Evaluating Computational Models of Vision with Functional Magnetic Resonance Imaging

Michael Eickenberg

UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE 427 :
INFORMATIQUE PARIS SUD

Laboratoire : *Equipe Parietal, Inria Saclay, CEA Saclay*

THÈSE DE DOCTORAT

INFORMATIQUE

par

Michael EICKENBERG

Evaluation de modèles computationnels de la vision humaine en imagérie par résonance magnétique fonctionnelle

Date de soutenance : 21/09/2015

Composition du jury :

Directeur de thèse :	**Bertrand** THIRION	DR, Inria Saclay, Palaiseau
Rapporteurs :	Marcel VAN GERVEN	DR, Donders Institute, Nijmegen, NL
	Nikolaus KRIEGESKORTE	DR, Cambridge University, UK
Examinateurs :	Stéphane MALLAT	Professeur, ENS, Paris
	Yves FREGNAC	DR, CRNS Gif-sur-Yvette, Gif-sur-Yvette
	Balazs KEGL	DR, Université de Paris Sud, Orsay
	Alexandre GRAMFORT	DR, CRNS, Télécom ParisTech, Paris



UNIVERSITÉ DE PARIS SUD

DOCTORAL SCHOOL OF COMPUTER SCIENCE

PREPARED AT PARIETAL TEAM - INRIA SACLAY

Evaluating Computational Models of Vision with Functional Magnetic Resonance Imaging

Michael Eickenberg

A dissertation submitted in partial fulfilment

of the requirements for the degree of doctor of science,

specialized in computer science.
Director: Bertrand Thirion
Co-Directors: Alexandre Gramfort, Gaël Varoquaux

Defended publicly the 21th of September 2015 in front of a jury consisting of

Advisor	Bertrand Thirion	INRIA / CEA, Saclay, France
Reviewers	Marcel Van Gerven	Donders Institute, Nijmegen, NL
	Nikolaus Kriegeskorte	MRC-CBU, Cambridge University, UK
Examiners	Stéphane Mallat	ENS, Paris, France
	Yves Frégnac	UNIC, CNRS, Gif-sur-Yvette, France
	Balász Kégl	CNRS, Université de Paris Sud, France
	Alexandre Gramfort	Telecom Paristech, Paris, France
	Gaël Varoquaux	INRIA, Saclay, France

UNIVERSITÉ DE PARIS SUD
ÉCOLE DOCTORALE INFORMATIQUE
ÉQUIPE PARIETAL - INRIA SACLAY

Evaluation de modèles computationnels de la vision en imagerie par résonance magnétique fonctionnelle

Michael Eickenberg

Thèse de doctorat pour obtenir le grade de
DOCTEUR de l'UNIVERSITÉ de PARIS-SUD

Dirigée par Bertrand Thirion.
Co-dirigée par Alexandre Gramfort et Gaël Varoquaux

Présentée et soutenue publiquement le 21 Septembre 2015 devant un jury composé de :

Directeur	Bertrand Thirion	INRIA / CEA, Saclay, France
Rapporteurs	Marcel Van Gerven	Donders Institute, Nijmegen, NL
	Nikolaus Kriegeskorte	MRC-CBU, Cambridge University, UK
Examinateurs	Stéphane Mallat	ENS, Paris, France
	Yves Frégnac	UNIC, CNRS, Gif-sur-Yvette, France
	Balász Kégl	CNRS, Université de Paris Sud, France
	Alexandre Gramfort	Telecom Paristech, Paris, France
	Gaël Varoquaux	INRIA, Saclay, France

Abstract

Computer vision studies and biological vision studies have evolved in parallel over the last century with mostly unilateral inspiration taken from biological vision and going into the engineering of computer vision systems. From the utility of edge and blob detection to the realization that a layered or hierarchical approach to abstraction can be very powerful, most of these phenomena are to be found in natural visual systems in some way or another.

With the successes in computer vision brought about during the last one and a half decades, which can be subdivided into different sub-eras (see chapter 4), it has become a highly intriguing question to assess whether these methods can help study brain function.

The main goal of this thesis is to confront a few more or less biologically inspired computational models of vision with actual brain data. The chosen brain data acquisition modality is fMRI, since it gives a good global overview of activity at a reasonable spatial resolution.

In recovering brain activity maps for the presentation of a preferably high number of visual stimuli, we shall attempt to relate the measured activity with the vectorial representations of the stimuli generated by the computational models.

Since these representations are typically very high in dimension, we need to resort to non classical statistical methods to establish a relationship between the model representations and the brain data. The method for functional translation from the computational model coefficients to brain activity data is kept linear. This is essential for evaluation, in order to keep most nonlinear complexity in the computational model under scrutiny, instead of adjusting for the lack of it through nonlinear estimators. However, due to the abundance of coefficients in typical computational models, the forward problem is ill-posed and calls for regularization as well as an evaluation on held-out data typical of the field of machine learning.

In chapter 2 of this thesis we will familiarize ourselves with the nature of the fMRI BOLD response by evaluating the utility of estimating the hemodynamic impulse response function (HRF) due to experimental stimulation. The evaluation is specifically geared towards assessing whether attention to estimating the shape of the HRF is merited in the context of machine learning forward and reverse models.

A focus on specific convex regularization techniques will be explored in chapter 3. We introduce a convex region-selecting penalty which segments smooth active sets from a uniform zero background. This spatial regularizer is applicable to the space of brain images. We will evaluate it in a reverse modelling setting - predicting an external variable from brain activity patterns. Here again, the number of voxels typically largely exceeds the number of observations, leading to an ill-posed problem necessitating regularization. We choose to regularize by taking into account the spatial neighborhood structure of brain images, because neighboring voxels tend to correlate in activity.

In chapter 5 we evaluate a first forward ("encoding") model with specific attention to the benefits of adding a layer to a convolutional filter model of vision. Indeed, one-layer filter models such as Gabor or Morlet filtering followed by rectification have been tried and tested successfully in numerous

experiments, including the notable application to fMRI by [Kay et al., 2008]. Since first layer filters represent an adequate model for lower-level brain activity such as center-surround calculations in LGN or edge detection in V1, the question is whether adding a second layer to the analysis could be beneficial to model fitting. Experiments were performed on two datasets: 1) Data acquired in the Parietal lab prior to this thesis - BOLD fMRI responses to the presentation of natural visual textures. 2) The dataset of [Kay et al., 2008] for a parallel analysis on natural images. We evaluated the second layer of the scattering transform [Mallat, 2012] against the first layer, which is a wavelet transform modulus (e.g. Gabor or Morlet modulus). In addition, the texture experiment also yielded itself to classical statistics: brain activation due to texture stimulation and differential brain activation between different texture classes are explored..

Recent breakthroughs in the field of convolutional networks have led to breathtaking progress in their capacity to perform tasks that were previously believed to be strongholds of human superiority over machines. Deep layered convolutional architectures create progressively abstract representations of the data they analyze with increasing layer number. The first layer of a convolutional network geared towards object recognition typically learns to detect edges, other first-order texture boundaries, color boundaries and blobs - similar in essence to the functionality one may find in earliest vision. At the end of the convolutional network there are indicator channels outputting probability estimates for a certain number of object categories. These are based on linear transformations of the penultimate layer, which can thus be declared to linearize object category. In [Cadieu et al., 2014] it is shown that populations of inferotemporal neurons behave similarly. Having pinpointed similarities to biological signal processing at the beginning and at the end of the convolutional net processing hierarchy, we proceed to investigate similarities of the representations along the layers of the network.

Chapter 1 introduces the reader to fMRI and standard analysis methods. Chapter 4 gives an overview of computer-vision models.

Acknowledgements

At some point during the meanderings around and into the adjacent possible, guided less and less by external force fields giving drift direction, the question becomes whether the movement reduces to isotropically random or picks up its own intrinsic momentum that can compensate dwindling gradient information. In this situation, at one and the same moment one can feel perfectly isolated and fully woven into the surrounding social fabric. Scaling a mountain alone through a thick layer of fog, in the hope that the peak is either attainable and high enough to be above the clouds or the clouds low enough to see from somewhere off the slope can be quite a solitary experience. On the other hand, flowing in the school-of-fish-like dynamics of the scientific community is a collective and immersive action. We also flow along with society at large, which acknowledges an essential part of the human condition - curiosity - and accommodates a formalized version of it, i.e. letting us do our work, not of course without expecting and reaping its own benefits. We are pretty lucky to be able to do research - to stand with one leg on the shoulder of the scientific giants that preceded us and with the other on those many other shoulders of giants that give us the space to do what we do right now. At that level of generality there would be millions of people to thank. Here I will restrict myself to people I know, who helped me along these last years and without who I would not have come this far. My first word of thanks goes out to my thesis director, Bertrand Thirion, who guided me through one year of internship and three years of PhD, without wavering, especially through the inevitable sticky moments this first true journey into research can have. The same can be said of my co-supervisors Alexandre Gramfort and Gaël Varoquaux. All three never failed to push me forward when it was necessary and always had the right words at the right moment. The reviewers of my thesis manuscript were Marcel van Gerven and Nikolaus Kriegeskorte, whom I thank for accepting to review and for the feedback they gave me. Naturally I would also like to thank the rest of my thesis committee, which consisted of Yves Frégnac, who acted as president, Bertrand Thirion, Alexandre Gramfort, Balazs Kégl, and Stéphane Mallat. I would like to address special thanks to Stéphane Mallat for having accompanied my work over all these years through our common ANR project and always giving very helpful feedback. This also goes for all his team members, past and present, who in addition to being extremely competent, are also the friendliest of people: Joan Bruna, Joakim Andén, Laurent Sifre, Edouard Oyallon. I'm glad to be able to continue my work in this lab alongside Edouard, Mathieu Andreux, Irène Waldspurger, Vincent Lostanlen, Sira Ferradans, Grégoire Sergeant-Perthuis and Carmine Cella. Towards the end of my internship year preceding my PhD, I was also lucky enough to be able to work on the signal processing topic of super-resolution for MEG with Alexandre Gramfort and Gabriel Peyré, who I thank for the time he took and the impressive supervision and feedback he gave at every discussion. I also greatly enjoyed interacting with Samuel Vaiter, Mohammed Golbabaee, Charles Deledalle, Hugo Raguet, both at the lab and at spars2013. Next, in a brief stint of 11 days in Berkeley, I was able to meet and work with Jack Gallant's lab, the team that made my PhD results possible by kindly providing the data of their immensely suc-

cessful encoding experiments to the public for download. I met the brightest and friendliest people ever and am very thankful to have been able to collaborate with Alex Huth and Natalia Bilenko and to have met Jack himself, Mike Oliver, Anwar Nuñez, James Gao, Fatma Imamoglu and the rest of the team during my stay. I would like to thank my collaborators for engaging and fruitful interactions: Fabian Pedregosa, Mehdi Senoussi, Philippe Ciuciu, Danilo Bzdok, Elvis Dohmatob, Olivier Grisel, Thomas Hannagan, Kyle Kastner, Konstantin Shmelkov. Further shout-outs have to go to all current and former lab mates who I enjoyed all those coffee breaks with: Benoît Da Mota for all the petaflops, Yannick Schwartz for the trolling, all the driving, the drive-trolling and the troll-driving as well as some good conversations that bridged any traffic jam, Virgile Fritsch who I congratulate for his part-time 32 hour week, Viviana Siless who I congratulate for absolutely making it and for some awesome food samples. Sergio Medina, Bernard Ng, Clément Moutard, Martin Perez Guevara, Esther Lin, Salma Bougacha, Laetitia Grabot, Arthur Mensch, Kamalakar Dadi, Philippe Gervais, Solveig Badillo. Les manip' radio and nurses and the whole Neurospin subject recruitment and scanning team for making that part easy for many of us. Jaques Grobler for throwing discs and all the latest memes and Jaques Grobler and Svenja Lach for the SA vibes and fun at parties, Alexandre Abraham for being the essence of kindness and taking care of everything, Valentina Borghesani for also being the essence of kindness and taking care of everything, Mehdi Rahim for the same - all three incarnations of incredible real-world efficiency. Régine Bricquet for being the real-life incarnation of at least nine people in parallel with incredible real-world efficiency. Alexandre Abraham, Nicolas Chauffert, Murielle Fabre, Sandrine Lefranc, Remi Magnin, Guillaume Radecki for Porquerolles at which what happens stays. Andrés Hoyos-Idrobo, Léonard Blier, Elvis Dohmatob, Pedro Pinheiro Chagas, Loïc Esteve and Aina Frau Pascual for being awesome at ping pong. Christophe Pallier, Thomas Hannagan, Evelyn Eger, Alexandre Vignaud, Alexis Amadon and Aaron Schurger for very useful insights along all these years. Valentina, Laetitia and Aaron for helping me organize the Unsupervised Decoding Club for a year. Baptiste Gauthier, Yannick Schwartz, Alexandre Abraham, Marie Amalric, Murielle Fabre and actually everybody else for awesome coffee break discussions. Pedro Pinheiro Chagas for beer and Pedro and Aina for breakfast coffees. Kyle Kastner for getting me motivated to look at neural networks and Theano and Olivier and Kyle for great discussions about these topics. Charles Ollion for deep learning meetups and workshops and Charles, Kyle, Olivier, Mike, Anwar, Gaël and Danilo for good company at NIPS. Denis Engemann and Danilo for convincing me that the future is well taken care of. Aina as well as Valentina and Fabian for being there in the right moments and Murielle, Aina, Valentina, Mehdi and Alex for making the thesis defense go smoothly. Of course Marian, Trevor, Dana and so many others that I never lost contact with through these last four years. And of course my parents and my sister. And Kaja, for sticking through this with me until the end.

Contents

1 Introduction to functional MRI

1.1 Imaging modalities

There exist a number of techniques for the acquisition of brain activity which rely on a very diverse set of possible observable signals. In general, any signal that is sufficiently immediately generated and modulated by brain activity can serve as a tap for a brain activity acquisition method. Useful dimensions by which to taxonomize a large number of these methods are the following:

- *invasiveness*, i.e. to what extent the body containing the brain of interest is manipulated during the acquisition,
- *spatial resolution*, characterized by a minimal characteristic spatial scale below which no details can be measured,
- *temporal resolution*, characterized by a time scale below which no details can be measured.

It is important to note that depending on the process by which brain signal is obtained, loss of resolution can be incurred at intermediate steps. Ideally, the final measurement should reflect the intrinsic resolution of the signal due to this process, but in principle, these can be uncoupled: For example, an inherently slow signal can be sampled at many time points, which can potentially reduce noise, but it cannot recover any high temporal frequencies previously lost. We present an incomplete overview of methods in order to be able to situate fMRI, which this thesis makes use of, better with respect to the others.

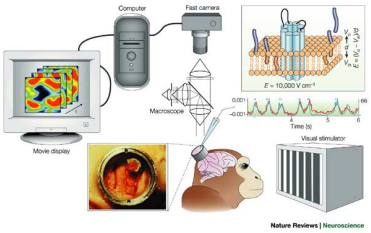

Figure 1.1: VSD setup. Taken from [Grinvald, 2004]

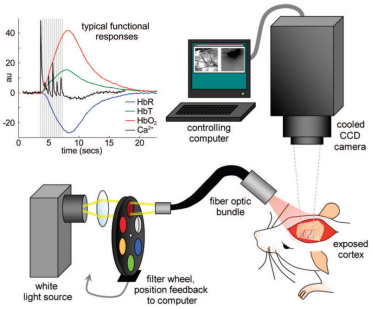

Figure 1.2: Typical general optical imaging setup. If there exist contrast agents (natural or not) that modulate light according to biological function, this setup can be used to acquire images. Taken from [Hillman, 2007]

1.1.1 Highly invasive methods

Highly invasive methods are characterized by requiring surgical intervention to enable acquisition.

Recently, several methods acquiring light images have had success. These methods include *voltage sensitive die* (VSD) [Tasaki et al., 1968, Orbach et al., 1985] methods, where a substance which changes color as a function of local potential electric energy is applied to the cortex, making electric brain activity visible to a camera.

While VSDs modulate with voltage change, *calcium imaging* [Smetters et al., 1999] is a technique by which so-called calcium indicators, molecules that become fluorescent on calcium binding, are used to assess the calcium content of neurons, which is directly related to their activity because it contributes to the polarization of the cell.

In general, any contrast agent creating modulation of fluorescence or reflectivity properties as a function of biological processes can be amenable to optical imaging methods [Hillman, 2008].

More traditionally, there is electrophysiology, for which a variety of techniques has been developed. The general setup requires the placement of an electrical conductor into or into the vicinity of neurons in order to measure the local voltage. Intracellular recordings are obtained when an electrode is placed inside a neuron. Local field potentials are obtained by placing an electrode at a sufficient and sufficiently similar distance from several neurons. The electrodes measure voltage fluctuations from neurons within a certain radius.

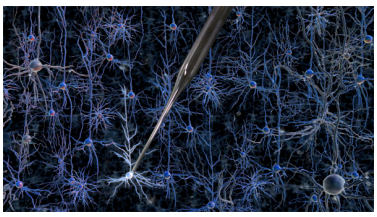

Figure 1.3: Schematic visualization of an intracellular recording.

It is possible to place arrays of many electrodes arranged in a grid to record from multiple locations at once. *Intracranial EEG* or *Electrocorticography* (ECOG) is a similar technique in which a plastic sheet containing more widely spaced electrodes is placed on the cortex.

The latter, along with depth electrodes, are also used in humans, for example to determine the focus points of epilepsy attacks that a patient may suffer.

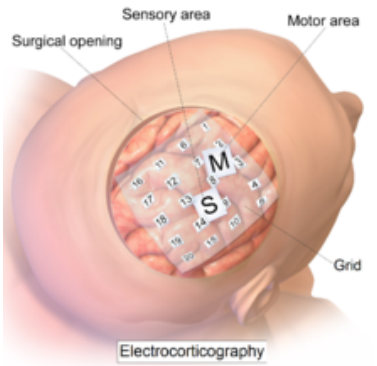

Figure 1.4: Electrocorticography electrode grid placed on the surface of the cortex of a human subject.

1.1.2 Less invasive methods

Comparing to the strongly invasive methods mentioned above, there are less invasive imaging modalities, whose degree of invasiveness owes to the use of radiation or radioactivity. Anatomical and functional brain imaging can be obtained by *computed tomography* (CT). Tomography is the measurement of projections of a 3-dimensional object with varying density onto a certain number of 2-dimensional planes. Reconstruction of the original 3-dimensional object can be done by solving an inverse problem, which is often linear. In *X-ray CT* the measurement projections onto planes are obtained using X-ray light which is partially absorbed depending on the local properties of the matter it traverses. Functional and other metabolic information can be imaged by injecting a contrast agent which the organism transports to specific sites and which change the way the X-rays are absorbed.

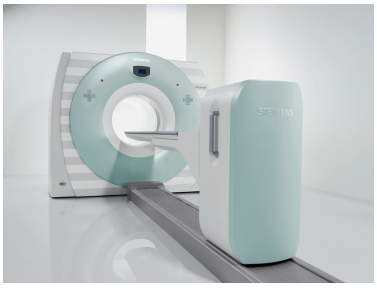

Figure 1.5: X-ray CT and PET scanner Siemens Biograph TruePoint

Another form of computed tomography is *Positron Emission Tomography* (PET), which is designed to track metabolic activity. A fast-decaying radioactive glucose is injected. At each radioactive decay, two gamma photons are emitted in opposite directions and captured outside the head. The emission of two photons is necessary to conserve momentum and makes it possible to localize brain activity.

1.1.3 Non-invasive methods

Non-invasive methods are techniques which measure brain signal and anatomy without any known potentially adverse side-effects. No surgery is required and no contrast agents are injected. Electromagnetic brain activity can be measured outside the head. Scalp electrodes can acquire an *electroencephalogram* (EEG), providing measurements at almost arbitrary temporal resolution. By the non-relativistic Maxwell equations, the measurements are a linear function of the total brain electric activity. Similarly, dynamic magnetic activity of the brain can be measured using superconducting *SQUID* sensors,

giving rise to a *magnetoencephalogram* (MEG). One can measure magnetic field intensity (using so-called *magnetometers*) as well as magnetic field gradient in two directions tangential to the surface (using so-called *gradiometers*). Both EEG and MEG acquisitions, by the simple fact that data acquisition is performed at a distance from the signal source, act as a spatial low-pass filter, where the kernel has heavy tails and decreases as $\sqrt{R^2 + x^2}^{-3}$, where R is a characteristic distance. [1] Due to the Maxwell equations this cannot be avoided. Even if measurements were taken continuously on a full sphere around the head, the reconstruction problem remains ill-conditioned. Additionally, for EEG, the scalp acts like a further spatial low-pass filter, aggravating the ill conditioning of the source reconstruction problem. In the MEG case the scalp does not act as an additional low-pass filter, making measurements more precise. Typically one has access to around 300 channels, all types taken together, which is more than using a normal EEG setup, which can use anywhere from very few to around 250 electrodes. The fact in both MEG and EEG that there can only be up to hundreds of electrodes due to space constraints on the scalp makes the source reconstruction problem ill-posed in addition to ill-conditioned (because there are many more candidate locations for sources than measurements). Both ill-conditioning and ill-posedness can be addressed by regularization.

A further non-invasive method which relates back to the optical imaging methods mentioned earlier is *fNIRS* (functional near-infrared spectroscopy). As many may have experienced, the light of a traditional flashlight, when covered by the hand, becomes a light red. This indicates permeability of tissue by light in the red spectrum. As it turns out, this permeability is most pronounced in the near infra-red spectrum (650nm to 1350nm). Skin, tissue and bone are almost transparent in the window of 700-900nm. However, oxy-hemoglobin and deoxy-hemoglobin have stronger absorption properties and can thus be identified. They can also be distinguished amongst each other because their absorption spectra differ. This phenomenon can be used for optical imaging. By using several light sources and several points of measurement, a forward model of light diffusion can be inverted, leading to spatial localization at a $\sim$ 1cm resolution.

Another method classified as non-invasive is *Magnetic Resonance Imaging* (MRI). Since this is the acquisition modality employed in this thesis, it will be described in a separate section.

1.2 MRI

Magnetic resonance imaging is based on *Nuclear Magnetic Resonance* (NMR). This non spatially specific effect is then exploited to obtain a spatial image.

1.2.1 Nuclear magnetic resonance

A proton [2] placed in a homogeneous magnetic field will align its spin with the magnetic field vector. Energy introduced in the form of a radio frequency (RF) pulse can cause the proton to be excited into precession around its axis. The proton dissipates this energy by emitting radio waves at its precession frequency until it has returned into alignment with the magnetic field. Cru-

[1] This kernel arises exactly in a stylized setting, where one studies the magnetic field evoked by sources on one straight line, measured on a parallel straight line. One observes a convolution of the source distribution with the kernel described here, which amounts to low-pass filtering. See [Eickenberg, SPARS 2013, Poster 141] for details.

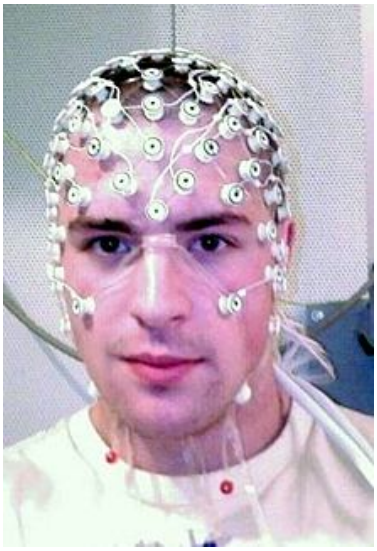

Figure 1.6: Placement of EEG electrodes on the head.

Figure 1.7: MEG machine in a magnetically shielded room (MSR).

[2] In general, an atom or molecule with non-zero net spin

cially, the precession frequency, called the *Larmor frequency*, is proportional to the magnetic field to which the proton is exposed.

The above description relies on concepts from classical physics, but the quantum mechanical description is similar: The application of a magnetic field splits the ground energy state of a proton into two possible states. The energy gap between the states is proportional to the magnetic field and thus an RF pulse containing photons of the corresponding frequency can excite the protons into the antiparallel state. The proton decays back into its ground state with a probability following an exponential law with a certain half-life, emitting the acquired energy in a photon with the Larmor frequency. This effect was first described in 1938 by Isidor Rabi, based on the Stern-Gerlach experiment. In the late 40s this technique was extended to liquids and solids, independently by Felix Bloch and Edward Mills Purcell.

Figure 1.8: NMR spectrometer for the study of the structure of molecules via quantum magnetic effects such as Zeeman energy level splitting.

1.2.2 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) exploits the fact that nuclear magnetic resonance frequency has such a simple dependence on the intensity of the magnetic field surrounding the nucleus. Producing a linear change in magnetic field intensity (not orientation) along a given axis in 3D space makes all hyperplanes perpendicular to the change direction have the same magnetic field intensity, leading to the same Larmor frequency for protons lying upon it. This fact permits selective excitation of 2D slices in 3D space, since it is possible to send an RF pulse in a prescribed frequency range, using a cardinal sine (sinc) waveform $\sin(\pi\omega t)/(\pi\omega t)$. In order to acquire an image of a 3D object, one may now consider the simpler problem of acquiring it slice by slice. However, the excitation of a full slice will lead to the simultaneous decay of excitation over all of the slice at the same Larmor frequency, rendering localization of activation impossible. To address this issue, spatial gradients along the slicing plane are put in place after the slice has been excited, leading to a variation of Larmor frequency across the slice. One can apply a linear gradient of different intensities in two directions along the slice. Measurement of the emitted RF signal for a given configuration of gradients yields a sum of signals over varying Larmor frequencies. When applying two linear gradients in perpendicular directions across the slice it is impossible to avoid identical Larmor frequencies on a set of parallel lines. These lines are in fact hyperplanes to the gradient intensity vector. Measuring the nuclear RF signal of a slice in many different constellations of linear gradient, usually on an equally spaced grid of gradient intensities, makes it possible to disambiguate the signal from specific spatial locations which all summed into these factors. The space of possible gradient intensity vectors is called *k-space*, and by virtue of the fact that the measured RF signal is a linear superposition of signal at different frequencies, the acquired k-space signal conveniently is equal to the Fourier transform of the spatial signal. By applying a simple inverse Fourier transform, the spatial structure of the slice can be recovered.

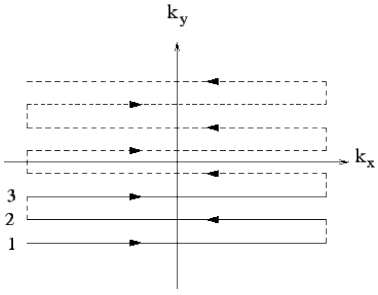

Figure 1.9: Example of a k-space trajectory for the measurement of an EPI image

- *T1* is the characteristic time scale on which excited protons fall back into the ground state, aligned with the homogeneous magnetic field. Also called *spin-lattice* decay, it indicates the time by which the longitudinal (z-axis) magnetization has decayed to $\exp(-1) \approx 31\%$ of its maximum

magnetization.

- After a so-called 90-degree-pulse, which excites around half the protons, leading to a net longitudinal magnetization of 0 and a synchronization in the xy-plane, due to material-intrinsic local field inhomogeneities the spins dephase, leading to an exponential decrease of the net xy-plane magnetization. *T2* is the characteristic time of this decay type.

- *T2** is similar to T2 decay but due to extrinsic magnetic field inhomogeneities, for example blood flow. This relaxation type gives rise to the BOLD signal.

Depending on the timing of the measurements and the gradient pulses applied, the measured image can be dominated by different types of signal decay. *T1-weighted* images are typically used for anatomical imaging and *T2-weighted* images are used to create the BOLD contrast.

Initially, typical acquisitions proceeded by exciting a slice via an RF pulse, placing a series of spatial gradients and possibly other RF pulses, followed by measurement of the emitted radio frequency signal after a certain time of evolution. Echo-planar imaging [Stehling et al., 1991] was introduced later and relies on the fact that several gradient positions can be measured with only one slice excitation, permitting a much faster sweep of k-space and an order of magnitude of acquisition speed improvement.

A typical MRI scanner consists of a superconducting hollow cylindrical magnetic coil carrying a magnetic field of between 1.5T and 7T parallel to the axis of the cylinder. For human MRI machines the cylinder is oriented horizontally. The axis of the cylinder is called the z-axis, elevation is the y axis and the remaining left-right axis is called x. The z-axis gradient is created by a pair of Helmholtz coils placed at either ends of the cylinder. The x- and y-gradients are created by pairs of half-cylinder coils acting in an opposing manner.

Figure 1.10: Clinical MRI machine

1.3 Functional MRI

Most modern functional MRI acquisitions rely upon the BOLD effect. It will be briefly introduced before a review of typical experimentation types done with this signal.

1.3.1 Blood Oxygen Level Dependent signal

In 1989, Ogawa and collaborators made a finding that should revolutionize the way brain function is measured [Ogawa and Lee, 1990]. They discovered hemoglobin as a natural contrast agent in T2*-weighted imaging. Since hemoglobin is directly involved in the oxygen supply within the body, this contrast can measure metabolic activity in the brain. Following neural activity and depletion of energy, the active region is supplied with fresh blood in a localized manner. Immediately following neural stimulation, the concentration of *deoxyhemoglobin* rises, followed by an onrush of more than necessary *oxyhemoglobin*. Deoxyhemoglobin is paramagnetic, whereas oxyhemoglobin is diamagnetic. These are different susceptibility properties which lead to

different behavior when these materials are placed in a magnetic field: diamagnetic materials give rise to a magnetic field which counteracts the one in which they are placed, leading to a local net-reduction of magnetic field. Paramagnets act the opposite way by contributing in the direction of an externally applied magnetic field.

Thus oxygen level as a metabolic indicator becomes visible to MRI. While the full mechanism of neuro-vascular coupling is still not fully understood, direct links to neural activity have been established [Logothetis et al., 2001].

Since however the BOLD response is a mixture of cerebral blood flow, cerebral blood volume and cerebral metabolic rate of oxygen changes, it is a relative measure without a baseline. In certain settings this necessitates the interpretation of a contrast of two images instead of the images themselves.

Apart from extreme cases, the BOLD response is approximately linear in the underlying neural activity. Given a train of stimulative or behavioral events it has also been put forward and tested that the BOLD response, within a certain range, acts like a *linear time-invariant system* (LTI) [Boynton et al., 1996]. As a consequence, the response of a given brain voxel to a function of neural activity and events is fully characterized by the convolution of the neural activity function with a causal finite impulse response function, generally named the *hemodynamic response function* (HRF).

With MRI scanners available in many hospitals and research institutions, the discovery of the BOLD signal gave cognitive scientist a relatively cheap, reliable and spatially well-resolved tool to examine brain function. The explosion in number of publications pertaining to fMRI studies confirms this [Poldrack et al., 2011a].

1.3.2 fMRI experiment types

Different ways of studying the brain with fMRI have been established and been given category names. A major difference in brain activity can be observed between the brain engaged in a specific task and the brain at rest. The latter type of experiment requires the subject to engage in nothing but possibly mind-wandering. This can be done eyes open or eyes closed without visual stimulation. A rich body of literature exists around resting state fMRI. A notable discovery using fMRI has been the default mode brain network, which seems to be strongly active during rest periods [Raichle et al., 2001].

The other type of experiment is task-related. Given a cognitive task or external stimulation, brain activity specific to it will be elicited and can be contrasted against appropriately chosen control conditions in which the studied brain function is presumably inactive.

Different types of experiments include the presentation of visual, auditory or sensory stimulation either in passive perception or with active tasks such as memory, attention, discrimination or decision tasks. Underlying the design of high-level cognitive tasks there is often a mechanism of testing one theory against another, by choosing situations in which they would yield different predictions.

An fMRI experiment with stimulation and/or tasks is typically set up either as an event-related or a block design. In an event-related design, neural activity is elicited in brief singular events which can be visualized on a graph

as spikes. They lead to a hemodynamic response consisting of the superposition of HRFs time-locked to the events (provided these are not spaced too closely, incurring a violation of the LTI model). In a block design, neural activity is elicited during extended periods of time, for example by showing different images of a same category. The stimulus may change in order to sustain neural activity, but the activation will be considered as one block and averaged. The neural activation functions can be visualized as boxcar functions (this is the case e.g. in [Haxby et al., 2001]). When convolved with the hemodynamic response function, they give rise to a shifted version of the boxcar with slightly smoothed edges.

1.3.3 Statistical methods for fMRI data analysis

Functional MRI is an imaging modality in which raw images are uninterpretable or at least very difficult to interpret by the human eye. Effect sizes are small with respect to baseline signal: BOLD signal changes typically amount to 5% of the mean image and the signal to noise ratio is low - typically around 0.1. This situation makes statistical analysis indispensible even for qualitative analysis as activation is indistinguishable from non-activation by looking at the raw signal.

Preprocessing

Before any analysis can be done on fMRI data, a number of preprocessing steps needs to be taken. A typical experimental acquisition session involves an anatomical scan, using a T1-weighted image. This anatomical image permits a segmentation into different tissue types, such as white matter and gray matter. Cerebro-spinal fluid (CSF) and skull or the rest of the head are also segmented. If coregistered with a functional acquisition, this permits the identification of gray-matter voxels in the functional images.

During functional scanning over a length of time, head movement is almost unavoidable. In *motion correction* or *realignment*, each acquired functional image is transformed to match a reference image (taken, e.g. from the middle of the acquisition), using strictly rigid body transforms, which can be parametrized by a translation and a 3D rotation. This assumption encodes the fact that we do not expect the brain to change shape or size during the acquisition. Since the alignment is done between images of the same type, one can expect that a well-aligned image should incur minimal ℓ_2 or correlation error. This is the cost function that is usually optimized in the transformation parameters.

An optional preprocessing step is slice-timing correction. In effect, due to the fact that the 3D images are acquired by slices, one by one, the time of the acquisition of each slice is different. Further, adjacent slices may not be acquired at adjacent points in time. In interleaved acquisition and in multiband acquisitions, this is not the case. It is straightforward to see that the slice timing delay relative to the neural event onset will cause a shift in the sampling of the hemodynamic response. If using methods that are rigid in their assumptions of the HRF, and brain volume acquisition takes more than TR=2s, then slice-timing correction may be a useful preprocessing step. It is done by temporal interpolation, where the type of interpolation needs to be

chosen. Temporal sinc interpolation yields the least biased results at the cost of needing more filter taps than e.g. a linear or quadratic interpolation.

It is also possible that fMRI acquisitions incur artefacts such as ghosting or spikes. Visual scrutiny or analysis using ICA or simple temporal differential analysis summaries can give indications as to the locations of spikes and corresponding volumes.

After these steps, which are often referred to as *minimal preprocessing*, we are ready to perform fMRI statistics, basing ourselves on the hypothesis or fact that now a given voxel refers to the same part of the brain throughout the analysis.

FMRI signal has several properties that need to be taken into account in order for analysis not to fail. First and foremost there are the low-frequency drifts which dominate the norm of the signal and have been essentially characterized as nuisance variables. Importantly, the information recoverable by an fMRI analysis resides in the high frequencies or must reside in the high frequencies, because otherwise it is confoundable with drifts and discarded when drifts are discarded. As a consequence, when designing an experiment, one must be careful not to include effects that are too slow. A typical cutoff for drift frequency is $1/128Hz$. Drifts can be removed by high-pass filtering through projecting onto high-frequency Fourier coefficients. One can also use global or local polynomials up to a certain degree, as they enforce slow variation. For local polynomial smoothing the Savitzky-Golay filter has turned out helpful. Drifts can be removed either before statistics or accommodated for in statistical estimation.

Another issue that needs to be taken into account is noise. Noise is generally so strong that it is impossible to see the BOLD-induced signal change in one image by eye. It must be included at least implicitly into any data analysis model put forward. One can choose a white Gaussian model, but an autoregressive model with a one-timepoint history has also been successfully employed.

GLM

In the *General Linear Model* (GLM), voxel activations due to experimental conditions are written as the noisy linear forward model

$$y = X\beta + \varepsilon,$$

where X represents in its columns the event or condition regressors, which, exploiting the LTI model, are indicators of neural activity convolved with the hemodynamic response. These different regressors are weighted by the entries of the β-vector and the noise vector is added. Assuming white Gaussian noise ε with zero mean and variance σ^2 and full column rank of X, the best unbiased estimator for β is

$$\hat{\beta} = X^+ y = \beta + X^+ \varepsilon.$$

The new noise term $X^+\varepsilon$ is still Gaussian with zero mean.

After performing the GLM, we are usually interested in establishing a relative difference measure between two or more conditions, in the form of a statistical contrast. Let $(e_i)_i$ be unit vectors and suppose we are interested in

contrasting condition i with condition j. Then with $c = e_i - e_j$, we would like to infer whether the null-hypothesis that $c^T\beta = 0$ can be rejected. We have

$$c^T\beta = c^T\hat{\beta} - c^T X^+ \varepsilon,$$

which is a Gaussian variable $\mathcal{N}(c^T\hat{\beta}, \sigma\sqrt{c^T X^+ X^{+,T} c}) = \mathcal{N}(c^T\hat{\beta}, \sigma\sqrt{c^T (X^T X)^{-1} c})$. It is then straightforward to determine the probability of reaching the mean of this distribution with a distribution of the same variance centered at zero. A succinct test statistic is $z = \frac{c^T\hat{\beta}}{\sqrt{c^T (X^T X)^{-1} c}\sigma}$, which is the z-score of the normal distribution.

It is to be noted that we normally do not have access to σ and have to estimate it in the model. This can be done by observing that an estimate can be obtained in the GLM residuals:

$$r = y - X\hat{\beta} = (\mathrm{Id} - XX^+)\varepsilon,$$

which leads to $\|r\|^2 = \varepsilon(\mathrm{Id} - XX^+)\varepsilon = (n-p)\hat{\sigma}^2$, where $X \in \mathbb{R}^{n\times p}$ and the $n - p$ scaling due to the loss in degrees of freedom incurred by the orthogonal projection. Using $\hat{\sigma}$ in the above equation

$$t = \frac{c^T\hat{\beta}}{\sqrt{c^T (X^T X)^{-1} c}\hat{\sigma}}$$

gives us a t-statistic on which we can perform the same inference.

Unsupervised methods

When there is no task, behavior or stimulation and the brain is scanned while resting or mind-wandering, there is still intriguing structure in the resulting signal. One method to obtain brain activation maps and their activations as time courses is *Independent Components Analysis* (ICA). This supposes that the signal is a linear combination A of underlying sources s, which are maximally statistically independent. In order to function, ICA requires at least as many samples as there are dimensions in the data. Since fMRI data are very high-dimensional and typically scarce in the sense that there are much more voxels per image than images, one resorts to discovering independent timecourses instead of independent activation maps. The matrix of timecourses has the correct shape proportions and the resulting ICA will show which latent timecourse components were active in which voxel. The map of each latent factor is usually spatially coherent even though by construction it contains no spatial information apart from a (crucial) smoothing before estimating the components. Another approach, this time with a focus on spatially contiguous and sparse activation maps is known as "TV-l1 multi-subject dictionary learning" (TV-MSDL), see [Abraham et al., 2013] for details. Both approaches give rise to clean maps than can be further segmented into regions if desired. A typical analysis performed on resting state data is the study of interactions between such regions. One can also obtain regions using an anatomical atlas. One can for example tell apart disease condition from normal condition by classifying the covariance matrices, where disease condition can be e.g. autism or schizophrenia.

Encoding and Decoding

Encoding and *Decoding*, as introduced in [Naselaris et al., 2011] describe a direction in which modeling is performed. *Encoding* models, also known as *forward* models, are aligned with the direction of causality as far as possible. In an fMRI experiment, this means that the brain response is predicted from the stimulus. The way it is advocated in [Naselaris et al., 2011] is to make use of simple linear models on top of an arbitrarily complicated and nonlinear representation of the stimulus. If an encoding model of this type can explain brain activity well, it is an indication of the usefulness of the underlying nonlinear representation. These models can advance the understanding of brain function. *Decoding* models, or *inverse* models, perform inference in the opposite direction: E.g. given a brain image, a decoding model attempts to infer information about the stimulus. Often the output is chosen to be categorical, e.g. an object category seen on the screen, but can also be continuous and potentially multi-dimensional. Uses of this modeling direction arise e.g. in brain-computer interfaces, where brain state is used to control a machine, or potentially for medical diagnosis for the prediction of a disease phenotype.

Ringing implicitly within the mention of *encoding* and *decoding* models is method of evaluation. While one may reasonably argue that an encoding model is nothing other than a potentially ill-posed way of performing a GLM, usually the evaluation criteria are quite different. While the classical GLM is amenable to classical statistics, due to the full column rank of the design matrix X, the evaluation of an encoding model is better seen as the evaluation of a modern machine learning method, which calculate predictive performance on unseen data. This way, even if the design matrix is singular due to an abundance of feature columns, modeling capacity can be quantified. Decoding models are evaluated analogously – also by an accuracy measure on held-out data.

In this thesis, the focus will be on encoding and decoding models.

2 Data-driven HRF estimation for encoding and decoding in fMRI

Despite the common usage of a canonical, data-independent, hemodynamic response function (HRF), it is known that the shape of the HRF varies across brain regions and subjects. This suggests that a data-driven estimation of this function could lead to more statistical power when modeling BOLD fMRI data. However, unconstrained estimation of the HRF can yield highly unstable results when the number of free parameters is large. We develop a method for the joint estimation of activation and HRF by means of a rank constraint, forcing the estimated HRF to be equal across events or experimental conditions, yet permitting it to differ across voxels. Model estimation leads to an optimization problem that we propose to solve with an efficient quasi-Newton method, exploiting fast gradient computations. This model, called GLM with Rank-1 constraint (R1-GLM), can be extended to the setting of GLM with separate designs which has been shown to improve decoding accuracy in brain activity decoding experiments. We compare 10 different HRF modeling methods in terms of encoding and decoding score on two different datasets. Our results show that the R1-GLM model outperforms competing methods in both encoding and decoding settings, positioning it as an attractive method both from the points of view of accuracy and computational efficiency.

In the next section, we provide an example motivating the study of HRF estimation techniques. The subsequent sections have been published in the Neuroimage journal.

Sections 2.2 to 2.7 have been published in

- F. Pedregosa, M. Eickenberg, P. Ciuciu, B. Thirion, A. Gramfort *"Data-driven HRF estimation for encoding and decoding models"*, NeuroImage, Volume 104, 1 January 2015, Pages 209-220.
- F. Pedregosa, M. Eickenberg, B. Thirion, and A. Gramfort, *"HRF estimation improves sensitivity of fMRI encoding and decoding models"*, Proc. 3rd International Workshop Pattern Recognition in NeuroImaging, 2013

2.1 Motivating Example

The BOLD hemodynamic response to a stimulus is a complicated mechanism, dependent on the oxygen consumption following energy release due to neural activity, but also mechanical properties of blood flow and the blood vessels by which it is transported. It is thus somewhat surprising that linear time invariant systems modeling does capture the BOLD response quite well, provided that certain conditions on the inter-stimulus interval are met. This property is studied in [Boynton et al., 1996]. However, when consecutive stimuli are placed too close together temporally, at e.g less than 2 seconds, then the system does not satisfy the superposition property. This can be seen e.g. by considering a higher order Volterra expansion of the hemodynamic response: In the quadratic term one observes nontrivial binary interactions when stimuli are very close [Friston et al., 2000].

In this chapter we focus on the modeling of BOLD response in the framework of a linear time-invariant system only, e.g. sytems equal to their own Volterra expansion of first order, where we assume stimulation as impulse-like input and BOLD signal is the filtered response (the convolution with the hemodynamic impulse response function). In this context it is crucial to be able to characterize the impulse response of the system since otherwise the estimation of activity can be completely misguided.

In figure 2.1 we can see a depiction of two impulse sequences describing stimulus events for two experimental conditions. Both stimulus event trains yield a hemodynamic response whose superposition yields the full BOLD response. If this signal is analyzed with the "wrong" impulse response function (peak shifted from 6 seconds to 4 seconds), then the estimated activations can become very wrong. In this specific case they do not even preserve order.

Slightly more formally, we can write the event sequences as trains of Dirac deltas

$$E_m(t) = \sum_{n=1}^{N_m} \delta(t - t_{m,n}), \quad m \in \{1, 2\}, \tag{2.1}$$

where m represents different experimental conditions and the $t_{m,n}$ indicate the event times for condition m.

Given an HRF $h(t)$ which is assumed to have finite temporal support $[0, L_h]$, the regressors used in a GLM are then the convolution of the E_m event trains with the HRF:

$$X_m(t) = E_m * h(t) = \sum_{n=1}^{N_m} h(t - t_{m,n}) \tag{2.2}$$

The BOLD signal in one voxel is then modeled as a linear combination of these regressors:

$$y(t) = \sum_{m=1}^{M} \beta_m X_m(t) \tag{2.3}$$

Writing $\langle f, g \rangle = \int_{-\infty}^{\infty} f(t)g(t)\mathrm{d}t$ and $\|f\|^2 = \langle f, f \rangle$ (for finite numbers of events these integrals clearly exist), given a BOLD signal $y(t)$ the least squares estimate is written

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2} \|y(t) - \sum_{m=1}^{M} \beta_m X_m(t)\|^2 \tag{2.4}$$

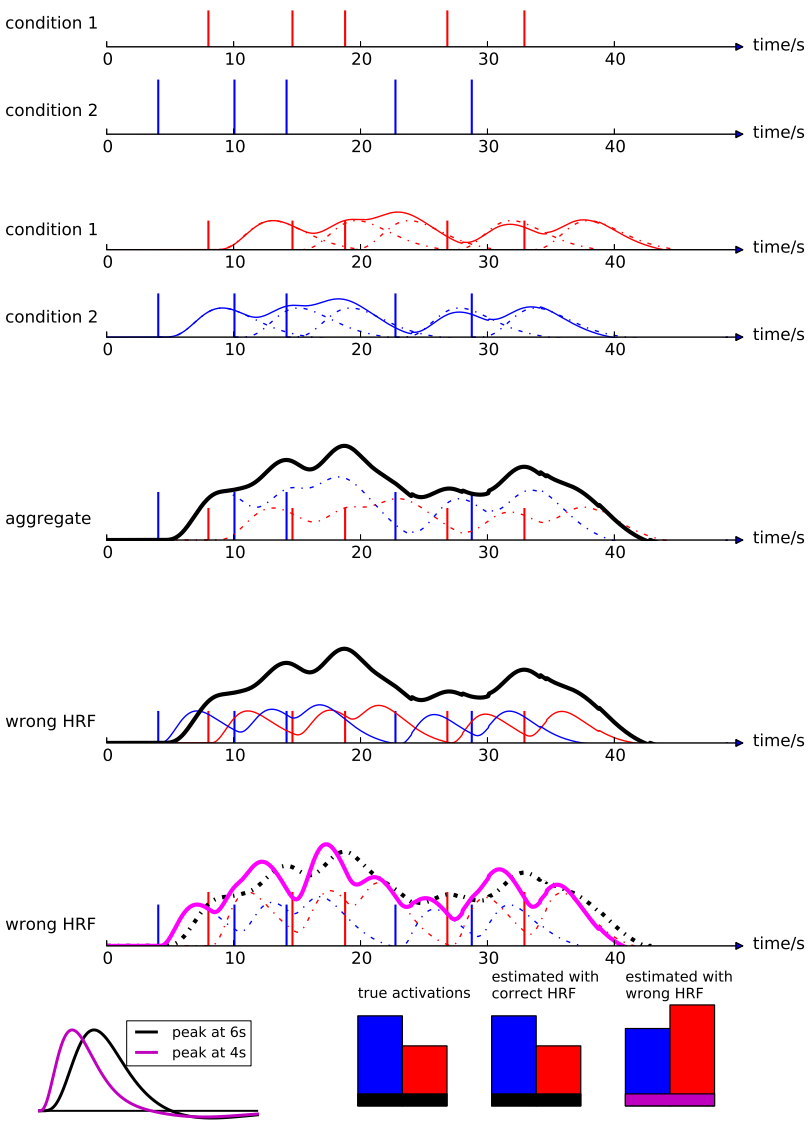

Figure 2.1: Time series of events convolved with an HRF. If a different HRF is used for activation estimation, then activation differences can flip signs. At the top we see event trains for two different conditions, which are differently activated (cond blue $>$ cond red) for a chosen voxel. The second set of plots shows the event trains convolved with a hemodynamic response function peaking at 6s. Dotted lines show responses of individual events. The third plot shows the total activity of the voxel due to the events (black line). The fourth plot shows the total activity and event responses using an HRF that peaks at 4s ("the wrong HRF"). The fifth plot shows in magenta the best fit obtainable with the HRF peaking at 4s. The last line shows that the estimated activation maps for condition blue $<$ condition red, inverting the order of the two.

Letting $G_{l,m} = \langle X_l, X_m \rangle$ be the Gram matrix and $c_m = \langle X_m, y \rangle$ be the inner product similarity between regressors and BOLD time course, the solution to the least squares problem can be written as

$$\hat{\beta} = G^{-1}c \tag{2.5}$$

Assuming now that we have two conditions and that the BOLD activity was generated using a "ground truth" HRF $h(t)$, we assess what happens if the activity is estimated using a different underlying HRF $g(t)$. Let

$$\begin{aligned} X_m^h(t) &= E_m * h(t) \\ X_m^g(t) &= E_m * g(t) \\ G_{l,m}^{g,g} &= \langle X_l^g, X_m^g \rangle \\ G_{l,m}^{g,h} &= \langle X_l^g, X_m^h \rangle \end{aligned}$$

and the BOLD signal generated as $y(t) = \sum_m \beta_m X_m^h(t)$. Then using the HRF $g(t)$ leads to the following estimation of activity:

$$\begin{pmatrix} \hat{\beta}_1^g \\ \hat{\beta}_2^g \end{pmatrix} = (G^{g,g})^{-1} G^{g,h} \beta = \begin{pmatrix} \langle X_1^g, X_1^g \rangle, \langle X_1^g, X_2^g \rangle \\ \langle X_2^g, X_1^g \rangle, \langle X_2^g, X_2^g \rangle \end{pmatrix}^{-1} \begin{pmatrix} \langle X_1^g, X_1^h \rangle, \langle X_1^g, X_2^h \rangle \\ \langle X_2^g, X_1^h \rangle, \langle X_2^g, X_2^h \rangle \end{pmatrix} \beta \tag{2.6}$$

In order to evaluate this estimation, we need to take a closer look at the scalar products involved. We exploit the fact that these can be written as a convolution evaluated at 0 and can then use associativity and commutativity properties of the convolution. Using the notation $\check{f} : x \mapsto f(-x)$ we can write:

$$\langle X_m^g, X_l^h \rangle = X_m^g * \check{X}_l^h(0) = (E_m * g) * (\check{E}_l * \check{h})(0) = (E_m * \check{E}_l) * (g * \check{h})(0) \tag{2.7}$$

The rule for the convolution of Diracs gives us $(E_m * \check{E}_l)(t) = \sum_{n,k} \delta(t - (t_{m,n} - t_{l,k}))$. Since the support of $g * \check{h}$ is $[-L_h, L_g]$, if the events are spaced at a larger inter-stimulus interval than $\max(L_g, L_h)$, the scalar product reduces to $\langle X_m^g, X_l^h \rangle = N_m \delta_{ml} (g * \check{h})(0) = N_m \delta_{ml} \langle g, h \rangle$. The estimated activations then become

$$\hat{\beta}_m = \frac{\langle g, h \rangle}{\langle g, g \rangle} \beta_m, \tag{2.8}$$

If the hemodynamic responses to events do not significantly overlap (i.e. the events are sufficiently temporally separated), then using the wrong HRF for estimation merely leads to the activation maps being scaled by a factor.

and we conclude that using the "wrong" hrf in the absence of response overlap merely results in a rescaling of activation maps. In the context of two different event types, let us assume that event 2 follows event 1 after half of the duration of the hemodynamic response and that event 1 occurs periodically with inter-stimulus interval equal to the length of the HRF. Then, with the shorthand $\langle g, h \rangle_t = (g * \check{h})(t)$, we obtain

$$\begin{aligned} \langle X_1^g, X_2^h \rangle = \langle X_2^g, X_1^h \rangle &= N \langle g, h \rangle_{\frac{L}{2}} \\ \langle X_1^g, X_1^h \rangle = \langle X_2^g, X_2^h \rangle &= N \langle g, h \rangle_0 \end{aligned}$$

We thus obtain

$$\begin{pmatrix} \hat{\beta}_1^g \\ \hat{\beta}_2^g \end{pmatrix} = \begin{pmatrix} \langle g, g \rangle_0, \langle g, g \rangle_{\frac{L}{2}} \\ \langle g, g \rangle_{\frac{L}{2}}, \langle g, g \rangle_0 \end{pmatrix}^{-1} \begin{pmatrix} \langle g, h \rangle_0, \langle g, h \rangle_{\frac{L}{2}} \\ \langle g, h \rangle_{\frac{L}{2}}, \langle g, h \rangle_0 \end{pmatrix} \beta$$

Assume for simplicity that the HRFs g, h are step functions, for example $g = \sqrt{\frac{2}{L}}\mathbb{1}_{[\frac{L}{2},L]}$ and $h = \sqrt{\frac{2}{L}}\mathbb{1}_{[0,\frac{L}{2}]}$. In this case $\langle g,g\rangle = \langle h,h\rangle = \langle g,h\rangle_{\frac{L}{2}} = 1$ and the other values are equal to 0, leading to

$$\begin{pmatrix} \hat{\beta}_1^g \\ \hat{\beta}_2^g \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \beta = \begin{pmatrix} \beta_2 \\ \beta_1 \end{pmatrix},$$

If the HRF used for estimation is radically different from the HRF generating the signal, and the events are unfortunately placed, then activation contrasts can flip sign. In the constructed example here, the two conditions exchange activation maps.

leading to an exact switching of activations. In practice the effect may not be as clear cut, but figure 2.1 shows an example with plausible HRFs, where the order of the strengths of the weights is inverted.

In the following, we will make the case for an HRF estimation per voxel in the context of encoding and decoding models. Indeed, most efforts of decoding brain state from fMRI data use a deconvolution step in the form of an event related GLM in order to extract the activation coefficients β, instead of learning predictive models directly on BOLD signal (some exceptions exist and will be mentioned). We will show that the estimation of the hemodynamic response function per voxel aids both forward and reverse modeling techniques, from stimulus to brain activity and back.

2.2 Data-driven HRF estimation for encoding and decoding models

The use of machine learning techniques to predict the cognitive state of a subject from their functional MRI (fMRI) data recorded during task performance has become a popular analysis approach for neuroimaging studies over the last decade [Cox and Savoy, 2003, Haynes and Rees, 2006]. It is now commonly referred to as *brain reading* or *decoding*. In this setting, the BOLD signal is used to predict the task or stimulus that the subject was performing. Although it is possible to perform decoding directly on raw BOLD signal [Mourão Miranda et al., 2007, Miyawaki et al., 2008], the common approach in fast event-related designs consists in extracting the activation coefficients (beta-maps) from the BOLD signal to perform the decoding analysis on these estimates. Similarly, in the voxel-based *encoding* models [Kay et al., 2008, Naselaris et al., 2011], the activation coefficients are extracted from the BOLD signal, this time to learn a model to predict the BOLD response in a given voxel, based on a given representation of the stimuli. In addition, a third approach, known as representational similarity analysis or RSA [Kriegeskorte et al., 2008a] takes as input the activation coefficients. In this case a comparison is made between the similarity observed in the activation coefficients, quantified by a correlation measure, and the similarity between the stimuli, quantified by a similarity measure defined from the experimental setting.

These activation coefficients are computed by means of the General Linear Model (GLM) [Friston et al., 1995]. While this approach has been successfully used in a wide range of studies, it does suffer from limitations [Poline and Brett, 2012]. For instance, the GLM commonly relies on a data-independent *canonical* form of the hemodynamic response function (HRF) to estimate the activation coefficient. However it is known [Handwerker et al., 2004, Badillo et al., 2013b] that the shape of this response function can vary substantially

across subjects and brain regions. This suggests that an adaptive modeling of this response function should improve the accuracy of subsequent analysis.

To overcome the aforementioned limitation, Finite Impulse Response (FIR) models have been proposed within the GLM framework [Dale, 1999, Glover, 1999]. These models do not assume any particular shape for the HRF and amount to estimating a large number of parameters in order to identify it. While the FIR-based modeling makes it possible to estimate the activation coefficient and the HRF simultaneously, the increased flexibility has a cost. The estimator is less robust and prone to overfitting, i.e. it may generalize badly to unseen data. In general, FIR models are most appropriate for studies focused on the characterization of the shape of the hemodynamic response, and not for studies that are primarily focused on detecting activation [Poldrack et al., 2011b]

Several strategies aiming at reducing the number of degrees of freedom of the FIR model - and thus at limiting the risk of overfitting - have been proposed. One possibility is to constrain the shape of the HRF to be a linear combination of a small number of basis functions. A common choice of basis is formed by three elements consisting of a reference HRF as well as its time and dispersion derivatives [Friston et al., 1998], although it is also possible to compute a basis set that spans a desired function space [Woolrich et al., 2004]. More generally, one can also define a parametric model of the HRF and estimate the parameters that best fit this function [Lindquist and Wager, 2007]. However, in this case the estimated HRF may no longer be a linear function of the input parameters.

Sensitivity to noise and overfitting can also be reduced through regularization. For example, temporal regularization has been used in the smooth FIR [Goutte et al., 2000, Ciuciu et al., 2003, Casanova et al., 2008] to favor solutions with small second order time derivative. These approaches require the setting of one or several hyperparameters, at the voxel or potentially at the parcel level (if several voxels in a pre-defined parcel are assumed to share some aspects of the HRF timecourse). Even if efficient techniques such as generalized cross-validation [Golub et al., 1979] can be used to choose the regularization parameters, these methods are inherently more costly than basis-constrained methods. Basis-constrained methods also require setting the number of basis elements; however, this parameter is not continuous (as in the case of regularized methods), and in practice only few values are explored: for example the 3-element basis set formed by a reference HRF plus derivatives and the FIR model. This paper focuses on basis-constrained regularization of the HRF to avoid dealing with hyperparameter selection with the goal of remaining computationally attractive. A different approach to increase robustness of the estimates consists in linking the estimated HRFs across a predefined brain parcel, taking advantage of the spatially dependent nature of fMRI [Wang et al., 2013]. However, hemodynamically-informed parcellations [Chaari et al., 2012, Badillo et al., 2013a] rely on the computation of a large number of estimations at the voxel or sub-parcel level. In this chapter we focus on voxel-wise estimation methods.

We propose a method for the simultaneous estimation of HRF and activation coefficients based on low-rank modeling. Within this model, and as in [Makni et al., 2008, Kay et al., 2008, Vincent et al., 2010, Degras and

Lindquist, 2014], the HRF is constrained to be equal across the different conditions, yet permitting it to be different across voxels. Unlike previous works, we formulate this model as a constrained least squares problem, where the vector of coefficients is constrained to lie within the space of rank one matrices. We formulate the model within the framework of smooth optimization and use quasi-Newton methods to find the vector of estimates. This model was briefly presented in the conference paper [Pedregosa et al., 2013]. Here we provide more experimental validation and a more detailed presentation of the method. We also added results using a GLM with separate designs [Mumford et al., 2012]. Ten alternative approaches are now compared on two publicly available datasets. The solver has also been significantly improved to scale to full brain data.

The contributions of this chapter are two-fold. First, we quantify the importance of HRF estimation in encoding and decoding models. While the benefit of data-driven estimates of the HRF have already been reported in the case of decoding [Turner et al., 2012] and encoding approaches [Vu et al., 2011], we here provide a comprehensive comparison of models. Second, we evaluate a method called *GLM with Rank-1 constraint (R1-GLM)* that improves encoding and decoding scores over state-of-the-art methods while remaining computationally tractable on a full brain volume. We propose an efficient algorithm for this method and discuss practical issues such as initialization. Finally, we provide access to an open source software implementation of the methods discussed in this chapter.

Notation: $\|\cdot\|$ and $\|\cdot\|_\infty$ denote the Euclidean and infinity norm for vectors. We use lowercase boldface letter to denote vectors and uppercase boldface letter to denote matrices. $\mathbf{I}$ denotes the identity matrix, $\mathbf{1}_n$ denotes the vector of ones of size n, $\otimes$ denotes the Kronecker product and $\mathrm{vec}(\mathbf{A})$ denotes the concatenation of the columns of a matrix $\mathbf{A}$ into a single column vector. $\mathbf{A}^\dagger$ denotes the Moore-Penrose pseudoinverse. Given the vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ with $\mathbf{a}_i \in \mathbb{R}^n$ for each $1 \leq i \leq k$, we will use the notation $[\mathbf{a}_1, \ldots, \mathbf{a}_k] \in \mathbb{R}^{n \times k}$ to represent the columnwise concatenation of the k vectors into a matrix of size $n \times k$. We will use Matlab-style colon notation to denote slices of an array, that is $\mathbf{x}(1:k)$ will denote the first k elements of $\mathbf{x}$.

2.3 Methods

In this section we describe different methods for extracting the HRF and activation coefficients from BOLD signals. We will refer to each different stimulus as *condition* and we will call *trial* a unique presentation of a given stimulus. We will denote by k the total number of stimuli, $\mathbf{y} \in \mathbb{R}^n$ the BOLD signal at a single voxel and n the total number of images acquired.

The General Linear Model

The original GLM model [Friston et al., 1995] makes the assumption that the hemodynamic response is a linear transformation of the underlying neuronal signal. We define the $n \times k$-matrix $\mathbf{X}_{\mathrm{GLM}}$ as the columnwise stacking of different regressors, each one defined as the convolution of a reference

HRF [Boynton et al., 1996, Glover, 1999] with the stimulus onsets for the given condition. In this work we used as reference HRF the one provided by the software SPM 8 [Friston et al., 2011]. Assuming additive white noise, $n \geq k$ and $\mathbf{X}_{\text{GLM}}$ to be full rank, the vector of estimates is given by $\hat{\boldsymbol{\beta}}_{\text{GLM}} = \mathbf{X}_{\text{GLM}}^{\dagger}\mathbf{y}$, where $\hat{\boldsymbol{\beta}}_{\text{GLM}}$ is a vector of size k representing the amplitude of each one of the conditions in a given voxel.

A popular modification of this setting consists in extending the GLM design matrix with the temporal and width derivatives of the reference HRF. This basis, formed by the reference HRF and its derivatives with respect to time and width parameters, will be used throughout this work. We will refer to it as the *3HRF basis*. In this case, each one of the basis elements is convolved with the stimulus onsets of each condition, obtaining a design matrix of size $n \times 3k$. This way, for each condition, we estimate the form of the HRF as a sum of basis functions that correspond to the first order Taylor expansion of the parametrization of the response function. Another basis set that will be used is the Finite Impulse Response (FIR) set. This basis set spans the complete vector space of dimension corresponding to the length of the impulse response and it is thus a flexible model for capturing the HRF shape. It consists of the canonical unit vectors for the given duration of the estimated HRF. Other basis functions such as FMRIB's Linear Optimal Basis Sets [Woolrich et al., 2004] are equally possible but were not considered in this work.

More generally, one can extend this approach to any set of basis functions. Given the matrix formed by the stacking of d basis elements $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_d]$, the design matrix $\mathbf{X}_{\mathbf{B}}$ is formed by successively stacking the regressors obtained by convolving each of the basis elements with the stimulus onsets of each condition. This results in a matrix of size $n \times dk$ and under the aforementioned conditions the vector of estimates is given by $\hat{\boldsymbol{\beta}}_{\mathbf{B}} = \mathbf{X}_{\mathbf{B}}^{\dagger}\mathbf{y}$. In this case, $\hat{\boldsymbol{\beta}}_{\mathbf{B}}$ is no longer a vector of size k: it has length $k \times d$ instead and can no longer be interpreted as the amplitude of the activation. One possibility to recover the trial-by-trial response amplitude is to select the parameters from a single time point as done by some of the models considered in [Mumford et al., 2012], however this procedure assumes that the peak BOLD response is located at that time point. Another possibility is to construct the estimated HRF and take as amplitude coefficient the peak amplitude of this estimated HRF. This is the approach that we have used in this paper.

GLM with rank constraint

In the basis-constrained GLM model, the HRF estimation is performed independently for each condition. This method works reliably whenever the number of conditions is small, but in experimental designs with a large number of conditions it performs poorly due to the limited conditioning of the problem and the increasing variance of the estimates.

At a given voxel, it is expected that for similar stimuli the estimated HRF are also similar [Henson et al., 2002]. Hence, a natural idea is to promote a common HRF across the various stimuli (given that they are sufficiently similar), which should result in more robust estimates [Makni et al., 2008, Vincent et al., 2010]. In this work we consider a model in which a common

HRF is shared across the different stimuli. Besides the estimation of the HRF, a unique coefficient is obtained per column of our event matrix. This amounts to the estimation of $k + d$ free parameters instead of $k \times d$ as in the standard basis-constrained GLM setting.

The novelty of our method stems from the observation that the formulation of the GLM model with a common HRF across conditions translates to a rank constraint on the vector of estimates. This assumption amounts to enforcing the vector of estimates to be of the form $\boldsymbol{\beta}_{\mathbf{B}} = [\mathbf{h}\beta_1, \mathbf{h}\beta_2, \cdots, \mathbf{h}\beta_k]$ for some HRF $\mathbf{h} \in \mathbb{R}^d$ and a vector of coefficients $\boldsymbol{\beta} \in \mathbb{R}^k$. More compactly, this can be written as $\boldsymbol{\beta}_{\mathbf{B}} = \mathrm{vec}(\mathbf{h}\boldsymbol{\beta}^T)$. This can be seen as a constraint on the vector of coefficients to be the vectorization of a rank-one matrix, hence the name *Rank-1 GLM (R1-GLM)*.

In this model, the coefficients no longer have a closed form expressions, but can be estimated by minimizing the mean squared error of a bilinear model. Given $\mathbf{X}_{\mathbf{B}}$ and $\mathbf{y}$ as before, $\mathbf{Z} \in \mathbb{R}^{n \times q}$ a matrix of nuisance parameters such as drift regressors, we define $F_{\mathrm{R1}}(\mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{X}_{\mathbf{B}}, \mathbf{y}, \mathbf{Z}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}_{\mathbf{B}}\,\mathrm{vec}(\mathbf{h}\boldsymbol{\beta}^T) - \mathbf{Z}\boldsymbol{\omega}\|^2$ to be the objective function to be minimized. The optimization problem reads:

$$\begin{aligned} \hat{\mathbf{h}},\ \hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\omega}} \ = \ & \arg\min_{\mathbf{h},\boldsymbol{\beta},\boldsymbol{\omega}} \ F_{\mathrm{R1}}(\mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{X}_{\mathbf{B}}, \mathbf{y}, \mathbf{Z}) \\ & \text{subject to } \|\mathbf{B}\mathbf{h}\|_\infty = 1 \text{ and } \langle \mathbf{B}\mathbf{h}, \mathbf{h}_{\mathrm{ref}} \rangle > 0 \ , \end{aligned} \tag{2.9}$$

The norm constraint is added to avoid the scale ambiguity between $\mathbf{h}$ and $\boldsymbol{\beta}$ and the sign is chosen so that the estimated HRF correlates positively with a given reference HRF $\mathbf{h}_{\mathrm{ref}}$. Otherwise the signs of the HRF and $\boldsymbol{\beta}$ can be simultaneously flipped without changing the value of the cost function. Omitting the norm constraint, which is always obtainable by appropriate rescaling, the optimization problem is *smooth* and is convex with respect to $\mathbf{h}$, $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$, however it is not *jointly convex* in variables $\mathbf{h}$, $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$.

From a practical point of view this formulation has a number of advantages. First, in contrast with the GLM without rank-1 constraint the estimated coefficients are already factored into the estimated HRF and the activation coefficients. That is, once the estimation of the model parameters from Eq. (2.9) is obtained, $\hat{\boldsymbol{\beta}}$ is a vector of size k and $\hat{\mathbf{h}}$ is a vector of size d that can be both used in subsequent analysis, while in models without rank-1 constraint only the vector of coefficients (equivalent to $\mathrm{vec}(\mathbf{h}\boldsymbol{\beta}^T)$ in rank-1 constrained models) of size $k \times d$ is estimated. In the latter case, the estimated HRF and the beta-maps still have to be extracted from this vector by methods such as normalization by the peak of the HRF, averaging or projecting to the set of Rank-1 matrices.

Second, it is readily adapted to prediction on unseen trials. While for classical (non rank-1 models) the HRF estimation is performed per condition with no HRF associated with unseen conditions, in this setting, because the estimated HRF is linked and equal across conditions it is natural to use this estimate on unseen conditions. This setting occurs often in encoding models where prediction on unseen trials is part of the cross-validation procedure.

This model can also be extended to a parametric HRF model. That is, given the hemodynamic response defined as a function $h : \mathbb{R}^{d_1} \to \mathbb{R}^d$ of some parameters $\boldsymbol{\alpha}$, we can formulate the analogous model of Eq. (2.9) as an optimization over the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with the design matrix $\mathbf{X}_{\mathrm{FIR}}$ given

by the convolution of the event matrix with the FIR basis:

$$\begin{aligned}\hat{\boldsymbol{\alpha}},\ \hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\omega}} \ &= \ \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\omega}} F_{\text{R1}}(h(\boldsymbol{\alpha}),\boldsymbol{\beta},\boldsymbol{\omega},\mathbf{X}_{\text{FIR}},\mathbf{y},\mathbf{Z}) \\ &\quad \text{subject to } \|h(\boldsymbol{\alpha})\|_\infty = 1 \text{ and } \langle h(\boldsymbol{\alpha}),\mathbf{h}_{\text{ref}}\rangle > 0\end{aligned} \tag{2.10}$$

In section 2.3 we will discuss optimization strategies for both models.

Extension to separate designs

An extension to the classical GLM that improves the estimation with correlated designs was proposed in [Mumford et al., 2012]. In this setting, each voxel is modeled as a linear combination of two regressors in a design matrix $\mathbf{X}_{\text{GLM}}$. The first one is the regressor associated with a given condition and the second one is the sum of all other regressors. This results in k design matrices, one for each condition. The estimate for a given condition is given by the first element in the two-dimensional array ${\mathbf{X}_{Si}}^{\dagger}\mathbf{y}$, where $\mathbf{X}_{Si}$ is the design matrix for condition i. We will denote this model GLM with separate designs (GLMS). It has been reported to find a better estimate in rapid event designs leading to a boost in accuracy for decoding tasks [Mumford et al., 2012, Schoenmakers et al., 2013, Lei et al., 2013].

This approach was further extended in [Turner et al., 2012] to include the FIR basis instead of the predefined canonical function. Here we employ it in the more general setting of a predefined basis set. Given a set of basis functions we construct the design matrix for condition i as the columnwise concatenation of two matrices $\mathbf{X}^0_{\text{BS}i}$ and $\mathbf{X}^1_{\text{BS}i}$. $\mathbf{X}^0_{\text{BS}i}$ is given by the columns associated with the current condition in the GLM matrix and $\mathbf{X}^1_{\text{BS}i}$ is the sum of all other columns. In this case, the vector of estimates is given by the first d vectors of $\mathbf{X}^{\dagger}_{\text{BS}i}\mathbf{y}$. See [Turner et al., 2012] for a more complete description of the matrices $\mathbf{X}^0_{\text{BS}i}$ and $\mathbf{X}^1_{\text{BS}i}$.

It is possible to use the same rank-1 constraint as before in the setting of separate designs, linking the HRF across conditions. We will refer to this model as *Rank-1 GLM with separate designs (R1-GLMS)*. In this case the objective function has the form $F_{\text{R1-S}}(\mathbf{h},\boldsymbol{\beta},\boldsymbol{\omega},\mathbf{r},\mathbf{X}_{\mathbf{B}},\mathbf{y},\mathbf{Z}) = \frac{1}{2}\sum_{i=1}^{k}\|\mathbf{y} - \beta_i\mathbf{X}^0_{\text{BS}i}\mathbf{h} - r_i\mathbf{X}^1_{\text{BS}i}\mathbf{h} - \mathbf{Z}\boldsymbol{\omega}\|^2$, where $\mathbf{r}\in\mathbb{R}^d$ is a vector representing the activation of all events except the event of interest and will not be used in subsequent analyses. We can compute the vector of estimates $\hat{\boldsymbol{\beta}}$ as the solution to the optimization problem

$$\begin{aligned}\hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\omega}},\ \hat{\mathbf{h}},\ \hat{\mathbf{r}} \ &= \ \arg\min_{\mathbf{h},\boldsymbol{\beta},\boldsymbol{\omega},\mathbf{r}} F_{\text{R1-S}}(\mathbf{h},\boldsymbol{\beta},\boldsymbol{\omega},\mathbf{r},\mathbf{X}_{\mathbf{B}},\mathbf{y},\mathbf{Z}) \\ &\quad \text{subject to } \|\mathbf{B}\mathbf{h}\|_\infty = 1 \text{ and } \langle \mathbf{B}\mathbf{h},\mathbf{h}_{\text{ref}}\rangle > 0\end{aligned} \tag{2.11}$$

Optimization

For the estimation of rank-1 models on a full brain volume, a model is estimated at each voxel separately. Since a typical brain volume contains more than 40,000 voxels, the efficiency of the estimation at a single voxel is of great importance. In this section we will detail an efficient procedure based on quasi-Newton methods for the estimation of R1-GLM and R1-GLMS models on a given voxel.

One approach to minimize (2.9) is to alternate the minimization with respect to the variables $\boldsymbol{\beta}$, $\mathbf{h}$ and $\boldsymbol{\omega}$. By recalling the Kronecker product iden-

tities [Horn and Johnson, 1991], and using the identity $\text{vec}(\mathbf{h}\boldsymbol{\beta}^T) = \boldsymbol{\beta} \otimes \mathbf{h}$ we can rewrite the objective function (2.9) to be minimized as:

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X_B}(\boldsymbol{\beta} \otimes \mathbf{h}) - \mathbf{Z}\boldsymbol{\omega}\|^2 = \tag{2.12}$$

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X_B}(\mathbf{I} \otimes \mathbf{h})\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\omega}\|^2 = \tag{2.13}$$

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X_B}(\boldsymbol{\beta} \otimes \mathbf{I})\mathbf{h} - \mathbf{Z}\boldsymbol{\omega}\|^2 \ . \tag{2.14}$$

Updating $\mathbf{h}$, $\boldsymbol{\beta}$ or $\boldsymbol{\omega}$ sequentially thus amounts to solving a (constrained) least squares problem at each iteration. A similar procedure is detailed in [Degras and Lindquist, 2014]. However, this approach requires computing the matrices $\mathbf{X_B}(\boldsymbol{\beta} \otimes \mathbf{I})$ and $\mathbf{X_B}(\mathbf{I} \otimes \mathbf{h})$ at each iteration, which are typically dense, resulting in a high computational cost per iteration. Note also that the optimization problem is not jointly convex in variables $\mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\omega}$, therefore we cannot apply convergence guarantees from convex analysis.

We rather propose a more efficient approach by optimizing both variables jointly. We define a global variable $\mathbf{z}$ as the concatenation of $(\mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\omega})$ into a single vector, $\mathbf{z} = \text{vec}([\mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\omega}])$, and cast the problem as an optimization with respect to this new variable. Generic solvers for numerical optimization [Nocedal and Wright, 2006] can then be used. The solvers that we will consider take as input an objective function and its gradient. In this case, the partial derivatives with respect to variable $\mathbf{z}$ can be written as $\partial F_{\text{R1}}/\partial \mathbf{z} = \text{vec}([\partial F_{\text{R1}}/\partial \mathbf{h}, \partial F_{\text{R1}}/\partial \boldsymbol{\beta}, \partial F_{\text{R1}}/\partial \boldsymbol{\omega}])$, whose expression can be easily derived using the aforementioned Kronecker product identities:

$$\begin{cases} \dfrac{\partial F_{\text{R1}}}{\partial \mathbf{h}} = -(\boldsymbol{\beta}^T \otimes \mathbf{I})\mathbf{X}^T(\mathbf{y} - \mathbf{X}\,\text{vec}(\mathbf{h}\boldsymbol{\beta}^T) - \mathbf{Z}\boldsymbol{\omega}) \\ \dfrac{\partial F_{\text{R1}}}{\partial \boldsymbol{\beta}} = -(\mathbf{I} \otimes \mathbf{h}^T)\mathbf{X}^T(\mathbf{y} - \mathbf{X}\,\text{vec}(\mathbf{h}\boldsymbol{\beta}^T) - \mathbf{Z}\boldsymbol{\omega}) \\ \dfrac{\partial F_{\text{R1}}}{\partial \boldsymbol{\omega}} = -\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\,\text{vec}(\mathbf{h}\boldsymbol{\beta}^T) - \mathbf{Z}\boldsymbol{\omega}) \end{cases}$$

If instead a parametric model of the HRF is used as in Eq. (2.10), the equivalent partial derivatives can be easily computed by the chain rule.

For the sake of efficiency, it is essential to avoid evaluating the Kronecker products naively, but rather reformulate them using the above mentioned Kronecker identities. For example, the matrix $\mathbf{M} = \mathbf{X}(\mathbf{I} \otimes \mathbf{h})$ should not be computed explicitly but should rather be stored as a linear operator such that when applied to a vector $\boldsymbol{\beta} \in \mathbb{R}^k$ it computes $M(\boldsymbol{\beta}) = \mathbf{X}(\boldsymbol{\beta} \otimes \mathbf{h})$, avoiding thus the explicit computation of $\mathbf{I} \otimes \mathbf{h}$.

Similar equations can be derived for the rank-1 model with separate designs of Eq. (2.11) (R1-GLMS), in which case the variable $\mathbf{z}$ is defined as the concatenation of $(\mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{r})$, i.e. $\mathbf{z} = \text{vec}([\mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{r}])$. The gradient of $F_{\text{R1-S}}$ with respect to $\mathbf{z}$ can be computed as

$$\partial F_{\text{R1-S}}/\partial \mathbf{z} = \text{vec}([\partial F_{\text{R1-S}}/\partial \mathbf{h}, \partial F_{\text{R1-S}}/\partial \boldsymbol{\beta}, \partial F_{\text{R1-S}}/\partial \boldsymbol{\omega}, F_{\text{R1-S}}/\partial \mathbf{r}]).$$

The partial derivatives read:

$$\begin{cases} \frac{\partial F}{\partial \mathbf{h}} &= \sum_i^k -(\mathbf{X}^0_{\text{BS}_i}\boldsymbol{\beta}_i - \mathbf{X}^1_{\text{BS}_i} r_i)^T(\mathbf{y} - \boldsymbol{\beta}_i\mathbf{X}^0_{\text{BS}_i} h - w_i\mathbf{X}^1_{\text{BS}_i}\mathbf{h}) \\ \frac{\partial F}{\partial \beta_i} &= -(\mathbf{X}^0_{\text{BS}_i}\mathbf{h})^T(\mathbf{y} - \boldsymbol{\beta}_i\mathbf{X}^0_{\text{BS}_i}\mathbf{h} - w_i\mathbf{X}^1_{\text{BS}_i}\mathbf{h}) \\ \frac{\partial F}{\partial \omega_i} &= -\mathbf{Z}^T(\mathbf{y} - \boldsymbol{\beta}_i\mathbf{X}^0_{\text{BS}_i}\mathbf{h} - w_i\mathbf{X}^1_{\text{BS}_i}\mathbf{h}) \\ \frac{\partial F}{\partial r_i} &= -(\mathbf{X}^1_{\text{BS}_i}\mathbf{h})^T(\mathbf{y} - \boldsymbol{\beta}_i\mathbf{X}^0_{\text{BS}_i}\mathbf{h} - w_i\mathbf{X}^1_{\text{BS}_i}\mathbf{h}) \end{cases}$$

A good initialization plays a crucial role in the convergence of any iterative algorithm. Furthermore, for non-convex problems a good initialization prevents the algorithm from converging to undesired local minima. We have used as initialization for the R1-GLM and R1-GLMS models the solution given by the GLM with separate designs (GLMS). Since the GLM with separate designs scales linearly in the number of voxels, this significantly reduces computation time whenever an important number of voxels is considered.

Whenever the design matrix $\mathbf{X_B}$ has more rows than columns (as is the case in both datasets we consider when $\mathbf{B}$ is the 3HRF basis), it is possible to find an orthogonal transformation that significantly speeds up the computation of the Rank-1 model. Let $\mathbf{Q}, \mathbf{R}$ be the "thin" QR decomposition of $\mathbf{X_B} \in \mathbb{R}^{n \times dk}$, that is, $\mathbf{QR} = \mathbf{X_B}$ with $\mathbf{Q} \in \mathbb{R}^{n \times dk}$ an orthogonal matrix and $\mathbf{R} \in \mathbb{R}^{dk \times dk}$ a triangular matrix. Because of the invariance of the Euclidean norm to orthogonal transformations, the change of variable $\mathbf{X_B} \leftarrow \mathbf{Q}^T\mathbf{X_B}$, $\mathbf{y} \leftarrow \mathbf{Q}^T\mathbf{y}$ yields a Rank-1 model in Eq. (2.9) with equivalent solutions. This reduces the size of the design matrix to a square triangular matrix of size $dk \times dk$ (instead of $n \times dk$) and reduces the explained variable $\mathbf{y}$ to a vector of size kd (instead of n). After this change of variable, the convergence of the Rank-1 model is significantly faster due to the faster computation of the objective function and its partial derivatives. We have observed that the total running time of the algorithm can be reduced by 30% using this transformation.

Some numerical solvers such as L-BFGS-B [Liu and Nocedal, 1989] require the constraints to be given as box constraints. While our original problem includes an equality constraint we can easily adapt it to use convex box constraints instead. We replace the equality constraint $\|\mathbf{Bh}\|_\infty = 1$ by the convex inequality constraint $\|\mathbf{Bh}\|_\infty \leq 1$, which is equivalent to the box constraint $-1 \leq (\mathbf{Bh})_i \leq 1$ supported by the above solver. However, this change of constraint allows solutions in which $\mathbf{h}$ can be arbitrarily close to zero. To avoid such degenerate cases we add the smooth term $-\|\mathbf{B}(:,1)h_1\|_2^2$ to the cost function. Since there is a free scale parameter between $\mathbf{h}$ and $\boldsymbol{\beta}$, this does not bias the problem, but forces $\mathbf{Bh}$ to lie as far as possible from the origin (thus saturating the box constraints). Once a descent direction has been found by the L-BFGS-B method we perform a line search procedure to determine the step length. The line-search procedure was implemented to satisfy the strong Wolfe conditions [Nocedal and Wright, 2006]. Finally, when the optimization algorithm has converged to a stationary point, we rescale the solution setting to ensure that the equality constraint holds. This still leaves a sign ambiguity between the estimated HRF and the associated beta-maps. To make these parameters identifiable, the sign of the estimated HRF will be chosen so that these correlate positively with the reference HRF.

We have compared several first-order (Conjugate Gradient), quasi-Newton (L-BFGS) and Newton methods on this problems and found that in general quasi-Newton methods performed best in terms of computation time. In our implementation, we adopt the L-BFGS-B as the default solver.

In Algorithm 1 we describe an algorithm based on L-BFGS that can be used to optimize R1-GLM and R1-GLMS models (a reference implementation for the Python language is described in subsection Software). Variable $\mathbf{r}$ is only used for the R1-GLMS method and its use is denoted within parenthesis,

i.e. $(, \mathbf{r})$, so that for the R1-GLM it can simply be ignored.

Algorithm 1: Optimization of R1-GLM and R1-GLMS models

Data: Given initial points $\boldsymbol{\beta}_0 \in \mathbb{R}^k, \mathbf{h}_0 \in \mathbb{R}^d, \boldsymbol{\omega}_0 \in \mathbb{R}^q$ $(, \mathbf{r}_0 \in \mathbb{R}^k)$, convergence tolerance $\epsilon > 0$, inverse Hessian approximation $\mathbf{H}_0$.

Result: $\boldsymbol{\beta}_m, \mathbf{h}_m$

(Optional): Compute the QR decomposition of $\mathbf{X_B}$, $\mathbf{QR} = \mathbf{X_B}$, and replace $\mathbf{X_B} \leftarrow \mathbf{Q}^T\mathbf{X_B}, \mathbf{y} \leftarrow \mathbf{Q}^T\mathbf{y}$;

Initialization. Set $m \leftarrow 0, \mathbf{z} \leftarrow \text{vec}([\mathbf{h}_0, \boldsymbol{\beta}_0, \boldsymbol{\omega}_0(, \mathbf{r}_0)])$;

while $\|\nabla f\| > \epsilon$ **do**

  Compute search direction. Set $\mathbf{p}_m \leftarrow -\mathbf{H}_m \nabla f(\mathbf{h}_m, \boldsymbol{\beta}_m, \boldsymbol{\omega}_m(, \mathbf{r}_m))$, where f is the objective function of the R1-GLM or R1-GLMS model.;

  Set $\mathbf{z}_{m+1} = \mathbf{z}_m + \gamma_m \mathbf{p}_m$, where γ_m is computed from a line search procedure subject to the box constraints $\|\mathbf{h}_m\|_\infty \leq 1$.;

  $m \leftarrow m + 1$;

Extract R1-GLM(S) parameters from $\mathbf{z}_m$. Set $\mathbf{h}_m \leftarrow \mathbf{z}_m(1:d), \boldsymbol{\beta}_m \leftarrow \mathbf{z}_m(d+1:m+d)$;

Normalize and set sign so that the estimated HRF is positively correlated with a reference HRF:

$q_m \leftarrow \|\mathbf{h}_m\|_\infty \text{sign}(\mathbf{h}_m^T \mathbf{h}_{\text{ref}}),\ \mathbf{h}_m \leftarrow \mathbf{h}_m / q_m,\ \boldsymbol{\beta}_m \leftarrow \boldsymbol{\beta}_m q_m$;

The full estimation of the R1-GLM model with 3HRF basis for one subject of the dataset described in section *Dataset 2: decoding of potential gain levels* (16×3 conditions, 720 time points, $41,622$ voxels) took 14 minutes in a 8-cores Intel Xeon 2.67GHz machine. The total running time for the 17 subjects was less than four hours.

Software

We provide a software implementation of all the models discussed in this section in the freely available (BSD licensed) pure-Python package hrf_estimation [1].

[1] https://pypi.python.org/pypi/hrf_estimation

2.4 Data description

With the aim of making the results in this paper easily reproducible, we have chosen two freely available datasets to validate our approach and to compare different HRF modeling techniques. Details on the datasets can be found in Appendix 8. In the following we explain the specific processing performed on these datasets for the purposes of this chapter.

2.4.1 Dataset 1: encoding of visual information

Detailed dataset descriptions are to be found in the Appendix.

We performed local detrending using a Savitzky-Golay filter [Savitzky and Golay, 1964] with a polynomial of degree 4 and a window length of 91 TR. The activation coefficients (beta-map) and HRF were extracted from the training set by means of the different methods we would like to compare. The training set consisted of 80% of the original session (4 out of 5 runs). This

resulted in estimated coefficients (beta-map) for each of the 70×4 images in the training set.

We proceed to train the encoding model. The stimuli are handled as local image contrasts, that are represented by spatially smoothed Gabor pyramid transform modulus with two orientations and four scales. Ridge regression (regularization parameter chosen by Generalized Cross-Validation [Golub et al., 1979], see chapter 8 for original work on the extenstion of leave-one-out to leave-k-out cross-validation) was then used to learn a predictor of voxel activity on the training set. By using this encoding model and the estimated HRF it is possible to predict the BOLD signal for the 70 images in the test set (20 % of the original session). We emphasize that learning the HRF on the training set instead of on the full dataset is necessary to avoid overfitting while assessing the quality of the estimated HRF by any HRF-learning method: otherwise, the estimation of the HRF may incorporate specificities of the test set leading to artificially higher scores.

Generalized cross-validation for k left out samples derived in appendix - chapter 8

In a first step, we perform the image identification task from [Kay et al., 2008]. From the training set we estimate the activation coefficients that will be used to compute the activation maps. We use an encoding model using Gabor filters that predicts the activation coefficient from the training stimuli. From the stimuli in the validation set we predict the activation coefficients that we then use to identify the correct image. The predicted image is the one yielding the highest correlation with the measured activity. This procedure mimics the one presented in [Kay et al., 2008, Supplementary material].

In a second step, we report score as the Pearson correlation between the measurements and the predicted BOLD signal on left out data. The prediction of BOLD signal on the test set is performed from conditions that were not present in the train set. In order to do this, an HRF for these conditions is necessary. As highlighted in the methods section, the construction of an HRF for these conditions is ambiguous for non Rank-1 methods that perform HRF estimation on the different stimuli. In these cases we chose to use the mean HRF across conditions as the HRF for unseen conditions. Finally, linear predictions on the left out fold were compared to the measured BOLD signals.

2.4.2 Dataset 2: decoding of potential gain levels

For all subjects three runs were recorded, each consisting of 240 brain images with a repetition time (TR) of 2 seconds and a stimulus presentation at every 4 seconds. In order to perform HRF estimation on more data than what is available on a single run, we performed the estimation on the three runs simultaneously. This assumes HRF consistency across runs, which was obtained by concatenating the data from the three runs and creating a block-diagonal design matrix correspondingly (each block is the design of one run).

After training a regression model on 90% of the data, we predict the gain level on the remaining 10%. As a performance measure we use Kendall tau rank correlation coefficient [Kendall, 1938] between the true gain levels and the predicted levels, which is a measure for the orderings of the data. We argue that this evaluation metric is better suited than a regression loss for this task because of the discrete and ordered nature of the labels. Also, this loss is less sensitive to shrinkage of the prediction that might occur when pe-

nalizing a regression model [Bekhti et al., 2014]. The Kendall tau coefficient always lies within the interval $[-1, 1]$, with 1 being perfect agreement between the two rankings and -1 perfect disagreement. Chance level lies at zero. This metric was previously proposed for fMRI decoding with ordered labels in [Doyle et al., 2013].

2.5 Results

In order to compare the different methods discussed previously, we ran the same encoding and decoding studies while varying the estimation method for the activation coefficients (beta-maps). The methods we considered are standard GLM (denoted GLM), GLM with separate designs (GLMS), Rank-1 GLM (R1-GLM) and Rank-1 GLM with separate designs (R1-GLMS). For all these models we consider different basis sets for estimating the HRF: a set of three elements formed by the reference HRF and its time and dispersion derivative, a FIR basis set (of size 20 in the first dataset and of size 10 in the second dataset) formed by the canonical vectors and the single basis set formed by the reference HRF (denoted "fixed HRF"), which in this case is the HRF used by the SPM 8 software.

It should be reminded that the focus of this study is not the study of the HRF in itself (such as variability across subjects, tasks or regions) but instead its possible impact on the accuracy of encoding and decoding paradigms. For this reason we report encoding and decoding scores but we do not investigate any of the possible HRF variability factors.

2.5.1 Dataset 1: encoding of visual information

In the original study, 500 voxels were used to perform image identification.

We first present the scores obtained in the image identification task for different variants of the GLM. This can be seen in Figure 2.2. The displayed score is the count of correctly identified images over the total number of images (chance level is therefore at 1/120). The identification algorithm here only uses the beta-maps obtained from the train and validation set. This makes the estimation of the HRF an intermediate result in this model. However, we expect that a correct estimation of the HRF directly translates into a better estimation of the activation coefficients in the sense of being able to achieve higher predictive accuracy. Our results are consistent with this hypothesis and in this task the rank-one (R1) and glm-separate (GLMS) models outperform the classical GLM model. The benefits range from 0.9% for R1-GLM in subject 2 to 8.2% for the same method and subject 1. It is worth noticing that methods with FIR basis obtain a higher score than methods using the 3HRF basis.

In order to test whether this increase is statistically significant we performed the following statistical test. The success of recovering the correct image can be modeled as a binomial distribution, with p_A the probability of recovering the correct image with method A and p_B the probability of recovering the correct image with method B. We define the null hypothesis H_0 as the statement that both probabilities are equal, $H_0 : p_A = p_B$, and the alternate hypothesis that both probabilities and not equal, $H_1 : p_1 \neq p_2$

(this test is sometimes known as the binomial proportion test [Röhmel and Mansmann, 1999]). The score test statistic for the one-tailed test is $T = (p_A - p_B)/\sqrt{p(1-p)\frac{2}{n}}$, where $p = (p_A + p_B)/2$ and n is the number of repetitions, in this case $n = 120$. This statistic is normally distributed for large n. The p-value associated with this statistical test when comparing every model (by order of performance) with the model "GLM with with fixed HRF" is $(0.10, 0.10, 0.15, 0.19, 0.21, 0.26, 0.5, 0.5, 0.82, 0.81)$ for the first subject and $(0.18, 0.18, 0.25, 0.34, 0.34, 0.44, 0.5, 0.5, 0.86, 0.93)$ for the second.

Figure 2.2: Image identification score (higher is better) on two different subjects from the first dataset. The metric counts the number of correctly identified images over the total number of images (chance level is $1/120 \approx 0.008$). This metric is less sensitive to the shape of the HRF than the voxel-wise encoding score. The benefits range from 0.9% points to 8.2% points across R1-constrained methods and subjects. The highest score is achieved by a R1-GLM method with a FIR basis set for subject 1 and by a R1-GLMS with FIR basis for subject 2.

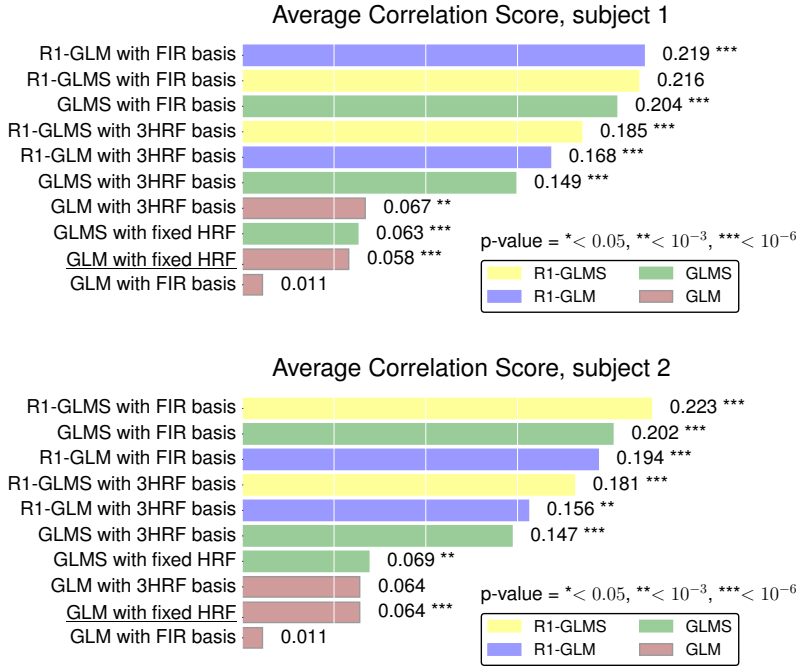

Figure 2.3: Average correlation score (higher is better) on two different subjects from the first dataset. The average correlation score is the Pearson correlation between the predicted BOLD and the true BOLD signal on left-out session, averaged across voxels and sessions. Methods that perform constrained HRF estimation significantly outperform methods that use a fixed reference HRF. As for the image identification performance, the best performing method for subject 1 is the R1-GLM, while for subject 2 it is the R1-GLMS model, both with FIR basis. In underlined typography is the GLM with a fixed HRF which is the method used by default in most software distributions. A Wilcoxon signed-rank test is performed between each method and the next one in the ordered result list by considering the leave-one-session out cross-validation scores for each method. We report p-values to assess whether the score differences are statistically significant.

We will now use a different metric for evaluating the performance of the encoding model. This metric is the Pearson correlation between the BOLD predicted by the encoding model and the true BOLD signal, averaged across voxels. We will compute this metric on a left-out session, which results in five scores for each method, corresponding to each of the cross-validation folds. Given two methods, a Wilcoxon signed-rank test can be used on these cross-validation scores to assess whether the score obtained by the two methods are significantly different. This way, irrespective of the variance across voxels, which is inherent to the study, we can reliably assess the relative ranking of the different models. In Figure 2.3 we show the scores for each method (averaged across sessions) and the p-value corresponding the Wilcoxon test between a given method and the previous one by order of performance.

We observed in Figure 2.3 that methods that learn the HRF together with some sort of regularization (be it Rank-1 constraint or induced by separate designs) perform noticeably better than methods that perform unconstrained HRF estimation, highlighting the importance of a robust estimation of the HRF as opposed to a free estimation as performed by the standard GLM model with FIR basis. This suggests that R1 and GLMS methods permit including FIR basis sets while minimizing the risk of overfitting inherent to the classical GLM model.

We also observed that models using the GLM with separate designs from [Mumford et al., 2012] perform significantly better on this dataset than the standard design, which is consistent with the purpose of these models. It improves estimation in highly correlated designs. The best performing model for both subjects in this task is the R1-GLMS with FIR basis, followed by the R1-GLM with FIR basis model for subject 1 and GLMS with FIR basis for subject 2. The difference between both models (Wilcoxon signed-rank test) was significant with a p-value $< 10^{-6}$. Since the results for both subjects are similar,

we will only use subject 1 for the rest of the figures.

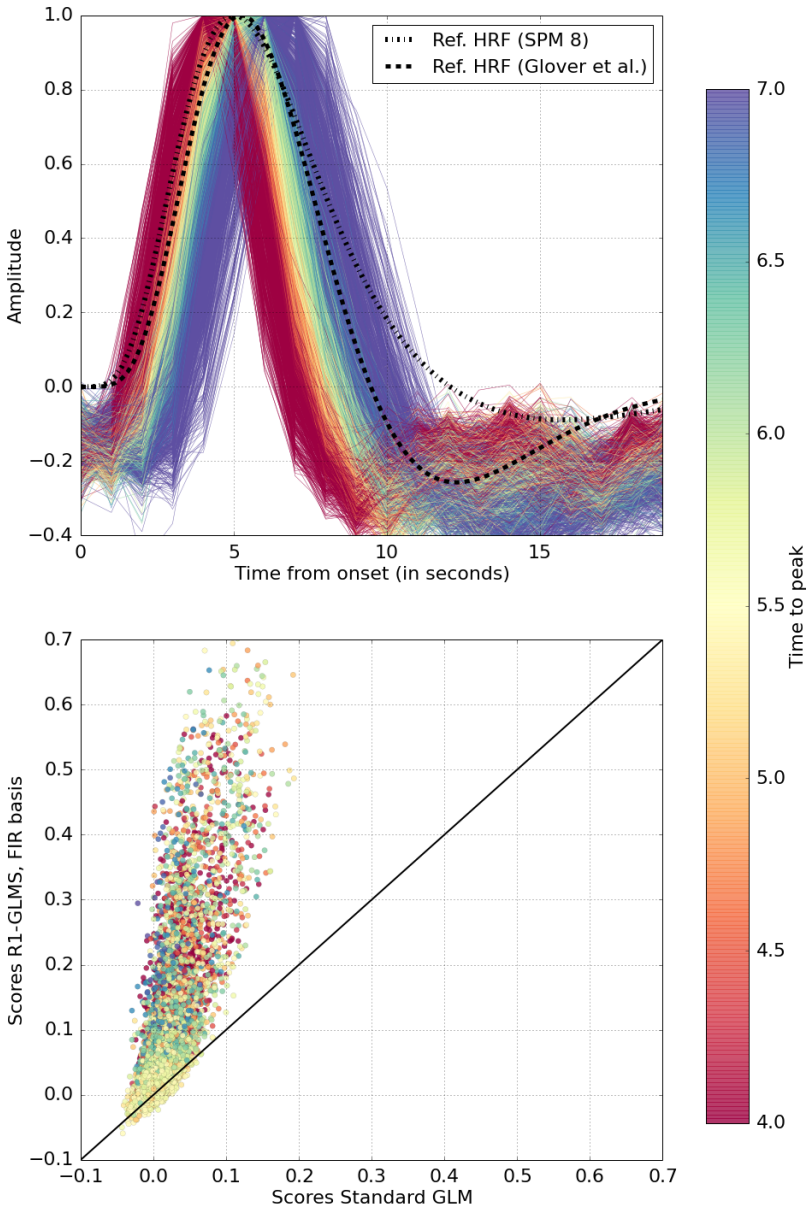

Figure 2.4: Top: HRF estimated by the R1-GLMS method on voxels for which the encoding score was above the mean encoding score (first dataset), color coded according to the time to peak of the estimated HRFs. The difference in the estimated HRFs suggests a substantial variability at the voxel level within a single subject and a single task. Bottom: voxel-wise encoding score for the best performing method (R1-GLMS with FIR basis) versus a standard GLM (GLM with fixed HRF) across voxels. The metric is Pearson correlation. Points above the black diagonal correspond to voxels that exhibit a higher score with the R1-GLMS method than with a standard GLM.

To further inspect the results, we investigated the estimation and encoding scores at the voxel level. This provides some valuable information. For example, parameters such as time-to-peak, width and undershoot of the estimated HRF can be used to characterize the mis-modeling of a reference HRF for the current study. Also, a voxel-wise comparison of the different methods can be used to identify which voxels exhibit a greater improvement for a given method. In the upper part of Figure 2.4 we show the HRF estimated for the first subject by our best performing method (the Rank-1 with separate designs and FIR basis). For comparison we also present two commonly used reference HRFs: one used in the software SPM and one defined in [Glover, 1999]and used by software such as NiPy[2] and fmristat[3]. Because the HRF estimation will fail on voxels for which there is not enough signal, we only show the

[2] http://nipy.org

[3] http://www.math.mcgill.ca/keith/fmristat/

estimated HRF for voxels for which the encoding score is above the mean encoding score. In this plot the time-to-peak of the estimated HRF is color coded. One can observe a substantial variability in the time to peak, confirming the existence of a non-negligeable variability of the estimated HRFs, even within a single subject and a single task. In particular, we found that only 50% of the estimated HRFs on the full brain volume peaked between 4.5 and 5.5 seconds.

In the lower part of Figure 2.4 we can see a scatter plot in which the coordinates of each point are the encoding scores with two different methods. The first coordinate (X-axis) is given by the score using a canonical GLM whilst the second coordinate (Y-axis) corresponds to the Rank-1 separate with FIR basis. Points above the black diagonal exhibit a higher score with our method than with a canonical GLM. As previously, the color represents the time to peak of the estimated HRF. From this plot we can see that voxels that have a low correlation score using a canonical GLM do not gain significant improvement by using a Rank-1 Separate FIR model instead. However, voxels that already exhibit a sufficiently high correlation score using a canonical GLM (> 0.05) see a significant increase in performance when estimated using our method.

These results suggest as a strategy to limit the computational cost of learning the HRF on an encoding study to perform first a standard GLM (or GLMS) on the full volume and then perform HRF estimation only on the best performing voxels.

The methods that we have considered for HRF estimation can be subdivided according to the design matrices they use (standard or separate) and the basis they use to generate the estimated HRF (3HRF and FIR). We now focus on the performance gains of each of these individual components. In the upper part of Figure 2.5 we consider the top-performing model, the Rank-1 GLMS, and compare the performance of two different basis sets: FIR with 20 elements in the Y-axis and the reference HRF plus its time and dispersion derivatives (3HRF) in the X-axis. The abundance of points above the diagonal demonstrates the superiority of the FIR basis on this dataset. The color trend in this plot suggests that the score improvement of the FIR basis with respect to the 3HRF basis becomes more pronounced as the time-to-peak of the estimated HRF deviates from the reference HRF (peak at 5s), which can be explained by observing that the 3HRF basis corresponds to a local model around the time-to-peak. In the bottom part of this figure we compare the different design matrices (standard or separate). Here we can see the voxel-wise encoding score for two Rank-1 models with FIR basis and different design matrices: separate design on the Y-axis and classical design on the X-axis. Although both models give similar results, a Wilcoxon signed-rank test on the leave-one-session-out cross-validation score confirmed the superiority of the separate designs model in this dataset with p-value $< 10^{-3}$.

In Figure 2.6 we can see the voxel-wise encoding score on a single acquisition slice. In the upper column, the score is plotted on each voxel and thresholded at a value of 0.045, which would correspond to a p-value < 0.05 for testing non-correlation assuming each signal is normally distributed, while in the bottom row the 0.055 contour (p-value < 0.001) for the same data is shown as a green line. Here it can be seen how the top performing voxels

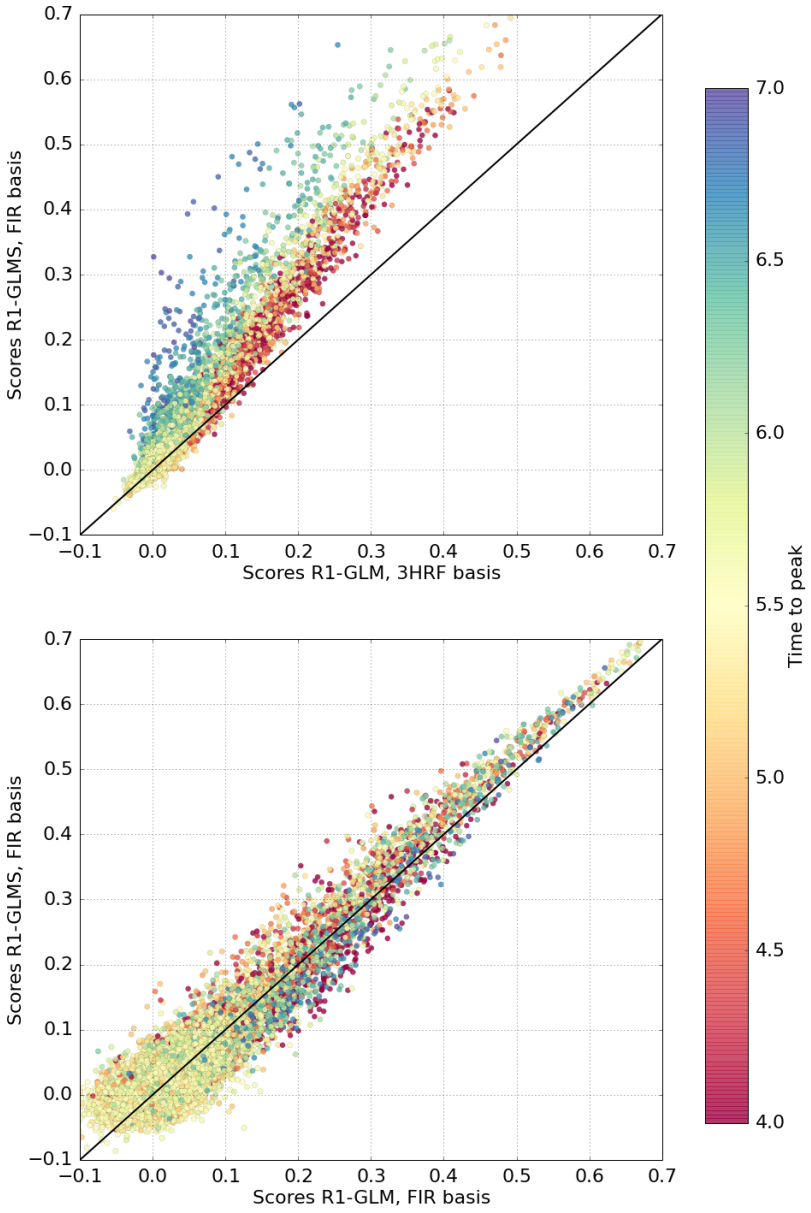

Figure 2.5: Voxel-wise encoding score for different models that perform HRF estimation (first dataset). As in figure 2.4, color codes for the time to peak of the estimated HRF at the given voxel. Top: two Rank-1 separate design models with different basis functions: FIR with 20 elements in the Y-axis and the reference HRF with its time and dispersion derivatives (3HRF) in the X-axis. The color trend in this plot suggests that the score improvement of the FIR basis with respect to the 3HRF becomes more pronounced as the time-to-peak of the estimated HRF deviates from the reference HRF (peak at 5s). This can be explained by taking into account that the 3HRF basis is a local model of the HRF around the peak time of the canonical HRF. Bottom: voxel-wise encoding score for two Rank-1 models with FIR basis and different design matrices: separate design on the Y-axis and classical design on the X-axis. Although both models give similar results, a Wilcoxon signed-rank test on the leave-one-session-out cross-validation score (averaged across voxels) confirmed the superiority of the separate designs model in this dataset with p-value $< 10^{-3}$.

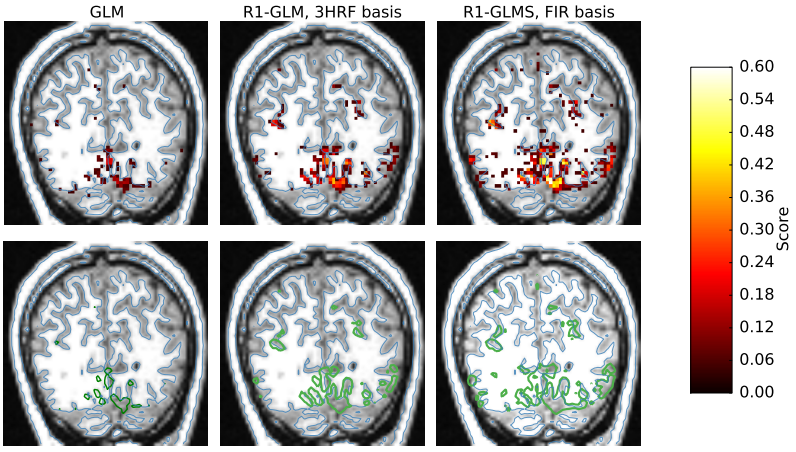

Figure 2.6: Voxel-wise encoding scores on a single acquisition slice for different estimation methods (first dataset). The metric is Pearson correlation. In the upper column, the voxel-wise score is thresholded at a value of 0.045 (p-value < 0.05), while in the bottom row the 0.055 contour (p-value < 0.001) for the same data is shown as a green line. Despite lacking proper segmentations of visual areas, the estimation methods produce results that highlight meaningful regions of interest around the calcarine fissure. This is particularly visible in the third column where our method R1-GLMS produces results with higher sensitiv-

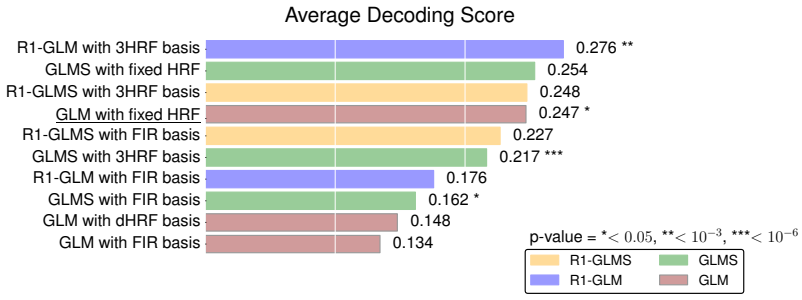

Figure 2.7: Averaged decoding score for the different method considered (higher is better) on the second dataset. The metric is Kendall tau. Methods that perform constrained HRF estimation significantly outperform methods that use a fixed (reference) HRF. In particular, the best performing method is the R1-GLM with 3HRF basis, followed by the R1-GLMS with 3HRF basis. In underlined typography is the GLM with a fixed HRF which is the method used by default in most software distributions. As in Figure 2.3, a Wilcoxon signed-rank test is performed and the p-value reported between a given method and the next method in the ordered result list to assess whether the difference in score is significant.

follow the gray matter. A possible hypothesis to explain the increase of the encoding score between the method R1-GLMS with FIR basis and the same method with 3HRF basis could be related either to the shape of the HRF deviating more from a canonical shape in lateral visual areas or to the higher signal-to-noise ratio often found in the visual cortex when compared to lateral visual areas.

2.5.2 Dataset 2: decoding of potential gain levels

The mean decoding score was computed over 50 random splittings of the data, with a test set of size 10%. The decoding regression model consisted of univariate feature selection (ANOVA) followed by a Ridge regression classifier as implemented in scikit-learn [Pedregosa et al., 2011]. Both parameters, number of voxels and amount of ℓ_2 regularization in Ridge regression, were chosen by cross-validation.

The mean score for the 10 models considered can be seen in Figure 2.7. Similarly to how we assessed superiority of a given method in encoding, we will say that a given method outperforms another if the paired difference of both scores (this time across folds) is significantly greater than zero. This is computed by performing a Wilcoxon signed rank test across voxels. For this reason we report p-values together with the mean score in Figure 2.7.

As was the case in encoding, Rank-1 constrained methods obtain the highest scores. In this case however, methods with 3HRF basis outperform methods using FIR basis. This can be explained by factors such as smaller sample size of each of the runs, smaller number of trials in the dataset and experimental design.

2.6 Discussion

We have compared different HRF modeling techniques and examined their generalization score on two different datasets: one in which the main task was an *encoding* task and one in which it was a *decoding* task. We compared 10 different methods that share a common formulation within the context of the General Linear Model. This includes models with canonical and separate designs, with and without HRF estimation constrained by a basis set, and with and without rank-1 constraint. We have focused on voxel-independent models of the HRF, possibly constrained by a basis set, and have omitted for efficiency reasons other possible models such as Bayesian models [Marr-

elec et al., 2003, Ciuciu et al., 2003, Makni et al., 2005] and regularized methods [Goutte et al., 2000, Casanova et al., 2008].

Other models such as spatial models [Vincent et al., 2010], and multi-subject methods [Zhang et al., 2012, 2013] that adaptively learn the HRF across several subjects are outside the scope of this work. The latter models are more relevant in the case of standard group studies and second level analysis.

Our first dataset consists of an encoding study and revealed that it is possible to boost the encoding score by appropriately modeling the HRF. We used two different metrics to assess the quality of our estimates. The first metric is the fraction of correctly identified images by an encoding model. For this we computed the activation coefficients on both the training and validation dataset. We then learned a predictive model of the activation coefficients from the stimuli. This was used to identify a novel image from a set of 120 potential images from which the activation coefficients were previously computed. The benefits range from 0.9% points to 8.2% points across R1-constrained methods and subjects. The best-performing model in this task is the R1-GLM with FIR basis. The second metric is the Pearson correlation. By considering the voxel-wise score on a full brain volume we observed that the increase in performance obtained by estimating the HRF was not homogeneous across voxels and more important for voxels that already exhibited a good score with a classical design (GLM) and a fixed HRF. The best-performing method is the Rank-1 with separate designs (R1-GLMS) and FIR basis model, providing a significant improvement over the second best-performing model. We also found substantial variability of the shape in the estimated HRF within a single subject and a single task.

The second dataset consists of a decoding task and the results confirmed that constrained (rank-1) estimation of the HRF also increased the decoding score of a classifier. The metric here is Kendall tau. However, in this case the best performing basis was no longer FIR basis consisting of ten elements but the three elements 3HRF basis (HRF and derivatives) instead, which can be explained by factors such as differences in acquisition parameters, signal-to-noise ratio or by the regions involved in the task.

A higher performance increase was observed when considering the correlation score within the encoding model. This higher sensitivity to a correct (or incorrect) estimation of the HRF can be explained by the fact that the estimation of the HRF is used to generate the BOLD signal on the test set. The metric is the correlation between the generated signal and the BOLD signal. It is thus natural to expect that a correct estimation of the HRF has a higher impact on the results.

In the decoding setup, activation coefficients (beta-map) are computed but the evaluation metric is the accuracy at predicting the stimulus type. The validation metric used for decoding is less sensitive to the HRF estimation procedure than the correlation metric from the encoding study, although it allowed us to observe a statistically significant improvement.

2.7 Conclusion

We have presented a method for the joint estimation of HRF and activation coefficients within the GLM framework. Based on ideas from previous lit-

erature [Makni et al., 2008, Vincent et al., 2010] we assume the HRF to be equal across conditions but variable across voxels. Unlike previous work, we cast our model as an optimization problem and propose an efficient algorithm based on quasi-Newton methods. We also extend this approach to the setting of GLM with separate designs.

We quantify the improvement in terms of generalization score in both encoding and decoding settings. Our results show that the rank-1 constrained method (R1-GLM and R1-GLMS) outperforms competing methods in both encoding and decoding settings.

2.8 Outlook

In the above contribution we presented a fast method for activation estimation, which is available as a software package [4]. It is shown that using the weight maps obtained by estimating the HRF, instead of keeping it fixed, lead systematically to higher scores in both the encoding and the decoding setting.

[4] https://pypi.python.org/pypi/hrf_estimation

For in-depth work on the estimation of the HRF as an object of study in itself, there is a body of work already available. For instance, [Ciuciu et al., 2003] devise a probabilistic model in which the hemodynamic response and its variance can be inferred. This model is refined to incorporate automatic parcellation and inter-subject studies in later contributions, as well as fast inference algorithms using a Variational Bayes approach.

An addition to the existing probabilistic models would be to frame the estimation of the HRF as a fully continuous function by means of Bayesian kernel methods such as Gaussian kernel Gaussian processes. This would fully eliminate the necessity to interpolate given non-discretely jittered event sequences, at the cost of needing to invert a covariance matrix the size of the number of sampling points of the HRF function.

It should also be considered to undertake a comprehensive evaluation of probabilistic models enabling HRF estimation towards the goal of a reliable and thorough comparison amongst them. These models should have different levels of detail and complexity, with the classical GLM with fixed HRF and i.i.d Gaussian noise assumption as baseline. As with other suggestions presented throughout this thesis, evaluation should be done, if possible, by evaluating the loglikelihood of the model on left-out data.

Take-Home messages

- Taking HRF shape into account generally improves model estimation for encoding and decoding;
- For the classic GLM, constraining an HRF to be the same across conditions for a given voxel is generally beneficial to estimation compared to not constraining;
- The separate GLM is competitive also without rank-1 contraint;
- Datasets rich in number of trials benefit from an HRF fit using a full basis, whereas less large datasets, while still benefitting from HRF estimation, may be better modeled with a less large HRF space, such as $3HRF$.

3 Combining Total Variation and Sparsity in a new way

In the *decoding* setting (cf. 1.3.3), we are interested in inferring brain state, as described by an external discrete or continuous variable, from images of brain activity. Given the usually high dimensional nature of brain images and the low number of samples (even if one restricts to a small subvolume, one typically has more voxels than samples), the naive least squares optimization problem as well as the logistic regression problem are ill-posed due to the non-trivial kernel of the design matrix.

This inverse problem can be regularized using convex penalties designed to make the problem well-posed in the kernel of the design matrix. In addition they can contribute to better conditioning, but not all of them do: Sparsity induced by the ℓ_1 norm can make the problem well-posed, but due to the known fact that brain activations tend to have spatial extent, hence inducing strong correlations in a design matrix containing them, the solutions can remain very unstable with respect to noise or resampling.

Here we introduce a spatial gradient regularizer based on an analysis sparse version of the group lasso, which can select spatially contiguous active regions together.

Sections 3.1 to 3.8 have been accepted to the MICCAI conference 2015.

- M. Eickenberg, E. Dohmatob, B. Thirion, G. Varoquaux *Sparsity meets Total Variation - Learning with Segmenting Penalties*, to appear in Proc. MICCAI 2015

Statistical learning with segmenting penalties

Prediction from medical images is a valuable aid to medical diagnosis if it is sufficiently reliable. For instance, anatomical MR images can reveal certain disease conditions, while their functional counterparts can predict behavior or neuropsychiatric phenotypes. However, a physician will not conclude from predictions by a black-box model: understanding the anatomical or functional features that underpin decision is critical. Generally however, the weight vectors of even a simple classifier such as an SVM are not easily amenable to such an examination: Often there is no visually apparent identifiable structure. Indeed, this is not only a prediction task, but also an

inverse problem that calls for adequate regularization. We address this challenge by introducing an efficient convex region-selecting penalty, that can be used to regularize linear model coefficient vectors. Our penalty combines the spatial-contiguity-enforcing discrete total variation regularization and the sparsity-enforcing ℓ_1 regularization into one group: Voxels are either active with non-zero spatial derivative or zero with inactive spatial derivative. This leads to the segmentation of contiguous spatial regions (inside which the signal can vary almost freely) against a background of zeros. This segmentation of medical images in a target-informed manner is another important tool for analysis. For example, functional MRI is used intensively to chart the functional organization of the brain. Given the size and the 3D nature of brain images, computational efficiency is key. Keeping this in mind, we contribute an efficient optimization scheme that leads to significant computational gains compared to existing schemes. On several MRI experiments involving predictable brain states, the penalty shows good segmentation capacity.

3.1 Introduction

For certain pathologies, medical images carry weak indicators of some external phenotype. For instance, in Magnetic Resonance images, a pattern of brain atrophy centered on the thalamus predicts the evolution in Alzheimer's disease for elderly patients [Stonnington et al., 2010]. Functional Magnetic Resonance Imaging (fMRI) can be used to infer the behavior of a subject from their brain activity [Haxby et al., 2001]. Machine learning methods are convenient tools for learning these biomarkers. With linear models, model parameters form a spatial map in the image domain. However, minimizing a prediction error gives little control on the fine details of the corresponding maps. Indeed, the prediction problem is usually an ill-posed inverse problem in the sense that there are often less samples than features available: In the case of limited observations of such high dimensional data, many different weight maps can generate exactly the same predictions. A default choice among these candidates is implictly taken by the type of estimator employed. Using the statistical learning framework of empirical risk minimization, this choice can be actively imposed via a penalty which favors maps according to certain criteria, which, with due caution, can be interpreted as a "prior", reflecting information one may already have or think plausible. Sparsity for instance, imposable in convex optimization via the ℓ_1 norm, is very useful as it leads to selection of a small number of voxels in the images for the prediction. It has been widely used in medical imaging, from fMRI [Yamashita et al., 2008] to regularizing diffeomorphic registration [Durrleman et al., 2011].

However, in many situations, imposing sparsity can lead to less stability in the estimated weight maps. Indeed, one often faces high correlations in neighboring voxels, which leads to selection of different voxels depending on which portion of the data one uses for estimation. Since the adjacent voxels contain similar information, only one of them is needed for estimation. The notion of spatial contiguity in activation patterns has led to several contributions which incorporate this information in an estimator. Using an ℓ_2 penalty on a finite differences operator which acts on the image, one can force adjacent voxels to have similar weights [Ng et al., 2010, Grosenick et al., 2013,

Kandel et al., 2013]: This is known as *GraphNet*.

An improvement upon this method is to impose true sparsity on the spatial derivative as in [Michel et al., 2011], or to combine sparsity of the derivative with sparsity of the weights [Baldassarre et al., 2012, Gramfort et al., 2013]. These penalties come with the mathematical property of positive homogeneity, which makes model selection easier. A drawback for these methods is that they tend to favor perfectly flat or staircased and blocky activation maps - a property that can be considered an artifact: One would tend to expect smooth variation within an active region.

3.1.1 Sparsity and segmentation

It is around this idea that we center our contribution: Our goal is to detect spatially contiguous patches –however variably active– in statistically estimated images and to inform the estimation of the image with these detections. In essence, our work draws from two bodies of literature: the aforementioned concept of sparsity and the field of segmentation.

Sparse penalties have remarkable theoretical recovery properties which have been extensively studied, see e.g [Fuchs, 2005, Candes and Romberg, 2005, Wainwright, 2009]. Their main effect is to promote estimates with a small non-zero support, but given sufficient incoherence properties on the design matrix and sparse ground truth activation, the true support of the signal can be recovered exactly. In fMRI, the sparsity property is very useful: specialized brain modules under study occupy only a small fraction of the image volume. Sparsity can thus be used in a foreground segmentation context: recovering non-zero functional regions from a noisy background. However, in many real-world applications, such as CT or medical imaging, the underlying measurement process leads to strong correlations in columns of the design matrix corresponding to neighboring pixels, rendering all recovery theorems non-applicable and making sparse support estimation highly unstable.

The other body of literature that we are concerned with is that of segmentation, with a specific interest in convex variational approaches, as they can be expressed as penalties in a risk minimizer. A central aspect is the Mumford-Shah functional that yields piecewise smooth approximations of images [Mumford and Shah, 1989]. Chan and Vese [Chan and Vese, 2001] introduced a variant for segmentation purposes computing piecewise constant approximations: the minimal partition problem. These variational formulations are not convex, but [Pock et al., 2009] have shown that good solutions to the minimal partition problem can be achieved with a similar but convex functional, based on total variation, *i.e.* the ℓ_1 norm of the image gradient. For our purposes, this approach is appealing, as TV can be used as a penalty – technically an analysis sparse penalty [Nam et al., 2013]– that imposes sparse gradients and has good properties for image denoising [Rudin et al., 1992] or estimation in a linear model [Candes and Romberg, 2005]. However, all these related segmentation approaches model an object as a homogeneous constant-valued domain, thus washing out internal structure. Here, in the context of foreground-background segmentation, we want to impose a flat structure on the background, but not in the selected image domain. In this setting, imposing flatness of only the zero background seems a better candi-

date for segmentation than imposing constant domains.

Our contribution is twofold: 1) We introduce a new penalty called *Sparse Variation*, based on the TV-ℓ_1 combination, which forces zero activity on coordinates and spatial derivative jointly, and smooth variation of coordinates and derivatives in spatially contiguous active zones. 2) We provide a novel optimization routine called *fast adaptive accuracy shrinkage thresholding algorithm*, which allows for very fast estimation up to a very high precision. It is important to stress that useful spatial maps can only be obtained by assuring that the optimizer has thoroughly converged [Dohmatob et al., 2014]. We empirically evaluate its properties in regression and classification on fMRI and structural MRI (voxel-based morphometry) data. In particular, we compare it to TV-ℓ_1 regularization and contrast it to GraphNet.

3.2 Sparse Variation: A new spatially regularizing penalty

In this section, we briefly introduce two existing spatially regularizing penalties, GraphNet and TV-ℓ_1, before introducing our new variant, *Sparse Variation*. Then, in a dedicated optimization section, we elaborate the algorithm type we use along with speed-up mechanisms to keep runtime as small as possible.

3.2.1 Penalized regression

Our framework encompasses generalized linear models of which we will describe and use two emblematic ones: Linear regression for continuous output regression problems and logistic regression for binary output classification problems. The following optimization problem encompasses these variants.

Let $n, p \in \mathbb{N}$ denote number of samples and number of feature dimensions respectively. Let $X \in \mathbb{R}^{n \times p}$ be the design matrix and $y \in \mathbb{R}^n$ the optimization target. Finally let w denote the weight vector and c and offset to be obtained by solving the optimization problem:

$$\arg\min_{w,c} \ell(Xw + c, y) + \Omega(w) \tag{3.1}$$

Here, ℓ is the so called loss function or data fidelity term and Ω is the regularizer. We suppose both $\ell(\cdot, y)$ and Ω convex. The mean squared error loss employed for regression reads $\ell_{\text{mse}}(Xw + c, y) = \frac{1}{2n}\|y - Xw - c\|_2^2 = \frac{1}{2n}\sum_{i=1}^n (y_i - \langle X_i, w\rangle - c)^2$, and, choosing $y_i \in \{-1, +1\}$, the logistic loss can be expressed as $\ell_{\log}(Xw + c, y) = \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(y_i(\langle X_i, w\rangle + c)))$.

3.2.2 Existing regularizers

The convex regularizer Ω imposes structure on the solution of the inverse problem. Two regularizers successfully applied to medical volume data are the GraphNet and TV-ℓ_1 penalties, which we introduce now.

In the following, ∇ will denote a finite differences spatial gradient operator acting upon an image. Generally, for a 3D grid of size $p = p_x p_y p_z$, which is ravelled into a long vector, we have $\nabla \in \mathbb{R}^{3p \times p}$. Whenever a true gradient is used in an optimization problem, it will contain the variable with

respect to which it is calculated in subscript, e.g. "∇_w". $\|\cdot\|_2$ denotes the euclidean norm. For a partition $\mathcal{G}$ of coordinates the $\ell_{2,1}$ group norm is written $\|v\|_{2,1} = \sum_{g\in\mathcal{G}} \|v_g\|_2$.

For all discussed penalties, $\lambda > 0$ regulates its strength and $\rho \in [0,1]$ is a parameter controlling the trade-off between coordinate sparsity and spatial regularity. The GraphNet penalty consists of the sum of an ℓ_1 penalty on all coordinates and a squared ℓ_2 penalty on the spatial gradient, whereas the TV-$\ell_1$penalty is the sum of an ℓ_1 penalty and an $\ell_{2,1}$ group penalty on the spatial derivative:

$$\begin{aligned}\Omega_{\mathrm{GN}}(w) &= \lambda((1-\rho)\|\nabla w\|_2^2 + \rho\|w\|_1)\\ \Omega_{\mathrm{TV}-\ell_1}(w) &= \lambda((1-\rho)\|\nabla w\|_{2,1} + \rho\|w\|_1),\end{aligned}$$

3.2.3 Sparse Variation

We propose a new penalty based on TV-ℓ_1, called *Sparse Variation*, which enforces contiguous zones of smooth activation against a background of exact zeros. Indeed, in TV-ℓ_1, the penalties for sparsity of the signal and sparsity of the gradient are separable in that they can be active and inactive independently. A non-zero constant block, for example, is active for the ℓ_1 penalty, but inactive for the gradient, except at the borders. This property can induce step functions and blockiness where one would expect smoothness. We address this issue in *Sparse Variation* by grouping coordinate activation with spatial derivative activation: Either a coordinate is active (nonzero) and its derivative is active (nonzero) as well - allowing for a smooth variation in active zones - or both are inactive (zero).

We define the composite linear operator $K = \begin{pmatrix} (1-\rho)\nabla \\ \rho\,\mathrm{Id}_p \end{pmatrix}$, where Id_p denotes the $p \times p$ identity matrix. For 3D grids, we have $K \in \mathbb{R}^{4p\times p}$. The *Sparse Variation* penalty can then be defined as follows

$$\Omega_{\mathrm{SV}}(w) = \lambda\|Kw\|_{2,1},$$

where the $\ell_{2,1}$ group norm consists of groups containing the coordinate and all derivatives at each coordinate.

3.3 Optimization strategy

All optimization problems mentioned in this manuscript - GraphNet, TV-$\ell_1$and *Sparse Variation*, in combination with either the logistic loss or the mean squared error loss - have a similar global structure, in that they consist of sums of two convex functions, one being smooth, the other nonsmooth. This structure can be exploited in so-called proximal splitting algorithms (see e.g. [Combettes and Pesquet, 2011]), of which we will present an optimized variant in detail. Let $\mathcal{L}(w) = F(w) + G(w)$ represent the cost function, where F is smooth and convex and G convex[1]. These algorithms rely on an implicit subgradient step in the non-smooth function called the *proximal operator*: $\mathrm{prox}_{tG}(y) := (\mathrm{Id} + t\partial G)^{-1}(y)$ is the unique solution to the strongly convex problem $\arg\min_x \frac{1}{2t}\|y-x\|_2^2 + G(x)$.

[1] We omit c here for notational ease. It can be seen as the last coordinate of w, or, in the case of linear regression, be entirely omitted after data centering and reconstructed at the end of optimization.

The simplest method, forward-backward splitting [Combettes and Pesquet, 2011], is known in the case of the ℓ_1-Lasso as Iterative Shrinkage-Thresholding Algorithm (ISTA) and will be referred to by this name in the following. At a given optimization step $k \in \mathbb{N}$ it consists in minimizing the following surrogate optimization problem $w_{k+1} = \arg\min_w F_w(w_k) + \langle \nabla F(w_k), (w - w_k)\rangle + \frac{L}{2}\|w - w_k\|_2^2 + G(w)$, an expansion around the current point, where $L > 0$ represents the Lipschitz constant of $\nabla_w F$. This amounts to iterations of $w_{k+1} = \text{prox}_{\frac{1}{L}G}\left(w_k - \frac{1}{L}\nabla_w F(w_k)\right)$.

In order to accelerate convergence, one can add a momentum term such as Nesterov momentum.Recently, this technique has been popularized as *fast iterative shrinkage-thresholding algorithm* or fISTA [Beck and Teboulle, 2009a]. In comparison to ISTA, the gradient steps are applied to a carefully chosen interpolation of the weight vectors w_k and w_{k-1}.

The often considerable acceleration brought about by this method comes at the cost that there is no guarantee that each step of fISTA actually decreases the objective function. Indeed, this can lead to large rebounds in global cost on the way towards convergence. This non-monotone behavior can be remedied by switching to ISTA-type iterations whenever an increase in global cost is detected. The *monotone fISTA* (mfISTA) algorithm was introduced in [Beck and Teboulle, 2009b] to address this issue.

3.3.1 Computing the proximal operator

For linear regression with *Sparse Variation*, we choose $F(w) = \frac{1}{2}\|Xw - y\|_2^2$ and $G(w) = \Omega_{\text{SV}}(w)$. Analogously, for TV-$\ell_1$, we use $G(w) = \Omega_{\text{TV}-\ell_1}(w)$. For logistic regression, we use $F(w) = \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(y_i\langle X_i, w\rangle))$. Note that for GraphNet, it is beneficial to incorporate the smooth spatial penalty in F in order to avoid inversion of a large regularized linear operator. For linear regression, this reads $F(w) = \frac{1}{2}\|Xw - y\|_2^2 + \lambda(1-\rho)\|\nabla w\|_2^2$ and $G(w) = \lambda\rho\|w\|_1$.

An advantage of GraphNet is that $\text{prox}_{G/L} = \text{prox}_{\frac{\lambda\rho}{L}\|\cdot\|_1}$ has the closed form $(\text{prox}_{\frac{\lambda\rho}{L}\|\cdot\|_1}(w))_i = (|w_i| - \frac{\lambda\rho}{L})_+ \,\text{sign}(w_i)$, which is componentwise soft-thresholding. The proximal operators for TV-$\ell_1$and *Sparse-Variation* do not exist in closed form and must be obtained via the solution of a second, "inner" optimization problem. Both *Sparse Variation* and TV-$\ell_1$penalties can be written as $\lambda\|K\cdot\|_\bullet$ for an appropriate norm $\|\cdot\|_\bullet$: For *Sparse Variation*, $\|\cdot\|_\bullet = \|\cdot\|_{2,1}$ and for TV-ℓ_1, $\|(u_x, u_y, u_z, u_0)\|_\bullet = \|(u_x, u_y, u_z)\|_{2,1} + \|u_0\|_1$. Let $\|v\|_\bullet^* = \max_{\|u\|_\bullet \le 1}\langle u, v\rangle$ denote its dual norm. Then we have a minimax problem permitting the inversion of minimum and maximum operators at the optimum.

$$
\begin{aligned}
\min_v \frac{1}{2}\|w - v\|_2^2 + \lambda\|Kv\| &= \min_v \frac{1}{2}\|w - v\|_2^2 + \lambda \max_{\|u\|_\bullet^* \le 1}\langle u, Kv\rangle \\
&= \min_v \max_{\|u\|_\bullet^* \le \lambda} \frac{1}{2}\|w - v\|_2^2 + \langle K^T u, v\rangle \\
&= \frac{1}{2}\|w\|_2^2 + \max_{\|u\|_\bullet^* \le \lambda} \min_v \frac{1}{2}\|v - w + K^T w\|_2^2 - \frac{1}{2}\|w - K^T u\|_2^2 \\
&= \frac{1}{2}\|w\|_2^2 + \max_{\|u\|_\bullet^* \le \lambda} -\frac{1}{2}\|w - K^T u\|_2^2
\end{aligned}
$$

At optimum we have $v = w - K^T u$ and $u = \arg\min_{\|u\|^*_\bullet \leq 1} \frac{1}{2}\|w - K^T u\|_2^2$. We can determine u using e.g. an mFISTA algorithm, with $F(u) = \frac{1}{2}\|w - K^T u\|_2^2$ and $G(u) = \chi_{\{\|\cdot\|^*_\bullet \leq \lambda\}}(u)$, where χ_B is the convex indicator function of a set B. Accuracy can be measured by evaluating the dual gap $\gamma(u, v) = \frac{1}{2}\|w - v\|_2^2 + \lambda\|Kv\| - (\frac{1}{2}\|w\|_2^2 - \frac{1}{2}\|w - K^T u\|_2^2 - \chi_{\{\|\cdot\|^*_\bullet \leq \lambda\}}(u))$, where u, v, the dual and primal candidates respectively, are linked by $v = w - K^T u$. When evaluating the dual gap, it is necessary that u respect the feasibility constraints $\|u\|^*_\bullet \leq \lambda$ in order for the result to be meaningful. Most algorithms ensure feasibility sometime during an iteration and it is then that the dual gap should be evaluated. The primal problem is not constrained, hence the choice $v := w - K^T$ as the primal candidate will work for any feasible u.

3.3.2 Fast Adaptively Accurate Shrinkage Thresholding Algorithm

It is important to note that evaluating $\text{prox}_{G/L}$ numerically is an inexact operation, which can easily lead to non-convergence of the outer loop. However, according to [Schmidt et al., 2011], the presented algorithms converge even if the proximal operator $\text{prox}_{G/L}$ is not calculated to infinite accuracy, but decreases sufficiently with the iteration number k of the outer loop (with proofs for both ISTA and fISTA). Accuracy can conveniently be captured by the dual gap value. Instead of relying on a fixed dual gap refinement strategy, we devise an adaptive method, which increases accuracy as needed, if energy fails to decrease during an ISTA step. Algorithm 2 describes this procedure in detail.

Algorithm 2: fAASTA

Data: w_0
$ISTA \leftarrow False$, $v_1 \leftarrow w_0$, $k \leftarrow 0$, $t_1 \leftarrow 1$, $dgtol \leftarrow 0.1$;
while *not converged* **do**
  $k \leftarrow k + 1$;
  $w_k \leftarrow \text{prox}_{G/L}(v_k - (1/L)\nabla F(v_k), dgtol)$;
  if $\mathcal{L}(w_k) > \mathcal{L}(w_{k-1})$ **then**
    $w_k \leftarrow w_{k-1}$;
    $v_k \leftarrow w_{k-1}$;
    if *ISTA* **then**
      $dgtol \leftarrow dgtol/2$;
      while
      $\mathcal{L}(\text{prox}_{G/L}(v_k - (1/L)\nabla_w F(v_k), dgtol)) > \mathcal{L}(w_{k-1})$ **do**
        $dgtol \leftarrow dgtol/2$
    $ISTA \leftarrow True$;
  else
    if *ISTA* **then**
      $v_k \leftarrow w_k$
    else
      $t_k \leftarrow \frac{1+\sqrt{1+4t_{k-1}^2}}{2}$;
      $v_k \leftarrow w_k + \frac{t_{k-1}-1}{t_k}(w_k - w_{k-1})$;
    $ISTA \leftarrow False$

3.4 Empirical Results

In order to develop an intuition on the properties of *Sparse Variation*, we first study a 1D problem in which we recover a signal from corrupted DCT measurements. Then we move on to study three 3D problems, namely the segmentation of activation patterns recovered from two fMRI experiments and the study of anatomical landmarks for age in structural MRI.

3.5 A simple 1D signal recovery problem

Here we study the properties of the proposed penalties on a 1D recovery from corrupted measurements problem. We mimic a spectroscopy setting in which a signal with a continuous spectrum on a small, spatially contiguous support is measured with additive noise. The spectrum is obtained by solving an inverse problem with a discrete cosine transform operator. The signal measurements are given by $y = X_{\text{DCT}}^{-1} w + \varepsilon$, where X_{DCT} is the DCT operator, w the spectrum and ε a noise vector. For our experiments we use a ground truth spectrum of size 200, with around 80% zeros and an activated region resembling that of a chemical compound signature: Two overlapping smooth peaks, which we create here using a lower-thresholded, downward pointing parabola. We add Gaussian noise of 40% signal norm. Figure 3.1 shows the ground truth, along with the best ℓ_2 recovery results for *Sparse Variation*, TV-$\ell_1$and GraphNet: Each method was evaluated on a grid of penalties λ and sparsity vs spatial contiguity ratios ρ. The couple λ, ρ which minimized mean squared error with the ground truth was selected. It is the closest reconstruction possible for the full parameter space of that method. This is a way to obtain an insight into the model space parameterized by λ, ρ: We are showing the outcome which is the closest possible to ground truth. As per its construction, the TV-$\ell_1$penalty promotes flat signals, whereas the *Sparse Variation* penalty allows better recovery of the smooth nature of the signal. GraphNet selects a very low regularization, thereby incurring the most noise.

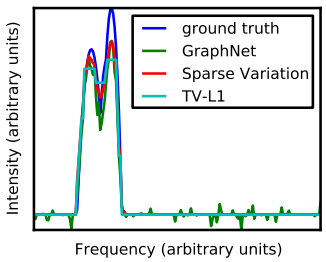

Figure 3.1: Recovery for 1D spectroscopy. Note the blocky nature of the TV-$\ell_1$solution, and the noise in the GraphNet estimation. The Sparse Variation solution follows the shape of the ground truth. Due to the ℓ_1-penalization, all solutions are shrunk towards 0.

The 1D example gives us an insight into the model spaces spanned by the parameter grid for the different methods.

3.6 Segmenting regions from MRI data

We analyse experiments in both fMRI and structural MRI and exhibit the use of both the regression and the classification settings. The strategy is to predict a continuous or categorial variable from brain images over a full parameter grid λ, ρ. For each penalty type, the weight maps of the best performing parameters in cross-validation on held out data are shown.

3.6.1 Classification example: Intra-subject study on object recognition

The human ventral temporal cortex exhibits specialization to certain recurrent concepts such as faces, but also several other object categories. We revisit the data from a seminal publication in this line of work [Haxby et al., 2001]: responses to visual stimuli of different categories - *faces, houses, chairs, scissors, bottles, shoes, cats*, and a control condition named *scrambledpix*, Fourier phase scrambled versions of the other stimuli. We test two classic contrasts, *faces versus houses* and *objects versus scramble* with the logistic loss.

The maps at optimally predictive parameter settings for the three maps overall detect similar regions. The top row of Figure 3.2 shows the segmented right-hand Fusiform Face Area. TV-$\ell_1$and *Sparse Variation* detect a somewhat similar region size, whereas GraphNet selects a stronger sparsity. For comparison, on the right we show an F-statistic, which indicates that good selection of regions is important in the context of interpretation. The bottom row represents the localization of the Lateral Occipital Complex (LOC). A similar description applies. It becomes apparent that *Sparse Variation* tends to select larger regions than the other two penalties. Note that the focality of these activation mappings is due to the single subject nature of the experiment.

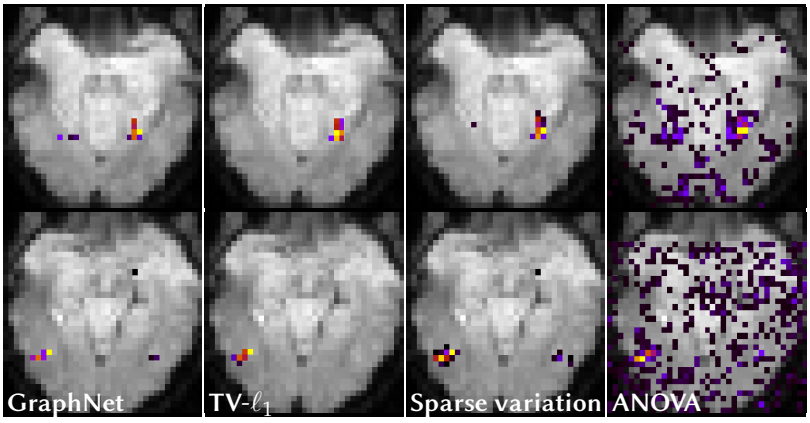

Figure 3.2: Weight maps obtained from discrimination tasks between two visual concepts on data from [Haxby et al., 2001]. **Top**: **FFA** (Fusiform Face Area) segmented in a face versus house discrimination. Axial cut at $z = -20$mm. around x=14mm, y=15mm. Accuracies on held-out data: GN: 95.5%, TV-ℓ_1 : 96.6%, SV: 97.7% **Bottom**: **LOC** (Lateral Occipital Complex) segmented in an object vs scramble discrimination. In this intra-subject analysis the maps are very well localized. Axial cut at $z = -16$mm.%. Accuracies: GN: 78.8%, TV-ℓ_1: 80.0%, SV: 80.0%

3.6.2 Regression example 1: Inter-subject analysis on gain prediction in gambling task

As an example for penalized linear regression using the proposed penalties in a multi-subject setting, we examined the fMRI gambling experiment by [Tom et al., 2007b]. Subjects were asked to decide whether they would enter a series of gambles with varying gains and losses. Here we attempt to estimate the gain of a given gamble from the fMRI activation it evokes on multiple different subjects.

At a fixed ratio of sparsity to spatial contiguity $\rho = 0.5$ we evaluated predictive power of models estimated by smooth lasso, TV-$\ell_1$and *Sparse Variation* on a grid of penalty values λ. The weight maps of the best predicting estimator for each penalty is shown in Figure 3.3. The strong noise in this multi-subject dataset makes the estimation difficult. At optimal predictive power the weight maps of TV-$\ell_1$and *Sparse Variation* show spatial contiguity and activation in expected regions, whereas the smooth lasso weights are scattered. Comparing TV-$\ell_1$to *Sparse Variation*, it becomes apparent that the main distinction is the "smoothness or zero" pattern enforced by the latter in comparison to more blocky activations for the former. Larger activated regions do justice to the multi-subject setting. Note the segmentation of the Insulae, which are duly mentioned in the original study.

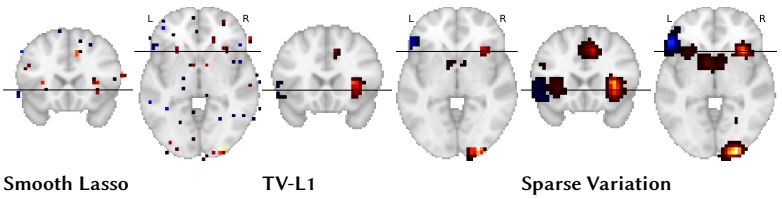

Figure 3.3: Weight vectors from estimating gain on the mixed gambles task [Tom et al., 2007b]. Prediction target is the gain proposed in a series of gambles proposed to the subjects. This inter-subject analysis shows broader regions of activation. Mean correlation scores on held out data:

3.6.3 Regression example 2: Estimating age from voxel-based morphometry

The Oasis database contains anatomical MRI data for 400 subjects [Marcus et al., 2007]. We extracted voxel-based morphometry images for these subjects and use *Sparse Variation* in a regression setting to estimate the ages of the subjects. Figure 3.4 shows the resulting weight maps for *Sparse Variation*, TV-$\ell_1$and GraphNet. All three regularizers stably identify the putamen, insula and para-hippocampal regions as descriptive. Note that TV-$\ell_1$selects contiguous regions where GraphNet associates small sparse clouds of voxels. Note further that the regions selected by TV-$\ell_1$are also found by *Sparse Variation* in a smoother version, in addition to several other regions not selected by TV-ℓ_1.

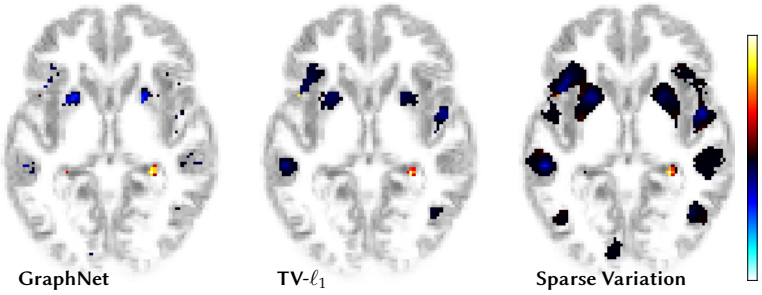

Figure 3.4: Weight vectors for age estimation from voxel-based morphometry maps from the Oasis dataset. *Sparse Variation* selects clearly defined regions which are easily amenable to further analysis. Mean correlation scores on held-out data: GN:0.805, TV-ℓ_1:0.793, SV:0.794

3.7 Convergence of the method

In data analysis, optimization speed is an important factor: The praticioner may often decide to use less accurate methods if others take too long to calculate. The adaptive refinement of the dual gap accuracy in the FAASTA setup leads to significant peformance gains with respect to other optimization methods. We compare this method to other ways of setting the dual gap accuracy in the inner loop. The other candidates are setting the dual gap tolerance to a constant, one strict (10^{-10}), one lax (0.1), and the dual gap tolerance refinement strategy according to [Schmidt et al., 2011] (decrease dual gap on the order of k^{-4}, where k is the iteration number). We also compare to the use of ISTA in the outer loop in a constant dual gap (0.1) setting and the adaptive refinement setting.

As can be seen in Figure 3.5, the results are striking. While the adaptive strategy always provides enough dual gap accuracy to ensure energy descent, the technique from [Schmidt et al., 2011] becomes too strict very quickly. Using a strict dual gap tolerance makes convergence very slow. Using a lax dual gap and fISTA as the outer algorithm leads to no energy decrease at all, and using the lax dual gap or the adaptive method with ISTA leads to stalling at insufficient accuracy rates. The proposed adaptive method provides by far the fastest convergence.

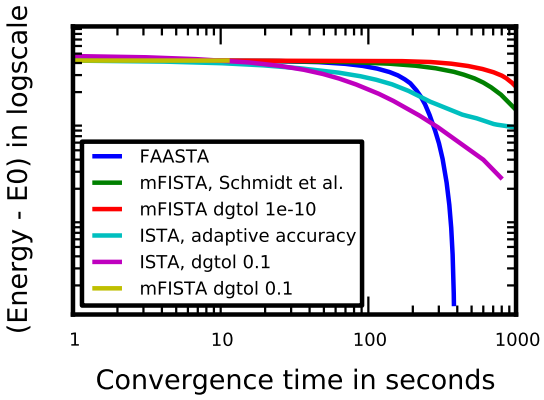

Figure 3.5: Convergence of several optimizers on *object vs scramble*. FAASTA takes around 400s to converge, whereas other methods take more than 15 minutes

3.8 Discussion

We introduced a new region-selective and sparsity-inducing convex penalty called *Sparse Variation*, in order to eliminate drawbacks of existing methods and combine their strengths. *Sparse Variation* forces large regions of an estimated image to zero, but allows smooth variation within spatially contiguous, active zones.

We use *Sparse Variation* in empirical risk estimators with mean squared error and logistic losses on three brain imaging problems, where we concentrate on the region segmenting properties of this penalty with respect to prior art, TV-ℓ_1 regularization and GraphNet. Indeed, it becomes apparent through all results that *Sparse Variation* tends to select smooth regions of interest. These regions of interest can be used in subsequent studies to obtain more refined results.

In order to obtain reliable spatial maps it is essential to ensure good convergence of the associated optimization problems. As with TV-ℓ_1 regularization, the optimization procedure via proximal splitting necessitates an inner optimization loop to evaluate the proximal operator. A linesearch strategy on dual gap accuracy is employed to refine the required dual gap accuracy only as much as needed to ensure fast convergence. On a benchmark with other accuracy setting strategies, our method converges in the least time.

In conclusion, *Sparse Variation* with *fAASTA* is the optimal choice of region segmenting optimizer, if analysis of estimator weight maps is envisaged.

3.9 Screening rules?

In extremely high-dimensional statistical problems with sparsity constraints, effort has been put into finding ways of reducing calculations by determining the non-zero support of the solution in advance to solving it. Specifically, so-called variable screening techniques should have lower computational complexity than the problem at hand. Somewhat surprisingly, this is very much possible in the Lasso and Group Lasso settings. El Ghaoui and others [El Ghaoui et al., 2012] proposed SAFE rules, which, in a computational step much less costly than solving the Lasso problem, can identify certain vari-

ables as inactive in the given problem. The later proposed STRONG rules are an inexact version which tendentially removes more variables but is liable to remove variables from the true support. Another type of screening is dual polytope projection (DPP, [Wang et al., 2012]) which also works for the group Lasso. It exploits firm nonexpansivity of the dual problem, which is a projection onto a convex set. Recently, Fercoq and colleagues refined Lasso screening rules to obtain the first set of rules that gives a true acceleration to most practical problems [Fercoq et al., 2015].

Do any of these results extend to analysis sparsity in a straightforward manner? This seems to remain an open question: The fact that the dual problem of an analysis sparsity primal problem implies a vector from the kernel of K^T seems to pose the main difficulty.

Indeed, a dual formulation of the TV-ℓ_1 and Sparse Variation problems treated here is as follows

$$\max_{\mu,\nu} -\frac{1}{2}\|\mu - y\|_2^2 + \frac{1}{2}\|y\|_2^2 - \chi_{\{\|\cdot\|_\bullet^* \leq 1\}}(\nu) - \chi_{\{\lambda K^T \nu = X^T \mu\}}(\nu, \mu), \quad (3.2)$$

where $\|\cdot\|_\bullet$ represents the norm used in the article ($\ell_{2,1}$ here), and $\|\cdot\|_\bullet^*$ its dual norm. Using a decomposition $\nu = K\xi + \eta$, where $\eta \in \ker K^T$, we obtain

$$\max_{\mu,\nu} -\frac{1}{2}\|\mu - y\|_2^2 + \frac{1}{2}\|y\|_2^2 - \chi_{\{\|\cdot\|_\bullet^* \leq 1\}}(K\xi + \eta) - \chi_{\{\lambda K^T K\xi = X^T \mu\}}(\nu, \mu), \quad (3.3)$$

Supposing that $K^T K$ is invertible, which is true in our case, due to the ℓ_1-component, we obtain $\xi = \frac{1}{\lambda}(K^T K)^{-1} X^T \mu$ and can rewrite

$$\max_{\mu,\nu} -\frac{1}{2}\|\mu - y\|_2^2 + \frac{1}{2}\|y\|_2^2 - \chi_{\{\|\cdot\|_\bullet^* \leq 1\}}\left(\frac{1}{\lambda} K^{+,T} X^T \mu + \eta\right) - \chi_{\ker K^T}(\eta). \quad (3.4)$$

The $\ell_{2,1}$ norm employed is a group-wise norm which allows separate consideration of variable groups in the dual. The dual norm of $\|\cdot\|_{2,1}$ is $\|\cdot\|_{2,\infty} = \max_{g \in \mathcal{G}} \|x_g\|_2$. If a variable group in the split variable (derivative space) does not saturate the bound of the dual norm, i.e.

$$\frac{1}{\lambda}\left\|(K^{+,T} X^T \mu + \eta)_g\right\|_2 < 1,$$

then it will be inactive in the primal and the coordinate associated with g equal to 0.

While it would be straightforward to evaluate this property while ignoring the potential effect of η, taking the latter into account is not easy. The variable η can vary freely in $\ker K^T$, thus making it possible to saturate coordinate groups which wouldn't be saturated with $\eta = 0$, or to desaturate coordinate groups which would saturate at $\eta = 0$. In our specific case, where K is a gradient-type operator, the value that η attributes to one coordinate group is immediately linked to that which it associates to its neighbors, thus tying the estimations of activity and inactivity spatially. Concretely, ignoring the identity part of K^T, the rest is a divergence-type operator, whose kernel contains the image of an associated curl-type operator. This intuition makes it possible to see that "closed loops" of 3D displacement steps are in this kernel and can be added to or subtracted from $\frac{1}{\lambda} K^{+,T} X^T \mu$.

In conclusion, bounding $\frac{1}{\lambda} K^{+,T} X^T \mu + \eta$ away from saturation is a difficult task.

3.10 Variation Lasso

For injective analysis operators K, a corresponding primal problem to (3.4) (and thus an equivalent formulation to the TV-$\ell_1$and Sparse Variation primal problems) is

$$\min_z \frac{1}{2}\|XK^+z - y\|_2^2 + \|z\|_{2,1} + \chi_{\mathrm{Im}\,K}(z), \tag{3.5}$$

which amounts to formulating the problem in the split variable $z \in \mathbb{R}^{4p}$, and constraining the split variable to be a feasible output of the analysis operator K (e.g. a spatial derivative). By making the dual problem more strict in supposing $\eta = 0 \in \ker K^T$, we relax the primal problem into omission of the linear constraint $\chi_{\mathrm{Im}\,K}(z)$. In doing so, we obtain a new optimization problem in the split variable z, which amounts to the following group lasso problem.

$$\min_z \frac{1}{2}\|XK^+z - y\|_2^2 + \lambda\|z\|_{2,1} \tag{3.6}$$

Applying K^T to its solution yields what we dub *Variation Lasso*. In this setting, potential long-range cross-talk between variable groups is broken, but short-range smoothness is reinstated by multiplication with K^T after optimization: $u = K^T z_{\mathrm{opt}}$.

While the group lasso screening rules of e.g. [Wang et al., 2012] are now applicable, the end result is less convincing owing to the lack of constraint in a split variable space several times larger than the space of interest.

We compare *Variation Lasso* to *Sparse Variation*, TV-L1 and the Lasso on the data of [Haxby et al., 2001]. For this, we frame the prediction of one of 8 brain states as a one-versus-rest multi-class classification problem. We obtain the best parameter settings for prediction by cross-validation on held-out data and present the average weight vectors over folds for the best parameters. Although predictive performance is a poor proxy for recovery, we would like our method to select plausible weight maps at optimal predictive power.

Take-Home Messages

- Typical fMRI decoding problems using linear classifiers are ill-posed and require regularization. The choice of regularizer impacts the shape of the resulting weight-map in a non-negligible way. All interpretation must be done in awareness of this fact;
- Choosing a convex foreground-segmenting penalty such as *Sparse Variation* regularizes the optimization problem and yields smooth weight maps against a zero background, along with an improvement in classification score over existing regularizers in several settings;
- We provide a novel optimization algorithm with adaptive accuracy in the inner proximal operator, which leads to fast convergence speeds.

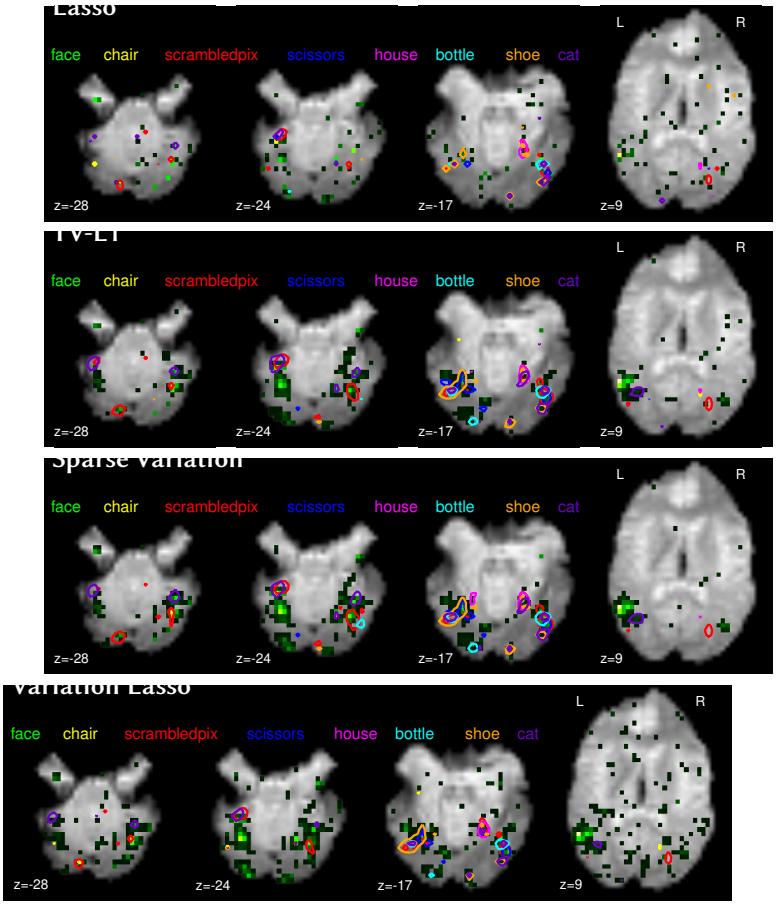

Figure 3.6: Averaged weight vectors of cross-validation folds at best performing parameter values (in the sense of classification accuracy) in multi-class classification on the data of Top: Lasso estimator. Sparse weight vector, no apparent spatial structure. Second: TV-L1. Sparsity and spatial contiguity of weight vectors. Third: Sparse Variation. Sparsity, spatial contiguity and smoothness of weight vectors. Indeed, with respect to TV-L1 are more extended and smoother. Last: Variation Lasso: Sparsity and spatial contiguity, but similar non-spatially-contiguous loadings as in Lasso, due to severing of connections mediated by $\ker K^T$

4 Computer-Vision models

Modern computer-vision is profoundly inspired by biological vision. It is probably fair to say that all of computer-vision is inspired by biological vision on a certain scale of analysis, but some, typically older ideas linked with symbolic approaches to artificial intelligence may have the tendency to seek workarounds to core object recognition as it is known and performed nowadays.

Computer-vision is a vast and rapidly growing field with enormous commerical interest driving its development. In the interest of concision we will restrict our brief introduction only to modern approaches with a strong emphasis on object recognition. Other fields of computer vision concern the video specific action recognition, optical flow estimation and scene understanding - the "where?" counterparts to object recognition as well as many technical applications of image processing such as optical character recognition, satellite and space imagery processing and many imaginable types of surveillance.

The field has seen several eras, where different concepts are predominant. A major transition has been taking place since 2012 with the arrival of large convolutional nets for object recognition. We provide an outline of these developments.

4.1 Classical Computer-Vision Pipelines for Object Recognition

Classically, object recognition pipelines have been modular. It has long been known and exploited that spatial gradients at several scales are relevant feature detectors for object recognition. Apart from the fact that they have also been found to be one of the main constituents of visual area V1 in mammals [Hubel and Wiesel, 1959], it also makes intuitive sense, since any boundary due to 3D occlusion will generate a sharp edge or texture boundary when projected onto a 2D plane. However, edges only do not permit the identification of an object: edges need to be understood in relation to each other. A possibility to integrate this information locally is to obtain e.g. histograms of orientations in patches. However, in general these descriptors vary more or less strongly under translation, rotation, change in pose and lighting of the object. Some of the variability can be addressed using "codebooks" in which one can "look up" a representant of the instance of the patch one found and thus have a comparison between code words instead of descriptors. In the following we give a brief overview of some of the descriptors and aggregators

employed.

4.1.1 Descriptors

Often, images descriptors are mentioned in one breath with the detection/extraction strategies that have originally been associated with them, but they can often be separated. As for detection methods, the main dichotomy is between dense methods, where every point is endowed with a descriptor or keypoint detection, for example scale-space laplacian maxima.

- *SIFT* or *scale-invariant feature transform* [Lowe, 1999] descriptors are histograms of image gradients around a given keypoint and at a selected scale. Image gradients in a patch around the keypoint are binned into a 4x4 spatial and 8 orientation bin histogram, where opposing directions of the gradient are identified, hence omitting phase information. In order to obtain rotation invariance, the descriptor is registered by a rotation to align the strongest orientation to a common angle. E.g. if the strongest orientation is at 45 degrees, then a rotation of -45 degrees is applied to the patch in order to have the strongest orientation at 0.

- *HOG* or *Histogram of Oriented Gradients* [Dalal and Triggs, 2005] is a dense descriptor, which creates histograms of gradients in image patches centered around each point, by binning in space and orientation. Sometimes the binning is smoothed across adjacent spatial and orientation bins. These descriptors are not made to be rotation invariant: Whereas SIFT descriptors are often used for image stitching, HOG features are mostly used for object detection, and objects do not generally occur at arbitrary orientations. Scale is often taken care of by a sliding window search algorithm, which will resize any window to a template size, giving rise to a registration in scale (but usually not orientation).

- *SURF* or *Speeded Up Robust Features* are similar to SIFT descriptors in spirit, but geared towards fast extraction while retaining robustness and specificity properties [Bay et al., 2006]. Around an interest point, determined by a Laplacian maximum computed via derivatives of Gaussians, the main orientation is determined using Haar wavelets. Oriented along this main orientation, a square is extracted and divided into 4x4 regions of 5x5 subregions each. In each of these subregions derivative descriptors are extracted in x and y direction. The responses of the subregions are accumulated over the regions 1) by summing and 2) by summing their absolute values. This results in 4 values per region and thus 64 values overall. Most computations can be carried out using integral images, leading to efficient implementations.

- *DAISY* descriptors are also similar to SIFT descriptors, but avoid histograms and gain speed to an extent where dense feature extraction is straightforward. In effect they replace local histogramming by simple local orientation filtering, implemented as derivatives of a Gaussian pyramid [Tola et al., 2010].

A large-scale comparison of performance between different types of descriptors can be found in [Mikolajczyk et al., 2005]

4.1.2 Agglomeration methods

As described in the introduction to this section, a typical classical computer vision pipeline needs to agglomerate low-level features into more stable representations. There are several inter-related ways of achieving this. They are centered around the notion of "bag of visual words".

- *K-Means* clustering. The archetype of agglomeration methods uses K-Means clustering on extracted features. Setting the number of cluster centers may be an issue, as well as the fact that cluster centers may flock to high-density regions, ignoring small but important regions of the space of descriptors. Nevertheless, a simple K-means clustering on descriptors to set up the system in an unsupervised manner and an association of any new patch to its closest cluster center has proven a good method for generating representations amenable to object recognition. *Soft K-Means* describes a manner of associating a new descriptor to existing centers. It amounts to a matching pursuit with one step: One associates the patch with its strongest scalar product against all cluster centers. This gives rise to a 1-sparse vector, indicating the cluster by the active coordinate and containing the correlation as activation.
- *Sparse Coding Dictionary learning.* Instead of K-Means, one can also perform ℓ_1- or matching pursuit sparse coding on the set of descriptors. This gives rise to a dictionary of reference descriptors such that any candidate descriptor can be well approximated by a weighted some of very few dictionary elements. The sparse weight vector can then be analyzed in a next step.
- *Fisher vectors.* Using a *Gaussian Mixture Model* for the visual word vocabulary, already fit to a training set, one can obtain more fine-grained appartenance measures than just the closest cluster for a given new descriptor. For a new point, taking the derivative of the loglikelihood of the model with respect to all parameters (for a spherical GMM the means and standard deviations) yields a measure of deviance with respect to the model: It encodes how the model can be modified to increase loglikelihood for this point. One can compare two of these vectors using a bilinear form with the Fisher information as Gram matrix, giving rise to the Fisher kernel. Multiplying the Cholesky square root of the Fisher information matrix against the derivative vectors yields the Fisher vector. These are naturally well scaled and amenable to classification [Perronnin et al., 2010].

4.1.3 Classification

After feature agglomeration, the hope is to be able to say that object category is sufficiently linearized in the new representation so that a linear classifier suffices to perform object recognition. To a certain extent this is the case, making it possible to apply a linear classifier such as the margin-maximizing support vector machine. Logistic regression yields similar results. When dataset size permits, support vector machines with non-linear kernels, e.g. the Gaussian RBF kernel are also employed, often leading to better results and thus indicating that the feature extraction method has not fully linearized object category.

4.2 Artificial Neural Networks

Artificial neural networks are layered structures performing very simple, stepwise calculations, usually an alternation between linear functions and pointwise or spatially localized nonlinearities.

The above description is intentionally devoid of any reference to biology and remains very general. However, it is in the study of biological systems that the intuitions for these systems arose. Despite being inherently endowed with temporal dynamics, certain types of neural functionality can be formulated in a static manner. Consider a spiking neuron which integrates over its dendritic input where the contribution of individual dendrites can be weighted differently, even negatively to produce an inhibitory effect, and which fires if a threshold is passed. Allowing for some Gaussian noise in the activations, the probability of a spike can be characterized as a sigmoidal function (a probit, to be specific) of a weighted sum. Replacing the probit sigmoid (the error function) by a logistic sigmoid $\sigma(t) = (1 + \exp(-x))^{-1}$, one obtains the generalized linear model with logistic link function, which, in a machine learning context, can be fit to data using logistic regression.

The constellation of linear functional and logistic sigmoid is the basis of many modern neural networks. Rosenblatt's *Perceptron* is a thresholded version of this building block, with a specialized learning rule to minimize error. General feedforward neural networks contain multiple layers of these building blocks, with several building blocks per layer. All outputs of a given layer serve as input to the building blocks of the next layer. Nonlinearities may vary: sigmoids, hyperbolic tangents, and most recently and very successfully, rectifier functions $x \mapsto x_+$ can be used.

In a learning framework, the goal is to adjust the parameters (the weight matrices and vectors of the linear functionals) such that the outputs correspond to what is expected of the network. In a classification task, the output of a neural network is generally designed to be a vector with dimensionality corresponding to the number of classes with a softmax activation

$$\mathrm{softmax}_i(x) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

. These outputs can be interpreted as probabilities and the error can be quantified as a multinomial cross-entropy: For each sample, the negative log probability attributed to the true class of the sample can be taken as the error and this quantity can be added up for several samples. Using the chain rule for differentiation, one obtains the gradient of the error in the set of weights for each layer and can make a small update to decrease this error. In the neural network literature the calculation of the network weight gradient step by step using the chain rule is called *backpropagation*. The weight updates are obtained by *stochastic gradient descent* (see e.g. [Bottou, 2010]), whereby the gradient of the weights is evaluated in one or a batch of few samples at a time and a small step is taken in the direction of the negative gradient in order to decrease the error. The step size is called *learning rate*. Stochastic gradient descent is justified by the fact that the quantity one would like to minimize is *expected error*, of which an unbiased estimator for any given sample size is the *mean error*. In many cases the gradient operator can be written inside the sum and the loss is separable sample by sample. In this situation stochastic

Artificial neural networks geared to object recognition can be seen as performing a joint optimization of all the steps of the pipeline mentioned previously. That is, if you manage to make it converge to performing operations equivalent to this pipeline. No useful object recognition neural network has been able to avoid restricting the first few linear layers to convolutions. This essentially enforces the extraction of local descriptors and reduces the number of parameters to estimate.

gradient descent can be used to attain a minimum of the error function.

With recent successes in real world applications with important economic interest, development in the field of artificial neural nets has dramatically increased and many variants on architecture and refinements of the learning algorithm have been proposed. Many changes in architecture boil down to constraining the linear transformations at each layer to adhere to a certain form. In [LeCun, 1985] the first network trained by gradient descent using convolutions as the linear operations was introduced to classify hand-written digits. Imposing that the linear transformations be convolutions is a translation of mathematical assumptions into the architecture: Convolutions imply spatial covariation of the output with respect to the input and carry the message that most translated versions of images are still images and to be treated in a similar way. A further, practical aspect is the typically very restricted size of the filter footprint, forcing the linear transformation to extract localized filter responses, while preserving spatial organization of the signal. For natural images, this typically results in the learning of edge, texture boundary and blob detectors. Having several layers of convolution operations followed by pointwise nonlinearities or localized nonlinearities like maximum pooling leads to a parallel treatment of all image patches by the same operations. This reflects biological processing in the sense that visual neurons perform mostly local computations and are organized in a retinotopic manner.

4.3 Biologically Inspired Models

While training by gradient descent is difficult to justify biologically, a hierarchy of levels of processing of visual information has been delineated in [Felleman and Van Essen, 1991]. Biologically inspired models of vision typically attempt to implement the functionality known to exist in some of these processing steps. Thus, LGN cells are modelled as center-on-surround-off cells, often implemented as a difference of Gaussian filters. V1 simple cells are modeled as edge detectors. If using LGN output, they are constructed from output from adjacent LGN neurons arranged along a straight line, otherwise as Gabor filters on image input directly. V1 complex cells pool over the output from spatially adjacent simple cells of the same orientation and scale. After this stage, modeling is much more difficult because the underlying functionality is unclear. Models taking into account correlations between different orientations exist [Freeman and Simoncelli, 2011, Portilla and Simoncelli, 2000]. A class of models attempting to implement visual processing from V1 to object detection along the ventral stream is the class of HMAX models [Riesenhuber and Poggio, 1999]. In architecture, these models are convolutional nets with maxpooling, but the filters used are fixed for the most part. In [Serre et al., 2007], an HMAX model with symmetric (cosine) Gabor filters in the first layer, max pooling across adjacent scale pairs and 2x2 pixel regions on the second layer, followed by convolutions with image templates chosen randomly from the outputs of the second layer but fixed thereafter, topped by a spatial average and an SVM, was able to obtain near state of the art object recognition scores in some metrics as well as near-human performance in a rapid presentation animate versus inanimate distinction task.

4.4 Scattering Transform

The scattering transform is a functional signal transformation developed by Stéphane Mallat and his team to address several known issues of existing signal transformation techniques with respect to invariance and stability of representation.

Indeed, creating a stable representation of a signal which is non-trivial is not an easy task. By *stability* we mean sufficiently regular behavior with respect to smooth deformations. This regularity is expressed as a Lipschitz condition on the size of the deformation. Given a 2D signal $x(u)$, for example an image, and a smooth function $\tau : \mathbb{R}^2 \to \mathbb{R}^2$, a deformation τx is defined by

$$\tau x(u) = x(u - \tau(u)).$$

We would like to obtain a representation Φ of the signal which preserves relevant information while being robust to deformations. *Relevant* is a term that needs to be defined. Indeed, in the case of images, object identity can be the information of interest, while its position may not be important for the analysis. Often, the transformations that are not of interest can be endowed with a group structure. Position, for instance, can be encoded by the group of translations. Let G be the group of translations and $g_v \in G$ act on a position $u \in \mathbb{R}^2$ as $g_v u = u + v$. Then we can define the signal transformation

$$g_v x(u) = x(g_v^{-1} u) = x(u - v).$$

A signal representation that is *invariant* to a group G has the property that

$$\Phi(gu) = \Phi(u) \quad \forall g \in G.$$

A signal representation that is *covariant* to a group G has the property that

$$\Phi(gu) = g\Phi(u) \quad \forall g \in G,$$

where the action of G on the image of Φ needs to be defined. For translation, if the image of Φ has spatial structure, then the elements of the group can act on that spatial structure.

While invariance and covariance properties for "small" groups such as translations and rotations can be very useful, a full invariance for smooth transformations is detrimental. Indeed, in the case of object recognition in images, a smooth deformation can transform certain object categories into others, thus losing crucial information. Almost only topological aspects of the object class can be preserved.

However, *stability* with respect to small deformations is vital. By *stability* we mean that the distance of representations between the signal and the deformed signal must be proportional to the "size" of the deformation: Small deformations must incur small representation change and arbitrarily large deformations may incur larger representation change. A meaningful way of quantifying the "size" of a smooth deformation is the norm of its gradient. Indeed, if $\|\nabla\tau\| = 0$, then $\nabla\tau = 0$ and $\tau = \text{const}$. Thus the deformation amounts to a rigid translation, whereas a nonzero norm of the deformation gradient can lead to local volume change and more generally other types of distortions. Stability with respect to deformation can thus be expressed by

$$\|\Phi(\tau u) - \Phi(u)\| \leq C(u)\|\nabla\tau\|. \tag{4.1}$$

For example, if Φ is linear in the gradient of the deformation field, then the stability property is immediate (given $\dim \operatorname{Im} \Phi < \infty$ or Φ bounded). Further, the representation should be continuous in the signal. A stronger property, which is usually imposed, is called the Lipschitz property and amounts to $C(u) = C\|u\|$ – a change in representation is bounded by a constant factor times a the change in signal.

Condition (4.1) also implies translation invariance immediately, because then $\nabla\tau = 0$ and thus $\Phi(\tau u) = \Phi(u)$ follows.

It is not straightforward to find nontrivial image representations satisfying the stability property and invariance to translations. In [Mallat, 2012], some examples are given. For example, the Fourier transform encodes translations as phase shifts in the representation. Thus, taking a complex modulus totally removes any position information. It is thus a translation-invariant representation. However, arbitrarily small scaling deformations cause an arbitrarily large difference in high frequencies, increasing proportionally to frequency. This behavior violates the stability criterion. On the other hand, sufficiently regular and localized wavelets satisfy a deformation stability property. However, wavelet transforms, by construction, are translation covariant, not invariant. A way to create a translation invariant representation out of a translation covariant one is to integrate the representation over all translations. This is true in general, for any group: to make a group-covariant representation group-invariant, calculate its integral over the group. For wavelets this operation is problematic, because it leads to trivial (constant) representations. When using wavelets, it is thus crucial to integrate over a nonlinear function of the wavelet transform. In [Bruna, 2013] it is shown that stability with respect to deformations necessitates the nonlinear function to be applied pointwise. If in addition one requires the transformation to conserve signal energy it is necessary that the pointwise nonlinearity be the complex modulus.

Satisfying the above properties gives rise to the scattering transform. Given a complex wavelet ψ, for example a Morlet-filter and a low-pass filter ϕ, the translation covariant version of the scattering transform can be written as a cascade of convolutions and complex moduli. By writing $\psi_{s,\gamma}$ for an orientation γ and a scale s[1], for the first layer we can write

$$U[s,\gamma]x = |x * \psi_{s,\gamma}|.$$

[1] The rotated and scaled version reads $\psi_{s,\gamma}(u) = 2^{-s}\psi(2^{-s}R_{-\gamma}u)$ for $R_{-\gamma}$ a rotation by angle $-\gamma$.

In order to make this translation covariant representation invariant, one can integrate it over space to obtain $S[s,\gamma]x = \int |x * \psi_{s,\gamma}| \mathrm{d}u$. One can attenuate global translation invariance to local translation invariance, which gives rise to an output of the form

$$S_J[s,\gamma]x = |x * \psi_{s,\gamma}| * \phi_J,$$

where ϕ_J is the low-pass filter from above at scale J. Both global and local translation invariant representations lose high frequencies, which can carry important information. In order to recover these, they are measured using a second wavelet transform, also followed by a modulus, which reads

$$U[s_1,\gamma_1,s_2,\gamma_2]x = ||x * \psi_{s_1,\gamma_1}| * \psi_{s_2,\gamma_2}|,$$

for s_1,γ_1 the parameters of the first layer wavelet transform and s_2,γ_2 the parameters of the second wavelet transform. Here again, the local translation invariance is obtained by low-pass filtering with ϕ_J:

$$S[s_1,\gamma_1,s_2,\gamma_2]x = ||x * \psi_{s_1,\gamma_1}| * \psi_{s_2,\gamma_2}| * \phi_J.$$

It becomes immediate to relate this architecture to convolutional networks. As a matter of fact, the scattering output of the low-pass filtered first and second layers is the output of a convolutional network, where the convolutions are performed with fixed, mathematically motivated wavelets.

5 Analyzing human visual responses to textures

This work was published in the proceedings of the Pattern Recognition in Neuroimaging conference

- M. Eickenberg, A. Gramfort, B. Thirion *Multilayer Scattering Image Analysis Fits fMRI Activity in Visual Areas* Proc. Pattern Recognition in Neuroimaging, 2012
- M. Eickenberg, F. Pedregosa, M. Senoussi, A. Gramfort, B. Thirion *Second order scattering descriptors predict fMRI activity due to visual textures* Proc. Pattern Recognition in Neuroimaging, 2013

5.1 Introduction

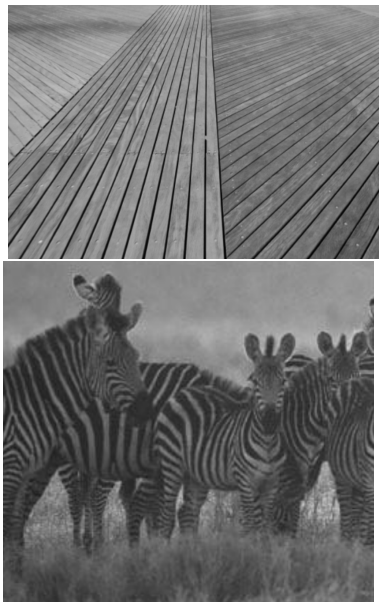

Figure 5.1: Top: Exerpt of picture of Coney Island Boardwalk taken from [Landy, 2013]. The boards of different orienations form a texture boundary by e.g. orientation contrast, not luminance contrast. Bottom: Zebras, taken from [Landy, 2002]. Individual zebras have texture boundaries with background and with other zebras, which are of different order.

This chapter is about *second order* models for image analysis. The term "Second Order" here refers to a second stage in processing after linear filtering a rectification of an image. "First order" models, i.e. linear filtering and rectification are very limited in capacity. They could not detect anything other than a linear template matching in an image and are not very selective: An orientation filter may still respond to a contour with orientation perpendicular to its preferred orientation or at a different scale. They cannot, by themselves, detect the beginning or the end of a contour. Even if it may remain implicit, object recognition and scene understanding rely heavily on *segmentation*, which should ultimately be semantically informative. By pure probability, a fair amount of occlusion borders are detectable on first order, by difference in luminance or possibly color. This is however not always the case, and more subtle texture boundaries can be just as important. These are generally called "higher order" if they cannot be identified by simple luminance contrast edge detection.

It is important to note that the term "order" is semantically overloaded and can mean different things. Another relevant meaning of *order* is the number of variables used for statistical analysis: Supposing that a texture is an instance of a 2D stationary process [1], one can see the probability distribution of "pixels" as first order statistics. Second order statistics are given by the joint distribution of all couples of points separated by a given distance vector. Third order statistics are characterized by the joint distributions of

[1] A stationary process is a continuous collection of random variables X_{st}, that are arranged spatially. Stationarity means that their distribution does not depend on absolute, but relative position: $(X_{s,t}, X_{s',t'}) \sim (X_{s-s',t-t'}, X_{0,0})$.

three points arranged in a specific triangle shape (given by two relative displacement vectors). Generally, *n-gon* statistics are characterized by the joint distributions of n points as a function of their relative positions to another. It is this notion of *order* which Julesz used to generate texture images that have regions differing on a certain statistical order [Julesz, 1981].

Here, unless otherwise stated, *order* will be taken to mean number of layers necessary for analysis.

There is a rich body of literature on the segregation of textures, mostly of psychophysical and computational nature, but also including electrophysiology and fMRI experiments. [Landy, 2013] provides an excellent overview.

In [Julesz, 1981], Julesz proposed the "texton" as an elementary unit of texture perception, from which the properties of "segregability" between two textures could be deduced. Textures differing in these properties could be effortlessly and pre-attentively segregated. Among these putative units could be edges, blobs, possibly end stops.

However, it was shown in [Nothdurft, 1991] that it is not so much the regional difference of textons that plays a role, but rather local phenomena such as orientation contrast.

As depicted in figure 5.2, they constructed an image that would be seen as a rhomb if the relevant features were absolute orientation and a square if the relevant features were local relative orientation. *Textons* subsequently disappeared from the study of perception, but the concept was revived for computer vision a decade later, e.g. by Malik [Leung and Malik, 2001]. The texton approach is to find "building blocks" or "atoms" of texture with the goal of ultimately being able to *synthesize* any texture. As stated above, our goal is to study putative *analysis* methods in the attempt to explain the way our visual systems process this texture information. The typically studied analysis methods attempt to model more or less in detail the properties of mammal visual systems, sometimes starting from LGN-type center-surround units [Thielscher and Neumann, 2003] and sometimes from V1-type edge detection, stylized using e.g. Gabor or Morlet filters. What is crucial to be able to accurately capture behavioral results from psychophyics experiments is not to stop at linear rectified edge detector type models, but to extract higher level information in a next step. This type of model has been named FRF (for *Filter-Rectify-Filter*) or LNL (for *Linear-Nonlinear-Linear*): Given a point-wise nonlinearity $n(z)$ and two filters ψ_1, ψ_2, its output can be written as $\mathrm{FRF}(x) = \psi_2 * n(\psi_1 * x)$. Crucially, this type of analysis model is able to segregate the above texture patterns just as it pops out to the eye. Usually, the nonlinearity is chosen to be a rectifier unit $n(z) = (z)_+$. Similarly, the scattering transform introduced in chapter 4 is of the FRF form for $n(z) = |z|$ and it is the transformation we shall use in our experiments. Pointwise nonlinearities can also be extended to simple local nonlinearities such as *max pooling*, which has also been used in biologically plausible models of vision [Riesenhuber and Poggio, 1999].

It is to be noted that the FRF models, while capable of capturing a variety of texture boundaries, cannot account for all texture segregation functionality attributable to the visual system. A texture devised by Ben-Yosef and Ben-Shahar [Ben-Yosef and Ben-Shahar, 2008] shows constant change in curvature across the image, but still gives rise to percepts of global contours (see

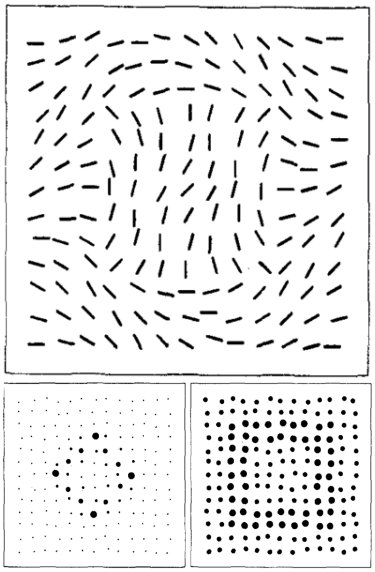

Figure 5.2: Line texture, taken from [Nothdurft1991]. If the visual system responded to equally oriented lines, the perceived figure would be a rhomb. If the visual system responded to orientation contrast, the figure would show a square. The latter is the case.

figure 5.3)

Many of the investigations performed to date are concerned with the capacity of visual systems to segregate images into different regions of texture and have been able to delineate interesting results as seen in this paragraph.

However, the contemplation of a uniform texture, not the border between two different ones, also evokes a visual representation, making it possible to say whether two textured regions are "made of the same stuff" or not. The representation may recruit functionality essentially overlapping with that which is used for segregation.

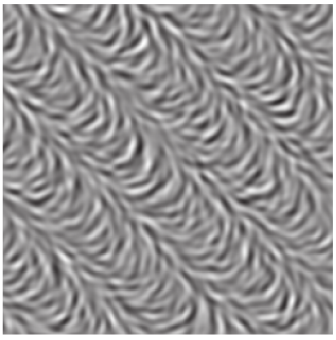

Figure 5.3: Texture of constant curvature change, taken from [Landy, 2013], originally from [Ben-Yosef, 2008] giving rise to percepts of global contours. An FRF model cannot identify these contours.

Most studies of biological vision relative to textures use highly controlled synthetic texture images built from very simple primitives. In [Portilla and Simoncelli, 2000], Portilla and Simoncelli were able to provide a minimal characterization of a wide range of natural texture types by extracting a certain number of statistical descriptors from the images. New texture images were generated from noise by applying gradient descent until the descriptor values matched those of a given other texture. This resulted in images of very similar appearance to the texture from which the descriptors had been extracted. Minimality of the representation was shown by removing each one of the descriptors in turn and re-synthesizing textures. Each omission had a strong perceptual effect on at least one class of textures, permitting the conclusion that at least the descriptors presented, or an equivalent set, are necessary to provide constraints that guide the generation of perceptually equivalent images.

Existing studies in fMRI do not employ natural textures or seemingly natural generated textures. Instead, they test the visual system with well-controlled synthetic texture images in order to analyze e.g. responses to second-order boundaries. When naturalistic textures are used, the analysis is focused on contrast maps and not fine-grained modeling, as we propose here.

[Cant et al., 2009] uses fMRI adaptation in an experiment to distinguish the effects of shape, texture and color of a stimulus, varying only one dimension at a time. Stimuli are four fake 3D objects endowed with four different types of textures and four different colors. They identify shape effects in lateral occipital complex and texture specific processing in collateral sulcus.

[Kastner et al., 2000] uses oriented line segment textures, oriented at 45 or 135 degrees in contrast to textures consisting of the same types of lines but forming texture boundaries giving rise to shapes of squares. While V1 did not show any difference in activity towards the two texture types, ventral V2, V3, V4 as well as dorsal V3A did show differential activity. The subjects' attention was diverted by a counting task of foveally presented letters while the texture presentation was restricted to the upper right quadrant of the visual field.

In [Larsson et al., 2006, Hallum et al., 2011] the sensitivity of human visual cortex to second order texture modulations is investigated using an fMRI adaptation paradigm. Grating-type textures as carriers which modulate smaller scale textures of contrast or orientation at different spatial frequencies are presented to subjects. While [Larsson et al., 2006] focuses on second order orientation selectivity, [Hallum et al., 2011] studies second order spatial frequency selectivity. The responses of visual areas V1, V2, V3 and V4 captured in [Hallum et al., 2011] by adding a normalization component to the classical

FRF model to obtain a filter-rectify-normalize-filter model. The study shows that V1 also is to some extent distinctive of second order texture variations.

In [Montaser-Kouhsari et al., 2007] it is shown that even texture-induced illusory contours are detected in V1 and that responses to these contours increase along the visual processing hierarchies.

In a different line, multi-modality of textures has also been studied, see [Whitaker et al., 2008] for a review on the topic.

Using fMRI, we propose to study second order image analysis models of the scattering transform type on two types of natural images: Every-day photos and images of uniform texture. The former permit the study of texture boundaries and representations in a natural context. The latter focuses on the way single uniform texture types are processed in the brain. Our main approach will be an encoding model, as described in chapter 1. The experiment on textures images is also amenable to classic analysis and reverse inference (decoding). We will compare two contrasts, one showing voxels responding in any way to texture and another showing voxels that respond differently to at least one of the textures. The decoding analysis will be used to predict texture class from different brain regions.

5.2 Experimental Setup

The fMRI BOLD response to visual stimuli was acquired during a visual comparison task, where subjects were asked to distinguish between images of six different texture classes. The study was of the rapid event-related type.

5.2.1 The experimental task

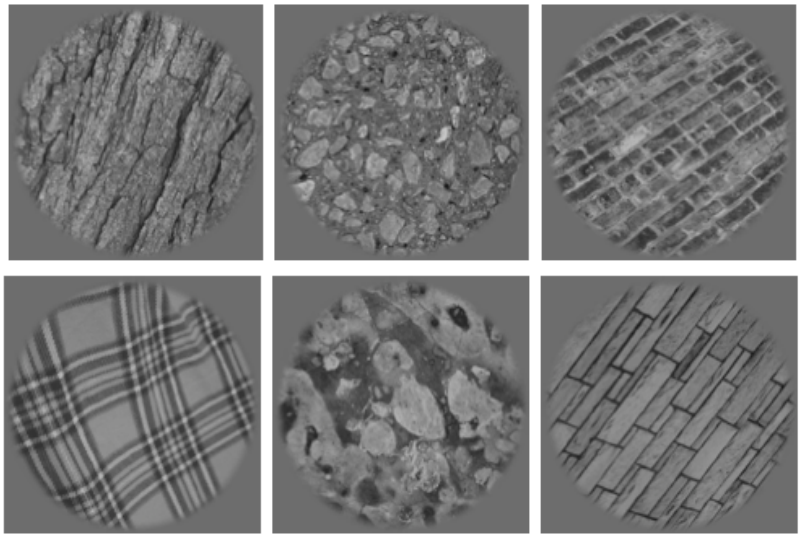

Figure 5.4: Sample stimuli used in the experiment: Extracts from the UIUC data set from [Lazebnik, 2005]. Representants of the six texture classes are shown inside the circular stimulus mask.

Stimulus images were circularly masked gray texture images of 120x120 pixels projected onto a screen situated behind the magnetic bore of the fMRI scanner and viewed by the subject via a mirror placed in their visual field. The circular stimulus spanned 14 degrees of field of view. Mean and standard

deviation of the pixel values of the images were fixed to 128 and 32 respectively for each image, where 0-255 is the full possible range.

Stimulus images were taken from the texture database of [Lazebnik et al., 2005] by taking random least overlapping crops of 120x120 pixels, applying a circular mask and a random rotation sampled uniformly from all angles of the circle.

Subjects were asked to compare two texture instances from the same class, presented one after the other, while fixating a central cross. One experimental block took 12 seconds: At second 0, the first image was presented for one second, flashed three times in an on-off-on-off-on sequence of 200ms duration each. At second 4 the second image was presented in the same manner. At second 8, a smaller image, centered around the fixation point, was presented, containing an extract of either the first, the second, or an unrelated image. The subject was asked to press a left-hand button if the first image had been repeated, the right-hand button if the second image had been repeated and no button if the image was unrelated.

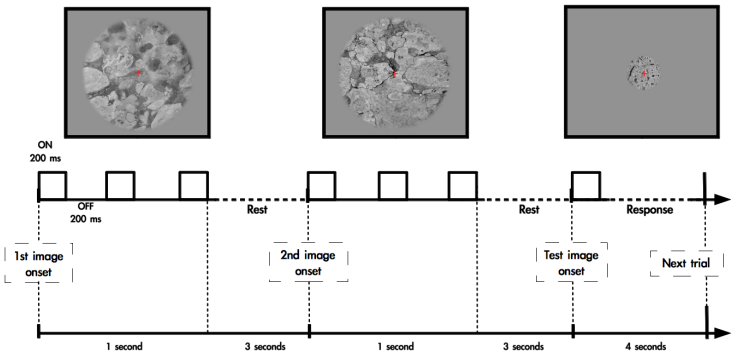

Figure 5.5: Visualization of one 12 second block presenting 2 full images and one task related image extract. At seconds 0 and 4, two different texture image of a given class are flashed to the screen. At second 8, a foveal exerpt is shown of image 1, image 2 or an unrelated image. The task is to decide which one it was.

One experimental session consisted of 36 such 12s blocks, which corresponds to the presentation of 72 images. All texture images were presented twice: once in first position, once in second position, in order to be able to account for effects due to ordering and in order to increase the signal to noise ratio in subsequent analyses.

One scanner session consisted of 6 experimental sessions. Thus a total of 216 distinct texture images were shown.

The task was presented to the subjects once before entering the scanner, with stimuli not used during the acquisition.

5.2.2 Measurements

Functional images were acquired on a 3T Siemens scanner (TR=2400ms, TE=30ms, matrix size 128×128, FOV 192mm$\times$192mm). Each volume consisted of 34 2mm-thick axial slices without gap, with an in-plane resolution of 1.5mm$\times$ 1.5mm. Anatomical T1 images were acquired on the same scanner with a spatial resolution of 1mm $\times$ 1mm $\times$ 1mm. EPI data acquisition was performed using the sequence in [Boegle et al., 2010] with IPAT=2; this sequence includes distortion and motion correction at the acquisition level. Slice timing correction and coregistration to the anatomy were performed using SPM8 software wrapped by `pypreprocess` [2].

Data were acquired from three subjects in two sessions each, using two

[2] http://github.com/neurospin/pypreprocess

different stimulus sets for the two sessions.

5.3 Data analysis methods

Several data analysis methods were performed. Classical statistics were used in order to estimate effects of texture type regressors and of differences in response between texture types. Reverse modelling (decoding) with texture class as target was employed globally and region by region in order to assess minimal information content of the acquisition pertaining to this target. An encoding model based on the scattering transforms of the texture images was evaluated and a contrast between layer 0, 1 and 2 performance versus layer 0 and 1 performance was computed[3].

[3] Scattering layer 1 contains smoothed wavelet moduli, layer 2 contains smoothed wavelet moduli of unsmoothed layer 1, layer 0 merely contains the smoothed input (low passed signal).

5.3.1 Data preprocessing

In order to disambiguate responses to consecutive image presentations, to slightly reduce the dimensionality and raise the signal to noise ratio, preprocessing of the time courses was performed in the form of a general linear model (GLM). This yielded an activity map for each unique image, implicitly averaging the responses to the two presentations. The two forms of GLM employed were a classic GLM with fixed HRF and an event by event GLM [Turner et al., 2012] - see also chapter 2) of this thesis. However, the rather large TR of 2.4s made it less interesting to model the HRF explicitly as done in chapter 2. Using the GLM approach we can restrict ourselves to predicting a one-dimensional activity per voxel and image instead of a whole time course.

5.3.2 Classical statistics

For the texture experiment, F-statistics were obtained for two different contrasts. The statistical map *fx_interest* identifies locations that show significant activation when a texture image is shown (versus baseline). The statistical map *fx_diff* indicates regions where the response to at least one texture class differs significantly from the mean response to textures. At the same significance level, the map *fx_diff* is a subset of the *fx_interest* map which indicates where texture information is encoded differently according to class.

5.3.3 Reverse modeling

In order to localize regions that encode different texture classes differently, we performed reverse modeling using machine learning techniques. After selecting the regions of interest V1, V2d, V2v, V3v, V3d, V4v, hV4, V3A/B, IPS0, using an ROI atlas on the *fsaverage* surface [Henriksson et al., 2012], we proceeded to fit a logistic regression classifier to predict texture class. The multiclass situation was handled using a one-vs-rest classification, and nested cross-validation was performed on a leave-one-session-out basis to set parameters and obtain mean scores across folds.

5.3.4 Forward modeling with scattering transform

Using Morlet wavelets, the two wavelet layers of scattering transform (and layer 0 - local averaging using a low pass filter) were applied to the texture

stimuli and the stimuli of [Kay et al., 2008]. The finest scale filter, the one with the highest spatial frequency, was chosen to have $\frac{3}{4}$ of the Nyquist rate. Eight orientations were used and five scales, each separated by an octave. A cross-validated ridge regression was performed to assess the predictive power of scattering layers 0, 1 and 2 combined versus only scattering layers 0 and 1. In addition to this, texture class regressors were added in the case of the texture experiments, since these were a strong confounding factor: Scattering coefficients permit easy linear separation of texture classes, hence regions selective only to texture class will be very well driven by these coefficients. Since we are interested in modeling low-level features, we strive to separate this out.

The regression models corresponding to the two scattering coefficient sets are evaluated using predictive r^2 as scoring on the outer loop of a nested cross validation, where the inner loop is used to select optimal parameters. Predictive r^2 represents explained variance on held out data, i.e. to which extent mean l_2 prediction error is smaller than the overall variance of the test set. Due to finite data, possible mis-estimation of the intercept, and bias in the model, the null distribution of this score can be centered around a negative value.

In order to evaluate the excess predictive power of the two-layer model, we calculate the differences of scores. Obtaining a meaningful threshold on this value is not obvious. Here we choose to construct an empirical null distribution and threshold according to a false discovery rate of 1% [Benjamini and Hochberg, 1995]. The distribution of score differences is unimodal and slightly right skewed, where the right tail represents points where the two-layer model has an excess score with respect to the one layer model. The null distribution is created by reflecting the left-hand side of the distribution around the mode. Using the ratio of the cumulative density functions of the score difference distribution and the empirical null, the cutoff threshold is determined at an FDR of 1%.

5.4 Results

We discuss the results of the described data analysis methods applied to the texture experiment or both the texture experiment and the natural images experiment. Starting with a classical statistical analysis using GLM contrasts on regressors representing texture class, we obtain an overview of which regions respond to texture image presentation and which regions respond differently to different textures. We follow up with the reverse or *decoding* model, which is applied to regions of interest as defined by a probabilistic atlas of the visual areas [Henriksson et al., 2012]. This analysis can show which regions contain straightforward representations (i.e. linearly separable and with relatively little noise) of the target variable texture class. Next we compare the more fine tuned forward models based directly on the stimulus images. They are conceived in a hierarchical manner, any given level containing all of the lower levels, and consist of texture class regressors, zeroth layer output (smoothed image), first layer scattering coefficients (smoothed Morlet filter moduli) and second layer scattering coefficients.

5.4.1 "fx interest" vs "fx diff"

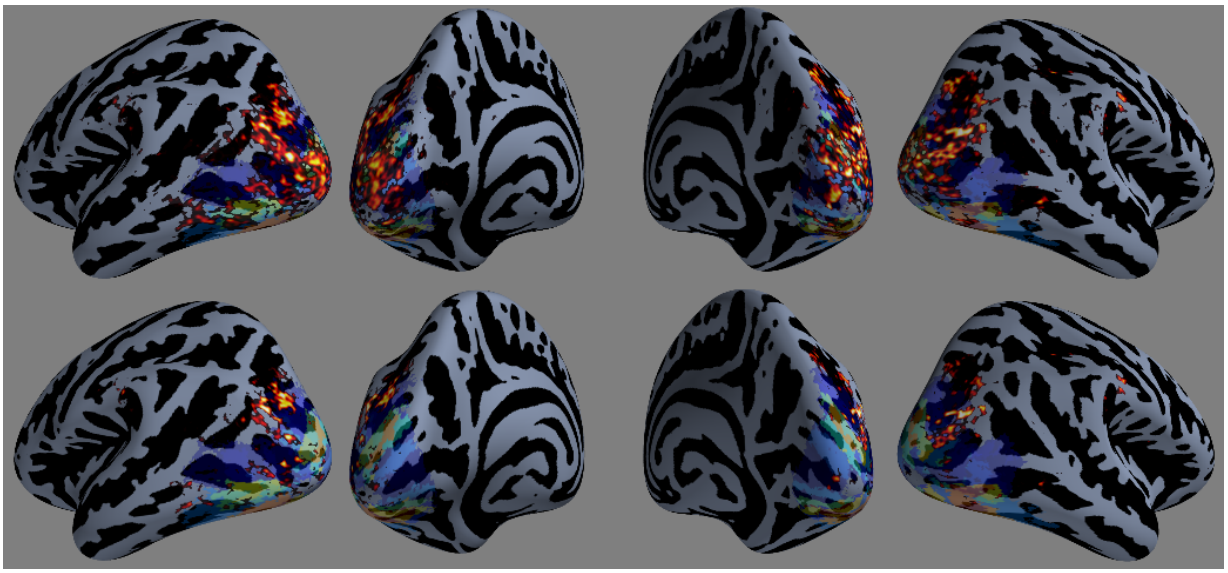

Figure 5.6: Two contrast map z-scores averaged across all sessions of all three subjects, thresholded at z=2 ($p=0.5 \times 10^{-6}$ uncorrected). Top: Contrast *fx_interest*, corresponding to significant activation elicited by texture images. Primary visual areas and both the ventral and dorsal visual stream are strongly activated. Bottom: Contrast *fx_difference* showing areas that responded significantly differently to the six texture classes. One observes that lower level visual areas are much less activated in this sense. Dorsal visual areas V3A/B and parts of the IPS seem to respond most differently to these visual cues.

The top part of Figure 5.6 shows significant activation of the contrast *fx_difference* on a group level fixed effects model (n=6) above a z-score threshold of 2 ($p=0.5 \times 10^{-6}$ uncorrected). We observe strong activation on the dorsal side of the visual stream and significant activation on the ventral side of the visual stream, in both hemispheres. Visual areas V1, V2d, V2v, V3v, V3d, V4v, hV4, V3A, as well as parts of IPS show significant responses to this visual stimulation.

However, the situation is completely different when evaluating which regions respond differently to the six texture classes. Most of the lowest level visual areas and most of the ventral stream do not exhibit significantly different activation across texture classes. However, the dorsal regions remain strongly active in the sense of this contrast.

5.4.2 Area specific reverse modeling

Reverse modeling results are shown in Figure 5.7. A logistic regression in a one versus rest setting decodes well above chance level ($\frac{1}{6}$) in accuracy score and slightly differently depending on the chosen regions. The highest accuracy scores are achieved in dorsal and lateral visual regions, reflecting the same tendency already observed in *fx_interest* and *fx_diff*.

5.4.3 Forward modelling

The forward modeling comparison of scattering transform layers was performed on the textures dataset as well as the natural images dataset. Apart from some patches in V1, a gain in prediction by using layer 2 is observable mostly in extrastriate areas, both along the ventral stream, but also dorsally, as far as IPS.

This is true for both the natural images dataset and the textures dataset, where it must be re-iterated that the texture data were analysed with a supplementary texture class regressor in order to factor out effects of texture

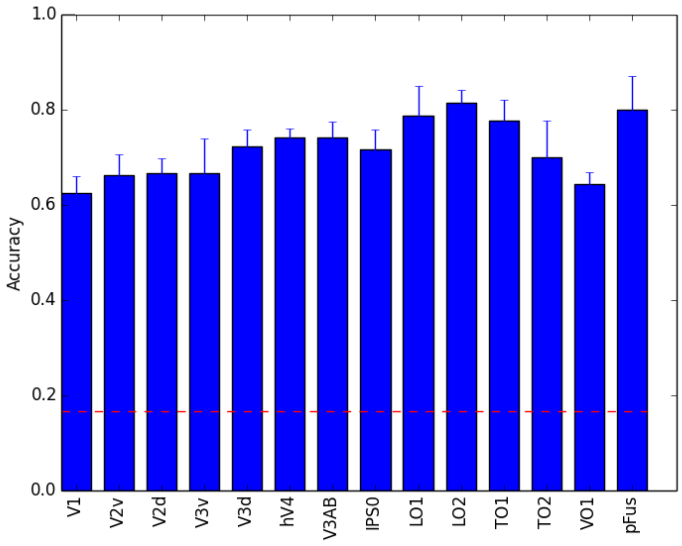

Figure 5.7: Bar diagram showing decoding scores with target variable texture class in a one vs rest setting. The decoding is performed on brain regions extracted using the predefined atlas [Henriksson et al., 2012]. The scoring function is prediction accuracy, with chance level at $\frac{1}{6} \approx 0.167$. Error bars indicate variance across subjects.

class alone and study the property of the local image descriptors. For the natural images the areas benefitting from the inclusion of layer 2 descriptors include transverse occipital sulcus and inferior IPS, both associated with scene perception, as well as specialized extrastriate areas such as probably the occipital face area and the extrastriate body area. Photos of scenes and photos of persons are abundant in the natural images dataset. Layer 2 seems to add specificity and permits a better fit of brain activity due to these complex concepts.

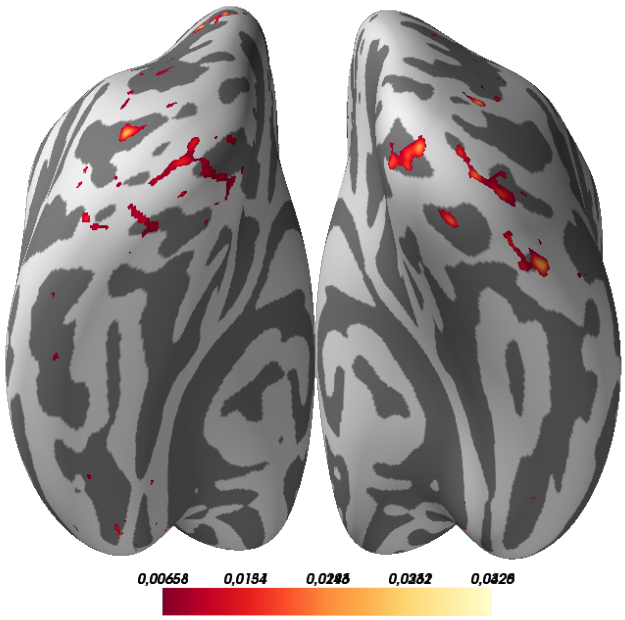

Figure 5.8: Scattering layers 0, 1 and 2 with texture class vs scattering layers 0 and 1 with texture class at FDR=1%. Excess predictive score is present in low level visual area V1, as well as higher level visual areas dorsally, ventrally and laterally.

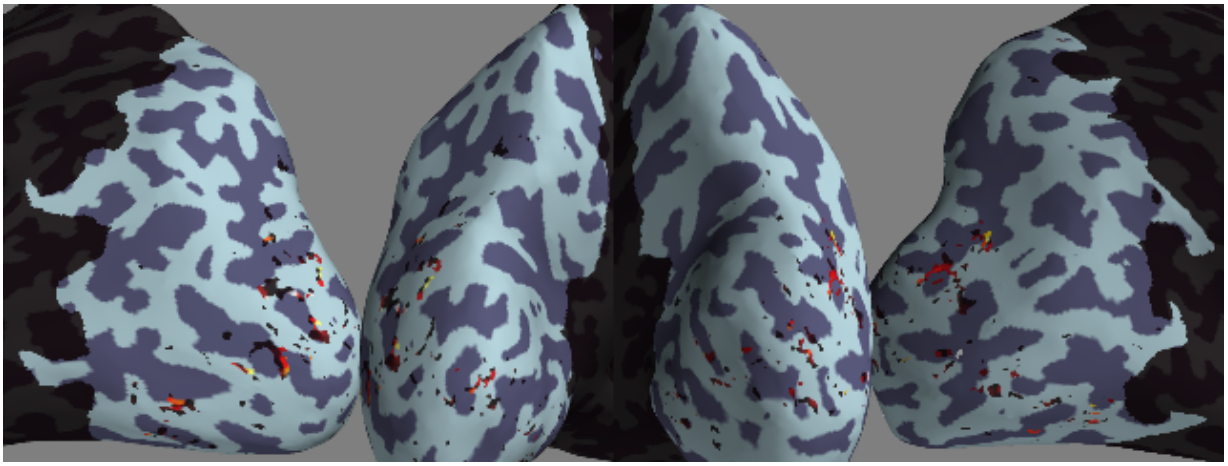

Figure 5.9: Scattering layers 0, 1 and 2 vs scattering layers 0 and 1 at FDR=1% for natural images data. Low level visual areas are well modeled by scattering layers 0 and 1 alone. The added value of layer 2 becomes visible in lateral and dorsal visual areas.

5.5 Discussion

Visual texture classification from neuroimaging data To the best of our knowledge, the presented work is the first one that presents a statistical analysis of the functional correlates of natural visual textures using fine-grained forward modeling. If we compare this setting to more classical object viewing, this entails a double challenge, related to the fact that the stimuli do not present clear outlines: *i)* the association of these images with semantic content is sometimes ambiguous, because two different textured patterns may look similar after some rescaling or the same texture type may appear in several contexts, and *ii)* the contour shape at luminance boundaries or perceivable texture boundaries is an essential clue (for instance, a trunk carrying a wood texture is typically elongated in one direction) which were not present here. Nevertheless, the classes were easily separable. Note that in the texture experiment, the subjects were naive to the existence of six latent categories.

While the luminance normalization, variations in position, and rotation angle could rule out the use of local linear mappings of image intensity (such as gradients) as a means to discriminate between these textures, the situation is actually more complex. Scattering layer 1 alone (smoothed wavelet moduli, which can capture local luminance changes) is actually discriminative of the different texture classes, showing that the first-order statistics are to some extent sufficient (see figure 5.10). However, the adjunction of layer 2 improves the classification accuracy significantly, showing that more complex image features actually help discriminating texture classes. Figure 5.10 shows classification performance for cross-validated logistic regression on scattering transforms of varying parameters as a function of training set size. Layer 0 only feature scores are colored blue. Since they only contain local averages of the normalized texture images, the scores do not lie above chance. The conjunction of layers 0 and 1 is shown in red and achieves well above chance linear separation of texture classes. Adding layer 2 to 0 and 1 yields the green lines, of which all except one yield strictly better classification accuracy than layers 0 and 1. The roto-translation invariant scattering transform, depicted in turquoise, which recombines layer 2 coefficients using a wavelet in the angular variable while also applying rotations to the image, is the best linearization of texture class - only very few samples are needed to properly

train a classifier.

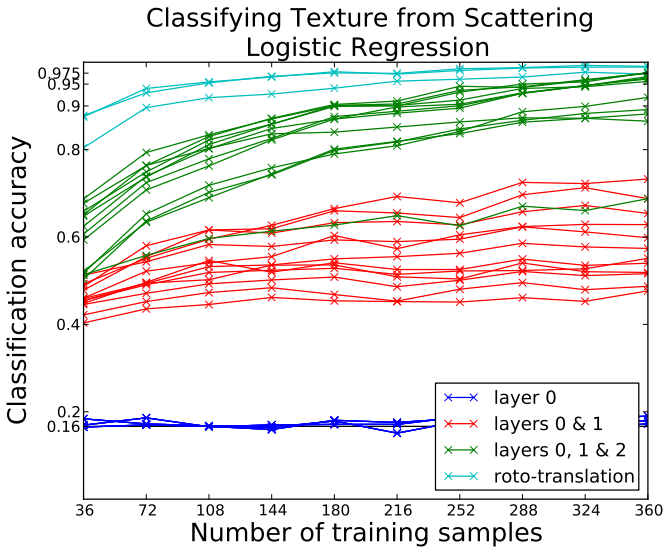

Figure 5.10: Classification of textures from their scattering transforms using logistic regression. For different combinations of orientation and scale, we show classification scores on several training set sizes of translation invariant (spatially averaged) scattering coefficients of layer 0 (spatial average), layers 0 and 1 (spatial average and first level wavelet modulus average), layers 0, 1 and 2 (second layer wavelet moduli average in addition) and roto-translation-invariant scattering (layer 2 integrated to be invariant to roto-translation.)

From the point of view of the FRF framework presented in the introduction, we must be careful about how to situate these classifying operations. Indeed, classifying linearly from a representation of a first order model, i.e. a filtering followed by a rectification, results in a linear-nonlinear-linear chain, where the last linear transformation is the dot product with classifier weights. It is not, however a filter-rectify-filter operation, since the last operation is not a spatial filtering. While the classifier weight dot product operation can yield functioning texture classfication, it collapses all of space and cannot make localized decisions on the presence or absence of a texture boundary. The second layer of the scattering transform is a true FRF model, followed by an output nonlinearity, the complex modulus. It is capable of detecting the texture boundaries an FRF model can detect. Adding the classification step on top collapses spatial information as before.

These distinctions may seem far-fetched at first, but it is crucial to make the difference between localized an global operations. Note that the first layers followed by classification would become an FRF operation if the classifier were applied locally at all locations using the same weight vector. This would result in a convolution (filtering) with a decision function and would make texture boundaries localizable by simple thresholding of the filter output. Achieving an *image* of a potential texture boundary is what distiguishes FRF from filter-rectify-classify.

Resulting topographies Figure 5.6 illustrates an expected effect, namely that the strong activations elicited by texture viewing are markedly reduced when only the differential effect of the textures classes is considered: low level visual areas (V1-V2) respond strongly to visual textures, but in a way that is not discriminative across classes. In contrast, the next areas in the dorsal pathway (V3A/B, IPS0) show different responses across texture classes. More surprising was the weakness of strong differential responses in the ventral visual cortex, that is known to respond strongly and discriminatively to different object categories. For higher level ventral regions this may in part

be explained by the absence of actual objects as defined by texture boundaries in the stimuli. However, Freeman and others in [Freeman et al., 2013] were able to show differential effects against Fourier scramble in early visual ventral regions. By most models of the early visual system these should also be expected, so further investigation into the absence of activation here is necessary.

Interestingly, a spatially confined yet significant discriminative response was observed bilaterally in postcentral sulcus, consistently across all subjects. Close inspection of subject pf120155, for whom the fMRI field of view contained the deepest part of the horizontal segment of the intra-parietal sulcus (the other two subjects' fMRI field of view did not contain any part of the horizontal IPS), revealed activation along the full length of the IPS into postcentral sulcus. This may be attributed to differential attentional effects across texture class. It is at least plausible since the difficulty of the discrimination task varied with texture type.

If we consider within-subject classification scores, the picture is slightly more complex. Texture-discriminative information can be found at the individual level in all brain regions, even though there is a tendency toward higher prediction accuracy along the visual hierarchy from V1 to dorsal (V3AB, IPS0), lateral (LO1/LO2) and, to a lesser extent fusiform regions.

More importantly for us, the topography of brain regions that are better linearly predicted by a model with two scattering layers than a model with one scattering layer displays a few clusters in the same dorsal, lateral and fusiform regions (Figure 5.8). Most importantly, this result is replicated in a completely different setting, where the subject views natural images that are not specifically tied to textures (Figure 5.9): IPS0, V3AB and LO1/LO2 contain again several clusters that are better modeled with a second layer of scattering, unlike lower-level regions. Ventral visual regions (V3v, V4) do not exhibit this effect in either of the datasets.

Recent findings After the acquisition and initial analysis of the present experiment, it was shown in [Freeman et al., 2013] that V1 could be reliably segmented from V2 by a simple contrast of natural texture images against phase scrambled versions of the same images endowed with the same spectral envelope. This finding is in support of the idea that among the range of functionality V2 are computations equivalent to the extraction of correlations between local orientations. It has been put forward that V2 responds to second order correlations of natural images much better than to synthetic images that do not exhibit a particular second order correlation structure.

Even more recently, Okazawa and others were able to employ the Portilla-Simoncelli texture synthesis algorithm, which was also used in [Freeman et al., 2013] above, to create texture images and analyze the responses of macaque V4 neurons in a dynamic manner [Okazawa et al., 2015]. After presenting a large number of images generated by the model for different parameter settings, the parameter space around stimuli which activated the neurons was studied more in detail by sampling the parameter space in the neighborhood. Thus a sparse dependency of the neural firing rates on the parameters could be estimated and neurons thus interpreted as responsive to combinations of very few of the parameters of the system.

In this work and in the literature we have found different depths of analysis systems to be adequate for the analysis of texture images. The detection of useful boundaries can be performed on several different layers. If these layers are similarly implemented in the visual system, then different parts of the visual system may be recruited to perform the task of segregation.

Problems with the experimental setup Analyzing the data from the texture experiment, we were able to compute a main effect of texture class and evaluate a forward model based on the scattering transform. In retrospect, there were several shortcomings we could identify which should be addressed in a follow-up experiment on this topic.

The nature of the experimental comparison task between visual textures raises the question of how strong the measured brain signal is conditional upon this particular format of stimulation.

The fact that attention is directed towards discriminating two images of the same texture may have had several unexpected consequences. Firstly, it will yield different responses than a visual system passively viewing and not attending to differences in the stimuli. Secondly, attention will vary across texture classes, because some are more difficult to discern instance-wise than others. Since we do not have an acquisition guiding attention away from the observations we are interested in, there no way of assessing to what extent attention is modulating responses. Evidence from other experiments suggests that this should be the case [Çukur et al., 2013]. Further, the task to discriminate between texture instances may have interfered with the instructions to fixate, at least with some subjects, since useful information for discrimination may have layn slightly removed from the central cross.

A future experiment on visual textures for the study of the human visual system should provide better control conditions. The aspect of control conditions is addressed in more detail in the summary of this thesis. With the recent successes in the training of neural networks for computer vision (see chapters 4 and 6), it should be feasible to create stimuli with very fine grained control on particular stimulus properties of interest to the study at hand (e.g. compuational levels of abstraction).

The number of presented images of 216 per scanner session and two scanner sessions may also be too small to study sufficiently large analysis models. Acquiring more data should generally improve confidence in results. Acquiring using adequately controlled stimuli should yield a complementary boost. Lastly, the adaptive sampling technique introduced in [Okazawa et al., 2015] could be applied to generate stimuli for the next scanner session based on the previous ones.

Take-Home Messages

- Two-layer analysis models based on linear filtering such as the scattering transform are good at classifying textures
- Texture types are easily classifiable from fMRI brain activity
- Early visual areas V1 and V2 respond very similarly to all 6 presented texture classes. Differential effects appear later in visual hierarchy.

- Scattering layer 2 adds significant predictivity to encoding models in extrastriate areas for texture image as well as natural image stimulation.

6 Mapping the visual hierarchy with convolutional nets

The understanding of human vision and computer vision have historically evolved with mutual inspirations, refining ideas such as hierarchical representations of images, invariance to transformations and feature encoding for object recognition. Convolutional networks used for computer vision, based on multiple layers of localized receptive fields, are achieving human-like capacity in core object recognition. As these networks represent candidate models for the computations performed in the mammalian visual system, we test whether they provide an accurate computational forward model of human fMRI data measured during the viewing of natural images. We construct a predictive model of brain activity for each brain voxel based on each of the layers of a convolutional net. Analyzing the predictive performance across layers yields characteristic fingerprints for each visual brain region: Our experimental results show that early visual areas are better described by lower level convolutional net layers and later visual areas are better described by higher level net layers, exhibiting a progression across ventral and dorsal streams. We validate the generalization capacity of our predictive model by synthesizing brain activity and performing classical analyses upon it, namely retinotopy and a contrast between face-selective and place-selective regions. The synthesis recovers the activations observed in fMRI studies of face and spatial visual processing, showing that this model captures representations of brain function that are universal across experimental paradigms.

This work has been submitted to PLoS Computational Biology

- M. Eickenberg, A. Gramfort, G. Varoquaux, B. Thirion, *Seeing it all: Computer-vision Neural Networks Map the Architecture of the Human Visual System* Submitted to PLoS Computational Biology

6.1 Introduction

Human and primate visual systems are highly performant in recognizing objects and scenes, providing the basis of an excellent understanding of the ambient 3D world. The visual cortex is hierarchically organized, which means that many functional modules have feedforward and feedback connections compatible with a global ordering from lower levels to higher levels [Felleman and Van Essen, 1991]. The concept of visual "pathways" or "streams" [Mishkin and Ungerleider, 1982, Goodale and Milner, 1992] is an established pattern which identifies principal directions of information flow for specific tasks, namely object representation in the "ventral stream" (from occipital cortex into temporal cortex) and localization and spatial computations in the "dorsal stream" (from occipital cortex into parietal cortex). They share much processing in the occipital early visual areas and less oustide of them. The ventral visual stream encompasses visual areas V1, V2, V3, V4 and several inferotemporal (IT) regions. Feedforward pathways from V1 to IT exist, and probably account for rapid object recognition [Thorpe et al., 1996, Fabre-Thorpe et al., 2001]. Many parts of the human and primate visual cortices exhibit retinotopic organization in so-called visual field maps: The image presented to the retina is kept topographically intact in the next processing steps on the cortical surface [Wandell et al., 2007]. This results in a one-to-one correspondence between a point on the retina and the "centers of processing" for that point in the visual field maps, such that neighboring points on the retina are processed nearby in the visual field maps as well.

The seminal work of [Hubel and Wiesel, 1959] showed that cat and other mammal V1 neurons selectively respond to edges with a certain location and orientation in the visual field. This discovery inspired a long line of research investigating what other visual regions do and how they do it. As an example, certain monkey V2 neurons were found to react to combinations of orientations, such as corners [Anzai et al., 2007]. Recently, it has been put forward that V2 may be an efficient encoder of expected natural image statistics arising from interactions of first-order edges [Freeman et al., 2013].

V4 is reported to respond to more complex geometric shapes, color, and a large number of other stimulus characteristics. Recently it has been posited that V4 performs mid-level feature extraction towards the goal of figure-ground segmentation, which can be modulated by top-down attention or bottom-up saliency [Roe et al., 2012]. Further down the ventral pathway, neurons in the IT cortex have been shown to be selective to parts of objects, objects and faces [Desimone et al., 1984, Logothetis et al., 1995]. Taken together, these findings indicate an increasing trend in abstractness of the representations formed along the ventral stream.

FMRI has been used with great success to identify and delineate the aforementioned visual field maps as well as brain regions that seem to specialize to certain tasks in the sense that their responses are particularly strong for specific types of stimuli. This type of result has typically been formulated as a statistical contrast map. See [Kanwisher et al., 1997, Downing et al., 2001, Epstein and Kanwisher, 1998] as examples for the localization of specialized regions using this technique. Finer models, known as "encoding" models or forward modeling techniques [Naselaris et al., 2011], have been used to study the brain response to stimuli in greater detail [Kay et al., 2008, Naselaris et al., 2009a, Nishimoto et al., 2011]. In this setting a rich model going beyond binary contrasts is employed. Using model prediction obtained from the stimulus, one tests how well brain activity can be linearly predicted. For example, in [Kay et al., 2008], almost 2000 naturalistic images were used as stimuli and the BOLD signal responses were then fit using a predictive model based on Gabor filterbank activations of the images shown. Primary visual cortex was very well modeled, but also extrastriate areas such as visual area V4 were well explained by the Gabor filter model.

The Gabor filter pyramid employed in the original work of [Kay et al., 2008] can be seen as an instance of a biologically inspired computer vision model. Indeed, all of modern computer vision, in its roots, has been inspired by biological vision. The basic filter extraction techniques at the beginning of the most successful computer vision pipelines are based on local image gradients or laplacians [Canny, 1986, Simoncelli and Freeman, 1995], which are operations that have been found in V1 as edge detection and in the LGN as center-surround features. The HMAX model was constructed to incorporate the idea of hierarchies of layers [Riesenhuber and Poggio, 1999]. It reached near state of the art object recognition capacities in [Serre et al., 2007].

The key question at stake here is *"What comes after the Gabor filter pyramid?"* in predictive modeling of BOLD fMRI in visual brain areas. The scattering transform model [Mallat, 2012, Bruna and Mallat, 2013] provided only one supplementary layer of which one cannot state much more than the existence of brain voxels which it models well [Eickenberg et al., 2013]. The layers C1 and C2 of HMAX as used in [Serre et al., 2007], obtained using random templates taken from the preceding pooling layer activation, were not geared optimally towards object recognition. This made the difference between layers difficult to evaluate (see e.g. [Kriegeskorte et al., 2008b]). Although quite similar in architecture, deep artificial neural networks are of much greater interest here. Indeed, they optimize intermediate layers to increase performance of object detection. This task or a representation equivalent to it is reportedly performed also in IT cortex in humans and primates.

Using these ideas of optimized feature hierarchies with layered architecture where single units of a layer compute a linear transformation of the activations of previous layers, followed by a simple pointwise nonlinearity, state of the art results have been obtained. Indeed, recent breakthroughs in the field of artificial neural networks have led to a series of unprecedented improvements in a variety of tasks, all achieved with the same family of architectures. Notably in domains previously considered to be the strongholds of human superiority over machines, such as object and speech recognition, these algorithms have gained ground, and, under certain metrics, have surpassed human performance.

On the neuroscience side, in [Cadieu et al., 2014] and [Yamins et al., 2014], it is impressively shown using electrophysiological data that IT neuron activity is similarly predictive of object category as the penultimate layer of a deep convolutional network which was not trained on the stimuli. What is even more interesting is that a deep neural net can predict the activity of IT neurons much better than either lower level computer vision models or object category predictors. Furthermore, deep convolutional networks trained on object categories and linearly fitted to neural activity are similarly predictive of neural activity as the same network fitted directly to neural data, suggesting that object category as "seen" by the network is a good proxy for the representation of neural activity. These two works inspired us to investigate the phenomenon with fMRI in order to obtain a global overview of the system.

Inspecting the first layer of a convolutional net yields filters that strongly resemble Gabor intensity filters, as well as color boundaries and color blob filters (shown at the top of Fig. 6.1). Inspecting the output of a convolutional net applied to an image often yields a correct object identification. We have thus pinned down similarities at the beginning and at the end of the ventral stream object recognition process and the artificial neural network computations. Evaluating its intermediate layers with respect to how well they can explain activity in visual areas of the brain becomes interesting.

In this contribution we assess the predictive capacity of the processing layers of the convolutional network OverFeat [Sermanet et al., 2013], which yielded state of the art object recognition scores on the ImageNet dataset in early 2014. In an encoding framework [Naselaris et al., 2011], we train a linear predictive model of brain activity for each of the layers on the datasets of [Kay et al., 2008] and [Huth et al., 2012] and compare the modeling capacity by evaluating the predictive score on held out data for every voxel. We compare these scores over different layers and obtain continuous progression profiles that are distinct in each visual area. To validate the model, we propose to investigate the generalization capacity of the predictive model that we have learnt. To do so we use previously unseen stimuli, of which some come from totally different experiments and follow largely different pixel statistics. The learnt predictive model, which can be seen as data-driven forward model to generate fMRI activations, is used to generate putative brain activation maps corresponding to these novel inputs. In treating the model as a synthesizer for fMRI brain activation, we can draw on the extensive literature of paradigm-driven fMRI research by reproducing classical experiments. We consider two benchmarks: retinotopy, i.e. the capturing of spatial informa-

tion to the point where visual field maps can be generated, and a faces/places contrast to capture high-level information.

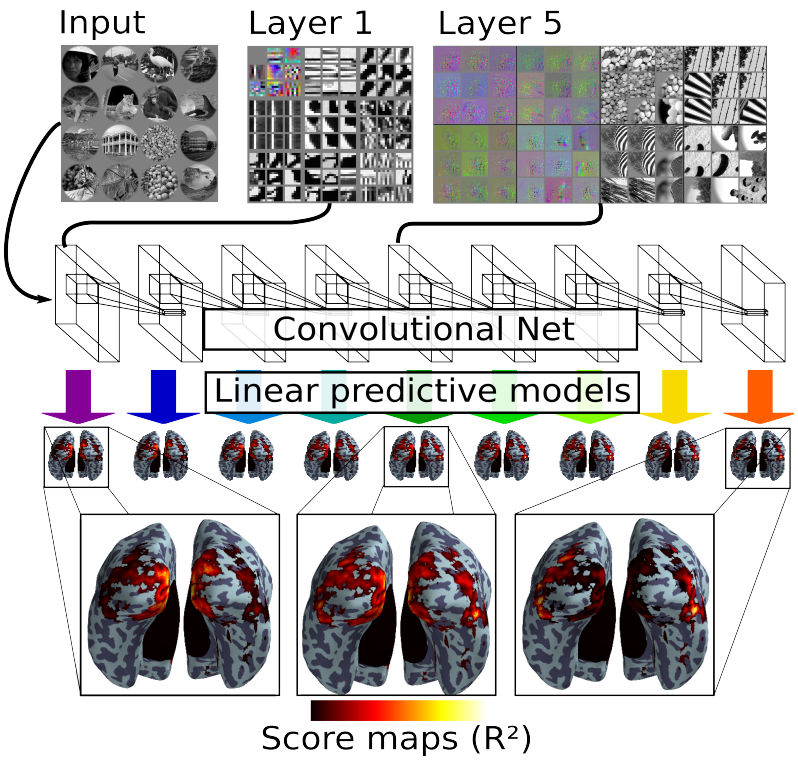

Figure 6.1: **The experimental setup.** Top left: 16 Examples of stimulus images (similar in content to the original stimuli presented to the subjects, and identical in masking) which are input to the convolutional network. Top middle: Selected features of first layer (top left of panel) and image patches activating these features (other eight panels). Top right: Image space gradients of selected feature maps from layer 5 (left panel) and example patches driving these feature maps. The gradients show which change in the image would lead to a stronger activation of the feature map (see [Simonyan et al., 2013]). Middle: Depicts convolutional net layers. Every layer is evaluated for its predictive capacity of all the voxels. For each layer, the corresponding predictive model is depicted by an arrow pointing downward from the convolutional net. It yields a score for each voxel, giving rise to a map of the brain, depicted below the arrow. Bottom: The close-up views are intended to highlight different areas that are well modeled: The first layer models best medial occipital regions close to the Calcarine, the last layer explains more variance in lateral and inferior occipital regions. The middle layer shows an intermediate score map between the two extremes.

Importantly, these synthetic experiments are a non-trivial step forward in several ways: They provide a new way of validating more open forward modeling techniques. By recovering actual activity patterns, they show that the underlying forward model is able to capture experimental results which until now had to be obtained in specific, dedicated experimental paradigms. Once sufficiently validated on known contrasts, they will provide a new tool for investigation of the effects of visual stimuli measurable by fMRI.

Related work

In [Khaligh-Razavi and Kriegeskorte, 2014] the authors evaluate a large number of computer vision models, including a convolutional network. They assess their representational capacity with respect to brain activity while subjects viewed images of objects. They find among other results that the last layers of the network exhibit similar representational similarities as IT neurons in the macaque as well as fMRI activation in humans.

Recent proof of concept work [Güçlü and van Gerven, 2014] uses a convolutional network (different from the one used here, see [Krizhevsky et al., 2012]), enabling the layer-wise analysis of voxel scores across layers. These results are restricted to one subject of [Kay et al., 2008], whereas we extend these results to both subjects. Moreover, we show that the mapping goes beyond a specific experimental paradigm by reproducing our analysis on a video-viewing experiment.

Also concurrent with the present work is [Khaligh-Razavi et al., 2014], in which different computer vision algorithms and all layers of the convolu-

tional network introduced in [Krizhevsky et al., 2012] are compared to the BOLD activity on the data of [Kay et al., 2008]. The analysis is mostly restricted to representational similarity analysis, but a form of "remixing" features with the weights of a predictive ridge regression is introduced. A score progression across layers and regions of interest is also shown.

While the previous work only describes a given subject or experiment, we bring an important novel step to the use of convolutional networks for the study of human vision: showing that results generalize across datasets and paradigms. First, we show the validity of the approach on a new dataset with videos rather than still images. Second, we synthesize plausible brain activity to new images from completely different experiments that rely on hand-crafted, well controlled stimuli. These results demonstrate that convolutional networks can capture and analyze specific cognitive processes that go beyond common studies of natural stimulation, generalizing to new experimental paradigms.

6.2 Methods

6.2.1 Datasets

We consider two different datasets of BOLD fMRI responses to visual stimulation of very different nature: still images and videos. The still images dataset [Kay et al., 2011a] originates from [Kay et al., 2008] and [Naselaris et al., 2009a]. It is described in detail in the Appendix 8.

The video stimulus was first presented in [Nishimoto et al., 2011] and used also in [Huth et al., 2012]. It consists of movie trailers and wildlife documentaries cut into blocks of 5-15 seconds and randomly shuffled. A train set of two hours duration and no repetition was separated from a test set in which around 10 minutes of unique stimulus were cut into blocks of around 3 minutes and repeated in random order 10 times each. Subjects fixated a central cross while passively viewing these stimuli. This dataset comprises one subject.

Both datasets provide functionally localized regions of interest. Visual areas V1, V2, V3, V4, V3A, V3B and LOC were determined using phase-coded retinotopic mapping. All surface projections were computed using `pycortex`[1]. Flatmap diagrams were created directly with pycortex and ROI boundaries were outlined according to localized maps, provided as volume maps in the dataset of [Kay et al., 2008] and as outlines for the data from [Huth et al., 2012]. Volume ROIs were projected to the surface using a nearest neighbor projection and outlines drawn along the borders of the projections.

[1] http://pycortex.org

6.2.2 The encoding pipeline

We chose the "large" version of the deep convolutional net "OverFeat" [Sermanet et al., 2013] to run our analyses. It features six convolutional layers and three fully connected ones. Details can be found in [Sermanet et al., 2013]. Here, we are interested in convolutional networks not to classify images, but as a means to transform them into successive intermediate representations: from Gabor-like features to abstract shapes (see Fig. 6.1). Using `sklearn-theano`[2], the network was applied to all stimulus images and the

[2] http://sklearn-theano.github.io

outputs of all neural network layers kept. Since the intermediate representations are rather large (e.g. $\sim 10^6$ features on the first layer), each channel of each layer was spatially smoothed and subsampled to achieve a number of features of around 25000 per layer. This was achieved by determining smallest integer subsampling necessary to obtain 25000 features or less: for instance, the first layer having $96 \times 113 \times 113 = 1225824$ features, a spatial subsampling of factor 8 per axis is necessary to bring the number of features down to 19154. The smoothing parameter for the Gaussian is set to $0.35 \times d$, where d is the downsampling factor (here 8). For the video data, sampled at 15Hz at an acquisition TR of 2s, temporal downsampling was additionally performed by calculating the temporal mean across 30 frames at a time. A compressive non-linearity, $\log(1 + x)$ was applied pointwise, similarly to the procedure described in [Naselaris et al., 2011]. Using only the stimuli from the training set, ℓ_2-penalized linear regression (ridge regression) was used to fit a forward model for the outputs of each layer for each brain voxel. For the video data, temporally lagged copies of the outputs at t-4, t-6 and t-8 seconds were used in order to account for hemodynamic lag.

We proceed by evaluating how well the activity of each brain voxel can be modeled by each of the OverFeat layers separately. The fitted model was evaluated in a K-Fold cross-validation scheme with bagging. The training data were themselves divided into train/test splits (in accordance with scanner sessions: *"leave one session out"*, K=5 for images, K=3 for videos) and a model trained on an inner train split was evaluated on the corresponding test split to select an optimal penalty. Model scores were obtained using predictive r^2 score for the dataset of [Kay et al., 2008]. This means that for a voxel v the activation y^v_{test} for the test set images was compared to the prediction by our model y^v_{pred} as follows: $r^2_v = 1 - \frac{\|y^v_{\text{test}} - y^v_{\text{pred}}\|^2}{\|y^v_{\text{test}} - \text{mean}(y^v_{\text{test}})\|^2}$, where $\text{mean}(y^v_{\text{test}})$ is the mean activation of voxel v on the test set. Video predictions were evaluated using correlation score $r_v = \frac{\langle y^v_{\text{pred}} - \text{mean}(y^v_{\text{pred}}), y^v_{\text{test}} - \text{mean}(y^v_{\text{test}})\rangle}{\|y^v_{\text{pred}} - \text{mean}(y^v_{\text{pred}})\| \|y^v_{\text{test}} - \text{mean}(y^v_{\text{test}})\|}$. The optimal models for each train/test split of the train data were averaged in order to gain stability of predictions. Mean scores over folds for the optimal penalty were kept as a quantitative measure of goodness of fit.

A schematic of the encoding model is provided in Fig. 6.1. All artificial neural network layers are depicted as being convolutional, although the last three are what is generally known as "fully connected" layers. However, all fully connected layers can be reformulated as convolutions and [Sermanet et al., 2013] takes advantage of this to perform detection and localization. The lowest level layer is depicted on the left and the highest level layer on the right. The brain images below each layer show an r^2 score map for the predictive model learnt on this layer. The scores are normalized per voxel such that the sum of scores across layers is 1. This is necessary due to differences in signal-to-noise ratio across brain regions and highlights the comparison of layers. As can be seen in the three close-up views of brain surfaces, the score maps look different across layers. This finding will be discussed in the results section.

Based on this result, we proceed with a per-ROI analysis of the cross-layer profile of reponses and a more systematic mapping of layer preferences across all voxels that are well-explained by the model.

6.2.3 Synthesis of visual experiments

Using the predictive models learnt on each convolutional network layer, we propose a very simple, yet powerful, summary model by averaging all layer model predictions for each voxel. We validate the predictive capacity of this averaged model by using it as a forward model able to synthesize brain activation maps: Using new stimuli and the coefficients learnt using ridge regression, our model predicts full brain activation maps ("beta maps").

These activation maps can serve a classical analysis purpose in which one evaluates a general linear model with relatively few condition regressors, e.g. by contrasting the activation maps between two different experimental conditions.

We propose to revisit two classic fMRI vision experiments, *retinotopy* and the *faces versus places* contrast, by generating them with our forward model. Since these are known experiments, they can be compared and interpreted in context. At the same time, they test different levels of complexity of our model, retinotopy being purely bound to receptive field location, the distinction of faces necessitating higher level features.

Note that retinotopic mapping was also used in the original study [Kay et al., 2008] to validate the forward model estimated using Gabor filters. In contrast to our setting, retinotopy was estimated by localizing receptive field maxima for each voxel instead of using the predictive model as a data synthesis pipeline.

Retinotopy

We created "natural retinotopy" stimuli (compare [Sereno et al., 1995]) by masking natural images with wedge-shaped masks. The wedges were 30° wide and placed at 15° steps, yielding 24 wedges in total. After creation of exact binary masks, they were slightly blurred with a Gaussian kernel of standard deviation amounting to 2% of the image width. We chose 25 random images from the validation set of [Kay et al., 2008] and masked each one with every wedge mask by pointwise multiplication.

The thus obtained set of 600 retinotopy stimuli were fed through the encoding pipeline to obtain brain images for each one of them. These brain images were then used for a subsequent retinotopy analysis. The design matrix for this analysis contains the cosine and the sine of the wedge angle of each stimulus and a constant offset. The retinotopic angle is calculated from the arising beta maps by computing the arctangent of the beta map values for the sine and cosine regressors. Responsiveness of the model to retinotopy was quantified by the F-statistic of the analysis. In order to obtain an easily interpretable retinotopic map, the beta maps were smoothed with a Gaussian kernel of standard deviation 1 voxel before the angle was calculated. Display threshold is set at $F > 1$.

Synthesizing a "Faces versus Places" contrast

Discriminating faces from places involves higher level feature extraction. It should be noted that with certain stimulus sets the distinction can also be done based on low level features such as edge detectors, but this is almost certainly untrue for the mechanism by which mammalian brains process faces

due to the strong invariance and selectivity properties with respect to non-trivial transformations that they can undergo (see [Pinto et al., 2008] for a discussion). In this sense, being able to replicate a "faces versus places" contrast with the proposed brain activity synthesis is a test for the ability to reproduce a higher level mechanism.

We compute a ground truth contrast against which we test our syntheses by selecting 45 close up images of faces and 48 images of scenes (outdoor landscapes as well as exteriors and interiors of buildings from the dataset of [Kay et al., 2008]). Examples similar to the original stimulus and identical in masking are depicted in Fig. 6.6 (A). Using a standard GLM, we compute a contrast map for "face $>$ place" and "place $>$ face", which are shown in Fig. 6.6 (C), thresholded at $t = 3.0$ in red tones and blue tones respectively.

Our first experiment is to synthesize brain activity using precisely the 93 images which produced the ground truth contrast. We trained our predictive model on the remaining 1657 training set images of [Kay et al., 2008] after removal of the 93 selected face and place stimuli. After computing the synthesized activation images for the latter, we proceeded to analyze them using the same standard GLM procedure as above for the ground truth.

Due to the fact that the noise structure of the synthetic model is different, the threshold of the generated contrast must be chosen in a different manner. We use a precision-recall approach that can be described in the following way: Having fixed the threshold of the ground truth contrast at $t = 3.0$, we define the support of the map as all the voxels that pass threshold. For a given threshold t on the synthesized map we define *recall* as the percentage of the support voxels from the ground truth contrast that are active in the thresholded synthesized map and *precision* as the percentage of active voxels in the thresholded synthesized map that are in the support of the ground truth map. We define the synthesized map threshold t_{R50} as the threshold guaranteeing a minimum of 50% recall while maximizing precision.

Our second experiment tests the generalization capacity of our model in a more extreme situation: In order to make sure that our feedforward model is not working with particularities of the stimulus set other than the features relevant to faces and scenes, we also evaluated the faces versus places part of the dataset from [Haxby et al., 2001]. This study showed distributed and overlapping representations of different classes of objects in ventral visual areas. Among the stimuli are 48 pictures of faces and 48 pictures of houses, both tightly segmented, on a light gray background. These stimuli are notably different in appearance from the ones used to train our model. We applied the same feedforward pipeline to obtain simulated activation maps for each of these images and the same GLM analysis and thresholding procedure as described in the preceding experiment.

6.3 Experimental results

On inspection of the three zoomed panels from Fig. 6.1 one observes that the score maps are different across layers. On the left, the model based on the first layer explains medial occipital regions well with respect to the others. It includes the calcarine sulcus, where V1 is situated, as well as its surroundings, which encompass ventral and dorsal V2 and V3. This contrasts to the score

map on the right, which represents the highest level model. The aforementioned medial occipital regions are relatively less well explained, but lateral occipital, ventral occipital and dorsal occipital regions exhibit comparatively higher scores.

6.3.1 Quantifying layer preference

For each voxel, we call the set of scores associated with the prediction of its activity from each layer the *score fingerprint* of that voxel. Given the fact that layer outputs are somewhat correlated (across layers) and each voxel contains many neurons, we do not expect sharp peaks in the score fingerprint for a specific "best" layer. Rather we expect a progression of scores over layers indicating a global trend towards simple, intermediate or more high-level representations. Using the ROI definitions provided by the datasets, we can study the mean score fingerprints per region of interest. The average score fingerprint per ROI was obtained using the 25% best predicted voxels within the region. For each region of interest, the mean score fingerprint was normalized by its maximum value. The resulting normalized progressions are shown in Fig. 6.2.

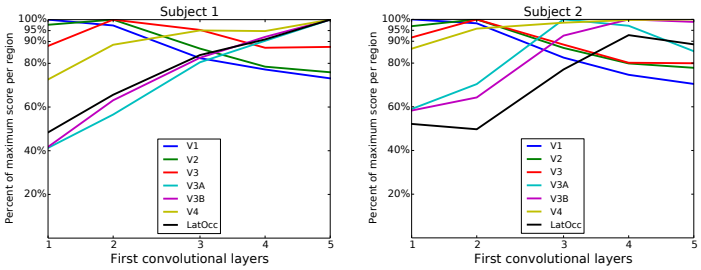

Figure 6.2: **Normalized average score fingerprints over ROIs.** Score progressions for two subjects averaged over regions of interest provided by the dataset. For each ROI, the score progression was normalized by its maximally predictive layer score. For V1 we observe peak score in layer 1 and a downward trend towards higher level layers. The V2 fingerprint peaks in the second layer and then decreases slightly slower than the V1 fingerprint. V3 fingerprint also peaks in layer 2 but decreases more slowly than V1/V2 fingerprints. V4 fingerprint peaks much later than the ones of V1/V2/V3 but is not much worse described by lower level layers. Fingerprints of V3A/B and LOC show a strong increase across layers.

We observe that for both subjects, the score fingerprint for V1 peaks at the first layer. It then decreases in relative accuracy as the layer index increases. For the mean fingerprint of V2, the peak lies on the second layer and the subsequent decrease is a little slower than that of the V1 fingerprint. This indicates a selectivity for a mix of higher level functionality less present in V1. The V3 mean score fingerprint also peaks at layer 2 and decreases less fast than the V2 fingerprint, indicating a layer selectivity mix of again slightly higher levels of representation than present in V2. The mean V4 fingerprint peaks significantly later than the first three, around layers 4 and 5, but at lower layers the representation is never extremely bad. The score fingerprint is constantly above 70% of its maximum score. In contrast, the dorsal areas V3A and V3B are much less well modeled by lower level layers than by higher level layers. Similarly, the lateral occipital complex (LOC) shows a strong increase in relative score with increasing representation layer number.

In Fig. 6.3 we show an "argmax" map over spatially smoothed scores ($\sigma = 1$ voxel). It is obtained by smoothing each score map and then associating each voxel with the layer which best fitted its activity. Despite the fact that the second strongest scores are sometimes only slightly below the maximum (cf. Fig. 6.2) it turns out that this marker provides compelling outlines of the organization of the visual system. It is indeed visible that the smoothing with argmax visualization is plausible with respect to network layer organization.

One observes that medial occipital regions are mostly in correspondence with the first layer, that there is a progression in layers along the ventral and dorsal directions, which is symmetric, and that there is a global symmetry across hemispheres.

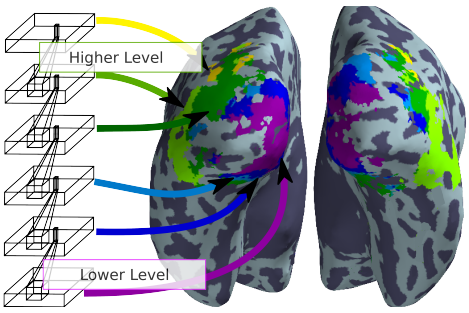

Figure 6.3: **Best model per voxel.** Among the voxels which are modeled by at least one of the convolutional network layers, we show which network layer models which region best. This is achieved by smoothing the layer score maps ($\sigma = 1$ voxel) and assigning each voxel to the layer of maximal score. One observes that the area around the Calcarine sulcus, where V1 lies, is best fit using the first layer. Further one observes a progression in layer selectivity in ventral and dorsal directions, as well as very strong hemispheric symmetry.

In order to better show the layer selectivity of each voxel as represented by its score fingerprint in a brain volume, we derived a summary statistic based on the following observation. As can be seen in Fig. 6.2, the average fingerprints of each region of interest have either an upward or a downward trend. It turns out that the first principal component of all score fingerprints over significantly well predicted voxels is a linear trend. Moreover, it explains over 80% of the variance of all fingerprints. The projection onto it can therefore be used as a summary of the voxel fingerprint. Here we use a fixed trend going from -1 at layer 1 to 1 at layer 9 in steps of 0.25. Projecting the score fingerprints onto this ascending trend, which amounts to evaluating the global slope, yields a summary of the voxel fingerprint. It is shown for subject 1 in Fig. 6.4 on the left. We observe that V1 fingerprints project almost entirely to the low level range of models, indicated by blue hues. V2 shows more presence of green, indicating intermediate level models. This trend continues in V3. V4 shows a clear preference for mid-level models. Subsequent regions show a tendency towards even higher level representations.

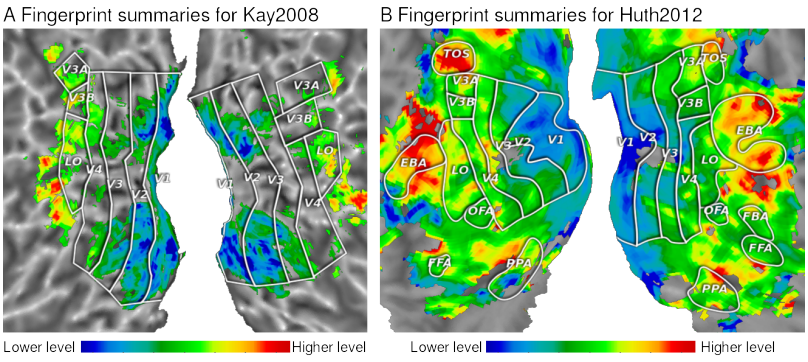

Figure 6.4: **Fingerprint summaries as brain map.** We compute a summary statistic for voxel fingerprints by evaluating their inner product with an ascending linear trend from -1 to 1 in nine steps of 0.25. This yields low values for low layer preference and high values for late layer preference. Observe the preference for low-level models in earlier visual areas V1 and V2. With increasingly higher layer selectivity for V3, V4 and ulterior visual areas, a trend from low level to high level representation across the ventral and dorsal visual pathways becomes apparent.

This progression is mirrored exactly on the second panel of Fig. 6.4. Applying an identical visualization technique to the score fingerprints obtained from modeling the video experiment, we observe a very similar progression of model selectivity across the early visual areas. As above, the fingerprint summary indicates lower level layer preference in V1 and V2, intermediate layers in V3 and V4 and high level layers in parts of lateral occipital and specialized areas such as the extrastriate body area (EBA, [Downing et al., 2001])

and the transverse occipital sulcus (TOS, [Bettencourt and Xu, 2013]).

Recall that the latter data were acquired in a completely different experiment, with videos instead of images. It is to be noted that the convolutional network was applied directly to the individual frames of the video, followed by a temporal aggregation in order to reach the temporal scale of the fMRI acquisition. No explicit motion processing or other video-specific processing was incorporated. The fact that the same underlying model obtains similar results is a strong demonstration of the reproducibility of our findings.

6.3.2 Synthesis of visual experiments

Retinotopy

The angular maps obtained by synthesizing fMRI activation from virtual wedge-shaped stimuli can be seen in Fig. 6.5. Comparison to existing literature shows that the model indeed captures the transitions of known retinotopic regions. For instance, one can observe the sign inversions of the gradient of the angle map at the transitions from ventral V1 to ventral V2 and ventral V3 to ventral V4. These transitions are very clear and in perfect correspondence with the outlines of the volume-based retinotopic regions of interest provided with the dataset. The transitions in dorsal primary visual areas are apparent but slightly less well delineated, possibly due to surface projection difficulties. In sum, the obtained virtual angle map is coherent with respect to the information available in the subject (see [Sereno et al., 1995] and [Wandell et al., 2007]).

Difficulties possibly due to distortion between available anatomical and functional images. Regions of interest were drawn according to projection of volume-based maps. Irregularities were observed in placement of dorsal areas V3A/B and left V4.

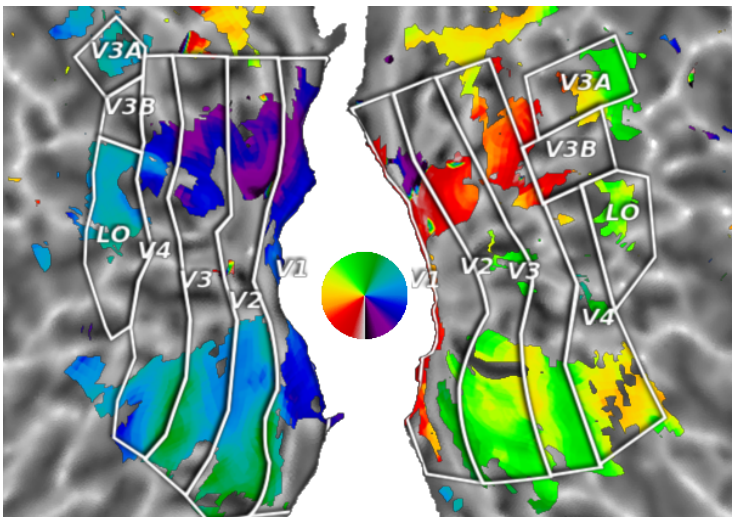

Figure 6.5: **Retinotopic map for subject 1.** Synthesizing the responses to retinotopic wedge stimuli and performing a classic phase-coding GLM analysis, we show the retinotopic angle map at display threshold $F = 1$. As can be seen in the ventral part of the brain map (lower half), the retinotopic mapping indicates visual angle inversions exactly at the locations previously identified by a localizer, aligning perfectly with the visual map borders traced on the surface. Dorsal areas (upper half) exhibit the same tendencies in a less pronounced manner.

Replicating the "Faces versus Places" contrast

We first synthesize the brain activity corresponding to the images which produced the ground truth contrast (but left out during model training). The results for the 93 held-out stimuli from [Kay et al., 2008] are shown in Fig. 6.6 (D) and the results of the transferral to the experiment of [Haxby et al., 2001] are to be seen in Fig. 6.6 (E). Observe the striking similarity of both simulated contrasts to the ground truth contrast in Fig. 6.6(C).

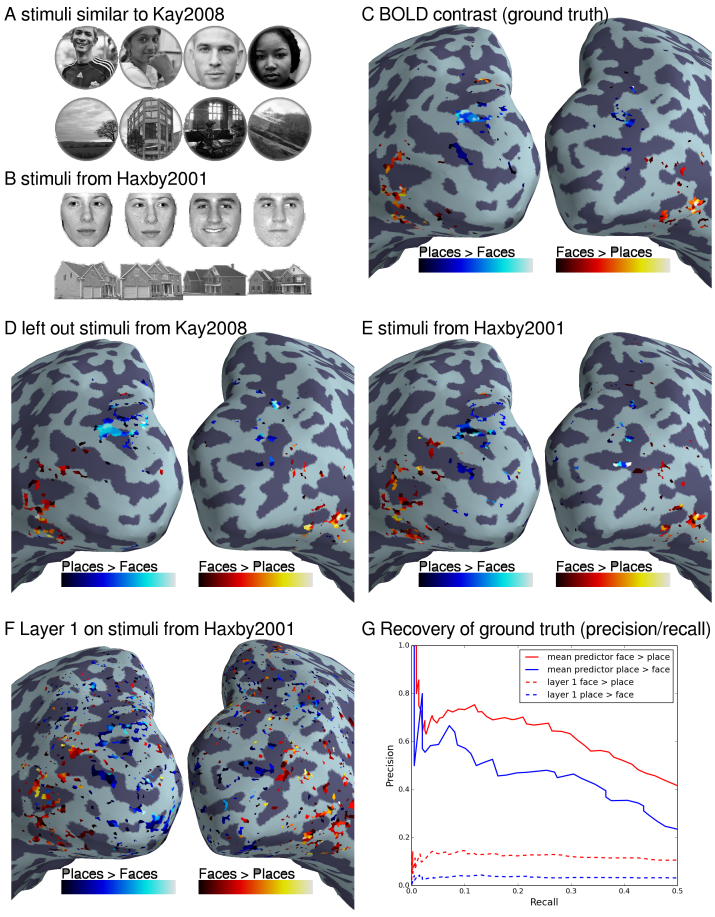

Figure 6.6: **Synthesizing Face versus Place contrast**. (A) Examples of the stimuli similar to those of [Kay et al., 2008] containing close up photos of faces (45 total) and places (48 total), removed from the train set of the synthesis model. (B) Examples of the stimuli from [Haxby et al., 2001] for faces and places (48 for each in total). (C) Contrast of BOLD activity from a GLM model of the held-out face and place stimuli. Referred to as ground truth in view of the synthetic data. (D) Predicted contrast for the 93 held out face and place stimuli from the training set of [Kay et al., 2008]. Thresholded at best precision given minimum recall of 50% of ground truth activation support. (E) Predicted contrast for the 96 face and house stimuli from [Haxby et al., 2001]. Thresholded as in D. (F) Predicted contrast for the 96 face and house stimuli from [Haxby et al., 2001] using only layer 1, i.e. a first order, edge-detector type feature map. Thresholded at 50% recall of ground truth as in D. Note the strong noise component in the map compared to D and E.

The areas that respond to faces are lateral occipital and inferior occipital. The Lateral Occipital Complex is known to have face selective subparts [Grill-Spector et al., 2001] and the inferior occipital Occipital Face Area is also known to be involved in face processing. It is possible that some more generally body part selective areas are active as well since the stimuli used to obtain the ground truth contrast may also contain a view on e.g. part of the torso [Taylor et al., 2007, Downing et al., 2001]. Note that both the fusiform face area and the fusiform body area are outside the field of view of the acquisition and thus invisible to the ground truth contrast and the synthesized contrast.

The areas responsive to places are mainly dorsal in the given field of view. We observe activation in regions that are most likely to be transverse occipital sulcus (TOS) and inferior intraparietal sulcus (IPS). Since these regions are typically close together anatomically and as no localizer for them was performed on the given brain, it is difficult to tell them apart. However, [Bettencourt and Xu, 2013] shows that TOS is strongly scene selective whereas inferior IPS may be more concerned with object individuation and localization. Note that the habitually mentioned place-selective Parahippocampal Place Area [Epstein and Kanwisher, 1998] is also not within the field of view of the acquisition.

In conclusion, the simulated face/place contrasts using stimuli from [Kay et al., 2008] and from the very different stimulus set of [Haxby et al., 2001] both create an activation contrast very close to the estimated ground truth contrast. The ground truth contrast yields coherent activation maps which fit well into existing literature.

In order to show that this type of synthesis is impossible with a first layer contour only model, we show the contrast using layer 1 from the model in Fig. 6.6 (F). The previously identified regions can no longer be distinguished from the strong noise in the surroundings. Fig. 6.6 (G) depicts the precision-recall curves for face and place selective areas for the averaged model and for the layer 1 model. Studying the high precision range at the left of the diagram, it becomes clear that the proposed average synthesis model shares its strongest activations exactly with the ground truth contrast, leading to 100% precision. This is never the case for the model obtained from layer 1.

6.4 Discussion

The study of the mammalian visual system has historically been led by crafting stimuli designed to trigger neural activation in various sub-systems of the visual cortex, from edges [Hubel and Wiesel, 1959], to abstract shapes and faces [Gallant et al., 1996, Desimone et al., 1984, Logothetis et al., 1995, Kanwisher et al., 1997, Bentin et al., 1996]. However, the visual system responds conditionally to the types of stimuli that it receives. Elicited neural responses from parametrically varied synthetic stimuli may be strongly related to the chosen stimulus ensemble, making generalizations difficult. Naturalistic stimuli provide experimental settings that are closer to real-life ecological settings, and evoke different responses [Gallant et al., 1998].

While most detailed understanding about neural computation has been pushed forward using electrophysiological experiments, the non-invasive methodology of fMRI offers the benefit of full-brain coverage. Many typical fMRI studies investigate binary hypotheses by devising stimuli specific to a question, whether they be naturalistic or not. In contrast, the dataset on which we rely [Kay et al., 2011a], is due to an investigation of the BOLD fMRI responses to a large number of not specifically chosen natural stimulus images, showing that it is possible to identify the stimulus among thousands of candidate images. Departing from studies based on manual crafting of specific stimuli and corresponding restrictive hypotheses, we propose to model brain responses due to pure natural image statistics. Indeed, capturing and modeling the rich statistics in images of the world that surrounds us must be a driving principle of the structure of visual cortex, as suggested by [Olshausen and Field, 1996] for the primary visual areas. Here, we rely on a very powerful computational model capturing these statistics: a deep convolutional network with enough representational capacity to approach human-level core object recognition.

Based on the contest-winning convolutional network *OverFeat*, we have built a feedforward model explaining brain activity elicited by visual stimulation from the image representations in the various layers of the convolutional network. We fitted a separate model to all brain activity for each layer and obtained prediction scores for each one of them. These prediction scores were analyzed in order to establish a comparison between the convolutional network feature hierarchy and brain regions. In an ROI analysis we show that early visual areas are better modeled with lower level layers from the convolutional network but that progressing ventrally and dorsally from the calcarine sulcus there is a clear increase in selectivity for complex representations as indicated by increasingly better scores for higher level layers. Furthermore, score fingerprint summaries obtained by projection of individual score fingerprints onto an ascending trend show a clear spatial gradient in affinity to higher level representations: Starting at V1 we observe a clear dominance of low level layers in the score fingerprint. Across subsequent extrastriate visual areas we observe a gradual and continuous increase in relative predictive power of the complex representations. The same result was obtained for a representation of score fingerprints due to a visual movie experiment. This yields a second indicator of the existence of a gradient in complexity coming from a completely different dataset. Finding the same global structure on such different stimuli is a strong confirmation that the uncovered structure is not spurious or due to experiment design.

It should be emphasized that this functional characterization does rely to some extent on the structural similarity between the functional organization of the visual cortex and that of the computational model. In a convolutional network, the linear transformation is restricted to the form of a convolution, which forces the replication of the same linear transformation at different positions in the preceding layer image. This forces similarity of processing across the 2D extent of the image and constrains the receptive fields of the units to be localized and spatially organized. This spatial sparsity saves computational resources and entails a strong inductive bias on the optimization by encoding locality and translation covariance. It is however important to note that biological visual systems generally do not exhibit linear translation

covariance. The retinotopic correspondence map allocates much more cortical surface to foveal regions than to peripheral regions. This is called cortical magnification (see e.g. [Schira et al., 2007] for details).

In this work, we introduce a new method for validating rich encoding models of brain activity. We generated simulated brain activation for known, standard fMRI experiments using a model-averaged predictor and analyzed them using well-known, classical task fMRI methods. We chose two experiments at different levels of complexity: Retinotopy, a low-level spatial organization property of the visual system, and the *faces versus places* contrast, an experiment necessitating high-level recognition capacity and complex representations. The results show that both experiments are well replicated. Angle gradient sign inversion lines indicating the bounds of visual areas are correctly identified. Face and place selective voxels as defined by a previously calculated contrast on true BOLD signal are correctly identified in the synthesized contrast in the sense that the voxels responding strongest to the simulated contrast are those that are the strongest in the BOLD contrast. This notion is visualized in a rigorous manner by presenting the synthetic maps at a threshold that recovers at least 50% of the supra-threshold area $t \geq 3.0$ of the original activation map.

Both for left-out face and place stimuli from the original experiment and the stimuli of faces and houses used from [Haxby et al., 2001], the model had never seen these images at training time. It had seen the same type of image as the held out set in the sense that they were taken from the same photo base, had the same round frame and the same mean intensity. The type of image coming from [Haxby et al., 2001] was segmented differently –tightly around the object– making the framing very different in addition to very different mean intensities and pixel dynamics. Our synthesis model for brain activation was robust to these differences and yielded very similar contrasts to the ground truth. Similarly, the retinotopy stimuli were constructed from previously unseen images, and the geometry of the retinopy wedges was entirely new to the system as well. Generalizing to such images, with different statistics from those of the experiment used to build the model, is clear evidence that our model captures the brain representations of high-level invariants and concepts in the images.

We have thus built a data-driven forward model able to synthesize visual cortex brain activity from an experiment involving natural images. This model transcends experimental paradigms and recovers neuroscientific results which would typically require the design of a specific paradigm and a full fMRI acquisition. In the current setting, any passive viewing task with central fixation can be simulated using this mechanism. After a validation of correspondence on many contrasts for which one has BOLD fMRI ground truth, one could use it in explorative mode to test new visual experimental paradigms. Discrepancies, i.e. the inability of the model to describe the response to a new stimulus adequately, would provide cues to refine this quantitative model of the visual cortex activity.

Take-Home Messages:

- Convolutional Neural Net layers fit the activity of fMRI brain voxels in

visual brain regions following stimulation with natural images;

- Mapping each voxel to the convolutional net layer which models it best yields a brain map of rising complexity along the visual processing hierarchy;
- Embedding score fingerprints into 1D PCA space yields smooth transition from low-level to high-level model preference across cortex;
- Averaging the predictions for all layers yields a forward model capable of predicting brain activity such that it generalizes to other experiments. We reproduce classical results when this brain activity is analyzed as if it were true brain activity. Examples: Low level: Retinotopy; High level: Faces vs Places.

7 Conclusion

7.1 Summary

- Through this thesis we aimed to advance the understanding of three fields by an increment: Understanding BOLD fMRI data, understanding appropriate analysis methods and understanding brain function.

- In order to better understand BOLD fMRI data, we performed a comprehensive analysis of the data from two event-related designs, with focus on the shape of the hemodynamic response function. We showed that taking the spatial variability of the hemodynamic response into account resulted in activation map estimation that yielded significant performance enhancement in both encoding and decoding models.

- We investigated the paradigm of *brain decoding* from fMRI data by approaching it from a machine learning and optimization perspective, with a strong emphasis on interpretability. Building on the existing set of spatially informed convex 1-homogenous penalties called TVℓ_1 norms, we added a modification leading to a more plausible prior: Global sparsity with contiguous regions of zero activation and contiguous regions of smoothly varying activation. Studying the properties of these penalties as priors on weight maps, we registered slight performance increases and plausible feature maps.

- In an fMRI experiment investigating the responses of the human visual system to images of natural textures, we observed that large portions of the early visual system responds to texture images. In studying significant differences in activation across the six texture classes it becomes apparent that V1 and V2 seem to activate very similarly to all classes. Only in later areas do the voxels respond significantly differently to different texture classes. This result is along the lines of [Freeman et al., 2013, Okazawa et al., 2015]. Further investigations are necessary in order to delineate these effects with higher confidence. Notably, a scale-up in number of stimuli and number of texture classes seems called for, along with the generation of synthetic stimuli yielding similar descriptors or potentially mixing between classes. This would give access to useful control conditions. Experimental design should modulate attention in order to be able to assess and discount its impact on estimation.

- Finally, using a pre-trained convolutional net which reached state of the art in the ImageNet object recognition challenge in 2013, we were able to

show that different convolutional net layers predict differently well well-delineated parts of the visual hierarchy, showing a progression of rising complexity (as represented by layer number) as one progresses down the ventral stream. This finding corroborates many models of vision in highlighting the correspondences of intermediate steps.

7.2 Outlook

In this thesis we touched upon several aspects of fMRI data analysis. From the inspection of the data themselves via the examination of relevant methods to the study of some neuroscientific aspects using encoding models.

Each of these topics has given rise to follow-up questions to be addressed in future work. In this outlook we concentrate on the main points, the investigation of which should lead to the largest gain in terms of insight.

7.2.1 Natural stimuli

One important debate can be centered around comparison of natural stimulation versus controlled environment. Specifically in order to test the responses of the visual systems with controlled stimuli, more and more complicated types of stimuli had to be devised in order to be able to drive neurons in higher level visual areas. These stimuli would often be parametrized by a low number of parameters and thus span a relatively low-dimensional space of images in a way that the experimenter has decided. Visual systems will likely respond to these stimuli, but conclusions can only be drawn taking to account the nature of that specific set of stimuli. By choosing this set of stimuli to be natural stimuli, one avoids this problem, since the visual system is then confronted with input that it can plausibly have seen in its environment, i.e. stimuli it is "made to deal with well". It still holds that the response of the visual system is conditional on the stimulus set, but this specific stimulus set is more easily justified. The issues arising from natural stimulation are immediate: It is largely impossible to sample the full space, so choices need to be made. Co-occurrences of concepts and correlations of more basic quantities are inevitable. It is in the interest of the experimenter to acquire as many data as possible to stand a chance at dissecting the correlation structure. Then again, this correlation may be "normal" and part of the statistics of natural images. This is certainly true for the correlation structure of lower level descriptors. But the fact that all images of boats that are in the stimulus ensemble only show boats in the middle of the ocean doesn't make boats outside the water irrecognizable to humans - although it may take a moment longer for us to recognize an object out of context. Here at the latest it should become very clear that the question the experimenter would like to ask of the data is crucial to and should precede the acquisition. Similar to this issue, there are non-negligible implicit biases within natural stimuli if they are taken to be photos. In general, photos are framed in very specific ways, placing prominent object in a very restricted number of positions and typically having the horizon horizontal and plumb line vertical, which are evidently not always placed that way on the retina. It is shown in [Pinto et al., 2008] how this type of framing bias can lead to classification performance that is

higher than expected for very simple models of vision on CalTech101. It is then shown that this bias can be removed by "creating natural stimuli" by placing natural objects on natural background such that position, orientation and potentially other variables are sampled uniformly. It is clear to see that this is a form of controlling stimuli, and that it is motivated relative to a specific question: in this case the evaluation of the capacity of invariant object recognition, a capability which primate brains tend to have. Natural stimulation is probably the only way to reach all visual areas and study them in detail. But specific questions should be asked, and control stimuli created to be able to answer these. Concretely, once models of the visual system become more and more accurate even in predicting BOLD response, they can be conveniently tested by creating stimuli that yield the same response according to the model, as has been done in [Freeman and Simoncelli, 2011, Freeman et al., 2013]. This approach should be the general way of attacking new models. A counterpart is to find stimuli that yield extremal responses of the model and to measure to what extent the response of the real system corresponds.

7.2.2 fMRI

This thesis has been fMRI-centered, touching on various aspects of statistical data analysis involving it. Many global studies concerning vision have been conducted and have yielded interesting results. Most of these results are of the "mapping" type, e.g. the discovery of a mosaic of seemingly specialized, modular brain units in higher level visual areas, such as FFA, PPA, OFA, TOS, EBA, VWFA and others, as defined by being differentially most active amongst a number of chosen stimulus classes. Additionally, mappings of spatial arrangment of functionality have been studied via retinotopy, to reveal spatial arrangement of brain areas from low level to high level. These are discoveries to which fMRI has been indispensable due to its global nature and sufficient spatial resolution. The studies of [Kay et al., 2008] and [Nishimoto et al., 2011] have shown that properties known to be true locally at the neuron level, e.g. the responsiveness of V1 to oriented edges, translate to the BOLD response of voxels: More contour energy leads to stronger BOLD response in the voxel. The analysis in chapter 6 shows that this can be pushed further by using features of mid and high-level complexity from object recognition neural networks. This population-level argumentation leaves the frustrating aftertaste of not being able to properly reach down and conclude on "what the neurons are doing". In V1, which is a relatively well understood visual area, it is clear that current fMRI methods, available for easy deployment in any laboratory, are not capable of spatially resolving cortical columns of any type (although specific research into pushing this frontier has been done [Yacoub et al., 2008]). If they were, it would become necessary to study the subtle local hemodynamics of these structures to be able to draw conclusions. Having sufficiently convinced oneself that one is able to thus properly study e.g. orientation selectivity in V1, one could then move on to the next visual areas and test hypotheses at the same level of resolution. Given the current techniques, the only way to see into subvoxel populations is by adaptation, which is an indirect technique that may not always permit the desired conclusions. On a higher level than orientation selectivity, what exactly would it

mean to adapt to a certain higher level feature? This discussion would lead back into that of optimal control of stimuli. For future developments in fine-grained forward modelling, it could thus be of interest to invest time on the acquisition side and assess the possibility of exceptionally high resolutions, ideally both spatially and temporally, on single slices of e.g. V1, after having formulated a clear hypothesis of what one expects to find.

7.2.3 Combination of modalities

The study of cognitive phenomena should not be intrinsically restricted to a certain modality. FMRI is situated at a certain trade-off in spatio-temporal resolution, which is arguably low on both sides. However, its spatial resolution is orders of magnitude better than that of e.g. MEG. In order to study the intricacies of information flow, especially that of visual feedback from higher to lower areas, one needs either some very cunning experimental design for fMRI to draw indirect conclusions, or sufficient temporal resolution to be able to resolve these phenomena. Combining fMRI for localization and MEG for temporal tracking of activation may in the best case permit conclusions in the best spatial and temporal resolutions of both modalities. The work of [Cichy et al., 2014] is an important step in this direction.

7.2.4 Concluding remarks

This thesis touched several aspects concerning the evaluation of computational models of vision with fMRI. It has been an attempt to settle in between the typical applications of machine learning and the field of neuroscience, with the goal to do justice to both. The result is the development of ideas for a set of tools that will probably shape the future of the studies of neuroscience: Hard bottom-up computational models will likely become capable of explaining more and more types of brain activity and with an extra effort, these models may remain interpretable. More fine-grained stimulus generation will result in a disection of populations of candidate models. A part from pushing neuroscientific understanding, these insights may also be used to create more and more easy to use brain machine interfaces.

8 Appendix

8.1 Dataset descriptions

8.1.1 Dataset 1: encoding of visual information

The first dataset we will consider is described in [Kay et al., 2008, Naselaris et al., 2009b, Kay et al., 2011b]. It contains BOLD fMRI responses in human subjects viewing natural images. Prediction of BOLD signal following the visual presentation of natural images is performed and compared against the measured fMRI BOLD signal. As the procedure consists of predicting the fMRI data from stimuli descriptors, it is an *encoding* model. This dataset is publicly available from `http://crcns.org`

Two subjects viewed 1750 training images, each presented twice, and 120 validation images, each presented 12 times, while fixating a central cross. Images were flashed 3 times per second (200 ms on-off-on-off-on) for one second every 4 seconds, leading to a rapid event-related design. The data were acquired in 5 scanner sessions on 5 different days, each comprising 5 runs of 70 training images –each image being presented twice within the run– and 2 runs of validation images showing 12 images, 10 times each. The images were recorded from the occipital cortex at a spatial resolution of $2\text{mm} \times 2\text{mm} \times 2.5\text{mm}$ and a temporal resolution of 1 second. Every brain volume for each subject has been aligned to the first volume of the first run of the first session for that subject. Across-session alignment was performed manually. Additionally, data were temporally interpolated to account for slice-timing differences. See [Kay et al., 2008] for further preprocessing details.

8.1.2 Dataset 2: decoding of potential gain levels

The second dataset described in [Tom et al., 2007a] is a gambling task where each of the 17 subjects was asked to accept or reject gambles that offered a 50/50 chance of gaining or losing money. The magnitude of the potential gain and loss was independently varied across 16 levels between trials. Each gamble has an amount of potential gains and potential losses that can be used as class label. In this experiment, we only considered gain levels. This leads to the challenge of predicting or *decoding* the gain level from brain images. The dataset is publicly available from `http://openfmri.org` under the name *mixed-gambles task* dataset.

The data preprocessing included slice timing, motion correction, coregistration to the anatomical images, tissue segmentation, normalization to MNI space and was performed using the SPM 8 software through the Pyprepro-

cess[1] interface.

[1] https://github.com/neurospin/pypreprocess

8.2 Analytical leave-k-out ridge regression

We present an analytical formula to perform cross-validation in kernel ridge regression.

To the best of our knowledge, the extension of analytical leave-one-out cross-validation to analytical leave-k-out cross-validation represents original work.

8.2.1 Notation

Let $\mathcal{X}$ be a set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. Let $n \in \mathbb{N}$ and $(x_i)_{i=1,...,n}$ be a finite number of samples from $\mathcal{X}$ and $K \in \mathbb{R}^{n \times n}$ be the Gram matrix with $K_{ij} = k(x_i, x_j)$. Further let $y_i \in \mathbb{R}, i \in \{1, ..., n\}$ be corresponding prediction target values (outcomes) and $w_i > 0$ weights indicating the importance of each sample.

For $1 \leq i \leq n$ we denote $K_i = K_{i,\cdot}$ the ith row of matrix K and $K_{\cdot,i}$ the ith column. For an index set $I \subset \{1, \ldots, n\}$, x_I denotes $(x_i)_{i \in I}$, K_I denotes all lines of K indexed by I, i.e. $(K_i)_{i \in I}$, $K_{\cdot,I}$ denotes all columns of K indexed by I, i.e. $(K_{\cdot,i})_{i \in I}$ and $K_{I,I}$ denotes the square submatrix of K indexed by I, i.e. $(K_{ij})_{i,j \in I}$.

We use the convention that $f(x_I) = (f(x_i))_{i \in I}$.

8.2.2 Refresher: Kernel Ridge regression with sample weights

We are interested in solving the following weighted Ridge regression problem:

$$\hat{f} = \arg\min_f \frac{1}{2} \sum_{i=1}^{n} w_i (f(x_i) - y_i)^2 + \frac{\lambda}{2} \|f\|_k^2. \tag{8.1}$$

Setting all $w_i = 1$ recovers the simple unweighted Ridge regression functional. The simplest representer theorem covers this case [Scholkopf and Smola, 2001] and guarantees that the solution can be written in the form $\hat{f} = \sum_{i=1}^{n} \hat{c}_i k(\cdot, x_i)$ for a certain dual solution $\hat{c} \in \mathbb{R}^n$. In particular, for $1 \leq j \leq n$ we have $\hat{f}(x_j) = \sum_{i=1}^{n} \hat{c}_i k(x_j, x_i) = (K\hat{c})_j = \langle K_j, \hat{c} \rangle$.

The regression problem can thus be reframed to finding the vector $\hat{c}$ such that

$$\hat{c} = \arg\min_c \frac{1}{2} \sum_{i=1}^{n} w_i (\langle K_i, c \rangle - y_i)^2 + \frac{1}{2} \sum_{i,j=1}^{n} c_i c_j K_{ij} \tag{8.2}$$

Defining $W = \mathrm{diag}((w_i)_i)$, equation (8.2) can be rewritten in matrix form as follows

$$\hat{c} = \arg\min_c \frac{1}{2} (Kc - y)^T W (Kc - y) + \frac{\lambda}{2} c^T K c. \tag{8.3}$$

Setting $\mathcal{L}(c) = \frac{1}{2}(Kc - y)^T W (Kc - y) + \frac{\lambda}{2} c^T K c$, its derivative with respect to c reads

$$\nabla \mathcal{L}(c) = KW(Kc - y) + \lambda K c = K(W(Kc - y) + \lambda c). \tag{8.4}$$

With constant (e.g. w.l.o.g. unit) sample weights, one can show that the solutions to $K((Kc - y) + \lambda c) = 0$ are exactly the same as those of $(Kc - y) + \lambda c = 0$ simply by using the eigendecomposition of the symmetric matrix K.

In order to establish a similar relation in presence of nontrivial sample weights, we set $\tilde{c} = \sqrt{W^{-1}}c$. Then the normal equations $\nabla\mathcal{L}(c) = 0$ become

$$0 = K(W(Kc - y) + \lambda c) = K\sqrt{W}(\sqrt{W}K\sqrt{W}\tilde{c} - \sqrt{W}y + \lambda\tilde{c}) \tag{8.5}$$

Setting $\tilde{K} = \sqrt{W}K\sqrt{W}$ and $\tilde{y} = \sqrt{W}y$ completes the transformation to yield

$$0 = \sqrt{W^{-1}}\tilde{K}((\tilde{K} + \lambda\,\mathrm{Id})\tilde{c} - y) \tag{8.6}$$

Since $\tilde{K}$ is symmetric and $\sqrt{W^{-1}}$ is bijective, we conclude that solving (8.6) is equivalent to solving

$$(\tilde{K} + \lambda\,\mathrm{Id})\tilde{c} = \tilde{y}, \tag{8.7}$$

which has exactly the same structure as the normal equations without sample weights and can thus be solved in the same way. We conclude that

$$\hat{c} = \sqrt{W}(\tilde{K} + \lambda\,\mathrm{Id})^{-1}\tilde{y} = \sqrt{W}(\sqrt{W}K\sqrt{W} + \lambda\,\mathrm{Id})^{-1}\sqrt{W}y. \tag{8.8}$$

8.2.3 Generalized generalized cross-validation

We now consider cross-validation with left out data. In the following, the train set will be indexed by $I \subset \{1, \ldots, n\}$ and the test set (held out data points) will be $I^C = \{1, \ldots, n\} \setminus I$. Following a reasoning similar to that used in [Rifkin, 2007], we proceed to establish an analytical expression for the prediction error on the test set of the model fitted on the train set.

Let $\hat{f}_I$ represent the model fitted on the train set, i.e.

$$\hat{f}_I = \arg\min_f \frac{1}{2}\sum_{i\in I} w_i(f(x_i) - y_i)^2 + \frac{\lambda}{2}\|f\|_k^2 \tag{8.9}$$

Now define z^I by

$$z_k^I = \begin{cases} y_k & \text{if } k \in I \\ \hat{f}_I(x_k) & \text{if } k \in I^C \end{cases} \tag{8.10}$$

Now let $\hat{g}_I$ be the unique[2] solution to the kernel ridge regression problem with target z^I:

[2] for $\lambda > 0$ the optimization problem is strictly convex

$$\hat{g}_I = \arg\min_f \frac{1}{2}\sum_{k=1}^{n} w_k(f(x_k) - z_k^I)^2 + \frac{\lambda}{2}\|f\|_k^2. \tag{8.11}$$

Then we have the following chain of inequalities

$$\begin{aligned}
\frac{1}{2}\sum_{k=1}^{n} w_k(\hat{g}_I(x_k) - y_k)^2 + \frac{\lambda}{2}\|\hat{g}_I\|_k^2 &\geq \frac{1}{2}\sum_{i\in I} w_i(\hat{g}_I(x_i) - y_i)^2 + \frac{\lambda}{2}\|\hat{g}_I\|_k^2 \\
&\geq \frac{1}{2}\sum_{i\in I} w_i(\hat{f}_I(x_i) - y_i)^2 + \frac{\lambda}{2}\|\hat{f}_I\|_k^2 \\
&= \frac{1}{2}\sum_{i\in I} w_i(\hat{f}_I(x_i) - y_i)^2 + \frac{1}{2}\sum_{j\in I^C} w_j(\hat{f}_I(x_j) - z_j^I)^2 + \frac{\lambda}{2}\|\hat{f}_I\|_k^2 \\
&\geq \frac{1}{2}\sum_{k=1}^{n} w_k(\hat{g}_I(x_k) - y_k)^2 + \frac{\lambda}{2}\|\hat{g}_I\|_k^2,
\end{aligned}$$

from which we deduce that g_I minimizes $f \mapsto \frac{1}{2}\sum_{i\in I} w_i(f(x_i)-y_i)^2 + \frac{\lambda}{2}\|f\|_k^2$, and by unicity of the minimizer must be identical to f_I. Equivalently, f_I is the solution to the Ridge functional with data $(x_j)_{j=1,\dots,n}$ and target z^I.

Let $\hat{f}$ be the minimizer of the Ridge functional using all data $I \cup I^C$. Then, for $x_j, 1 \leq j \leq n$ we have $\hat{f}(x_j) = (K\hat{c})_j$ and similarly, $f_I(x_j) = (K\hat{c}_I)_j$. By setting $\tilde{R} = \sqrt{W}(\tilde{K} + \lambda \,\mathrm{Id})^{-1}\sqrt{W}$, we can write $\hat{c} = \tilde{R}y$ and $\hat{c}_I = \tilde{R}z^I$. Thus we obtain $\hat{f}(x_j) = (K\tilde{R}y)_j$ and $\hat{f}_I(x_j) = (K\tilde{R}z^I)_j$. Since y and z^I differ only on I^C, we compare the two and obtain

$$\begin{aligned}\hat{f}(x_j) - \hat{f}_I(x_j) &= K\tilde{R}(y - z^I) \\ &= (K\tilde{R})_{\cdot,I^C}(y - z^I)_{I^C} \\ &= (K\tilde{R})_{\cdot,I^C}(y_{I^C} - \hat{f}_I(x_{I^C}))\end{aligned}$$

For the held out data vector $(x_j)_{j\in I^C}$, we find

$$\hat{f}(x_{I^C}) - \hat{f}_I(x_{I^C}) = (K\tilde{R})_{I^C,I^C}(y_{I^C} - \hat{f}_I(x_{I^C}))\hat{f}_I(x_{I^C}) \tag{8.12}$$

Reordering yields

$$\hat{f}(x_{I^C}) - (K\tilde{R})_{I^C,I^C}y_{I^C} = (\mathrm{Id}_{I^C} - (K\tilde{R})_{I^C,I^C})\hat{f}_I(x_{I^C}) \tag{8.13}$$

Observing that $\hat{f}_I(x_{I^C})$ represent the predictions on the held out data, we are interested in solving this linear system for it. This results in

$$\begin{aligned}\hat{f}_I(x_{I^C}) &= (\mathrm{Id}_{I^C} - (K\tilde{R})_{I^C,I^C})^{-1}(\hat{f}(x_{I^C}) - (K\tilde{R})_{I^C,I^C}y_{I^C}) \\ &= (\mathrm{Id}_{I^C} - (K\tilde{R})_{I^C,I^C})^{-1}(\hat{f}(x_{I^C}) - y_{I^C}) + y_{I^C}\end{aligned}$$

Defining the cross-validation error as $e_{I^C} = y_{I^C} - \hat{f}(x_{I^C})$ and using $\hat{f}(x_{I^C}) = (K\tilde{R})_{I^C}y$, we obtain

$$e_{I^C} = (\mathrm{Id}_{I^C} - (K\tilde{R})_{I^C,I^C})^{-1}(\mathrm{Id} - (K\tilde{R}))_{I^C}y \tag{8.14}$$

Further, observing that $\mathrm{Id} - K\tilde{R} = \lambda W^{-1}\tilde{R}$, we can write

$$\begin{aligned}e_{I^C} &= ((\lambda W^{-1}\tilde{R})_{I^C,I^C})^{-1}(\lambda W^{-1}\tilde{R})_{I^C}y \\ &= (\tilde{R}_{I^C,I^C})^{-1}\tilde{R}_{I^C}y,\end{aligned}$$

which establishes a very compact expression for the held-out error vector, making strong use of the resolvent operator $\tilde{R}$.

8.2.4 Performance evaluation

We propose to thoroughly evaluate the established analytical formula in a benchmark against the "traditional" approach for cross-validation on held-out data.

Comparing to an approach that inverts the regularized kernel matrix on only the train set, we expect to see an offset in preparatory calculation by the necessity of our model to invert the full regularized kernel matrix. However, for sufficiently small test set sizes, the evaluation of held-out error can be done much faster, since for each test set no new train set kernel matrix needs to be inverted.

9 Bibliography

Alexandre Abraham, Elvis Dohmatob, Bertrand Thirion, Dimitris Samaras, and Gael Varoquaux. Extracting brain regions from rest fMRI with total-variation constrained dictionary learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8150 LNCS:607–615, 2013. ISSN 03029743. DOI: 10.1007/978-3-642-40763-5_75.

Akiyuki Anzai, Xinmiao Peng, and David C Van Essen. Neurons in monkey visual area V2 encode combinations of orientations. *Nature neuroscience*, 10(10):1313–1321, 2007. ISSN 1097-6256. DOI: 10.1038/nn1975.

Solveig Badillo, Gael Varoquaux, and Philippe Ciuciu. Hemodynamic Estimation Based on Consensus Clustering. *2013 International Workshop on Pattern Recognition in Neuroimaging*, pages 211–215, June 2013a. DOI: 10.1109/PRNI.2013.61.

Solveig Badillo, Thomas Vincent, and Philippe Ciuciu. Group-level impacts of within- and between-subject hemodynamic variability in fMRI. *NeuroImage*, 82:433–448, November 2013b. ISSN 10538119. DOI: 10.1016/j.neuroimage.2013.05.100.

Luca Baldassarre, Janaina Mourao-Miranda, and Massimiliano Pontil. Structured sparsity models for brain decoding from fMRI data. In *PRNI*, page 5, 2012.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3951 LNCS:404–417, 2006. ISSN 03029743. DOI: 10.1007/11744023_32.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm with application to linear inverse problems, 2009a.

Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 2009b.

Yousra Bekhti, Nicolas Zilber, Fabian Pedregosa, Philippe Ciuciu, Virginie Van Wassenhove, and Alexandre Gramfort. Decoding perceptual thresholds from MEG/EEG. In *Pattern Recoginition in Neuroimaging (PRNI) (2014)*, Tubingen, Germany, 2014.

Guy Ben-Yosef and Ohad Ben-Shahar. Curvature-based perceptual singularities and texture saliency with early vision mechanisms. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 25(8):1974–1993, 2008. ISSN 1084-7529. DOI: 10.1364/JOSAA.25.001974.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

Shlomo Bentin, Truett Allison, Aina Puce, Erik Perez, and Gregory McCarthy. Electrophysiological studies of face perception in humans. *Journal of cognitive neuroscience*, 8:551, 1996.

Katherine C Bettencourt and Yaoda Xu. The role of transverse occipital sulcus in scene perception and its relationship to object individuation in inferior intraparietal sulcus. *Journal of cognitive neuroscience*, 25(10):1711–22, 2013. DOI: 10.1162/jocn. URL http://www.ncbi.nlm.nih.gov/pubmed/23662863.

Rainer Boegle, Julian Maclaren, and Maxim Zaitsev. Combining prospective motion correction and distortion correction for epi: towards a comprehensive correction of motion and susceptibility-induced artifacts. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 23(4):263–273, 2010. ISSN 0968-5243. DOI: 10.1007/s10334-010-0225-8. URL http://dx.doi.org/10.1007/s10334-010-0225-8.

León Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. *Proceedings of COMPSTAT'2010*, pages 177–186, 2010.

Geoffrey M Boynton, Stephen A Engel, Gary H Glover, and David J Heeger. Linear Systems Analysis of Functional Magnetic Resonance Imaging in Human V1. 16(13):4207–4221, 1996.

Joan Bruna. *Scattering Representations for Recognition*. PhD thesis, Ecole Polytechnique, 2013.

Joan Bruna and Stephane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013. ISSN 01628828. DOI: 10.1109/TPAMI.2012.230.

Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan a. Solomon, Najib J. Majaj, and James J. DiCarlo. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *Arxiv*, 10(12):35, 2014. ISSN 15537358. DOI: 10.1371/journal.pcbi.1003963. URL http://arxiv.org/abs/1406.3284.

Emmanuel Candes and Justin Romberg. Signal recovery from random projections. In *Proc. SPIE*, volume 5674, page 76, 2005.

John Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6), 1986. ISSN 0162-8828. DOI: 10.1109/TPAMI.1986.4767851.

Jonathan S. Cant, Stephen R. Arnott, and Melvyn a. Goodale. fMR-adaptation reveals separate processing regions for the perception of form and texture in the human ventral stream. *Experimental Brain Research*, 192:391–405, 2009. ISSN 00144819. DOI: 10.1007/s00221-008-1573-8.

Ramon Casanova, Srikanth Ryali, John Serences, Lucie Yang, Robert Kraft, Paul J Laurienti, and Joseph a Maldjian. The impact of temporal regularization on estimates of the BOLD hemodynamic response function: a comparative analysis. *NeuroImage*, 40(4):1606–18, May 2008. ISSN 1053-8119. DOI: 10.1016/j.neuroimage.2008.01.011.

Tolga Çukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature neuroscience*, 16(April): 763–70, 2013. ISSN 1546-1726. DOI: 10.1038/nn.3381. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3929490&tool=pmcentrez&rendertype=abstract`.

L Chaari, F Forbes, T Vincent, and P Ciuciu. Hemodynamic-informed parcellation of fMRI data in a joint detection estimation framework. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 15(Pt 3):180–8, January 2012.

Tony F Chan and Luminita A Vese. Active contours without edges. *Image processing, IEEE transactions on*, 10:266, 2001.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3): 455–62, 2014. ISSN 1546-1726. DOI: 10.1038/nn.3635. URL `http://www.ncbi.nlm.nih.gov/pubmed/24464044`.

Philippe Ciuciu, Jean-Baptiste Poline, Guillaume Marrelec, Jérôme Idier, Christophe Pallier, and Habib Benali. Unsupervised robust nonparametric estimation of the hemodynamic response function for any fMRI experiment. *IEEE transactions on Medical Imaging*, 22(10):1235–51, October 2003. ISSN 0278-0062. DOI: 10.1109/TMI.2003.817759.

Patrick L. Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing, 2011.

David D Cox and Robert L Savoy. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2):261–270, June 2003. ISSN 10538119. DOI: 10.1016/S1053-8119(03)00049-1.

Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 886–893, 2005. ISSN 1063-6919. DOI: 10.1109/CVPR.2005.177. URL `citeulike-article-id:3047126$\delimiter"026E30F$nhttp://dx.doi.org/10.1109/CVPR.2005.177`.

a M Dale. Optimal experimental design for event-related fMRI. *Human brain mapping*, 8(2-3):109–14, January 1999. ISSN 1065-9471.

David Degras and Martin A. Lindquist. A hierarchical model for simultaneous detection and estimation in multi-subject fMRI studies. *NeuroImage*, 98C:61–72, 2014.

R Desimone, TD Albright, CG Gross, and C Bruce. Stimulus-selective Properties of Inferior Temporal Neurons in the Macaque. *Journal of Neuroscience*, 4(8):2051–2062, 1984.

Elvis Dohmatob, Alexandre Gramfort, Bertrand Thirion, Gaël Varoquaux, et al. Benchmarking solvers for tv-l1 least-squares and logistic regression in brain imaging. *Pattern Recognition in Neuroimaging (PRNI)*, 2014.

P E Downing, Y Jiang, M Shuman, and N Kanwisher. A cortical area selective for visual processing of the human body. *Science (New York, N.Y.)*, 293 (September):2470–2473, 2001. ISSN 0036-8075. DOI: 10.1126/science.1063414.

O M Doyle, J Ashburner, F O Zelaya, S C R Williams, M a Mehta, and a F Marquand. Multivariate decoding of brain images using ordinal regression. *NeuroImage*, 81:347–57, November 2013. ISSN 1095-9572. DOI: 10.1016/j.neuroimage.2013.05.036.

S. Durrleman, Ma. Prastawa, G. Gerig, and S. Joshi. Optimal data-driven sparse parameterization of diffeomorphisms for population analysis. In *Information Processing in Medical Imaging*, page 123, 2011.

Michael Eickenberg, Fabian Pedregosa, Mehdi Senoussi, Alexandre Gramfort, and Bertrand Thirion. Second order scattering descriptors predict fMRI activity due to visual textures. In *Pattern Recognition in NeuroImaging, IEEE International Workshop on*, pages 5–8, 2013.

Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 8: 667–698, 2012. ISSN 13489151.

R Epstein and N Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(April):598–601, 1998. ISSN 0028-0836. DOI: 10.1038/33402.

M Fabre-Thorpe, A Delorme, C Marlot, and S Thorpe. A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of cognitive neuroscience*, 13:171–180, 2001. ISSN 0898-929X. DOI: 10.1162/089892901564234.

DJ Felleman and DC Van Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1991. URL `http://cercor.oxfordjournals.org/content/1/1/1.1.short`.

Olivier Fercoq, Institut Mines-t, Joseph Salmon, and Telecom Paristech. Mind the duality gap : safer rules for the Lasso. *Journal of Machine Learning Research*, 37, 2015.

Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011. ISSN 1097-6256. DOI: 10.1038/nn.2889.

Jeremy Freeman, Corey M Ziemba, David J Heeger, Eero P Simoncelli, and J Anthony Movshon. A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, 16(7):974–81, 2013. ISSN 1546-1726. DOI: 10.1038/nn.3402. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3710454&tool=pmcentrez&rendertype=abstract.

K J Friston, A Mechelli, R Turner, and C J Price. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, 12:466–477, 2000. ISSN 1053-8119. DOI: 10.1006/nimg.2000.0630.

Karl J Friston, A. P Holmes, and J. P. Poline. Statistical parametric maps in functional imaging : A general linear approach. 1995.

Karl J Friston, Oliver Josephs, Geraint Rees, and Robert Turner. Nonlinear event-related responses in fMRI. *Magnetic Resonance in Medicine*, 39(1): 41–52, 1998.

Karl J Friston, John T Ashburner, Stefan J Kiebel, Thomas E Nichols, and William D Penny. *Statistical parametric mapping: The analysis of functional brain images: The analysis of functional brain images*. Academic Press, 2011.

Jean-Jacques Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Trans. on I.T*, page 3601, 2005.

J L Gallant, C E Connor, S Rakshit, J W Lewis, and D C Van Essen. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *Journal of neurophysiology*, 76(4):2718–2739, 1996. ISSN 0022-3077.

J L Gallant, C E Connor, and D C Van Essen. Neural activity in areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing. *Neuroreport*, 9:2153–2158, 1998. ISSN 0959-4965. DOI: 10.1097/00001756-199806220-00045.

G H Glover. Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9(4):416–29, April 1999. ISSN 1053-8119.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

M Goodale and D Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(I):20–25, 1992. ISSN 01662236. DOI: 10.1016/0166-2236(92)90344-8.

C Goutte, F a Nielsen, and L K Hansen. Modeling the haemodynamic response in fMRI using smooth FIR filters. *IEEE transactions on Medical Imaging*, 19(12):1188–201, December 2000. ISSN 0278-0062. DOI: 10.1109/42.897811.

A. Gramfort, B Thirion, and G Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *PRNI*, pages 17–20, 2013.

Kalanit Grill-Spector, Zoe Kourtzi, and Nancy Kanwisher. The lateral occipital complex and its role in object recognition. *Vision Research*, 41:1409–1422, 2001. ISSN 00426989. DOI: 10.1016/S0042-6989(01)00073-6.

Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E Taylor. Interpretable whole-brain prediction analysis with graphnet. *NeuroImage*, 72:304–321, 2013.

Umut Güçlü and Marcel a. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the brain's ventral visual pathway. *arXiv preprint arXiv:1411.6422*, 2014. URL `http://arxiv.org/abs/1411.6422v1`.

Luke E Hallum, Michael S Landy, and David J Heeger. Human primary visual cortex (V1) is selective for second-order spatial frequency. *Journal of neurophysiology*, 105:2121–2131, 2011. ISSN 0022-3077. DOI: 10.1152/jn.01007.2010.

Daniel a Handwerker, John M Ollinger, and Mark D'Esposito. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, 21(4):1639–51, April 2004. ISSN 1053-8119. DOI: 10.1016/j.neuroimage.2003.11.029.

J V Haxby, M I Gobbini, M L Furey, a Ishai, J L Schouten, and P Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science (New York, N.Y.)*, 293(5539):2425–30, September 2001. ISSN 0036-8075. DOI: 10.1126/science.1063736.

John-Dylan Haynes and Geraint Rees. Decoding mental states from brain activity in humans. *Nature reviews. Neuroscience*, 7(7):523–34, July 2006. ISSN 1471-003X. DOI: 10.1038/nrn1931.

Linda Henriksson, Juha Karvonen, Niina Salminen-Vaparanta, Henry Railo, and Simo Vanni. Retinotopic maps, spatial tuning, and locations of human visual areas in surface coordinates characterized with multifocal and blocked fmri designs. *PLoS ONE*, 7(5):e36859, 05 2012. DOI: 10.1371/journal.pone.0036859. URL `http://dx.doi.org/10.1371`.

R N a Henson, C J Price, M D Rugg, R Turner, and K J Friston. Detecting latency differences in event-related BOLD responses: application to words versus nonwords and initial versus repeated face presentations. *NeuroImage*, 15(1):83–97, January 2002. ISSN 1053-8119. DOI: 10.1006/nimg.2001.0940.

Elizabeth M C Hillman. Optical brain imaging in vivo: techniques and applications from animal to man. *Journal of biomedical optics*, 12(5):051402, 2008. ISSN 10833668. DOI: 10.1117/1.2789693.

Roger A Horn and Charles R Johnson. *Topics in matrix analysis*. Cambridge university press, 1991.

D H Hubel and T N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148:574–591, 1959. ISSN 1469-7793. DOI: 10.1113/jphysiol.2009.174151.

Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, 76:1210–1224, 2012. ISSN 08966273. DOI: 10.1016/j.neuron.2012.10.014.

B Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290:91–97, 1981. ISSN 0028-0836. DOI: 10.1038/290091a0.

Benjamin M Kandel, David A Wolk, James C Gee, and Brian Avants. Predicting cognitive data from medical images using sparse linear regression. In *Information Processing in Medical Imaging*, page 86. Springer, 2013.

Nancy Kanwisher, Josh Mcdermott, and Marvin M Chun. The Fusiform Face Area : A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, 17(11):4302–4311, 1997.

S Kastner, P De Weerd, and L G Ungerleider. Texture segregation in the human visual cortex: A functional MRI study. *Journal of neurophysiology*, 83:2453–2457, 2000. ISSN 0022-3077.

K N Kay, T Naselaris, and J Gallant. fmri of human visual areas in response to natural images. crcns.org., 2011a.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185): 352–5, March 2008. ISSN 1476-4687. DOI: 10.1038/nature06713.

Kendrick N. Kay, Naselaris, and Jack L. Gallant. fMRI of human visual areas in response to natural images. *CRCNS.org*, 2011b. DOI: http://dx.doi.org/10.6080/K0QN64NG.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2): 81–93, 1938.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised , but Not Unsupervised , Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11), 2014. DOI: 10.1371/journal.pcbi.1003915.

Seyed-Mahdi Khaligh-Razavi, Linda Henriksson, Kendrick Kay, and Nikolaus Kriegeskorte. Explaining the hierarchy of visual representational geometries by remixing of features from many computational vision models. *bioRxiv*, 2014. DOI: 10.1101/009936.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 2008a.

Nikolaus Kriegeskorte, Marieke Mur, Douglas A. Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A. Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, Dec 2008b. DOI: 10.1016/j.neuron.2008.10.043. URL `http://dx.doi.org/10.1016/j.neuron.2008.10.043`.

a Krizhevsky, I Sutskever, and Ge Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pages 1097–1105, 2012.

Ms Landy. Texture analysis and perception. *The New Visual Neurosciences*, 10003(212):639–652, 2013. URL `http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Texture+analysis+and+perception#0`.

Jonas Larsson, Michael S Landy, and David J Heeger. Orientation-selective adaptation to first-and second-order patterns in human visual cortex. *Journal of Neurophysiology*, 95:862–881, 2006. DOI: 10.1152/jn.00668.2005. URL `http://jn.physiology.org/content/95/2/862.short`.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1265–1278, 2005.

Y LeCun. Une procédure d'apprentissage pour réseau a seuil asymmetrique (a Learning Scheme for Asymmetric Threshold Networks). In *Proceedings of Cognitiva 85*, pages 599–604, Paris, France, 1985.

Yu Lei, Li Tong, and Bin Yan. A mixed L2 norm regularized HRF estimation method for rapid event-related fMRI experiments. *Computational and mathematical methods in medicine*, 2013:643129, January 2013. ISSN 1748-6718. DOI: 10.1155/2013/643129.

Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001. ISSN 09205691. DOI: 10.1023/A:1011126920638.

Martin A. Lindquist and Tor D Wager. Validity and power in hemodynamic response modeling: A comparison study and a new approach. *Hum Brain Mapp*, 28(8):764–784, 2007. DOI: 10.1002/hbm.20310.Validity.

Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

N K Logothetis, J Pauls, and T Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current biology : CB*, 5(5):552–563, 1995. ISSN 0960-9822. DOI: 10.1016/S0960-9822(95)00108-4.

N K Logothetis, J Pauls, M Augath, T Trinath, and A Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412: 150–157, 2001. ISSN 0028-0836. DOI: 10.1038/35084005.

D.G. Lowe. Object Recognition fromLocal Scale-Invariant Features. *IEEE International Conference on Computer Vision*, 1999. ISSN 0-7695-0164-8. DOI: 10.1109/ICCV.1999.790410.

S. Makni, P. Ciuciu, J. Idier, and J.-B. Poline. Joint detection-estimation of brain activity in functional MRI: a Multichannel Deconvolution solution. *IEEE Transactions on Signal Processing*, 53(9):3488–3502, September 2005. ISSN 1053-587X. DOI: 10.1109/TSP.2005.853303.

Salima Makni, Christian Beckmann, Steve Smith, and Mark Woolrich. Bayesian deconvolution of fMRI data using bilinear dynamical systems. *NeuroImage*, 42(4):1381–96, October 2008. ISSN 1095-9572. DOI: 10.1016/j.neuroimage.2008.05.052.

Stéphane Mallat. Group Invariant Scattering. *Communications on Pure and Applied Mathematics*, 65:1331–1398, 2012. ISSN 00103640. DOI: 10.1002/cpa.21413.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *J Cogn Neurosci*, 19(9):1498–1507, Sep 2007.

Guillaume Marrelec, Habib Benali, Philippe Ciuciu, Mélanie Pélégrini-Issac, and Jean-Baptiste Poline. Robust bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. *Human Brain Mapping*, 19(1):1–17, 2003.

Vincent Michel, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, and Bertrand Thirion. Total variation regularization for fMRI-based prediction of behavior. *Medical Imaging, IEEE Transactions on*, 30:1328, 2011.

K. Mikolajczyk, K. Mikolajczyk, C. Schmid, and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005. ISSN 0162-8828. DOI: 10.1109/TPAMI.2005.188. URL `http://doi.ieeecomputersociety.org/10.110910.1109/TPAMI.2005.188`.

M. Mishkin and L. G. Ungerleider. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural Brain Research*, 6:57–77, 1982. ISSN 01664328. DOI: 10.1016/0166-4328(82)90081-X.

Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–29, December 2008. ISSN 1097-4199. DOI: 10.1016/j.neuron.2008.11.004.

Leila Montaser-Kouhsari, Michael S Landy, David J Heeger, and Jonas Larsson. Orientation-selective adaptation to illusory contours in human visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(9):2186–2195, 2007. ISSN 0270-6474. DOI: 10.1523/JNEUROSCI.4173-06.2007.

Janaina Mourão Miranda, Karl J Friston, and Michael Brammer. Dynamic discrimination analysis: a spatial-temporal SVM. *NeuroImage*, 36(1):88–99, May 2007. ISSN 1053-8119. DOI: 10.1016/j.neuroimage.2007.02.020.

David Mumford and Jayant Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42:577, 1989.

Jeanette a Mumford, Benjamin O Turner, F Gregory Ashby, and Russell a Poldrack. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, 59(3):2636–43, February 2012. ISSN 1095-9572. DOI: 10.1016/j.neuroimage.2011.08.076.

Sangnam Nam, Mike E Davies, Michael Elad, and Rémi Gribonval. The cosparse analysis model and algorithms. *App Comp Harmonic Analysis*, 34: 30, 2013.

Thomas Naselaris, Ryan J. Prenger, Kendrick N. Kay, Michael Oliver, and Jack L. Gallant. Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, 63(6):902–915, 2009a. ISSN 08966273. DOI: 10.1016/j.neuron.2009.09.006. URL http://dx.doi.org/10.1016/j.neuron.2009.09.006.

Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009b.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–10, May 2011. ISSN 1095-9572. DOI: 10.1016/j.neuroimage.2010.07.073.

Bernard Ng, Arash Vahdat, Ghassan Hamarneh, and Rafeef Abugharbieh. Generalized sparse classifiers for decoding cognitive states in fmri. In *Machine Learning in Medical Imaging*, pages 108–115. Springer, 2010.

Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21:1641–1646, 2011. ISSN 09609822. DOI: 10.1016/j.cub.2011.08.031.

Jorge Nocedal and S Wright. Numerical optimization, series in operations research and financial engineering. *Springer, New York*, 2006.

H. C. Nothdurft. Texture segmentation and pop-out from orientation contrast. *Vision Research*, 31(6):1073–1078, 1991. ISSN 00426989. DOI: 10.1016/0042-6989(91)90211-M.

S Ogawa and Tm Lee. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the …*, 87(December):9868–72, 1990. ISSN 0027-8424. DOI: 10.1073/pnas.87.24.9868. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1262394&tool=pmcentrez&rendertype=abstract$\delimiter"026E30F$nhttp://www.pnas.org/content/87/24/9868.short.

Gouki Okazawa, Satohiro Tajima, and Hidehiko Komatsu. Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proceedings of the National Academy of Sciences*, 112:E351–E360, 2015. ISSN 0027-8424. DOI: 10.1073/pnas.1415146112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1415146112.

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

H S Orbach, L B Cohen, and A Grinvald. Optical mapping of electrical activity in rat somatosensory and visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 5:1886–1895, 1985. ISSN 0270-6474.

Fabian Pedregosa, Olivier Grisel, Ron Weiss, Alexandre Passos, and Matthieu Brucher. Scikit-learn : Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Fabian Pedregosa, Michael Eickenberg, Bertrand Thirion, and Alexandre Gramfort. HRF estimation improves sensitivity of fMRI encoding and decoding models. *Proceedings of the 3nd International Workshop on Pattern Recognition in NeuroImaging (2013)*, pages 3–6, 2013.

Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6314 LNCS:143–156, 2010. ISSN 03029743. DOI: 10.1007/978-3-642-15561-1_11.

Nicolas Pinto, David D. Cox, and James J. DiCarlo. Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4(1):0151–0156, 2008. ISSN 1553734X. DOI: 10.1371/journal.pcbi.0040027.

Thomas Pock, Antonin Chambolle, Daniel Cremers, and Horst Bischof. A convex relaxation approach for computing minimal partitions. In *CVPR*, page 810, 2009.

Russell A Poldrack, Jeanette Mumford, and Thomas Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, Cambridge, 2011a. ISBN 9780511895029. DOI: 10.1017/cbo9780511895029. URL `http://dx.doi.org/10.1017/cbo9780511895029`.

Russell A. Poldrack, Jeanette A. Mumford, and Thomas E. Nichols. *Handbook of Functional MRI Data Analysis*. Cambridge University Press, 2011b.

Jean-Baptiste Poline and Matthew Brett. The general linear model and fMRI: does love last forever? *NeuroImage*, 62(2):871–80, August 2012. ISSN 1095-9572. DOI: 10.1016/j.neuroimage.2012.01.133.

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):49–71, 2000.

M E Raichle, a M MacLeod, a Z Snyder, W J Powers, D a Gusnard, and G L Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):676–682, 2001. ISSN 0027-8424. DOI: 10.1073/pnas.98.2.676.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2:1019–1025, 1999.

R Rifkin. Notes on regularized least squares. *Massachusetts Institute of Technology*, 2007. URL `http://citeseerx.`

ist.psu.edu/viewdoc/download?doi=10.1.1.169.8519&rep=rep1&type=pdf$\delimiter"026E30F$npapers3://publication/uuid/9F0A259E-0179-4532-8962-0324B1D7987E.

Anna W. Roe, Leonardo Chelazzi, Charles E. Connor, Bevil R. Conway, Ichiro Fujita, Jack L. Gallant, Haidong Lu, and Wim Vanduffel. Toward a Unified Theory of Visual Area V4. *Neuron*, 74(1):12–29, 2012. ISSN 08966273. DOI: 10.1016/j.neuron.2012.03.011. URL http://dx.doi.org/10.1016/j.neuron.2012.03.011.

Joachim Röhmel and Ulrich Mansmann. Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal*, 41(2): 149–170, 1999.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259, 1992.

Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8): 1627–1639, 1964.

Mark M Schira, Alex R Wade, and Christopher W Tyler. Two-dimensional mapping of the central and parafoveal visual field to human visual cortex. *Journal of neurophysiology*, 97(March 2007):4284–4295, 2007. ISSN 0022-3077. DOI: 10.1152/jn.00972.2006.

Mark Schmidt, Nicolas L. Roux, and Francis R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1458–1466. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4452-convergence-rates-of-inexact-proximal-gradient-methods-for-convex-optimization.pdf.

Sanne Schoenmakers, Markus Barth, Tom Heskes, and Marcel van Gerven. Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83:951–961, July 2013. ISSN 1095-9572. DOI: 10.1016/j.neuroimage.2013.07.043.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

MI Sereno, AM Dale, and JB Reppas. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 1995. URL http://www.sciencemag.org/content/268/5212/889.short.

Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. OverFeat : Integrated Recognition , Localization and Detection using Convolutional Networks. *arXiv preprint arXiv:1312.6229*, pages 1–15, 2013. URL http://arxiv.org/abs/1312.6229.

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007. ISSN 01628828. DOI: 10.1109/TPAMI.2007.56.

Eero P Simoncelli and William T Freeman. The Steerable Pyramid: A Flexible Multi-Scale Derivative Computation. *Conference, Ieee International Processing, Image (Rochester, N.Y.)*, III:444–447, 1995.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv preprint arXiv:1312.6034*, pages 1–8, 2013. URL `http://arxiv.org/abs/1312.6034`.

D Smetters, A Majewska, and R Yuste. Detecting action potentials in neuronal populations with calcium imaging. *Methods (San Diego, Calif.)*, 18: 215–221, 1999. ISSN 1046-2023. DOI: 10.1006/meth.1999.0774.

M. Stehling, R Turner, and P Mansfield. Echo-planar imaging: magnetic resonance imaging in a fraction of a second. *Science*, 254(NOVEMBER 1991): 43–50, 1991. ISSN 0036-8075. DOI: 10.1126/science.1925560.

C. Stonnington, C. Chu, S. Klöppel, C. Jack, J. Ashburner, and R. Frackowiak. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *Neuroimage*, 51:1405, 2010.

I Tasaki, A Watanabe, R Sandlin, and L Carnay. Changes in fluorescence, turbidity, and birefringence associated with nerve excitation. *Proceedings of the National Academy of Sciences of the United States of America*, 61(Rca 4463):883–888, 1968. ISSN 0027-8424. DOI: 10.1073/pnas.61.3.883.

John C Taylor, Alison J Wiggett, and Paul E Downing. Functional MRI analysis of body and body part representations in the extrastriate and fusiform body areas. *Journal of neurophysiology*, 98:1626–1633, 2007. ISSN 0022-3077. DOI: 10.1152/jn.00012.2007.

a. Thielscher and H. Neumann. Neural mechanisms of cortico-cortical interaction in texture boundary detection: A modeling approach. *Neuroscience*, 122:921–939, 2003. ISSN 03064522. DOI: 10.1016/j.neuroscience.2003.08.050.

S Thorpe, D Fize, and C Marlot. Speed of processing in the human visual system., 1996. ISSN 0028-0836.

Engin Tola, Vincent Lepetit, Pascal Fua, and Senior Member. DAISY: An efficient dense descriptor applied to wide-baseline stereo. 32(5):815–830, 2010.

Sabrina M Tom, Craig R Fox, Christopher Trepel, and Russell a Poldrack. The neural basis of loss aversion in decision-making under risk. *Science (New York, N.Y.)*, 315(5811):515–8, January 2007a. ISSN 1095-9203. DOI: 10.1126/science.1134239.

Sabrina M Tom, Craig R Fox, Christopher Trepel, and Russell A Poldrack. The neural basis of loss aversion in decision-making under risk. *Science*, 315 (5811):515–8, Jan 2007b. DOI: 10.1126/science.1134239.

Benjamin O Turner, Jeanette a Mumford, Russell a Poldrack, and F Gregory Ashby. Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage*, 62(3):1429–38, September 2012. ISSN 1095-9572. DOI: 10.1016/j.neuroimage.2012.05.057.

Thomas Vincent, Laurent Risser, and Philippe Ciuciu. Spatially adaptive mixture modeling for analysis of fMRI time series. *IEEE Transactions on Medical Imaging*, 29(4):1059–1074, 2010.

Vincent Q Vu, Pradeep Ravikumar, Thomas Naselaris, Kendrick N Kay, Jack L Gallant, and Bin Yu. Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models. *The annals of applied statistics*, 5(2B):1159, 2011.

M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming. *Trans Inf Theory*, 55:2183, 2009.

Brian Wandell, Serge O. Dumoulin, and Alyssa Brewer. Visual field maps in human cortex. *Neuron*, 56(1893):366–383, 2007. ISSN 08966273. DOI: 10.1016/j.neuron.2007.10.012.

Jiaping Wang, Hongtu Zhu, Jianqing Fan, Kelly Giovanello, and Weili Lin. Multiscale adaptive smoothing models for the hemodynamic response function in fMRI. *The Annals of Applied Statistics*, 7(2):904–935, June 2013. ISSN 1932-6157. DOI: 10.1214/12-AOAS609.

Jie Wang, Binbin Lin, Pinghua Gong, Peter Wonka, and Jieping Ye. Lasso screening rules via dual polytope projection. *CoRR*, abs/1211.3966, 2012.

T. Aisling Whitaker, Cristina Simões Franklin, and Fiona N. Newell. Vision and touch: Independent or integrated systems for the perception of texture? *Brain Research*, 1242:59–72, 2008. ISSN 00068993. DOI: 10.1016/j.brainres.2008.05.037. URL `http://dx.doi.org/10.1016/j.brainres.2008.05.037`.

Mark W Woolrich, Timothy E J Behrens, and Stephen M Smith. Constrained linear basis sets for HRF modelling using variational bayes. *NeuroImage*, 21(4):1748–61, April 2004. ISSN 1053-8119. DOI: 10.1016/j.neuroimage.2003.12.024.

Essa Yacoub, Noam Harel, and Kâmil Ugurbil. High-field fMRI unveils orientation columns in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105(30):10607–10612, 2008. ISSN 0027-8424. DOI: 10.1073/pnas.0804110105.

Okito Yamashita, Masa-aki Sato, Taku Yoshioka, Frank Tong, and Yukiyasu Kamitani. Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns. *NeuroImage*, 42(4):1414–1429, 2008.

Daniel L K Yamins, Ha Hong, Charles F Cadieu, Ethan a Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*,

111:8619–24, 2014. ISSN 1091-6490. DOI: 10.1073/pnas.1403112111. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4060707&tool=pmcentrez&rendertype=abstract.

Tingting Zhang, Fan Li, Lane Beckes, Casey Brown, and James A. Coan. Nonparametric inference of the hemodynamic response using multi-subject fMRI data. *NeuroImage*, 63(3):1754–65, November 2012. ISSN 1095-9572.

Tingting Zhang, Fan Li, Lane Beckes, and James a Coan. A semi-parametric model of the hemodynamic response for multi-subject fMRI data. *NeuroImage*, 75:136–45, July 2013. ISSN 1095-9572. DOI: 10.1016/j.neuroimage.2013.02.048.