by forcing short-term adaptation mechanisms one can modify the internal prior of an observer that can be measured subsequently using psychophysics or electrophysiology. Inspired by [104], the goal is to run experiments both in psychophysics and electrophysiology in unified protocols. The protocols must involve stimulation known to provoke visual illusions (*eg* motion after effect) followed by a classical stimulation. Such a protocol allows to study the effects of induced adaptation on the classical stimulus perception and whether or not it affects the observer prior.

References

[1] B. Abraham, O. I. Camps, and M. Sznaier. Dynamic texture with fourier descriptors. In *Proceedings of the 4th international workshop on texture analysis and synthesis*. Volume 1, 2005, pages 53–58.

[2] D. E. Acuna, M. Berniker, H. L. Fernandes, and K. P. Kording. Using psychophysics to ask if the brain samples or maximizes. *Journal of vision*, 15(3):7–7, 2015.

[3] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of Optical Society of America, A.*, 2(2):284–99, February 1985.

[4] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Josa a*, 2(2):284–299, 1985.

[5] E. D. Adrian. The basis of sensation. 1928.

[6] A. Afgoustidis. Orientation maps in v1 and non-euclidean geometry. *The journal of mathematical neuroscience (jmn)*, 5(1):1–45, 2015.

[7] I. Ayzenshtat, A. Gilad, G. Zurawel, and H. Slovin. Population response to natural images in the primary visual cortex encodes local stimulus attributes and perceptual processing. *The journal of neuroscience*, 32(40):13971–13986, 2012.

[8] I. Ayzenshtat, E. Meirovithz, H. Edelman, U. Werner-Reiss, E. Bienenstock, M. Abeles, and H. Slovin. Precise spatiotemporal patterns among visual cortical areas and their relation to visual stimulus processing. *The journal of neuroscience*, 30(33):11232–11245, 2010.

[9] D. Barbieri, G. Citti, G. Sanguinetti, and A. Sarti. An uncertainty principle underlying the functional architecture of v1. *Journal of physiology-paris*, 106(5):183–193, 2012.

[10] H. B. Barlow. Possible principles underlying the transformations of sensory messages, 1961.

[11] P. Baudot, M. Levy, O. Marre, C. Monier, M. Pananceau, and Y. Frégnac. Animation of natural scene by virtual eye-movements evokes high precision and low noise in v1 neurons. *Frontiers in neural circuits*, 7:206, 2013.

[12] A. Benucci, R. A. Frazor, and M. Carandini. Standing waves and traveling waves distinguish two circuits in visual cortex. *Neuron*, 55(1):103–117, 2007.

[13] R. Ben-Yishai, R. L. Bar-Or, and H. Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the national academy of sciences*, 92(9):3844–3848, 1995.

[14] R. Ben-Yishai, D. Hansel, and H. Sompolinsky. Traveling waves and the processing of weakly tuned inputs in a cortical network module. *Journal of computational neuroscience*, 4(1):57–77, 1997.

[15] G. G. Blasdel, G. Salama, et al. Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature*, 321(6070):579–585, 1986.

[16] T. Bonhoeffer and A. Grinvald. Optical imaging based on intrinsic signals: the methodology. *Brain mapping: the methods*:55–97, 1996.

[17] R. T. Born and D. C. Bradley. Structure and function of visual area MT. *Annual review of neuroscience*, 28(1):157–89, July 2005.

[18] W. H. Bosking, Y. Zhang, B. Schofield, and D. Fitzpatrick. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *The journal of neuroscience*, 17(6):2112–2127, 1997.

[19] C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering. *Computational statistics & data analysis*, 52(1):502–519, 2007.

[20] C. Bouveyron, S. Girard, and C. Schmid. High-dimensional discriminant analysis. *Communications in statistics-theory and methods*, 36(14):2607–2623, 2007.

[21] J. S. Bowers and C. J. Davis. Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, 138(3):389, 2012.

[22] G. E. Box and G. C. Tiao. *Bayesian inference in statistical analysis.* Volume 40. John Wiley & Sons, 2011.

[23] T. Briand, J. Vacher, B. Galerne, and J. Rabin. The heeger & bergen pyramid based texture synthesis algorithm. *Image processing on line*, 4:276–299, 2014.

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

[25] P. J. Brockwell and A. Lindner. Existence and uniqueness of stationary lévy-driven carma processes. *Stochastic processes and their applications*, 119(8):2660–2681, 2009.

[26] P. Brockwell, R. Davis, and Y. Yang. Continuous-time gaussian autoregression. *Statistica sinica*, 17(1):63, 2007.

[27] K. R. Brooks, T. Morris, and P. Thompson. Contrast and stimulus complexity moderate the relationship between spatial frequency and perceived speed: implications for mt models of speed perception. *Journal of vision*, 11(14):19–19, 2011.

[28] J. Bullier. Integrated model of visual processing. *Brain research reviews*, 36(2):96–107, 2001.

[29] R. Caruana, N. Karampatziakis, and A. Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on machine learning*. ACM, 2008, pages 96–103.

[30] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning*. ACM, 2006, pages 161–168.

[31] B. Cessac, B. Doyon, M. Quoy, and M. Samuelides. Mean-field equations, bifurcation map and route to chaos in discrete time neural networks. *Physica d: nonlinear phenomena*, 74(1-2):24–44, 1994.

[32] F. Chavane, D. Sharon, D. Jancke, O. Marre, Y. Frégnac, and A. Grinvald. Lateral spread of orientation selectivity in v1 is controlled by intracortical cooperativity. *Frontiers in systems neuroscience*, 5:4, 2011.

[33] S. Chemla and F. Chavane. Voltage-sensitive dye imaging: technique review and models. *Journal of physiology-paris*, 104(1):40–50, 2010.

[34] M. Colombo and P. Seriès. Bayes in the brain – on bayesian modelling in neuroscience. *The british journal for the philosophy of science*, 63(3):697–723, 2012.

[35] R. Costantini, L. Sbaiz, and S. Süsstrunk. Higher order svd analysis for dynamic texture synthesis. *Image processing, ieee transactions on*, 17(1):42–52, 2008.

[36] J. Csicsvari, D. A. Henze, B. Jamieson, K. D. Harris, A. Sirota, P. Barthó, K. D. Wise, and G. Buzsáki. Massively parallel recording of unit and local field potentials with silicon-based electrodes. *Journal of neurophysiology*, 90(2):1314–1323, 2003.

[37] H. V. Davila, B. M. Salzberg, L. Cohen, and A. Waggoner. A large change in axon fluorescence that provides a promising method for measuring membrane potential. *Nature*, 241(109):159–160, 1973.

[38] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

[39] R. L. De Valois, D. G. Albrecht, and L. G. Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vision research*, 22(5):545–559, 1982.

[40] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. i. general characteristics and postnatal development. *Journal of neurophysiology*, 69(4):1091–1117, 1993.

[41] A. Desolneux, L. Moisan, and J.-M. Morel. *From gestalt theory to image analysis: a probabilistic approach.* Volume 34. Springer Science & Business Media, 2007.

[42] J. J. J. DiCarlo, D. Zoccolan, and N. C. C. Rust. How Does the Brain Solve Visual Object Recognition? *Neuron*, 73(3):415–434, 2012.

[43] D. Dong. Maximizing causal information of natural scenes in motion. In U. J. Ilg and G. S. Masson, editors, *Dynamics of visual motion processing*, pages 261–282. Springer US, 2010.

[44] D. W. Dong and J. J. Atick. Statistics of natural time-varying images. *Network: computation in neural systems*, 6(3):345–358, 1995.

[45] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *International journal of computer vision*, 51(2):91–109, February 2003.

[46] K. Doya. *Bayesian brain: probabilistic approaches to neural coding.* MIT press, 2007.

[47] J. Du, I. H. Riedel-Kruse, J. C. Nawroth, M. L. Roukes, G. Laurent, and S. C. Masmanidis. High-resolution three-dimensional extracellular recording of neuronal activity with microfabricated electrode arrays. *Journal of neurophysiology*, 101(3):1671–1678, 2009.

[48] C. Ekanadham, D. Tranchina, and E. P. Simoncelli. A unified framework and method for automatic neural spike identification. *Journal of neuroscience methods*, 222:47–55, 2014.

[49] N. El Karoui, S. Peng, and M. C. Quenez. Backward stochastic differential equations in finance. *Mathematical finance*, 7(1):1–71, 1997.

[50] O. Faugeras, J. Touboul, and B. Cessac. A constructive mean field analysis of multi population neural networks with random synaptic weights and stochastic inputs. *Arxiv preprint arxiv:0808.1113*, 2008.

[51] T. Fekete, D. B. Omer, S. Naaman, and A. Grinvald. Removal of spatial biological artifacts in functional maps by local similarity minimization. *Journal of neuroscience methods*, 178(1):31–39, 2009.

[52] D. Ferster and K. D. Miller. Neural mechanisms of orientation selectivity in the visual cortex. *Annual review of neuroscience*, 23:441, 2000.

[53] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394, 1987.

[54] J. Filip, M. Haindl, and D. Chetverikov. Fast synthesis of dynamic colour textures. In *Pattern recognition, 2006. icpr 2006. 18th international conference on*. Volume 4. IEEE, 2006, pages 25–28.

[55] J. Fournier, C. Monier, M. Pananceau, and Y. Frégnac. Adaptation of the simple or complex nature of v1 receptive fields to visual statistics. *Nature neuroscience*, 14(8):1053–1060, 2011.

[56] R. F. Fox. Stochastic versions of the hodgkin-huxley equations. *Biophysical journal*, 72(5):2068–2074, 1997.

[57] J. Freeman and E. P. Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011.

[58] J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, and J. A. Movshon. A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, 16(7):974–981, 2013.

[59] Y. Frégnac and B. Bathellier. Cortical correlates of low-level perception: from neural circuits to percepts. *Neuron*, 88(1):110–126, 2015.

[60] K. Friston. The history of the future of the bayesian brain. *Neuroimage*, 62(2):1230–1233, 2012.

[61] R. D. Frostig. *In vivo optical imaging of brain function*. CRC press, 2009.

[62] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[63] B. Galerne. Stochastic image models and texture synthesis. PhD thesis. ENS de Cachan, 2011.

[64] B. Galerne, Y. Gousseau, and J. M. Morel. Micro-Texture synthesis by phase randomization. *Image processing on line*, 1, 2011.

[65] B. Galerne, Y. Gousseau, and J. M. Morel. Random Phase Textures: theory and synthesis. *Ieee t. image. process.*, 2010.

[66] D. Ganguli and E. P. Simoncelli. Efficient sensory encoding and bayesian inference with heterogeneous neural populations. *Neural computation*, 2014.

[67] W. S. Geisler and D. Kersten. Illusions, perception and bayes. *Nature neuroscience*, 5(6):508–510, 2002.

[68] W. S. Geisler. Visual perception and the statistical properties of natural scenes. *Annu. rev. psychol.*, 59:167–192, 2008.

[69] I. M. Gel'fand, N. Y. Vilenkin, and A. Feinstein. Generalized functions. vol. 4. 1964.

[70] E. Giné and R. Nickl. Mathematical foundations of infinite-dimensional statistical models. *Cambridge series in statistical and probabilistic mathematics*, 2015.

[71] R. L. Goris, E. P. Simoncelli, and J. A. Movshon. Origin and function of tuning diversity in macaque visual cortex. *Neuron*, 88(4):819–831, 2015.

[72] A. B. Graf, A. Kohn, M. Jazayeri, and J. A. Movshon. Decoding the activity of neuronal populations in macaque primary visual cortex. *Nature neuroscience*, 14(2):239–245, 2011.

[73] A. Gramfort, B. Thirion, and G. Varoquaux. Identifying predictive regions from fmri with tv-l1 prior. In *Pattern recognition in neuroimaging (prni), 2013 international workshop on.* IEEE, 2013, pages 17–20.

[74] R. L. Gregory. Perceptions as hypotheses. *Philosophical transactions of the royal society b: biological sciences*, 290(1038):181–197, July 1980.

[75] A. Grinvald and R. Hildesheim. Vsdi: a new era in functional imaging of cortical dynamics. *Nature reviews neuroscience*, 5(11):874–885, 2004.

[76] A. Grinvald, D. Shoham, A. Shmuel, D. Glaser, I. Vanzetta, E. Shtoyerman, H. Slovin, C. Wijnbergen, R. Hildesheim, and A. Arieli. In-vivo optical imaging of cortical architecture and dynamics. In, *Modern techniques in neuroscience research*, pages 893–969. Springer, 1999.

[77] R. W. Guillery. Observations of synaptic structures: origins of the neuron doctrine and its current status. *Philosophical transactions of the royal society of london b: biological sciences*, 360(1458):1281–1307, 2005.

[78] B. Haider, M. R. Krause, A. Duque, Y. Yu, J. Touryan, J. A. Mazer, and D. A. McCormick. Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation. *Neuron*, 65(1):107–121, 2010.

[79] O. Hassan and S. T. Hammett. Perceptual biases are inconsistent with bayesian encoding of speed in the human visual system. *Journal of vision*, 15(2):9–9, 2015.

[80] O. Hassan, P. Thompson, and S. T. Hammett. Perceived speed in peripheral vision can go up or down. *Journal of vision*, 16(6):20–20, 2016.

[81] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction, second edition*. Of *Springer Series in Statistics*. Springer New York, 2009.

[82] D. J. Heeger and D. Ress. What does fmri tell us about neuronal activity? *Nature reviews neuroscience*, 3(2):142–151, 2002.

[83] J. A. Hirsch and L. M. Martinez. Laminar processing in the visual cortical column. *Current opinion in neurobiology*, 16(4):377–384, 2006.

[84] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The journal of physiology*, 117(4):500, 1952.

[85] S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001.

[86] J. Huang and D. Mumford. Statistics of natural images and models. In *Computer vision and pattern recognition, 1999. ieee computer society conference on*. Volume 1. IEEE, 1999.

[87] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The journal of physiology*, 148(3):574–591, 1959.

[88] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of neurophysiology*, 28(2):229–289, 1965.

[89] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The journal of physiology*, 160(1):106–154, 1962.

[90] C. P. Hung, G. Kreiman, T. Poggio, and J. J. DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005.

[91] M. Hyndman, A. D. Jepson, and D. J. Fleet. Higher-order autoregressive models for dynamic textures. In *Bmvc*, 2007, pages 1–10.

[92] J. Ito. Spike triggered average. *Encyclopedia of computational neuroscience*:2832–2835, 2015.

[93] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*. Volume 6. Springer.

[94] M. Jazayeri and J. A. Movshon. Optimal representation of sensory information by neural populations. *Nature neuroscience*, 9(5):690–696, 2006.

[95] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, and B. Thirion. Multiscale mining of fmri data with hierarchical structured sparsity. *Siam journal on imaging sciences*, 5(3):835–856, 2012.

[96] M. Jogan and A. A. Stocker. Signal integration in human visual speed perception. *The journal of neuroscience*, 35(25):9381–9390, 2015.

[97] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions, vol. 1-2.* New York: John Wiley & Sons, 1994, pages 211–213.

[98] J. P. Jones and L. A. Palmer. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1187–1211, 1987.

[99] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1233–1258, 1987.

[100] M. Kaschube, M. Schnabel, and F. Wolf. Self-organization and the selection of pinwheel density in visual cortical development. *New journal of physics*, 10(1):015009, 2008.

[101] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

[102] D. Kersten, P. Mamassian, and A. Yuille. Object perception as bayesian inference. *Annu. rev. psychol.*, 55:271–304, 2004.

[103] D. C. Knill and A. Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in neurosciences*, 27(12):712–719, 2004.

[104] A. Kohn and J. A. Movshon. Adaptation changes the direction tuning of macaque mt neurons. *Nature neuroscience*, 7(7):764–772, 2004.

[105] M. Köppen. The curse of dimensionality. In *5th online world conference on soft computing in industrial applications (wsc5)*, 2000, pages 4–8.

[106] K. P. Kording. Bayesian statistics: relevant for the brain? *Current opinion in neurobiology*, 25:130–133, 2014.

[107] S. B. Kotsiantis. Supervised machine learning: a review of classification techniques. *Informatica*, 31:249–268, 2007.

[108] V. A. F. Lamme. Blindsight: the role of feedforward and feedback corticocortical connections. *Acta psychologica*, 107(1):209–228, 2001.

[109] V. A. F. Lamme, H. Super, and H. Spekreijse. Feedforward, horizontal, and feedback processing in the visual cortex. *Current opinion in neurobiology*, 8(4):529–535, 1998.

[110] C.-B. Liu, R.-S. Lin, N. Ahuja, and M.-H. Yang. Dynamic textures synthesis as nonlinear manifold learning and traversing. In *Bmvc*. Citeseer, 2006, pages 859–868.

[111] W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438, 2006.

[112] J. H. Macke, S. Gerwinn, L. E. White, M. Kaschube, and M. Bethge. Gaussian process methods for estimating cortical maps. *Neuroimage*, 56(2):570–581, 2011.

[113] S. Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

[114] H. A. Mallot. An overall description of retinotopic mapping in the cat's visual cortex areas 17, 18, and 19. *Biological cybernetics*, 52(1):45–51, 1985.

[115] V. Mante, R. A. Frazor, V. Bonin, W. S. Geisler, and M. Carandini. Independence of luminance and contrast in natural scenes and in the early visual system. *Nature neuroscience*, 8(12):1690–1697, 2005.

[116] V. Markounikau, C. Igel, A. Grinvald, and D. Jancke. A dynamic neural field model of mesoscopic cortical activity captured with voltage-sensitive dye imaging. *Plos comput biol*, 6(9):e1000919, 2010.

[117] G. Marmont. Studies on the axon membrane. i. a new method. *Journal of cellular and comparative physiology*, 34(3):351–382, 1949.

[118] O. Marre, D. Amodei, N. Deshmukh, K. Sadeghi, F. Soo, T. E. Holy, and M. J. Berry. Mapping a complete neural population in the retina. *The journal of neuroscience*, 32(43):14859–14873, 2012.

[119] L. M. Martinez, J.-M. Alonso, R. C. Reid, and J. A. Hirsch. Laminar processing of stimulus orientation in cat visual cortex. *The journal of physiology*, 540(1):321–333, 2002.

[120] J. H. Maunsell and D. C. Van Essen. Functional properties of neurons in middle temporal visual area of the macaque monkey. i. selectivity for stimulus direction, speed, and orientation. *Journal of neurophysiology*, 49(5):1127–1147, 1983.

[121] J. M. McFarland, Y. Cui, and D. A. Butts. Inferring nonlinear neuronal computation based on physiologically plausible inputs. *Plos comput biol*, 9(7):e1003143, 2013.

[122] R. Meidan. On the connection between ordinary and generalized stochastic processes. *Journal of mathematical analysis and applications*, 76(1):124–133, 1980.

[123] S. Menard. *Applied logistic regression analysis.* (106). Sage, 2002.

[124] A. I. Meso and C. Simoncini. Towards an understanding of the roles of visual areas mt and mst in computing speed. *Frontiers in computational neuroscience*, 8, 2014.

[125] A. I. Meso and J. M. Zanker. Speed encoding in correlation motion detectors as a consequence of spatial structure. *Biological cybernetics*, 100(5):361–370, 2009.

[126] V. Michel, C. Damon, and B. Thirion. Mutual information-based feature selection enhances fmri brain activity classification. In *2008 5th ieee international symposium on biomedical imaging: from nano to macro.* IEEE, 2008, pages 592–595.

[127] L. Moisan. Periodic plus smooth image decomposition. *Journal of mathematical imaging and vision*, 39(2):161–179, 2011.

[128] V. Mountcastle, A. Berman, and P. Davies. Topographic organization and modality representation in first somatic area of cat's cerebral cortex by method of single unit analysis. *Am j physiol*, 183(464):10, 1955.

[129] M. C. Murphy, A. J. Poplawsky, A. L. Vazquez, K. C. Chan, S.-G. Kim, and M. Fukuda. Improved spatial accuracy of functional maps in the rat olfactory bulb using supervised machine learning approach. *Neuroimage*, 2016.

[130] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[131] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

[132] P. Neri and D. J. Heeger. Spatiotemporal mechanisms for detecting and identifying image features in human vision. *Nature neuroscience*, 5(8):812–816, 2002.

[133] P. Neri and D. M. Levi. Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision research*, 46(16):2465–2474, 2006.

[134] P. Neri, A. J. Parker, and C. Blakemore. Probing the human stereoscopic system with reverse correlation. *Nature*, 401(6754):695–698, 1999.

[135] O. Nestares, D. Fleet, and D. Heeger. Likelihood functions and confidence bounds for total-least-squares problems. In *Ieee conference on computer vision and pattern recognition. cvpr 2000.* Volume 1. IEEE Comput. Soc, Hilton Head Island, SC, USA, 2000, pages 523–530.

[136] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

[137] F. Oberhettinger. *Tables of mellin transforms*. Springer Science & Business Media, 2012.

[138] G. C. Ohlmacher and J. C. Davis. Using multiple logistic regression and gis technology to predict landslide hazard in northeast kansas, usa. *Engineering geology*, 69(3):331–343, 2003.

[139] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:13, 1996.

[140] B. A. Olshausen and D. J. Field. How close are we to understanding v1? *Neural computation*, 17(8):1665–1699, 2005.

[141] S. Onat, P. König, and D. Jancke. Natural scene evoked population dynamics across cat primary visual cortex captured with voltage-sensitive dye imaging. *Cerebral cortex*, 21(11):2542–2554, 2011.

[142] S. Onat, N. Nortmann, S. Rekauzke, P. König, and D. Jancke. Independent encoding of grating motion across stationary feature maps in primary visual cortex visualized with voltage-sensitive dye imaging. *Neuroimage*, 55(4):1763–1770, 2011.

[143] L. Paninski. Convergence properties of three spike-triggered analysis techniques. *Network: computation in neural systems*, 14(3):437–464, 2003.

[144] L. Paninski. Nonparametric inference of prior probabilities from bayes-optimal behavior. In *Advances in neural information processing systems*, 2005, pages 1067–1074.

[145] M. Park and J. W. Pillow. Receptive field inference with localized priors. *Plos comput biol*, 7(10):e1002219, 2011.

[146] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[147] J. A. Perrone and A. Thiele. Speed skills: measuring the visual speed analyzing properties of primate MT neurons. *Nat. neurosci.*, 4(5):526–532, May 2001.

[148] M. Piccolino. Luigi galvani and animal electricity: two centuries after the foundation of electrophysiology. *Trends in neurosciences*, 20(10):443–448, 1997.

[149] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.

[150] A. Pouget, J. M. Beck, W. J. Ma, and P. E. Latham. Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170–1178, 2013.

[151] A. Pouget, P. Dayan, and R. S. Zemel. Inference and computation with population codes. *Annual review of neuroscience*, 26(1):381–410, 2003.

[152] N. Priebe, C. Cassanello, and S. Lisberger. The neural representation of speed in macaque area MT/V5. *J. neurosci.*, 23:5650–5661, 2003.

[153] N. J. Priebe, S. G. Lisberger, and J. A. Movshon. Tuning for spatiotemporal frequency and speed in directionally selective neurons of macaque striate cortex. *The journal of neuroscience*, 26(11):2941–2950, 2006.

[154] Z. Qiao, L. Zhou, and J. Z. Huang. Effective linear discriminant analysis for high dimensional, low sample size data. In *Proceeding of the world congress on engineering*. Volume 2. Citeseer, 2008, pages 2–4.

[155] H. Raguet. A signal processing approach to voltage-sensitive dye optical imaging. PhD thesis. Paris 9, 2014.

[156] H. Raguet, C. Monier, L. Foubert, I. Ferezou, Y. Fregnac, and G. Peyré. Spatially structured sparse morphological component separation for voltage-sensitive dye optical imaging. *Journal of neuroscience methods*, 257 :76–96, 2016.

[157] A. Rahman, M. Murshed, et al. Dynamic texture synthesis using motion distribution statistics. *Journal of research and practice in information technology*, 40(2):129, 2008.

[158] B. Renshaw, A. Forbes, and B. R. Morison. Activity of isocortex and hippocampus: electrical studies with micro-electrodes. *Journal of neurophysiology*, 3(1):74–105, 1940.

[159] A. Reynaud, S. Takerkart, G. S. Masson, and F. Chavane. Linear model decomposition for voltage-sensitive dye imaging signals: application in awake behaving monkey. *Neuroimage*, 54(2):1196–1210, 2011.

[160] R. V. Rikhye and M. Sur. Spatial correlations in natural scenes modulate response reliability in mouse visual cortex. *The journal of neuroscience*, 35(43):14661–14680, 2015.

[161] C. Rossant, S. N. Kadir, D. F. M. Goodman, J. Schulman, M. L. D. Hunter, A. B. Saleem, A. Grosmark, M. Belluscio, G. H. Denfield, A. S. Ecker, et al. Spike sorting for large, dense electrode arrays. *Nature neuroscience*, 2016.

[162] G. A. Rousselet, S. J. Thorpe, and M. Fabre-Thorpe. How parallel is visual processing in the ventral pathway? *Trends in cognitive sciences*, 8(8):363–370, 2004.

[163] D. L. Ruderman and W. Bialek. Statistics of natural images: scaling in the woods. *Physical review letters*, 73(6):814, 1994.

[164] N. C. Rust and J. A. Movshon. In praise of artifice. *Nature neuroscience*, 8(12):1647–1650, 2005.

[165] N. C. Rust, O. Schwartz, J. A. Movshon, and E. P. Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46(6):945–956, 2005.

[166] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology, 1990.

[167] B. M. Salzberg, H. V. Davila, and L. B. Cohen. Optical recording of impulses in individual neurones of an invertebrate central nervous system. *Nature*, 246(5434):508–509, 1973.

[168] I. Samengo and T. Gollisch. Spike-triggered covariance: geometric proof, symmetry properties, and extension beyond gaussian stimuli. *Journal of computational neuroscience*, 34(1):137–161, 2013.

[169] P. Sanz-Leon, I. Vanzetta, G. S. Masson, and L. U. Perrinet. Motion clouds: model-based stimulus synthesis of natural-like random textures for the study of motion perception. *Journal of neurophysiology*, 107(11):3217–3226, March 2012.

[170] P. R. Schrater, D. C. Knill, and E. P. Simoncelli. Mechanisms of visual motion detection. *Nature neuroscience*, 3(1):64–68, 2000.

[171] P. R. Schrater, D. C. Knill, and E. P. Simoncelli. Perceiving visual expansion without optic flow. *Nature*, 410(6830):816–819, 2001.

[172] O. Schwartz, J. W. Pillow, N. C. Rust, and E. P. Simoncelli. Spike-triggered neural characterization. *Journal of vision*, 6(4):13–13, 2006.

[173] J. Shao, Y. Wang, X. Deng, S. Wang, et al. Sparse linear discriminant analysis by thresholding for high dimensional data. *The annals of statistics*, 39(2):1241–1265, 2011.

[174] D. Sharon and A. Grinvald. Dynamics and constancy in cortical spatiotemporal patterns of orientation processing. *Science*, 295(5554):512–515, 2002.

[175] E. P. Simoncelli. Vision and the statistics of the visual environment. *Current opinion in neurobiology*, 13(2):144–149, 2003.

[176] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

[177] E. P. Simoncelli, L. Paninski, J. Pillow, and O. Schwartz. Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences*, 3:327–338, 2004.

[178] C. Simoncini, L. U. Perrinet, A. Montagnini, P. Mamassian, and G. S. Masson. More is not always better: adaptive gain control explains dissociation between perception and action. *Nature neurosci*, 15(11):1596–1603, November 2012.

[179] A. T. Smith and G. K. Edgar. The influence of spatial frequency on perceived temporal frequency and perceived speed. *Vision res.*, 30:1467–1474, 1990.

[180] M. A. Smith, N. Majaj, and J. A. Movshon. Dynamics of pattern motion computation. In G. S. Masson and U. J. Ilg, editors, *Dynamics of visual motion processing: neuronal, behavioral and computational approaches*, pages 55–72. Springer, Berlin-Heidelberg, first edition, 2010.

[181] P. L. Smith. Stochastic dynamic models of response time and accuracy: a foundational primer. *Journal of mathematical psychology*, 44(3):408–463, 2000.

[182] G. Sotiropoulos, A. R. Seitz, and P. Seriès. Contrast dependency and prior expectations in human speed perception. *Vision research*, 97:16–23, 2014.

[183] G. Sotiropoulos, A. R. Seitz, and P. Seriès. Contrast dependency and prior expectations in human speed perception. *Vision research*, 97:16–23, 2014.

[184] A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, 9(4):578–585, 2006.

[185] N. V. Swindale. Orientation tuning curves: empirical description and estimation of parameters. *Biological cybernetics*, 78(1):45–56, 1998.

[186] F. E. H. Tay and L. Cao. Application of support vector machines in financial time series forecasting. *Omega*, 29(4):309–317, 2001.

[187] M. Teplan. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002.

[188] B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J.-B. Poline, D. Lebihan, and S. Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116, 2006.

[189] P. Thompson. Perceived rate of movement depends on contrast. *Vision research*, 22(3):377–380, 1982.

[190] P. Thompson, K. Brooks, and S. T. Hammett. Speed can go up as well as down at low contrast: implications for models of motion perception. *Vision research*, 46(6):782–786, 2006.

[191] M. Unser, P. D. Tafti, A. Amini, and H. Kirshner. A unified formulation of gaussian versus sparse stochastic processes - part II: Discrete-Domain theory. *Ieee transactions on information theory*, 60(5):3036–3051, 2014.

[192] M. Unser and P. Tafti. *An introduction to sparse stochastic processes.* Cambridge University Press, Cambridge, UK, 2014. 367 p.

[193] M. Unser, P. D. Tafti, and Q. Sun. A unified formulation of gaussian versus sparse stochastic processes-part i: continuous-domain theory. *Information theory, ieee transactions on*, 60(3):1945–1962, 2014.

[194] J. Vacher, A. I. Meso, L. U. Perrinet, and G. Peyré. Biologically inspired dynamic textures for probing motion perception. In *Advances in neural information processing systems*, 2015, pages 1909–1917.

[195] A. Van der Schaaf and J. Van Hateren. Modelling the power spectra of natural images: statistics and information. *Vision research*, 36(17):2759–2770, 1996.

[196] N. G. Van K. *Stochastic processes in physics and chemistry.* Volume 1. Elsevier, 1992.

[197] V. N. Vapnik. An overview of statistical learning theory. *Ieee transactions on neural networks*, 10(5):988–999, 1999.

[198] V. N. Vapnik. *Statistical learning theory*. Volume 1. Wiley New York, 1998.

[199] V. N. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 1999.

[200] G. Varoquaux, A. Gramfort, and B. Thirion. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. *Arxiv preprint arxiv:1206.6447*, 2012.

[201] W. E. Vinje and J. L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

[202] B. Vintch and J. L. Gardner. Cortical correlates of human motion perception biases. *The journal of neuroscience*, 34(7):2592–2604, 2014.

[203] B. Vintch, J. A. Movshon, and E. P. Simoncelli. A convolutional subunit model for neuronal responses in macaque v1. *The journal of neuroscience*, 35(44):14829–14841, 2015.

[204] V. Q. Vu, P. Ravikumar, T. Naselaris, K. N. Kay, J. L. Gallant, and B. Yu. Encoding and decoding v1 fmri responses to natural images with sparse nonparametric models. *The annals of applied statistics*, 5(2B):1159, 2011.

[205] A. S. Waggoner. Dye indicators of membrane potential. *Annual review of biophysics and bioengineering*, 8(1):47–68, 1979.

[206] M. J. Wainwright and E. P. Simoncelli. Scale mixtures of gaussians and the statistics of natural images. In *Nips*. Citeseer, 1999, pages 855–861.

[207] S. Watanabe. Karhunen-loeve expansion and factor analysis, theoretical remarks and applications. In *Proc. 4th prague conf. inform. theory*, 1965.

[208] L. Y. Wei, S. Lefebvre, V. Kwatra, and G. Turk. State of the art in example-based texture synthesis. In *Eurographics 2009, state of the art report, eg-star*. Eurographics Association, 2009.

[209] X. X. Wei and A. A. Stocker. Efficient coding provides a direct link between prior and likelihood in perceptual bayesian inference. In *Nips*. P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, 2012, pages 1313–1321.

[210] Y. Weiss and D. J. Fleet. Velocity likelihoods in biological and machine vision. In *In probabilistic models of the brain: perception and neural function*, 2001, pages 81–100.

[211] Y. Weiss, E. P. Simoncelli, and E. H. Adelson. Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604, June 2002.

[212] W. Wu, P. H. Tiesinga, T. R. Tucker, S. R. Mitroff, and D. Fitzpatrick. Dynamics of population response to changes of motion direction in primary visual cortex. *The journal of neuroscience*, 31(36):12767–12777, 2011.

[213] G. S. Xia, S. Ferradans, G. Peyré, and J. F. Aujol. Synthesizing and mixing stationary gaussian texture models. *Siam journal on imaging sciences*, 7(1):476–508, 2014.

[214] Y. Xiao, R. Rao, G. Cecchi, and E. Kaplan. Improved mapping of information distribution across the cortical surface with the support vector machine. *Neural networks*, 21(2):341–348, 2008.

[215] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[216] E. Yavuz. Source separation analysis of visual cortical dynamics revealed by voltage sensitive dye imaging. PhD thesis. Université Pierre et Marie Curie-Paris VI, 2012.

[217] R. A. Young and R. M. Lesperance. The gaussian derivative model for spatial-temporal vision: II. cortical data. *Spatial vision*, 14(3):321–390, 2001.

[218] L. Yuan, F. Wen, C. Liu, and H.-Y. Shum. Synthesizing dynamic texture with closed-loop linear dynamic system. In, *Computer vision-eccv 2004*, pages 603–616. Springer, 2004.