Feature extraction and supervised learning on fMRI: from practice to theory

Fabian Pedregosa-Izquierdo

UNIVERSITÉ PIERRE ET MARIE CURIE

DOCTORAL SCHOOL OF COMPUTER SCIENCE

PREPARED AT PARIETAL TEAM - INRIA SACLAY

Feature extraction and supervised learning on fMRI: from practice to theory

Fabian Pedregosa-Izquierdo

A dissertation submitted in partial fulfillement

of the requirements for the degree of doctor of science,

specialized in computer science.

Defended publicly the 20th of February 2015 in front of a jury composed of

Advisors	Francis Bach	INRIA / ENS, Paris, France
	Alexandre Gramfort	Telecom Paristech, Paris, France
Reviewers	Dimitri Van de Ville	Univ. Geneva / EPFL, Geneva, CH
	Alain Rakotomamonjy	University of Rouen, Rouen, France
Examiners	Ludovic Denoyer	UPMC, Paris, France
	Bertrand Thirion	INRIA / CEA, Saclay, France
	Marcel Van Gerven	Donders Instute, Nijmegen, NL

UNIVERSITÉ PIERRE ET MARIE CURIE

ÉCOLE DOCTORALE INFORMATIQUE,
TÉLÉCOMMUNICATIONS ET ÉLECTRONIQUE

ÉQUIPE PARIETAL - INRIA SACLAY

Estimation de variables et apprentissage supervisé en IRMf: de la pratique à la théorie

Fabian Pedregosa-Izquierdo

Thèse de doctorat pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Dirigée par Francis Bach et Alexandre Gramfort.

Présentée et soutenue publiquement le 20 Février 2015 devant un jury composé de :

Directeurs	Francis Bach	INRIA / ENS, Paris, France
	Alexandre Gramfort	Telecom Paristech, Paris, France
Rapporteurs	Dimitri Van de Ville	Univ. Geneva / EPFL, Geneva, CH
	Alain Rakotomamonjy	University of Rouen, Rouen, France
Examinateurs	Ludovic Denoyer	UPMC, Paris, France
	Bertrand Thirion	INRIA / CEA, Saclay, France
	Marcel Van Gerven	Donders Instute, Nijmegen, NL

Abstract

Until the advent of non-invasive neuroimaging modalities the knowledge of the human brain came from the study of its lesions, post-mortem analyses and invasive experimentations. Nowdays, modern imaging techniques such as fMRI are revealing several aspects of the human brain with progressively high spatio-temporal resolution. However, in order to answer increasingly complex neuroscientific questions the technical improvements in acquisition must be matched with novel data analysis methods. In this thesis we examine different applications of machine learning to the processing of fMRI data. We propose novel extensions and investigate the theoretical properties of different models.

Often the data acquired through the fMRI scanner follows a *feature extraction* step in which time-independent activation coefficients are extracted from the fMRI signal. The first contribution of this thesis is the introduction a model named Rank-1 GLM (R1-GLM) for the joint estimation of time-independent activation coefficients and the hemodynamic response function (HRF). We quantify the improvement of this approach with respect to existing procedures on different fMRI datasets.

The second part of this thesis is devoted to the problem of fMRI-based *decoding*, i.e., the task of predicting some information about the stimuli from brain activation maps. From a statistical standpoint, this problem is challenging due to the high dimensionality of the data, often thousands of variables, while the number of images available for training is small, typically a few hundreds. We examine the case in which the target variable consist of discretely ordered values. The second contribution of this thesis is to propose the following two metrics to assess the performance of a decoding model: the absolute error and pairwise disagreement. We describe several models that optimize a convex surrogate of these loss functions and examine their performance on different fMRI datasets.

Motivated by the success of some ordinal regression models for the task of fMRI-based decoding, we turn to study some theoretical properties of these methods. The property that we investigate is known as *consistency* or *Fisher consistency* and relates the minimization of a loss to the minimization of its surrogate. The third, and most theoretical, contribution of this thesis is to examine the consistency properties of a rich family of surrogate loss functions that are used in the context of ordinal regression. We give sufficient conditions for the consistency of the surrogate loss functions considered. This allows us to give theoretical reasons for some empirically observed differences in performance between surrogates.

Keywords: fMRI, BOLD, HRF, feature extraction, supervised learning, ranking, ordinal regression, decoding, encoding.

Acknowledgements

My first words of gratitude will be for my advisors. I would like to thank Alexandre Gramfort for sharing with me during these three years his passion, his expertise and his time. You have introduced me into the world of research while at the same time living me the freedom to pursue my goals, and I will always be in debt for this. Also a great thanks to Francis Bach, who always had time for my quesions. Thanks you for your patience, for encouraging me when I drifted me into areas that were new to me and for enlightening remarks on several aspects of my work. I would also like to thank Bertrand Thirion, my "advisor in the shadow", for hosting me within the Parietal team and for sharing with many of the ideas that are developed within this thesis. Your honesty, patience and thoroughness are a continuous source of inspiration. I would also like to thank Bertrand and Gael Varoquaux for creating a unique work environment at the Parietal team: it has been a pleasure to be a part of this lab.

This work would not have been possible without the help of my coauthors. I would first like to thank Michael Eickenberg for the work we did together and for being always enthusiastic about new ideas. I would also like to thank Philippe Ciuciu for sharing his expertise on HRF estimation with me. The rest of the Parietal team also deserves a mention for coping with me during so much time: Gael Varoquaux (for bringing me to France five years ago), Régine Bricquet (a dedicated assistant makes a big difference), Elvis "amigo amigo computador" Dohmatob, Danilo Bzdok, Vincent "comme ta soeur" Michel, Aina "sobrasada" Frau, Fernando Yepes, Mehdi Rahim, Alexandre Abraham, Virgile Fritsch, Jean Kossaifi, Andres Hoyos, Loic Esteve, Yannick "horrible personne" Schwarz, Olivier Grisel, Salma Bougacha, Philippe Gervais, Benoit "petaflop" Da Mota, Bernard Ng, Viviana "reghishtrashion" Siless, Solveig Badillo, Nicolas Chauffert and Matthieu Kowalski. I've also had the pleasure to interact with people from the Unicog team, from which I would like to mention Valentina Borghesani, Manuela Piazza, Christophe Pallier, Elodie Cauvet, Evelyn Eger, Lucie Charles and Pedro Pinhero Chagas. I'm also grateful to the scikit-learn crowd for teaching me so much about programming and machine learning: Andreas Mueller, Vlad Niculae, Lars Buitinck, Mathieu Blondel, Jake VanderPlas, Peter Prettenhofer and many others.

I would equally like to thanks Alain Rakotomamonjy and Dimitri Van de Ville for accepting to review this manuscript and to Ludovic Denoyer, Marcel Van Gerven and Bertrand Thirion for accepting to be part of the thesis defense jury.

Mis últimas palabras son para mi familia: para mi madre, mi padre y mi abuela. Nada de esto habría sido posible sin vuestro apoyo. Un agradecimento especial para Vale, por todo y por llegar en el momento adecuado. También para los amigos de toda la vida, aquellos con los que no existe la distancia: Aitor Frías, Hugo Romero y Ángel Soler. Un grazie anche alla mia nuova famiglia italiana, per avermi fatto sentire il benvenuto nella vostra vita.

Notation

Notation	Name	Definition
$\Gamma(x)$	Gamma function	$\Gamma(x) = \int_0^\infty x^{t-1}e^{-x}dx$
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ	
$\|\mathbf{x}\|$ or $\|\mathbf{x}\|_2$	Euclidean norm for vectors	$\sqrt{\sum_i x_i^2}$
$\|\mathbf{x}\|_\mathcal{F}$	Frobenius norm of a matrix	$\sqrt{\sum_i \sum_j \mathbf{X}_{ij}^2}$
$\mathbf{I}_n$	Identity matrix of size n	$I_{ij} = \delta_{ij},\ \forall 1 \leq i, j \leq n$
$\mathbf{1}_n$	Vector of ones of size n	$1_i = 1,\ \forall 1 \geq i \geq n$
$\text{tr}(\mathbf{A})$	Trace of a matrix	$\sum_i A_{ii}$
$\mathbf{X}^\dagger$	Moore-Penrose pseudoinverse	Generalized inverse matrix
$\mathbf{A} \otimes \mathbf{B}$	Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$	
$\text{vec}(\mathbf{A})$	Vectorization of a matrix	Concatenation of the columns of a matrix into a single column vector
$\mathbb{E}(X)$	Expectation of the random variable X	$\int XdP$
$\mathcal{R}_\ell(h)$	Risk of the estimator h	$\mathbb{E}_{X\times Y}(\ell(Y, h(X)))$
$\mathcal{H}(x)$	Heaviside function	$\mathcal{H}(x) = 1$ if $x \geq 0$ and 0 otherwise
$[k]$	Integers from 1 to k	$[k] = \{1,\ 2, \ldots,\ k\}$

Contents

1 Organization and contributions of this thesis

The first two chapters of this thesis introduce and define concepts that will be developed later on. The other chapters can be read independently of each other. Original contributions and their relative published material are referenced at the beginning of each chapter.

Chapter 2 - Introduction to Functional MRI

In this chapter we introduce functional magnetic resonance imaging (fMRI) as a non-invasive functional imaging modality with good spatial resolution and whole brain coverage. We start by presenting briefly the main human brain structures and then reviewing the principal brain imaging techniques in use nowadays, with special emphasis on fMRI.

The primary form of fMRI measures the oxygen change in blood flow. This is known as the the Blood-oxygen-level dependent (BOLD) contrast. We present a *feature extraction* model known as the *general linear model* (GLM) [Friston et al., 1995] that allows to extract time-independent activation coefficient given the BOLD signal and an experimental paradigm. The difficulty of this process stems from the fact that the BOLD signal does not increase instantaneously after the stimuli onset nor does it return to baseline immediately after the stimulus offset. Instead, the BOLD signal peaks approximately 5 seconds after stimulation, and is followed by an undershoot that lasts as long as 30 seconds. The idealized, noiseless response to an infinitesimally brief stimulus is known as the *Hemodynamic Response Function* (HRF).

In order to estimate the activation coefficients, the GLM assumes a *linear time invariant* (LTI) relationship between the BOLD signal and the neural response. This relationship has been reported in a number of studies and can be summarized as *i*) *Multiplicative scaling*. If a neural response is scaled by a factor of α, then the BOLD response is also scaled by a factor of α. *ii*) *Additivity*. If the response of two separate events is known, the signal for those same events is the sum of the independent signals (Fig. 1.1). *iii*) *Time invariant*. If the stimulus is shifted by t seconds, the BOLD response will also be shifted by this same amount.

The GLM in its classical formulation assumes a known form for the hemodynamic response function (HRF). In Chapter 4 we will present an extension of the GLM model that estimates jointly the activation coefficients and the hemodynamic response function.

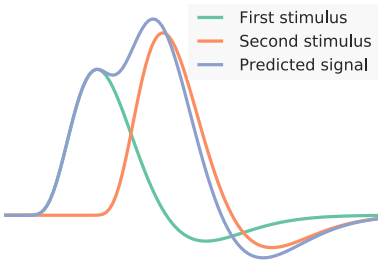


Figure 1.1: The general linear model (GLM) predicts that the expected BOLD response to two overlapping stimuli is the sum of the two independent stimuli. In green, the response to the first stimulus that is located at 1 second. In orange, the response to the second stimulus that appears at 6 seconds. In blue, the predicted BOLD response.

Chapter 3 - Statistical Inference in fMRI

In this chapter we present the statistical methods that will be used for drawing conclusions from fMRI experiments in further chapters. The chapter is divided into two sections. The first section summarizes the basics of statistical hypothesis testing while the second section describes the basics of supervised machine learning.

Statistical tests can be broadly divided into *parametric* and *nonparametric* tests. Parametric test assume a known probability distribution under the null hypothesis for the distribution parameter that is under consideration. Nonparametric tests do not assume a known form of this probability distribution although they might require some regularity conditions on the distribution such as symmetry. In this chapter we describe two parametric statistical tests: the t-test and the F-test. We will also present a nonparametric test: the Wilcoxon signed-rank test. These tests will be used at different parts of the manuscript. The t and F-test will be used to perform voxel-wise inference in section 3.1.3 and the Wilcoxon test will be used to compare the performance of different supervised learning models in Chapter 4, 5 and 6.

A notable application of parametric statistical tests to fMRI is the creation of Statistical Parametric Maps (SPMs) (Fig.). These are images with values that are, under the null hypothesis, distributed according to a known probability density function, usually Student t or the F distribution. To create such maps, one proceeds by performing a parametric test at each voxel. The resulting statistics are assembled into an image - the SPM.

In the second part of this chapter we introduce different supervised learning models that will be used in subsequent chapters. We will consider models that can be estimated as the solution to a minimization problem of the form

$$\arg\min_{f \in \mathcal{F}} \mathcal{R}_n^{\psi}(f) + \lambda \Omega(f) \quad ,$$

where $\mathcal{R}_n^{\psi}(f)$ is a data-fitting term that minimizes a surrogate of the loss function term and $\Omega(f)$ is a regularization term that biases the estimated model towards a set of desired solutions. This way, the model is a trade-off between a data fidelity term and a regularization term.

We describe different surrogate loss functions and penalties that have found applications in the context of fMRI analysis. The surrogate loss functions that we describe here are Support Vector Machines, Logistic Regression, Support Vector Regression and Least Squares. The penalties that we consider here are squared ℓ_2, ℓ_1, elastic-net ($\ell_2^2 + \ell_1$) and total-variation (TV).

Finally, we present two applications of supervised learning to reveal cognitive mechanisms in fMRI studies. The first application is commonly known as *decoding* or *mind reading* and consist in predicting the cognitive state of a subject from the activation coefficients. The neuroscientific questions that decoding is able to address are commonly shaped within the statistical hypothesis testing framework. The inference that we want to establish is whether the classifier trained on data from a given brain area of one subject is accurate enough to claim that the area encodes some information about the stimuli. In this setting, the null hypothesis is that a given brain area does not contain stimuli-related information. The ability of the

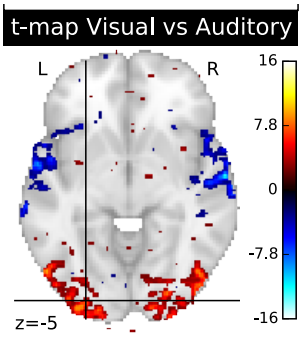


Figure 1.2: Statistical Parametric Maps (SPMs) are images with values that are, under the null hypothesis, distributed according to a known probability density function. In the figure, a t-map (i.e. the values are distributed according to a Student t distribution) for a contrast of a Visual vs an Auditory task. Thresholded at p-value $< 10^{-3}$. It can be seen how the voxels that exhibit a higher significance of this contrast belong to visual areas (red) and auditory areas (blue)

classifier to correctly predict some information about the stimuli is considered a positive evidence in support of the alternate hypothesis of presence of stimuli-related information within the brain activity. As an application of decoding, we present the dataset [Borghesani et al., 2014], in which we used decoding techniques to establish in which regions of the brain it is possible to decode different aspects of words representing real-world objects.

A different application is known as *encoding*. Here, the activation coefficients are predicted from some information about the stimuli. The success of an encoding model depends in great measure on our ability to construct an appropriate representation of the stimuli, a transformation that is often nonlinear. For example, Naselaris et al. [2009] constructed two different models for each voxel: a model based on phase-invariant Gabor wavelets, and a semantic model that was based on a scene category label for each natural scene. The authors showed that the Gabor wavelet model provided good predictions of activity in early visual areas, while the semantic model predicted activity at higher stages of visual processing.

Encoding and decoding can be seen as complementary operations: while encoding uses stimuli to predict activity, decoding uses activity to predict information about the stimuli. Furthermore, encoding offers the advantage over decoding models that they can naturally cope with unseen stimuli. For example, [Kay et al., 2008] used an encoding model to identify natural images that the subject had not seen before. In this case, the predicted activation coefficients were used to select the image that matched most closely the measured activation coefficients.

Chapter 4 - Data-driven HRF estimation for encoding and decoding models

As pointed in Chapter 2, prior to its use statistical inference procedures, the fMRI data usually goes through *feature extraction* process that converts the BOLD time course into time-independent activation coefficient. This is commonly achieved using a model known as Linear General Model (GLM). While this approach has been successfully used in a wide range of studies, it does suffer from limitations. For instance, the GLM commonly relies on a data-independent *reference* form of the hemodynamic response function (HRF) to estimate the activation coefficient (also known as *canonical* or *reference* HRF). However it is known that the shape of this response function can vary substantially across subjects, age and brain regions. This suggests that an adaptive modeling of this response function can improve the accuracy of subsequent analysis.

In this work we propose a model in which a common HRF is shared across the different stimuli that we denote *Rank-1 GLM* (R1-GLM). The novelty of our method stems from the observation that the formulation of the GLM 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 $\boldsymbol{\beta}_{\mathbf{B}} = \text{vec}(\mathbf{h}\boldsymbol{\beta}^T)$. 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)*.

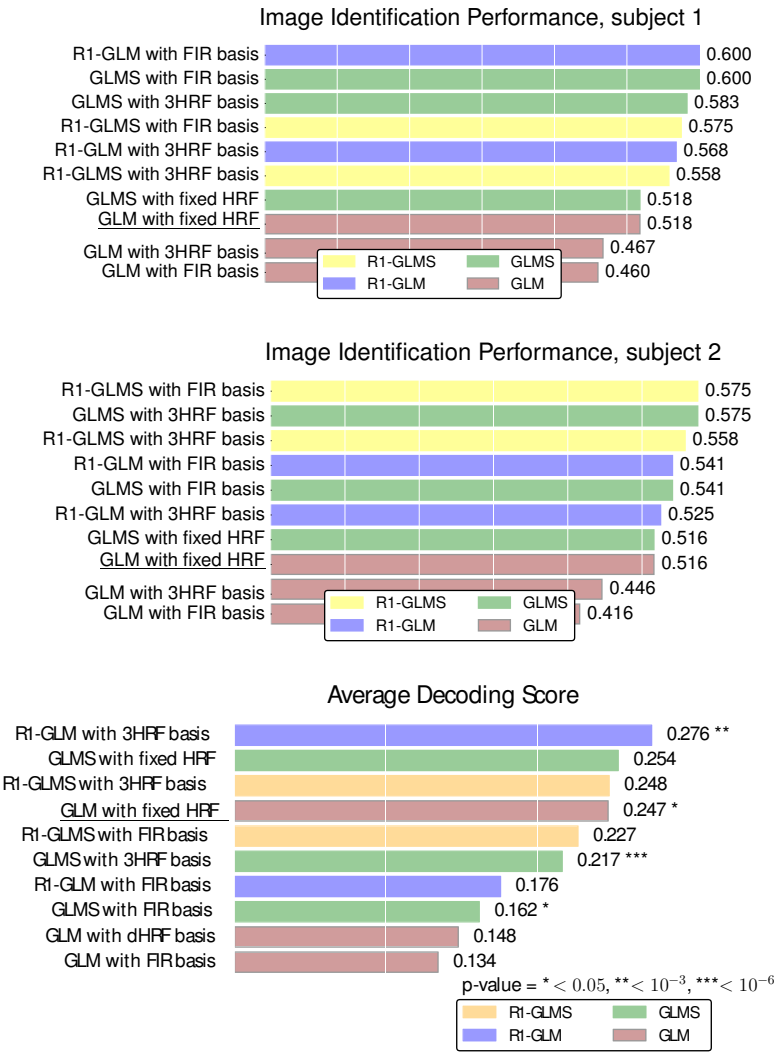


Figure 1.3: Image identification score (higher is better) on two different subjects from the first dataset and average decoding score on the second dataset. In the first dataset the metric counts the number of correctly identified images over the total number of images (chance level is $1/120 \approx 0.008$). 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.

The metric in the second dataset (decoding task) 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 this model, the coefficients have no longer a closed form expression, but can be estimated by minimizing the following loss function. Given $\mathbf{X_B}$ and $\mathbf{y}$ as before, $\mathbf{Z} \in \mathbb{R}^{n \times q}$ a matrix of nuisance parameters such as drift regressors, the optimization problem reads:

$$\begin{aligned} \hat{\mathbf{h}},\ \hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\omega}} \ &= \ \underset{\mathbf{h},\boldsymbol{\beta},\boldsymbol{\omega}}{\arg\min}\ \frac{1}{2}\|\mathbf{y} - \mathbf{X_B}\,\mathrm{vec}(\mathbf{h}\boldsymbol{\beta}^T) - \mathbf{Z}\boldsymbol{\omega}\|^2 \\ &\text{subject to } \|\mathbf{Bh}\|_\infty = 1 \text{ and } \langle \mathbf{Bh}, \mathbf{h}_{\text{ref}} \rangle > 0 \ , \end{aligned} \tag{1.1}$$

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 $\mathbf{h}_{\text{ref}}$. This ensures the identifiability of the parameters. The optimization problem (Eq. (1.1)) 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 $\boldsymbol{\omega}$.

We compare different methods for the joint estimation of HRF and activation coefficients in terms of their score for an encoding and an encoding task. The methods we considered are standard GLM (denoted GLM), a variant of the GLM that improves estimation in highly correlated settings

known as 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.

We consider two different datasets. The first dataset is presented in [Kay et al., 2008] where it is investigated using an encoding paradigm. The second dataset has been presented in [Jimura and Poldrack, 2011] and is naturally investigated using a decoding paradigm. The scores obtained in both datasets can be seen in Figure 1.3. In both cases, the proposed method (R1-GLM or its variant R1-GLMS) achieve the highest scores. The results presented in this chapter have been published in [Pedregosa et al., 2014].

Chapter 5 - Decoding with Ordinal Labels

We have presented in Chapter 3 the decoding problem in fMRI. In this setting it is often the case that the target variable consist of discretely ordered values. This occurs for example when target values consists of human generated ratings, such as values on a Likert scale, size of objects (Fig. 1.4), the symptoms of a physical disease or a rating scale for clinical pain measurement.

Figure 1.4: In [Borghesani et al., 2014], the authors investigated the possibility to predict different aspects of the words subjects were reading while undergoing an fMRI acquisition. One of the problems we investigated is to predict the size of the objects that the words refer to. In this case, the different stimuli are ordered according to their relative size, i.e. hammer is smaller than cow which is smaller than a whale, etc. In this case, the target variable is of *ordinal nature*.

In this chapter we propose the usage of two loss functions to assess the performance of a decoding model when the target consist of discretely ordered values: the absolute error and pairwise disagreement. These two loss functions emphasize different aspects of the problem: while the absolute error gives a measure of the distance between the predicted label and the true label, the pairwise disagreement gives a measure of correct ordering of the predicted labels. These loss functions lead to two different supervised learning problems. The problem in which we seek to predict a label as close as possible to the correct label is known as *ordinal regression* while the problem of predicting ordering as close as possible to the true ordering of a sequence of labels is traditionally known as *ranking*.

The first models specifically tailored for the problem of ordinal regression date back to the proportional odds and proportional hazards models of [McCullagh, 1980]. We present three different surrogate loss functions of the absolute error: least absolute error, ordinal logistic regression and cost-sensitive multiclass classification.

Ranking models were introduced chronologically later than ordinal regression but its popularity has grown in recent years thanks to its applica-

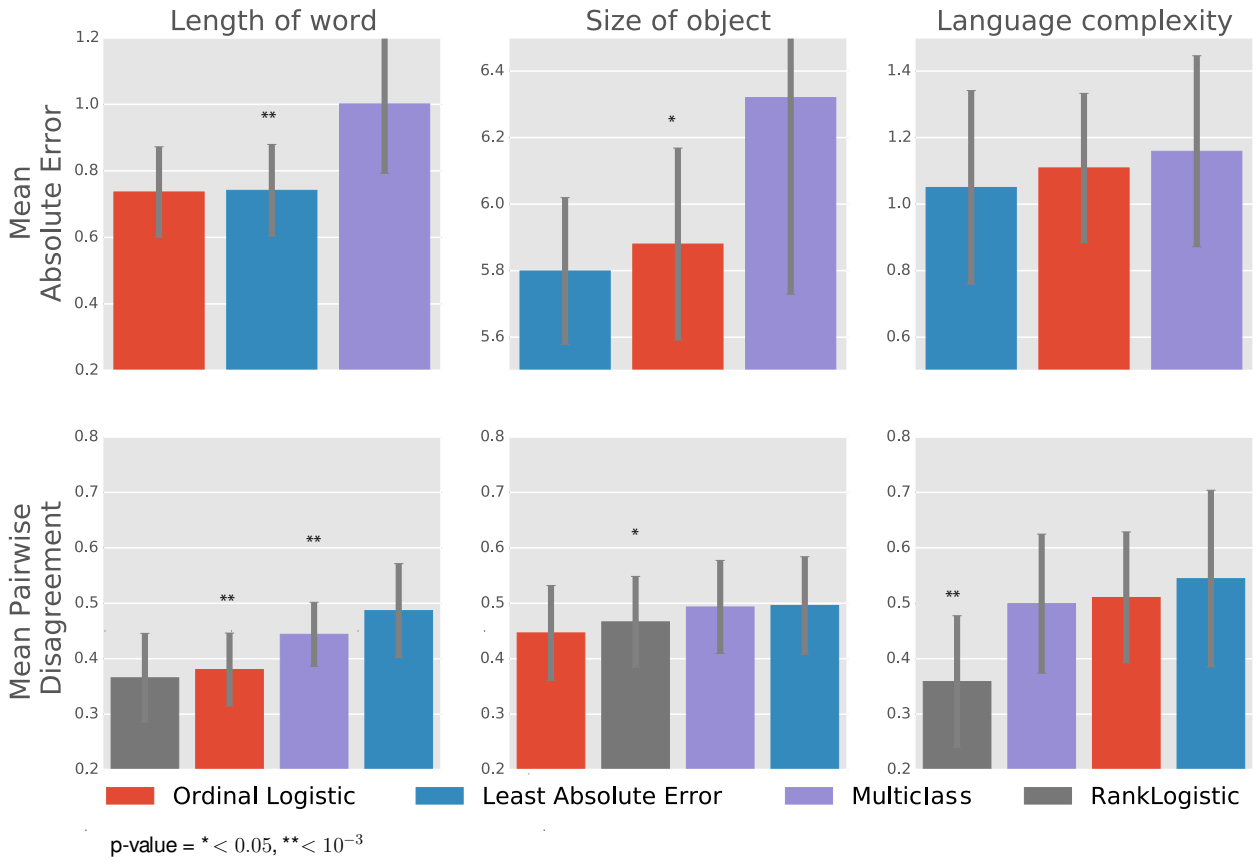


Figure 1.5: Generalization errors (lower is better) for three fMRI decoding problems. Two different metrics are used corresponding to the rows in the figure: mean absolute error and mean pairwise disagreement. The $*$ symbol represents the p-value associated with a Wilcoxon signed-rank test. This test is used to determine whether a given method outperforms significantly the next best-performing method.

bility to information retrieval (and search engines in particular) [Liu, 2009]. To the best of my knowledge, the first attempt to minimize a convex surrogate of the pairwise disagreement appears is due to Herbrich et al. [1999]. We consider a model that minimizes a surrogate of the pairwise disagreement: the RankLogistic model. This model can be viewed as a modification of the popular RankSVM model [Herbrich et al., 1999, Joachims, 2002].

The choice of either metric (absolute error or pairwise disagreement) will depend on the problem at hand. For example, in clinical applications it is often desirable to predict a label as close as possible to the true label, in which case the absolute error is the appropriate metric. If however, the purpose of the decoding study is to perform a statistical hypothesis test to claim that the area encodes some information about the stimuli, then the pairwise disagreement can be an appropriate measure [Pedregosa et al., 2012, Borghesani et al., 2014, Bekhti et al., 2014].

We examine their generalization error on both synthetic and two real world fMRI datasets and identify the best methods for each evaluation metric (Fig. 1.5). Our results show that when considering the absolute error as evaluation metric, the least absolute error and the logistic ordinal model are the best performing methods. On the other hand, when considering the mean pairwise disagreement the RankLogistic was the best performing method. For neuroimaging studies, this contribution outlines what in our opinion are the best models to choose when faced with a decoding problem in which the target variables are naturally ordered.

Chapter 6 - Fisher Consistency of Ordinal Regression Methods

Ordinal regression is the supervised learning problem of learning a rule to predict labels from an ordinal scale. Some ordinal regression models have been used in Chapter 5 to model the decoding problem when the target variable consist of ordered values.

Motivated by its applicability to decoding studies we turn to study some theoretical properties of these methods. The aim is that a theoretical approach can provide a better understanding the methods at hand. For example, Chu and Keerthi [2005] proposed two different models for the task of ordinal regression: SVOR with explicit constraints and SVOR with implicit constraints. In that work, the second approach obtained better generalization error in terms of the absolute error loss function. Similar results were obtained by Lin and Li [2006] replacing the hinge loss by an exponential loss. Yet again, Rennie and Srebro [2005] arrived to similar conclusions considering the logistic loss instead. Given these results, it seems natural to ask: is this coincidence or are there theoretical reasons for this behavior? We will use the result of this chapter to provide an affirmative answer to this last question.

Many of the ordinal regression models that have been proposed in the literature can be viewed as methods that minimize a convex surrogate of the zero-one, absolute (as outlined in Chapter 5), or squared errors. In this chapter we investigate some theoretical properties of ordinal regression methods. The property that we will investigate is known as *Fisher consistency* and relates the minimization of a given loss to the minimization of its surrogate.

We consider a rich family of loss functions for ordinal regression. We follow [Li and Lin, 2007] and determine as admissible loss functions of ordinal regression those that verify the V-shape property, a condition that includes to the best of my knowledge all popular loss functions that have been used in the context of ordinal regression: absolute error, squared error and 0-1 loss.

In order to introduce the notion of consistency we must fix some notation. In the supervised learning setting, we are given a set of n observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ drawn i.i.d. from $X \times Y$ and a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. The goal is to learn from the training examples a measurable mapping called a *classifier* $h : \mathcal{X} \to \mathcal{Y}$ so that the *risk* given below is as small as possible:

$$\mathcal{R}_\ell(h) = \mathbb{E}_{X \times Y}(\ell(Y, h(X))) \quad . \tag{1.2}$$

Attempting to directly minimize the risk is not feasible in practice. First, the probability distribution P is unknown and the risk must be minimized approximately based on the observations. Second, due to the non-convexity and discontinuity of ℓ, the risk is difficult to optimize and can lead to an NP-hard problem. It is therefore common to approximate ℓ by a function $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$, called a *surrogate loss function*, which has better computational properties. The goal becomes to find the *decision function* f that minimizes instead the ψ-risk, defined as

$$\mathcal{R}_n^\psi(f) = \mathbb{E}_{X \times Y}(\psi(Y, f(X))) \quad . \tag{1.3}$$

Fisher consistency is a desirable property for surrogate loss functions. It implies that in the population setting, i.e., if the probability distribution P were available, then optimization of the ψ-risk would yield a function (not necessarily unique) with smallest possible risk, known as *Bayes predictor* and denoted by h^*. This implies that within the population setting, the minimization of the ψ-risk and the minimization of the risk both yield solutions with same risk.

In this chapter we provide a theoretical analysis of the Fisher consistency properties of a rich family of ordinal regression surrogate loss functions, including proportional odds and support vector ordinal regression. The loss functions that we considered can be divided into three categories: regression-based, threshold-based and classification-based.

Regression-based loss function. The *regression-based approach* treats the labels as real values. It uses a standard regression algorithm to learn a real-valued function, and then predicts by rounding to the closest label. In this setting we will examine consistency of two different surrogate loss functions, the absolute error (that we will denote $\psi_{\mathcal{A}}$) and the squared error (denoted $\psi_{\mathcal{S}}$), which are convex surrogates of $\ell_{\mathcal{A}}$ and $\ell_{\mathcal{S}}$, respectively. Given $\alpha \in \mathbb{R}$, $y \in [k]$, these are defined as

$$\psi_{\mathcal{A}}(y, \alpha) = |y - \alpha|, \quad \psi_{\mathcal{S}}(y, \alpha) = (y - \alpha)^2 \quad .$$

We prove that the $\psi_{\mathcal{A}}$ surrogate is consistent with respect to the absolute error and that the $\psi_{\mathcal{S}}$ surrogate is consistent with respect to the squared error. Consistency of $\psi_{\mathcal{A}}$ was already proven by [Ramaswamy and Agarwal, 2012] for the case of 3 classes. Here we give an alternate simple proof that extends beyond $k > 3$.

Threshold-based loss function. While the regression-based loss functions may lead to optimal predictors when no constraint is placed on the regressor function space as we will see, in practice only simple function spaces are explored such as linear or polynomial functions. In these situations, the regression-based approach might lack flexibility. *Threshold-based loss functions* provide greater flexibility by seeking for both a mapping $f : \mathcal{X} \to \mathbb{R}$ and a non-decreasing vector $\boldsymbol{\theta} \in \mathbb{R}^{k-1}$, often referred to as *thresholds*, that map the class labels into ordered real values. In this context of we consider two different families of surrogate loss functions: the *cumulative link* surrogates and the *margin-based* surrogates. The first family of surrogate loss function that we will consider is the *cumulative link* surrogates. In such models the posterior probability is modeled as $P(Y \leq i | X = x) = \sigma(g_i(x))$, where σ is an appropriate link function. We will prove consistency for the case where σ is the sigmoid function, i.e., $\sigma(t) = 1/(1 + \exp(-t))$. This model is known as the *proportional odds* model or *cumulative logit* model [McCullagh, 1980]. For $x \in \mathcal{X}, y \in [k]$ and $\alpha_i = g_i(x)$, the proportional odds surrogate (denoted ψ_C) is defined as

$$\psi_C(y, \boldsymbol{\alpha}) = \begin{cases} -\log(\sigma(\alpha_1)) & \text{if } y = 1 \\ -\log(\sigma(\alpha_y) - \sigma(\alpha_{y-1})) & \text{if } 1 < y < k \\ -\log(1 - \sigma(\alpha_{k-1})) & \text{if } y = k. \end{cases} \tag{1.4}$$

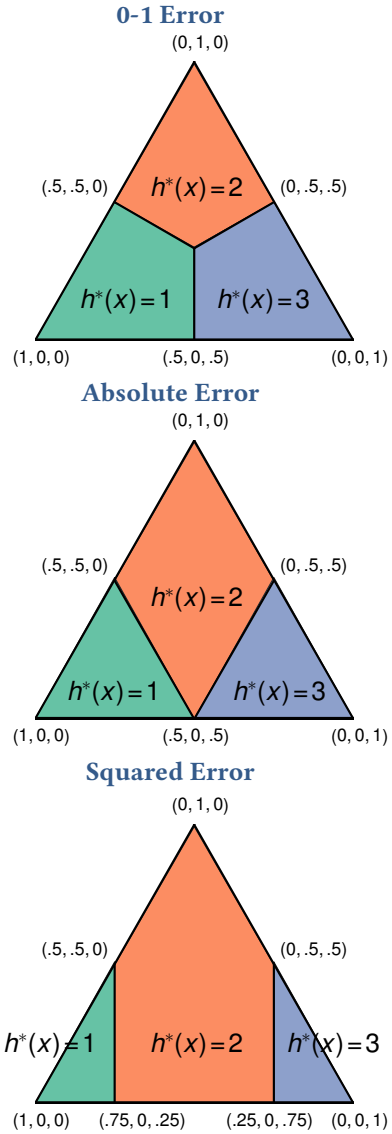


Figure 1.6: Bayes predictor can be visualized on the probability simplex. Bayes predictor is a function of the conditional probability $\eta_i(x) = P(y = i | X = x)$. The vector $(\eta_1, \ldots, \eta_k)$ belongs to the probability simplex of R^n, which is contained within an hyperplane of dimension $k - 1$. In the figure, probability simplex in $\mathbb{R}^3$ is colored according to the output of Bayes predictor.

The other family of surrogates, the margin-based surrogates (denoted $\psi^{\ell}_{\mathcal{M}}$) depends on a V-shaped loss function ℓ and is given by

$$\psi^{\ell}_{\mathcal{M}}(y, \boldsymbol{\alpha}) = \sum_{i=1}^{y-1} \Delta\ell(y, i)\phi(\alpha_i) - \sum_{i=y}^{k-1} \Delta\ell(y, i)\phi(-\alpha_i) \ .$$

where $\Delta\ell(y, i)$ is the forward difference with respect to the second parameter, defined as $\Delta\ell(y, i) = \ell(y, i + 1) - \ell(y, i)$. This formulation parametrizes several popular approaches to ordinal regression. For example, let ϕ be the hinge loss and ℓ the zero-one loss, then ψ^{T}_{ℓ} coincides with the Support Vector Ordinal Regression ("explicit constraints" variant) of [Shashua and Levin, 2003, Chu and Keerthi, 2007]. If instead the mean absolute loss is considered, this approach coincides with the "implicit constraints" formulation of the same reference. For other values of ϕ and ℓ this loss includes the approaches proposed in [Shashua and Levin, 2003, Chu and Keerthi, 2005, Rennie and Srebro, 2005, Lin and Li, 2006].

Classification-based loss function Since we aim at predicting a finite number of labels with a specific loss functions, it is also possible to use generic multiclass formulations such as the one proposed in [Lee et al., 2004] which can take into account generic losses. Given ϕ a real-valued function, this formulations considers the following surrogate

$$\psi^{\ell}_{\mathcal{L}}(y, \boldsymbol{\alpha}) = \sum_{i=1}^{k} \ell(y, i)\phi(-\alpha_i) \tag{1.5}$$

for $\boldsymbol{\alpha} \in \mathbb{R}^k$ such that $\sum_{i=1}^{k} \alpha_i = 0$. The prediction function in this case is given by $\text{pred}(\boldsymbol{\alpha}) = \arg\max_{i \in [k]} \alpha_i$. Note however that this method requires the estimation of $k - 1$ decision functions. For this reason, in practical settings threshold-based are often preferred as these only require the estimation of one decision function and $k - 1$ thresholds.Consistency results of this surrogate was proven by Zhang [2004], so we will limit ourselves to compare their results to our findings of consistency for threshold-based surrogates in Section 6.3.6.

For all the surrogates considered, we either prove consistency or provide sufficient conditions under which these approaches are consistent. Finally, we illustrate our findings by comparing the performance of two methods on 8 different datasets. Although the conditions for consistency that are required by the underlying probability distribution are not necessarily met, we observed that methods that are consistent w.r.t a given loss often outperform other methods that are not consistent with respect to that loss.

Chapter 7 - Conclusion and Perspectives

We summarize the contributions of this thesis and point out possible extensions that can be considered in the future. These are:

1. For the R1-GLM model introduced in Chapter 4 we outline a possible direction to improve its computational properties by means of tensor factorizations.

2. For the R1-GLM we outline an approach to consider a common HRF at the parcel level. This would allow the model to take advantage of the spatially dependent nature of fMRI.

3. The R1-GLM model, being non-convex, comes with no guarantees of convergence to a global optimum for the algorithms considered. We propose to study conditions under which the model is guaranteed to have a unique global optimum.

4. Some of the results presented in Chapter 6 are valid under restrictive conditions on the probability distribution that generates the data. We propose to extend these results to a more general setting by relaxing some of the conditions imposed to achieve consistency of some models.

5. We report the possibility to apply ordinal regression methods to 0-1 multiclass classification. Although ordinal regression methods have been initially developed for loss functions that minimize a distance between the labels (typically the absolute error loss), our theoretical results show that some popular models are instead consistent with the 0-1 loss. This suggest that these methods might be competitive within a multiclass classification setting. A potential advantage of these methods compared to other multiclass classification methods is the lower amount of parameters to estimate.

Bibliography

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)*, page p00, Tubingen, Germany, June 2014.

Valentina Borghesani, Fabian Pedregosa, Evelyn Eger, Marco Buiatti, and Manuela Piazza. A perceptual-to-conceptual gradient of word coding along the ventral path. In *4th International Workshop on Pattern Recognition in Neuroimaging*, pages 3–6, 2014.

Wei Chu and S Sathiya Keerthi. New Approaches to Support Vector Ordinal Regression. In *Proceedings of the 22th International Conference on Machine Learning (ICML)*, 2005.

Wei Chu and S Sathiya Keerthi. Support Vector Ordinal Regression. *Neural computation*, 815(2001):792–815, 2007.

Karl J. Friston, A. P Holmes, and J. P. Poline. Statistical Parametric Maps in Functional Imaging : A General Linear Approach. *Human Brain Mapping*, 2(4), 1995.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. *Large margin rank boundaries for ordinal regression*, pages 115–132. MIT; 1998, 1999.

Koji Jimura and Russell A Poldrack. Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, pages 1–9, 2011.

Thorsten Joachims. Optimizing Search Engines using Clickthrough Data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002.

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.

Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99 (465):67–81, 2004.

Ling Li and Hsuan-tien Lin. Ordinal Regression by Extended Binary Classification. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2007.

Hsuan-Tien Lin and Ling Li. Large-margin thresholded ensembles for ordinal regression: Theory and practice. In *Algorithmic Learning Theory*, pages 319–333. Springer, 2006.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

Peter McCullagh. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society*, 42(2):109–142, 1980.

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, 2009.

Fabian Pedregosa, Elodie Cauvet, Gaël Varoquaux, Christophe Pallier, Bertrand Thirion, and Alexandre Gramfort. Learning to rank from medical imaging data. In *Machine Learning in Medical Imaging*, pages 234–241. Springer Berlin Heidelberg, 2012.

Fabian Pedregosa, Michael Eickenberg, Philippe Ciuciu, Bertrand Thirion, and Alexandre Gramfort. Data-driven HRF estimation for encoding and decoding models. *NeuroImage*, pages 209–220, November 2014.

Harish G Ramaswamy and Shivani Agarwal. Classification Calibration Dimension for General Multiclass Losses. In *Advances in Neural Information Processing Systems*, pages 1–15, 2012.

Jason D. M. Rennie and Nathan Srebro. Loss Functions for Preference Levels : Regression with Discrete Ordered Labels. In *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, 2005.

Amnon Shashua and Anat Levin. Ranking with large margin principle : Two approaches. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2003.

Tong Zhang. Statistical Behavior and Consistency of Classification Methods based on Convex Risk Minimization. *The Annals of Statistics*, 32:56–85, 2004.

2 Introduction to Functional MRI

In this chapter we introduce functional magnetic resonance imaging (fMRI). We will start by providing some insight into human brain structure and function. Then, we will introduce the principal brain imaging techniques in use nowadays. Different imaging techniques can be used to answer different neuroscientific questions. Functional MRI, due to its good spatial resolution and whole brain coverage is specially well suited to answer questions relating the localization of brain activity for a given task.

Before the data acquired through fMRI can be used in statistical analysis it has to go through a preprocessing pipeline. In the last part of this chapter we detail the different steps of this pipeline, with special emphasis on the general linear model (GLM), a model that allows to extract time-independent activation coefficients from the fMRI time series in event-related designs. These activation coefficients will form the basis of statistical studies presented in later chapters.

Contents

2.1 General brain structures

The human brain has a volume of around 1200 cm^3 and an average weight of 1.5 kg. It is composed of neurons, glia cells and blood vessels. Glia cells are responsible for the structural and metabolic support of neurons. About 86 billion neurons [Azevedo et al., 2009] process and transmit information through electrical and chemical signals. The information is transmitted along the neuron by *action potentials* (also called *spikes*), that are short-lasting electrical events in which the electrical membrane potential of a cell rapidly rises and falls.

A neuron (Fig. 2.1) has a cell body (called the *soma*), many regions for receiving information from other neural cells (called *dendrites*), and often an *axon* (*nerve fiber*) for transmitting information to other cells. Neurons communicate with one another via chemical synapses, where the axon terminal of one cell impinges upon another neuron's dendrite, soma or, less commonly, axon. Neurons can have over 1000 dentritic branches, making connections with tens of thousands of other cells. Synapses can be excitatory or inhibitory and either increase or decrease activity in the target neuron. Some neurons also communicate via electrical synapses, which are direct, electrically conductive junctions between cells.

The human brain can be decomposed in two parts: the *white matter*, constituted by the nerve fibers, and the *gray matter* constituted by the neural cell bodies. The surface of the human brain is a highly circonvoluted 6-layered structure called *neocortex* (or more simply *cerebral cortex*). This layer is folded in a way that increases the amount of surface that can fit into the volume available. A cortical fold is called *sulcus*, and the area between two *sulci* is called a *gyrus*.

The human cortex is often divided into four "lobes", called the frontal lobe, parietal lobe, temporal lobe and occipital lobe (see Figure 2.3). The left and right side hemispheres of the cortex are broadly similar in shape, and most cortical areas are replicated on both sides. Some areas, however, show strong lateralization, such as areas that are involved in language, located in the vast majority of subject in the left hemisphere.

How the different anatomical structures of the brain correspond to the neural substrate of cognitive functions is one of the oldest debates in neuroscience, defining an entire field: cognitive neuroscience. The idea of linking a given cognitive function to a specific brain region can be traced back to the work of nineteen century phrenologists, who based their localizationist attempts on the shape of the skull. In the 20th century, a group of neuropsychologists, in absence of direct means to investigate brain activity, studied patients with cortical damages observing that some focal lesions were associated with relatively global effects on behavior. This lead them to argue against a strictly localizationist view of brain organization. Nowadays it is widely recognized that the activity of specific brain regions underlie many cognitive functions (e.g.vision, in occipital areas). At the same time, the relevance of *brain networks* encompassing different anatomical regions for the multimodal integration of features necessary for higher level cognitive functions (e.g.attention in the fronto-parietal network) [Gazzaniga, 2004] has been acknowledged.

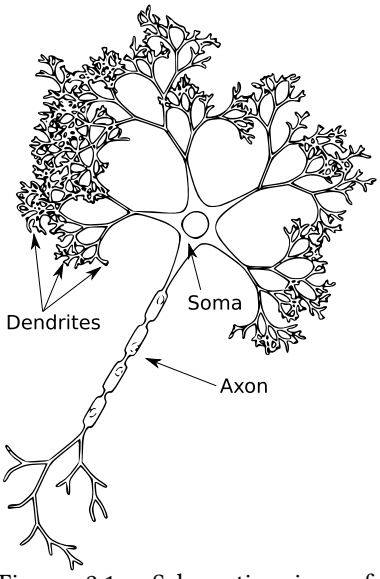


Figure 2.1: Schematic view of a neuron, in scale 10^5 : 1. A neuron has a cell body (*soma*), many regions for receiving information from other neural cells (*dentrites*) and often a nerve fiber called *axon*. Adapted from http://commons.wikimedia.org/.

Figure 2.2: Santiago Ramón y Cajal (Navarre, Spain 1852 – Madrid, Spain 1934) is widely regarded as the father of modern neuroscience. Cajal and italian anatomist Camillo Golgi impersonated the dispute between neuron and reticular theory at the turn of the 20th century. They received a joint Nobel Prize in Physiology and Medicine in 1906.

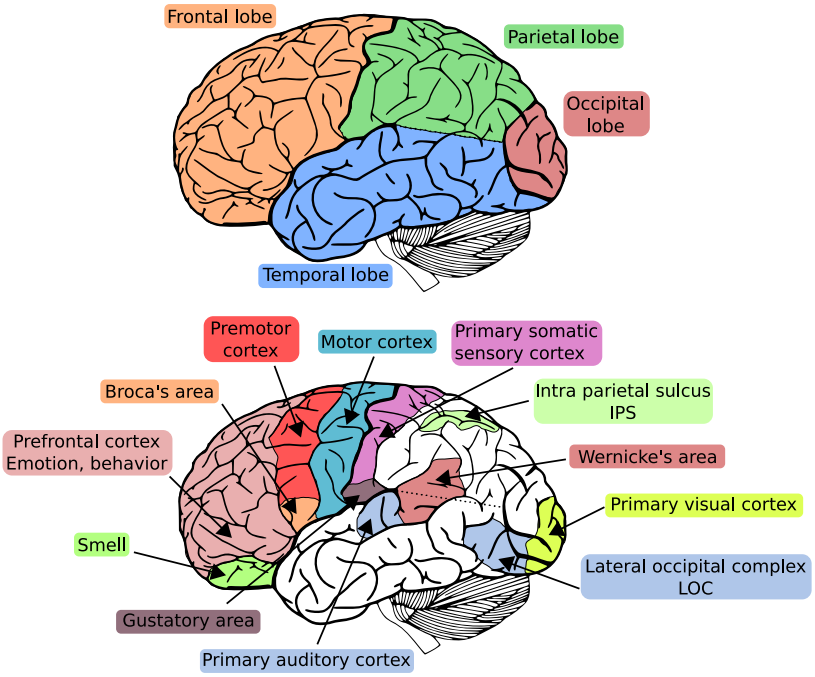


Figure 2.3: Lobes and some functional regions of the human brain (left hemisphere). Within each lobe are numerous cortical areas, each associated with a particular function such as sensory areas (*e.g. visual cortex, auditory cortex*) that receive and process information from sensory organs, motors areas (*e.g. primary motor cortex, premotor cortex*) that control the movements of the subject, and associative areas (*e.g. Broca's area, Lateral Occipital Complex – LOC – or Intra Parietal Sulcus – IPS –*) that process the high-level information related to cognition. The experiments detailed in this thesis are related to object recognition (*visual cortex* and *LOC*) and number processing (*parietal cortex* and *IPS*). Source: adapted from [Michel, 2010].

.

2.2 Functional neuroimaging modalities

Until the advent in the 1920s of non-invasive neuroimaging modalities, most of the accumulated knowledge of the brain came from the study of lesions, post-mortem analysis and invasive experimentations. With the advent of modern, non-invasive imaging techniques, several aspects of the human brain are revealed in vivo with high degree of precision.

Several brain imaging techniques are available today. These can be divided into *structural* or *anatomical* and *functional* imaging techniques. While structural imaging provides details on morphology and structure of tissues, functional imaging reveals physiological activities such as changes in metabolism, blood flow, regional chemical composition, and absorption. In this section we will discuss briefly the main functional neuroimaging modalities available today.

- **Electroencephalography - EEG** is a widely used modality for functional brain imaging. *EEG* measures electrical activity along the scalp. EEG activity reflects the synchronous activity of a population of neurons that have similar spatial orientation. If the cells do not have similar spatial orientation, their ions do not line up and thus do not create detectable waves. Pyramidal neurons of the cortex are thought to produce most of the EEG signals because they are well-aligned and fire together. Because voltage fields fall off with the square of distance, activity from deep sources is more difficult to detect than currents near the skull. Due to the ill-posed problem of volumetric data reconstruction from surface measurements, *EEG* has a poor spatial resolution compared to other modalities such as *fMRI*.

- **Stereotactic electroencephalography - sEEG** is an invasive version

of *EEG*, based on intra-cranial recording. It measures the electrical currents within some regions of the brain using deeply implanted electrodes, localized with a stereotactic technique. This approach has the good temporal resolution of *EEG* and enjoys an excellent spatial resolution. However, *sEEG* is very invasive and is only performed for medical purpose (*e.g* localization of epilepsy foci) and has a limited coverage (only the regions with electrodes). A close approach is *Electrocorticography – ECog –* that uses electrodes placed directly on the exposed surface of the brain. Even in this case its usage is restricted to medical purposes.

- **Magnetoencephalography - MEG** measures the magnetic field induced by neural electrical activity. The synchronized currents in neurons create magnetic fields of a few hundreds of femto Tesla (fT) that can be detected using specific devices. Although EEG and MEG signals originate from the same neurophysiological processes, there are important differences. Magnetic fields are less distorted than electric fields by the skull and scalp, which results in a better spatial resolution of the MEG. Whereas EEG is sensitive to both tangential and radial components of a current source in a spherical volume conductor, MEG detects only its tangential components. Because of this EEG can detect activity both in the sulci and at the top of the cortical gyri, whereas MEG is most sensitive to activity originating in sulci. EEG is, therefore, sensitive to activity in more brain areas, but activity that is visible in MEG can be localized with more accuracy. Note that EEG and MEG can be measured simultaneously.

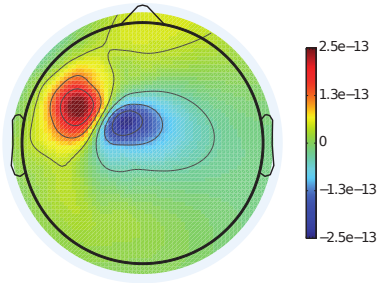


Figure 2.4: Magnetic field measured with MEG on a somatosensory experiment. It is a 2D topography 20 ms after stimulation. Source: [Gramfort, 2009]

- **Positron emission tomography - PET** is an imaging modality based on the detection of a radioactive tracers introduced in the body of the subject. The tracers (or *radionuclide* decay) emit a positron which can in turn emit, after recombination with an electron, a pair of photons that are detected simultaneously. PET therefore provides a quantitative measurement of the physiological activity. It can also be used for functional imaging, by choosing a specific tracer. In particular, the *fluorodeoxyglucose* (or *FDG*), is used for imaging the metabolic activity of a tissue. This is based on the assumption that areas of high radioactivity are associated with brain activity. *PET* has two major limitations: the tracers required for *PET* are produced by cyclotrons (a type of particle accelerator), which implies an heavy logistic. Furthermore, the use of radio-tracers is not harmless for the health of the subjects so *PET* is now used for medical purpose only.

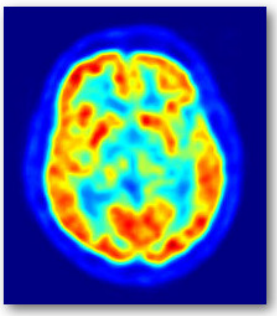


Figure 2.5: PET scan of a human brain. PET measures indirectly the flow of blood to different parts of the brain, which is, in general, believed to be correlated with neural activity. Souce: wikipedia.org

- **Single photon emission computed tomography - SPECT** is an imaging modality based on the detection of a radioactive tracer. SPECT is similar to *PET* in its use of radioactive tracer material. However, the measure in *SPECT* is the direct consequence of the tracer (the tracer emits gamma radiation), where *PET* is based on an indirect consequence of the tracer (positron then gamma radiation). The spatial resolution is slightly worse than *PET*. *SPECT* can be used for functional brain imaging, by using a specific tracer which will be assimilated by the tissue in an amount proportional to the cerebral blood flow.

- **Near-infrared spectroscopy - nIRS** is a recent modality for medical imaging. *nIRS* is based on the fact that the absorption of the light in the near-infrared domain contains information on the blood flow and blood oxygenation level. It is non-ionizing (harmless), and the instruments are not too expensive. However, the spectra obtained by *nIRS* can be difficult to interpret, and this technique, which requires a complex calibration, measures signals only close to the outer layer of the cortex.

- **Functional MRI – fMRI** is a widely used method for functional brain imaging, because it is non-invasive, has a good spatial resolution ($1mm^3$), and provides access, albeit indirectly, to the neural activity. Moreover, in standard acquisitions, *fMRI* yields a full-brain coverage, as it does not restrict the study to superficial layers or predefined regions of the cortex.

Different modalities have different trade offs in terms of spatial and temporal resolution. For example, EEG and MEG enjoy temporal resolutions of the order of few miliseconds and are thus well suited for studies of temporal dynamics of information processing but have limited spatial resolution. On the other hand, fMRI enjoys a better spatial resolution but the temporal resolution is around 1 second. Furthermore, as we will see in the next section, temporal resolution in fMRI is further limited by the slow spread of hemodynamic response, which lasts around 20 seconds after the stimuli presentation.

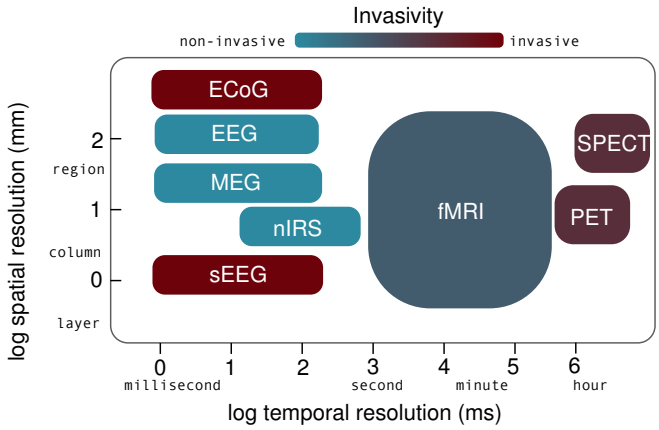


Figure 2.6: Spatial and temporal resolutions of different modalities commonly used for functional imaging. A typical fMRI acquisition (as of 2014) enjoys spatial resolution of the order of $1 - 3mm^3$ and temporal resolution of the order of 1-3 seconds.

Certain imaging techniques are more adapted than other to answer certain neuroscientific questions. Due to its good spatial resolution and whole brain coverage, fMRI is particularly well adapted to *localize* the effect of a certain experimental condition. This task is not reduced to the construction of brain maps, but also involves the understanding of the underlying brain connectivity [Johansen-Berg et al., 2005, Behrens et al., 2006] and the effects regions exert on each other in a certain experimental context [Pessiglione et al., 2007, Behrens et al., 2007]. One of the main hopes in functional imagining is that it might be used as an objective diagnosis tool for several diseases. In particular, the aim is to find some *biomarkers* for psychiatric diseases by comparing different population of patients: this is the case for autism, schizophrenia or Alzheimer's disease.

2.3 Functional MRI and BOLD signal

The primary form of fMRI measures the oxygen change in blood flow. This is known as the Blood-oxygen-level dependent (BOLD) contrast. Other increasingly popular functional MRI method is arterial spin labeling (ASL) [Detre et al., 1994, Alsop and Detre, 1998, Williams et al., 1992], which uses arterial water as tracer to measure cerebral blood flow. Compared to fMRI, ASL has a lower signal to noise ratio [Detre and Wang, 2002]. However, ASL provides reliable absolute quantification of cerebral blood flow with higher spatial and temporal resolution than other techniques [Borogovac and Asllani, 2012]. This thesis specifically considers BOLD functional MRI and through the manuscript we use the name functional MRI (fMRI) to denote functional MRI based on the BOLD signal.

The *BOLD* contrast can be explained by considering a protein present in the blood cells, called hemoglobin. Hemoglobin can bind with oxygen in order to bring it into the different cells of the organism, this link being reversible and unstable. Thus, it can be found in two different forms: *oxyhemoglobin* ($Hb - O_2$ - giving a bright red color to the blood), its oxygenated form, and *deoxyhemoglobin* (Hb - giving a blue-purple color to the blood), its deoxygenated form. When the *oxyhemoglobin* loses its oxygen atoms and becomes the *deoxyhemoglobin*, it becomes more affected by an externally applied magnetic field (due to the iron oxides). The presence of *deoxyhemoglobin* in the blood modifies the magnetic resonance signal of the protons of the water molecules surrounding the blood vessels.

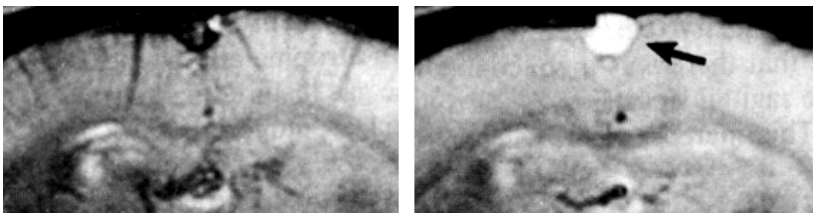


Figure 2.7: Illustration of the effect of the CO_2 on the *BOLD* contrast. Left - Coronal slice showing the *BOLD* contrast of an anesthetized rat which has breathed pure O_2. Right - Coronal slice of the same rat, showing the *BOLD* contrast after respiration of a mixture of 90% of O_2 and 10% of CO_2 (this mixture increases the oxygenation of the venous blood). The arrow shows the sagittal sinus, which is one of the major veins of the brain. This picture shows a strong increase of intensity in this vein, that illustrates that the variation of blood oxygenation is visible in *BOLD* contrast. Adapted from [Ogawa et al., 1990a].

The difference of magnetic susceptibility between the blood vessel and the surrounding tissues creates inhomogeneities in the magnetic field [Thulborn et al., 1982, Ogawa et al., 1990b] that are quantified by the magnetic resonance scanner. In the seminal paper [Ogawa et al., 1990a] studied the variations of *BOLD* contrast in the brain of an anesthetized rat during the inhalation of a gas that increases the *cerebral blood flow* (*CBF*), and thus blood oxygenation (see Figure 2.7).

The spatial resolution is given by the size of a *voxel*, a three-dimensional rectangular cuboid given by a single measure of the scanner. Voxel sizes range from 4mm to 1mm. Smaller voxels contain fewer neurons on average, incorporate less blood flow and hence have less signal to noise ratio than larger voxels. Smaller voxel size also makes up for longer acquisition time since this is proportional to the number of voxels per slice and the number of slices to scan.

The time resolution of an fMRI scanner is given by the repetition time (TR) of successive image acquisitions. A slice of the volume acquisition has an acquisition window that is about 20-30ms in duration. For example, in

the study [Borghesani et al., 2014] we used voxel sizes of $1.5 \times 1.5 \times 1.5$mm, 82 slices and a repetition time (TR) of 2.3 seconds for a full-brain coverage. These number are for routine fMRI, however it is possible to change the tradeoff between spatial and temporal resolution. With the advent of compressed sensing techniques for faster acquisition times [Lustig et al., 2007, Zong et al., 2014, Chauffert et al., 2014] and the deployment of scanners with fields of 7-Tesla and beyond [Hanke et al., 2014] these numbers are likely to change in the near future.

2.4 Estimation of activation coefficients

In this section we present a model that allows to extract time-independent activation coefficients relative to a given task given the BOLD time course and an experimental design. This model is known as the *general linear model* [Friston et al., 1995]. We start by describing the *hemodynamic response function* (Section 2.4.1) and then describe an assumption behind the general linear model, the linear-time-invariant property (Section 2.4.2) between the BOLD signal and the neural response. The general linear model is then presented in Section 2.4.3.

The concepts presented in this section will form the basis of the contribution presented in Chapter 4, where we present an extension of the general linear model that performs the joint estimation of HRF and activation coefficients.

Because it will not be referenced in later chapters we do not mention several *preprocessing* steps that can be applied to the BOLD signal in order to remove artifacts that might have occurred during acquisition or to enhance the signal to noise ratio. These include slice-timing correction, motion correction, spatial normalization and spatial smoothing.

2.4.1 Hemodynamic response function (HRF)

One of the difficulties associated with fMRI studies is that BOLD signal does not increase instantaneously after the stimulus presentation nor does it return to baseline immediately after the stimulus ends. Instead, the BOLD signal peaks approximately 5 seconds after stimulation, and is followed by an undershoot that lasts as long as 30 seconds.

The *Hemodynamic Response Function* (HRF) represents an ideal, noiseless response to an infinitesimally brief stimulus. Most software packages represent the HRF as a sum of two gamma probability density functions, where the first gamma probability density function models the shape of the initial stimulus response and the second gamma probability density functions models the undershoot. Its analytical form is

$$h(t) = \frac{t^{\alpha_1 - 1} \beta_1^{\alpha_1} e^{-\beta_1 t}}{\Gamma(\alpha_1)} - c \frac{t^{\alpha_2 - 1} \beta_2^{\alpha_2} e^{-\beta_2 t}}{\Gamma(\alpha_2)} \tag{2.1}$$

where Γ is the gamma function and $\alpha_1, \alpha_2, \beta_1, \beta_2$ control the shape and scale, respectively, and c determines the ratio of the response to undershoot.

All the packages that we have considered model the HRF as the different of two gamma probability density functions but other models are equally

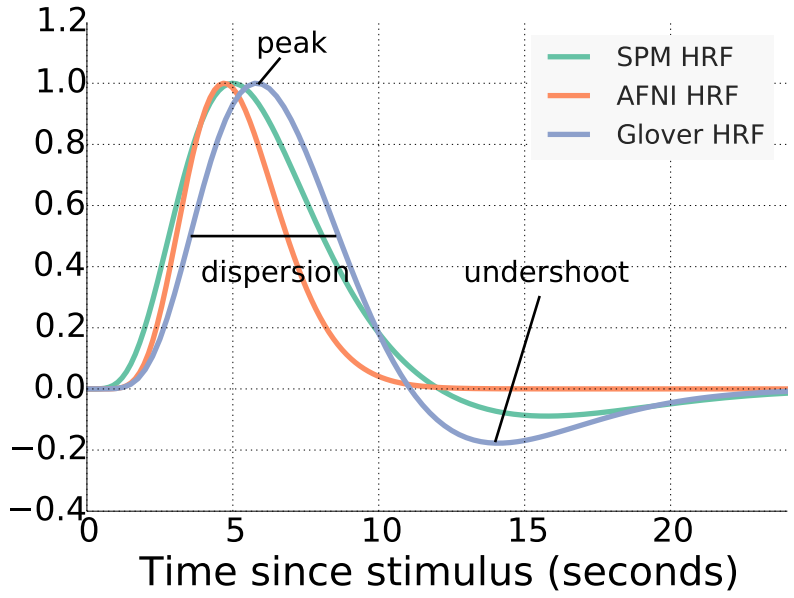


Figure 2.8: Hemodynamic Response Function (HRF) as implemented in different software packages. AFNI provides an HRF with no undershoot, i.e. modeled as a single gamma probability density function and where the peak is situated at 4.6 seconds. The software SPM provides an HRF that peaks at 5 seconds. Glover [1999] proposes two models of the HRF, one based on a motor task and another based on an auditory task. Here we show the HRF corresponding to the auditory task since this is the one that is used in the software NiPy.

possible. For instance, [Lindquist et al., 2009] proposes the use of a model based on the superposition of three inverse logit functions.

Glover [1999] proposed two different sets of parameters based on the shape of the HRF on two different experiments. The parameters that are commonly used in statistical software such as FMRISTAT[1] and NIPY[2] correspond to the HRF estimated in the auditory task. Its first gamma function peaks at 5.2 seconds, while the second gamma function (the undershoot) peaks at 12.2 seconds and has an amplitude of 35% of the first gamma function.

In the SPM[3] software, the reference HRF has its peak at 6 seconds and the delay of undershoot has its minima at 16 seconds. AFNI[4] on the other hands uses $c = 0$, that is, uses a model with a single gamma distribution. A comparison of these different HRF models can be seen in Figure 2.8. Because of its widespread use, we will use the HRF present in SPM 8 unless otherwise specified.

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

[2] http://nipy.org

[3] http://www.fil.ion.ucl.ac.uk/spm/

[4] http://afni.nimh.nih.gov/afni/

2.4.2 The linear-time-invariant assumption

In this section we present the main assumption behind the general linear model, the *linear time invariance* assumption.

A number of studies have reported that in certain regimes the relationship between the neural response and the BOLD signal exhibits *linear time invariant* (LTI) properties [Boynton et al., 1996, Cohen, 1997, Dale and Buckner, 1997]. These property can be summarized as

- **Multiplicative scaling**. If a neural response is scaled by a factor of α, then the BOLD response is also scaled by a factor of α.

- **Additivity**. If the response to two separate events is known, the signal for those events if they were to occur close in time is the sum of the independent signals.

- **Time invariant**. If the stimulus is shifted by t seconds, the BOLD response will also be shifted by this same amount.

While the LTI assumption is commonplace in practice, there is evidence for non-linearity in the amplitude of the BOLD response. For example, it is known that there is a saturation effect for stimuli that occur less than 2 seconds apart [Wager et al., 2005]. It has also been reported that very brief stimuli exhibit a larger BOLD response than would be expected based on longer stimuli [Yeşilyurt et al., 2008]. However, while these nonlinearities are important, there is a general consensus that for the range in which most cognitive fMRI studies occur, they will have relatively small impact.

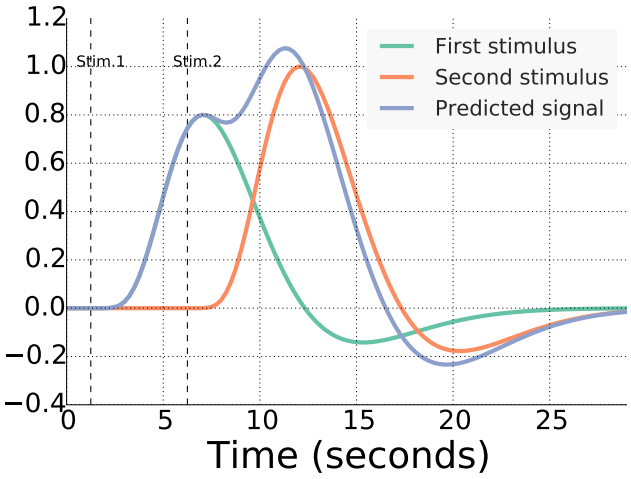


Figure 2.9: The linear time invariant (LTI) assumption implies that if the response to two separate events is known, the signal for those events if they were to occur close in time is the sum of the independent signals. In green, the response to the first stimulus that is located at 1 second. In orange, the response to the second stimulus that appears at 6 seconds. In blue, the predicted BOLD response.

Let $x(t)$ represent the predicted BOLD arising from neuronal activity as a function of time t and $h(\tau)$ be some reference HRF. The LTI assumption allows to easily construct the predicted BOLD response for a given stimulus function $u(t)$ which encodes the presence or absence of a stimulus (defined as one whenever the stimulus is present and zero otherwise). Then we can express the predicted BOLD (up to a constant factor) as the convolution of the stimulus function $u(t)$ with the HRF:

$$x(t) = \int_0^T u(t-\tau)h(\tau)\mathrm{d}\tau \tag{2.2}$$

2.4.3 The general linear model (GLM)

The General Linear Model (GLM) makes use of the knowledge of the hemodynamic response function and linear-time-invariant assumption to model the observed BOLD signal. This model states that the BOLD signal can be expressed in terms of a linear combination of the predicted fMRI responses for different stimuli (also denoted conditions) plus a noise term. Let $\{x_1(t), x_2(t), \ldots, x_k(t)\}$ be the predicted response for k different stimulus functions computed from Equation (2.2). We define the design matrix $\mathbf{X}$ as the columnwise stacking of different regressors, each one defined as the discretization of $x_i(t)$ to match the acquisition time of a given BOLD signal. The GLM in its basic form can be expressed as:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}(0, \sigma^2\mathbf{I}) \end{aligned} \tag{2.3}$$

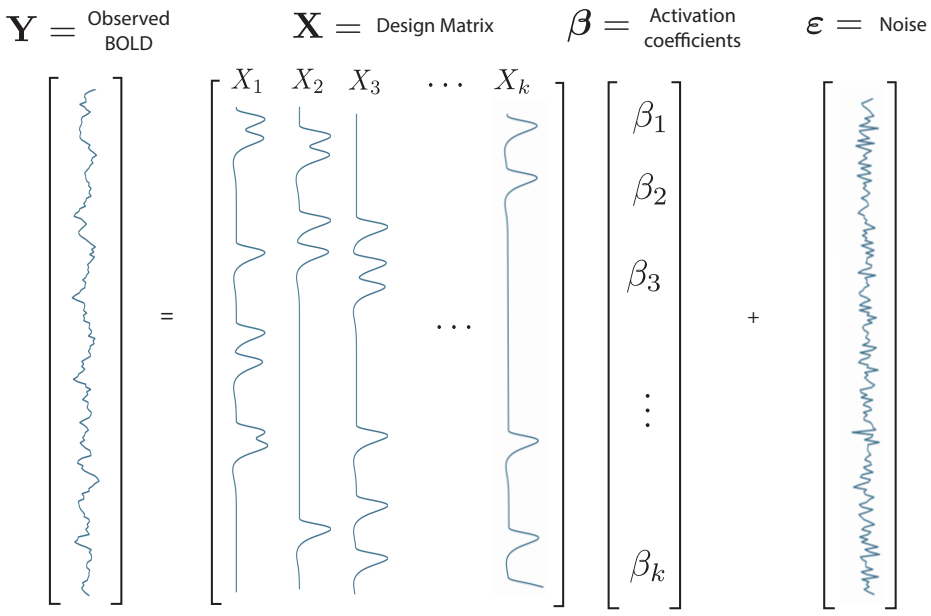


Figure 2.10: The GLM expresses the observed BOLD signal as a linear combination of regressors plus an error term. Each regressor of the design matrix is the convolution of a reference HRF and the stimulus function, a function that is 1 when the stimulus is present and zero otherwise. Each element of the (unknown) activation coefficients represent the relative amplitude of a given condition.

where $\mathbf{y} \in \mathbb{R}^n$ is the observed time course at a single voxel, $\boldsymbol{\beta} \in \mathbb{R}^k$ is the activation coefficients that represent the amplitude of the response for a given condition and $\boldsymbol{\varepsilon}$ is a noise term that we assume Gaussian for now (we will see in Section 2.4.4 how to take into account temporal autocorrelation).

Assuming Gaussian i.i.d noise, the maximum likelihood estimation of the activation coefficients is then given by $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \mathbf{X}^{\dagger}\mathbf{y}$. To estimate the activation coefficients in a full brain volume this procedure is repeated independently for each voxel. Since the design matrix is the same across voxels, a matrix decomposition of $\mathbf{X}$ such as SVD or QR can be computed once and then used to compute the least squares solution at every voxel.

In this setting we have considered the HRF to be known and fixed across the different conditions. We can easily generalize this setting to accommodate the case in which the HRF is generated by a given basis set. We will call this method *basis-constrained GLM*.

2.4.4 High-pass filtering and prewhitening

The BOLD signal contains low frequency trends that are usually removed before or during the estimation of activation coefficients. One popular approach of high-pass filtering is to add a discrete cosine transform (DCT) basis set to the design matrix. When using this basis set, the highest frequency that is desired to be removed from the data has to be chosen to avoid removing the frequency of the experimental task that is also being modeled. Another approach that is becoming increasingly popular, is to fit a local regression model to the time series and remove the estimated trend from the data. The software FSL uses LOWESS (locally weighted scatterplot smoothing) [Cleveland, 1979] while recent studies have successfully used a Savizky-Golay filter [5] [Barry and Gore, 2014, Çukur et al., 2013]. In the stud-

[5] Savitzky-Golay filter are available in Matlab under the name `sgolayfilt` and in Python's Scipy module under the name `scipy.signal.savgol_filter`

ies presented in Chapter 3 we will use this last filter. In [Çukur et al., 2013], the authors used a Savizky-Golay filter to estimate the low-frequency drifts with window length of 240 seconds and polynomial of degree 3. We have found that parameters close to these work well in practice.

The GLM specified in (2.3) assumes the noise $\boldsymbol{\varepsilon}$ follows a Gaussian random variable with covariance $\sigma^2\mathbf{I}$. However, it is known that the BOLD signal is temporally autocorrelated. Several authors [Bullmore et al., 1996, Kruggel et al., 2000] consider the BOLD noise as an autoregressive model *AR(1)*. This assumes each time point is correlated with the previous time point. The distribution of the error in this case is given by $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{V})$, where $\mathbf{V}$ is the symmetric correlation matrix and σ^2 is the variance. The correlation matrix and variance are commonly estimated from the residuals after fitting the GLM.

The most common solution to take this special structure into account is to *prewhiten* the data, that is, to remove the temporal correlation. Since the correlation matrix $\mathbf{V}$ is symmetric and positive definite, the Cholesky decomposition can be used to find a matrix $\mathbf{K}$ such that $\mathbf{V}^{-1} = \mathbf{K}^T\mathbf{K}$. To prewhiten the data, $\mathbf{K}$ is premultiplied on both sides of the GLM (Eq. (2.3)) to give $\mathbf{Ky} = \mathbf{KX}\boldsymbol{\beta} + \mathbf{K}\boldsymbol{\varepsilon}$. This makes the errors be independent, i.e., $\mathbf{K}\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$.

2.5 Conclusion

In this first chapter we have presented the principal structures of the human brain. We have then presented the principal functional imaging modalities in use today, with special emphasis on functional MRI. We have seen that functional MRI is an attractive modality for functional imaging with good spatial resolution for a whole brain coverage modality. The signal measured in fMRI studies is the BOLD signal, given in the form of a succession of scans in intervals of 1-4 seconds. The extraction of time-independent activation maps from the BOLD signal relies on the linear-time-invariant property between neural response and the BOLD signal. These can be estimated by solving a least-squares problem, a setting commonly referred to in neuroimaging as the *general linear model* (GLM). The GLM is usually formulated using a known form of the Hemodynamic Response Function.

Bibliography

David C. Alsop and John A. Detre. Multisection cerebral blood flow mr imaging with continuous arterial spin labeling. *Radiology*, 208(2):410–416, 1998.

Frederico A.C. Azevedo, Ludmila R.B. Carvalho, Lea T. Grinberg, José Marcelo Farfel, Renata E.L. Ferretti, Renata E.P. Leite, Wilson Jacob Filho, Roberto Lent, and Suzana Herculano-Houzel. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *The Journal of Comparative Neurology*, 513(5):532–541, 2009.

Robert L. Barry and John C. Gore. Enhanced phase regression with savitzky-golay filtering for high-resolution bold fmri. *Human Brain Mapping*, 35(8):3832–3840, 2014.

Timothy E.J. Behrens, M. Jenkinson, M.D. Robson, S.M. Smith, and H. Johansen-Berg. A consistent relationship between local white matter architecture and functional specialisation in medial frontal cortex. *Neuroimage*, 30(1): 220–227, 2006.

Timothy E.J. Behrens, Mark W. Woolrich, Mark E. Walton, and Matthew F.S. Rushworth. Learning the value of information in an uncertain world. *Nature neuroscience*, 10(9):1214–1221, 2007.

Valentina Borghesani, Fabian Pedregosa, Evelyn Eger, Marco Buiatti, and Manuela Piazza. A perceptual-to-conceptual gradient of word coding along the ventral path. In *Pattern Recognition in Neuroimaging*, Tubingen, Germany, June 2014. IEEE.

Ajna Borogovac and Iris Asllani. Arterial spin labeling (ASL) fMRI: Advantages, theoretical constrains and experimental challenges in neurosciences. *International Journal of Biomedical Imaging*, 2012, 2012.

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

Edward Bullmore, Michael Brammer, Steve C. R. Williams, Sophia Rabe-Hesketh, Nicolas Janot, Anthony David, John Mellers, Robert Howard, and Pak Sham. Statistical methods of estimation and inference for functional mr image analysis. *Magnetic Resonance in Medicine*, 35(2):261–277, 1996.

Nicolas Chauffert, Philippe Ciuciu, Jonas Kahn, and Pierre Weiss. Variable density sampling with continuous trajectories. application to MRI. *SIAM J. Imaging Science*, 7(4), 2014.

William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

Mark S. Cohen. Parametric analysis of fMRI data using linear systems methods. *NeuroImage*, 6(2):93–103, 1997.

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(6):763–770, 2013.

Anders M. Dale and Randy L. Buckner. Selective averaging of rapidly presented individual trials using fMRI. *Human Brain Mapping*, 5(5):329–40, 1997.

John A. Detre and Jiongjiong Wang. Technical aspects and utility of fMRI using BOLD and ASL. *Clinical Neurophysiology*, 113(5):621–634, 2002.

John A. Detre, Weiguo Zhang, David A. Roberts, Afonso C. Silva, Donald S. Williams, Donald J. Grandis, Alan P. Koretsky, and John S. Leigh. Tissue specific perfusion imaging using arterial spin labeling. *NMR in Biomedicine*, 7(1-2):75–82, 1994.

Karl J. Friston, A. P. Holmes, and J. B. Poline. Statistical Parametric Maps in Functional Imaging : A General Linear Approach. *Human Brain Mapping*, 2(4), 1995.

Michael S. Gazzaniga. *The cognitive neurosciences*. MIT press, 2004.

Gary H. Glover. Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9(4):416 – 429, 1999.

Alexandre Gramfort. *Mapping, timing and tracking cortical activations with MEG and EEG: Methods and application to human vision*. PhD thesis, 2009.

Michael Hanke, Florian J Baumgartner, Pierre Ibe, Falko R Kaule, Stefan Pollmann, Oliver Speck, Wolf Zinke, and Jörg Stadler. A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie. *Scientific Data*, 1, 2014.

Heidi Johansen-Berg, Timothy E.J Behrens, Emma Sillery, Olga Ciccarelli, Alan J Thompson, Stephen M Smith, and Paul M Matthews. Functional–anatomical validation and individual variation of diffusion tractography-based segmentation of the human thalamus. *Cerebral cortex*, 15(1):31–39, 2005.

F. Kruggel, S. Zysset, and D.Y. Von Cramon. Nonlinear regression of functional mri data: an item recognition task study. *NeuroImage*, 12(2):173–183, 2000.

Martin A Lindquist, Ji Meng Loh, Lauren Y Atlas, and Tor D Wager. Modeling the hemodynamic response function in fmri: efficiency, bias and mis-modeling. *Neuroimage*, 45(1):S187–S198, 2009.

Michael Lustig, David Donoho, and John M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. ISSN 1522-2594. doi: 10.1002/mrm.21391.

Vincent Michel. *Understanding the visual cortex by using classification techniques*. PhD thesis, Paris 11, 2010.

Seiji Ogawa, Tso-Ming Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24): 9868–9872, 1990a.

Seiji Ogawa, Tso-Ming Lee, Asha S. Nayak, and Paul Glynn. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, 14(1):68–78, 1990b.

Mathias Pessiglione, Liane Schmidt, Bogdan Draganski, Raffael Kalisch, Hakwan Lau, Ray J Dolan, and Chris D Frith. How the brain translates money into force: a neuroimaging study of subliminal motivation. *Science*, 316(5826): 904–906, 2007.

Keith R. Thulborn, John C. Waterton, Paul M. Matthews, and George K. Radda. Oxygenation dependence of the transverse relaxation time of water protons in whole blood at high field. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 714(2):265–270, 1982.

Tor D. Wager, Alberto Vazquez, Luis Hernandez, and Douglas C. Noll. Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. *NeuroImage*, 25(1):206 – 218, 2005.

Donald S Williams, John A Detre, John S Leigh, and Alan P Koretsky. Magnetic resonance imaging of perfusion using spin inversion of arterial water. *Proceedings of the National Academy of Sciences*, 89(1):212–216, 1992.

Bariş Yeşilyurt, Kâmil Uğurbil, and Kâmil Uludağ. Dynamics and nonlinearities of the BOLD response at very short stimulus durations. *Magnetic Resonance Imaging*, 26(7):853 – 862, 2008. Proceedings of the International School on Magnetic Resonance and Brain Function.

Xiaopeng Zong, Juyoung Lee, Alexander John Poplawsky, Seong-Gi Kim, and Jong Chul Ye. Compressed sensing fmri using gradient-recalled echo and EPI sequences. *NeuroImage*, 92(0):312 – 321, 2014. ISSN 1053-8119.

3 Statistical Inference in fMRI

In chapter 2, we have presented fMRI as functional imaging modality that is non-invasive and enjoys good spatial resolution and full brain coverage. In this chapter we present the statistical methods that will be used for drawing conclusions from fMRI experiments in further chapters.

The chapter is divided into two sections. The first section summarizes the basics of statistical hypothesis testing. We present two parametric test: the t-test and the F-test and one non-parametric test: the signed-rank Wilcoxon test. We discuss the voxel-wise parametric testing of the activation coefficients computed by the GLM. The result can be assembled into an image or map, a setting known as statistical parametric maps (SPMs).

The second section describes the basics of supervised machine learning. We introduce the supervised learning problem in the context of empirical risk minimization. We describe different surrogate loss functions and penalties that have found applications in the context of fMRI analysis. Finally, we present two applications of supervised learning to reveal cognitive mechanisms in fMRI studies. The first application is commonly known as *decoding* or *mind reading* and consist in predicting some information about the stimuli from the activation coefficients. The second application is known as *encoding* and can be seen as the complementary operation of decoding: here, the activation coefficients are predicted from some information about the stimuli.

Section 3.2.5 uses material from the following publication:

- V. Borghesani, F. Pedregosa, E. Eger, M. Buiatti, and M. Piazza, *"A perceptual-to-conceptual gradient of word coding along the ventral path"* Proceedings of the 4th International Workshop on Pattern Recognition in Neuroimaging, 2014.

Contents

3.1 Hypothesis testing

A *statistical hypothesis* is a statement about the parameter of a given distribution. The two complementary hypotheses in a hypothesis testing problem are called the *null hypothesis* and the *alternative hypothesis*. They will be denoted by H_0 and H_1, respectively.

Given a random samples $\{x_1, \ldots, x_n\}$ drawn from a probability space $(\mathcal{X}, \mathcal{A}, P_\theta)$, the goal of *statistical hypothesis testing* is to decide, based on the random sample, whether it is possible to reject the presumed *null hypothesis* H_0 for pre-specified level of significance. Let θ denote a distribution parameter, the general format of the null and alternative hypothesis is $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$, where Θ_0 is some subset of the parameter space and Θ_0^c is its complement. For example, if θ denotes the average activation of a voxel for a given condition, we might be interested in testing $H_0 : \theta_0 = 0$ versus $H_1 : \theta_0 \neq 0$ (or $H_1 : \theta_0 > 0$).

The *p-value* is a numerical quantity that serves to quantify the strength of the evidence against the null hypothesis and in favor of the alternative. Formally, the p-value is the probability of observing samples at least as favorable to the alternative hypothesis as the current samples, if the null hypothesis is true. Given a subset of the population, the p-value associated with a statistical test is usually computed by means of a function of these samples known as *test statistic*.

Statistical tests can be broadly divided into *parametric* and *nonparametric* tests. Parametric test assume a known probability distribution for the distribution parameter that is under consideration. Nonparametric tests do not assume a known form of this probability distribution although they might require some regularity conditions on the distribution such as symmetry. In the following subsection we will describe two parametric statistical tests: the t-test and the F-test. In this thesis, the t and F-test will be used to perform voxel-wise inference in section 3.1.3. We will also present the Wilcoxon signed-rank test, a nonparametric test that will be used to compare the performance of machine learning models in Chapter 4 and Chapter 5. The derivation of these tests is omitted but can be found in statistical textbooks such as [Casella and Berger, 2002, Rice, 2006].

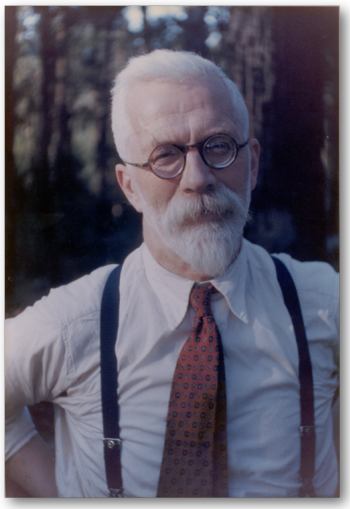


Figure 3.1: Sir R. A. Fisher (London, England 1908 - Adelaide, Australia 1962) made important contributions to the field of statistics. Among many notions in statistic, he coined the terms "test of significance", "Fisher consistency" (which we will develop in Chapter 5) and "null hypothesis" [Fisher, 1925].

Student was the pseudonym of Willian Sealy Gosset (England 1876 - England 1937). As a worker of the brewery Arthur Guiness & Son he was forbidden to publish under his real name to protect the firm from its competitors. Gosset made important contributions to the field of small sample statistics. In the seminal paper *The probable error of a mean* [Student, 1908], he introduced small sample estimation by means of the (Student) t-distribution family.

3.1.1 Parametric tests: t-test and F-test.

The t-test is any statistical hypothesis test in which the test statistic follows a Student t distribution under the null hypothesis. Most t-test statistics are of the form $t = Z/s$, where Z and s are functions of the samples, in which case the assumptions are: Z follows a standard normal distribution, s^2 follows a χ^2 distribution with p degrees of freedom, and Z, s are mutually independent. Once the t statistic is determined, a p-value can be found from the values of a Student t distribution with p degrees of freedom.

The statistical test that has as null hypothesis that the population mean is equal to a specified value μ_0 can be evaluated with a t-test known as the *one-sample t-test*. Given a sample $\{x_1, \ldots, x_n\}$ of size n, the hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0 \quad .$$

can be tested by performing a test that uses the test statistic

$$t = \frac{\bar{\mathbf{x}} - \mu_0}{s\sqrt{n}} \quad ,$$

where $\bar{\mathbf{x}}$ is the sample mean, s is the sample standard deviation of the sample and n is the sample size. Once the test statistic t has been computed, the test specifies to reject H_0 with significance level α if $t \geq t_d(1-\alpha)$, where $t_d(1-\alpha)$ is the $100(1-\alpha)$ percentile of the t distribution with $d = n-1$ degrees of freedom.

A different test based on the t distribution can be used to test the coefficients of a linear regression model. Given the equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + b + \boldsymbol{\varepsilon} \quad ,$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a given design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ and $b \in \mathbb{R}$ are terms to be estimated and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the error which follows a Gaussian $\mathcal{N}(0, \sigma^2\mathbf{I})$ distribution. It is desired to test that some linear combination of coefficients, $\mathbf{c}^T\boldsymbol{\beta}$ with $\mathbf{c} \in \mathbb{R}^p$, is significantly different from zero, i.e., $H_0 : \mathbf{c}^T\boldsymbol{\beta} = 0$ $H_1 : \mathbf{c}^T\boldsymbol{\beta} \neq 0$. In this case, the statistic

$$t = \frac{\mathbf{c}^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}} \tag{3.1}$$

follows a Student's distribution with $n-(p+1)$ degrees of freedom, where (n,p) are the dimensions of the design matrix and $\hat{\sigma}^2$ is the estimate of the variance.

The F-test can be seen as a generalization of the one-sample t-test to several groups. It can be used to asses whether the means of several pre-defined groups differ from each other. Given a total of n observations, divided into k groups of samples $\mathbf{x}_1, \ldots, \mathbf{x}_k$ with respective sizes $n_1, \ldots n_k$, a null hypothesis is of the form

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{versus} \quad H_1 : \text{at least one } \mu_i \neq \mu_j \quad ,$$

then the test statistic to test this hypothesis is calculated as the ratio between the between-group variability and the within-group variability:

$$F = \frac{\sum_i n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^2/(k-1)}{\sum_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^2/(n-k)}. \tag{3.2}$$

This statistic follows the F-distribution (also known as Snedecor's F distribution or the Fisher–Snedecor distribution) with $(k-1, n-k)$ degrees of freedom under the null hypothesis, i.e. the null hypothesis can be rejected according to this test with significance level α if the F statistic is greater than $F_{(k-1,n-k)}(1-\alpha)$, where $F_{(k-1,n-k)}$ denotes the F-distribution with $(k-1, n-k)$ degrees of freedom.

As done previously for the t-test, a variant of the F-test can be used to test the coefficients of a linear regression model. In this case, instead of testing that a given contrast is significantly different from zero, we will test that a *set of contrasts* are all simultaneously different from zero. In this case the contrast $\mathbf{C}$ is a matrix with k columns describing the possible linear combinations to be tested. For example, using a model with four parameters, to test whether all of them are equal to 0, $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, one

would specify a contrast of the form $\mathbf{C} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix of size 4×4.

For an arbitrary contrast $\mathbf{C}$, the F-statistic for this test is given by

$$F = \frac{\mathrm{Tr}(\mathbf{C}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{C}^T)}{\hat{\sigma}^2\mathrm{Tr}(\mathbf{C}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C})},$$

where the square root is taken element-wise. This expression follows an F distribution with r numerator and $n-(p+1)$ denominator degrees of freedom ($F_{r,n-(p+1)}$), where r is the rank of $\mathbf{C}$.

3.1.2 Nonparametric tests: Wilcoxon signed-rank test.

The *Wilcoxon signed-rank* test can be used to asses whether two population means differ. That is, given the samples $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$, we would like to test the following hypothesis $H_0 : \bar{\mathbf{x}} = \bar{\mathbf{y}}$, $H_1 : \bar{\mathbf{x}} \neq \bar{\mathbf{y}}$, where $\bar{\mathbf{x}}$ is the sample mean, $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Because of this, it can be seen as a nonparametric alternative to the two-sample t-test. We will use the Wilcoxon signed-rank test to replace the two-sample t-test when the normality assumptions of the last are not met. The assumptions behind Wilcoxon signed-rank are that *(a)* the two samples are paired (paired samples imply that each individual observation of one sample has a unique corresponding member in the other sample), and *(b)* the distribution of the difference between the values within each pair must be symmetrical, i.e., the median difference must be identical to the mean difference. Beginning with a set of paired values $\mathbf{x}_1$ and $\mathbf{x}_2$, each of size n, the test statistic W can be computed following the steps:

- calculate $|x_{1,i} - x_{2,i}|$ and $\mathrm{sign}(x_{1,i} - x_{2,i})$ for every $1 \geq i \geq n$. Exclude pairs which have zero difference.
- order the remaining n_r pairs from smallest absolute difference to largest absolute difference $|x_{1,i} - x_{2,i}|$.
- rank the pairs, starting with the smallest as 1. Ties receive a rank equal to the average of the ranks they span. Let r_i denote the rank.
- calculate the test statistic $W = |\sum_{i=1}^{n_r} \mathrm{sgn}(x_{1,i} - x_{2,i})r_i|$

As n_r increases, the distribution of W converges to a normal distribution. For small samples ($n_r < 10$), W is compared to a critical value from a reference table.

3.1.3 Voxel-wise hypothesis testing: Statistical Parametric Maps

Statistical Parametric Maps (SPMs)[1] are images with values that are, under the null hypothesis, distributed according to a known probability density function, usually Student t or the F distribution. To create such maps, one proceeds by performing a parametric test at each voxel. The resulting statistics are assembled into an image - the SPM.

Given the activation coefficients for a single voxel $\beta \in \mathbb{R}^k$ (cf. Section 2.4.3), with k being the number of conditions, it is possible to use a t-test to test whether a given linear combination of the conditions is significantly different from zero. As we did in section 3.1.1 we introduce the

[1] Statistical Parametric Mapping (SPM) can refer both to the set of techniques detailed in this section and to the SPM software distributed by the *Wellcome Department of Imaging NeuroScience* at University College London.

contrast $\mathbf{c} \in \mathbb{R}^k$ so that $\mathbf{c}^T\boldsymbol{\beta}$ is a linear combination of the conditions. The hypothesis can then be written as $H_0 : \mathbf{c}^T\boldsymbol{\beta} = 0$, $H_1 : \mathbf{c}^T\boldsymbol{\beta} \neq 0$. In this case, under the assumptions of the t-test for the coefficients of a linear regression model (Gaussian and i.i.d noise in the GLM), equation 3.2 gives the expression of the statistic for this test. Assigning the statistic to every voxel creates an image with the same dimensions as the input brain images, in this case a the image is called a t-map because of the t-test used to generate it.

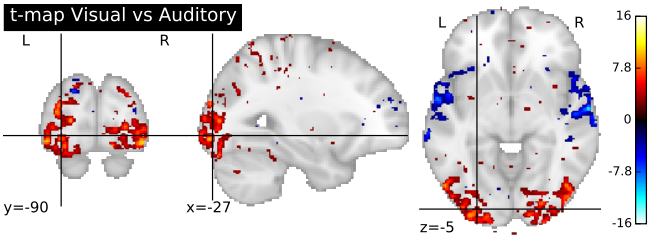


Figure 3.2: t-map for a contrast of a Visual vs an Auditory task. Thresholded at p-value $< 10^{-3}$. It can be seen how the voxels that exhibit a higher significance of this contrast belong to visual areas (red) and auditory areas (blue).

Figure 3.2 plots the t-map resulted from a functional localizer [Pinel et al., 2007] performed as part of the acquisition in Borghesani et al. [2014] dataset. For this, the conditions 'Visual' and 'Auditory' were compared. Since only two conditions were compared, the contrast is of the form $\mathbf{c} = [+1, -1, 0, \ldots, 0]$ where the entry $+1$ is for the Visual condition and -1 for the auditory condition. The image is thresholded so that only voxels with a p-value smaller than 10^{-3} are displayed. It can be seen how the voxels that exhibit a higher significance of this contrast belong to visual areas (red) and auditory areas (blue) [see Figure 2.3 for a localization of some brain regions]. This example involves the testing of a single contrasts using a t-statistic. In similar fashion, the test in which we consider d contrasts, i.e. $H_0 : \mathbf{c}_1^T\boldsymbol{\beta} = \mathbf{c}_2^T\boldsymbol{\beta} = \cdots = \mathbf{c}_d^T\boldsymbol{\beta} = 0$ and H_1 : at least one $\mathbf{c}_i^T\boldsymbol{\beta} \neq 0$ can be performed using an F-test as described in section 3.1.1.

3.1.4 Multiple comparisons issues

One major drawback of statistical parametric maps is the multiple comparisons issue. This occurs when multiple hypothesis tests are performed simultaneously and one must account for the possibility of errors occurring on each of these tests [Toothaker, 1993, Miller, 1966, Westfall, 1993]. In fMRI, due to the huge amount of voxels (on the order of 4×10^4 at $3mm^3$ resolution), some tests can lead to a large amount of false positive results, i.e., some voxels are found to be significant while in reality they were not. As a result, it is necessary to consider other types of error rates which account for the multiple comparisons issue.

A simple procedure to control the rate of false positives is through the *Bonferroni correction* method. This approach consists in dividing the threshold α by the number of tests p, which yields the new threshold $\alpha_b = \alpha/p$. The maps of voxels selected by thresholding the *p-values* for the object recognition task (subject 1), are given in Fig.3.3, for different threshold values (0.05, 0.01 and 0.05 corrected by *Bonferroni*). We notice that *Bonferroni correction* is very severe, and that it keeps very few significant voxels.

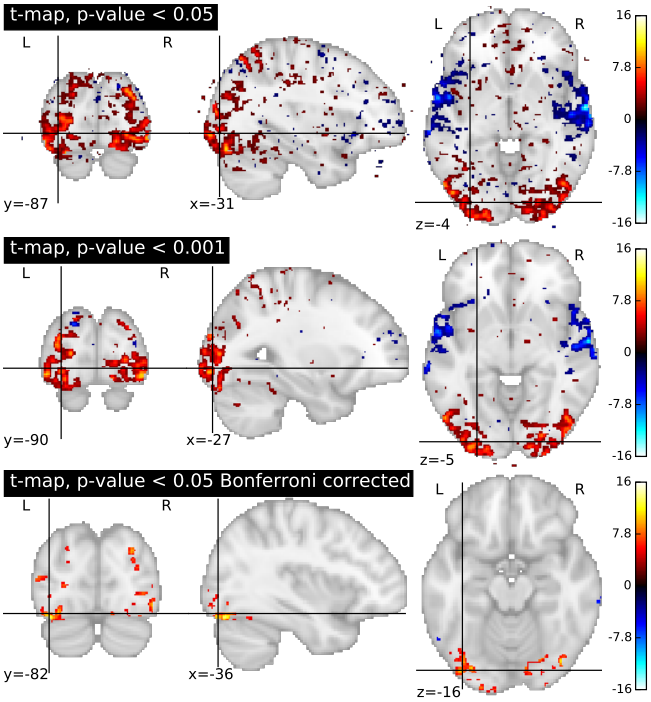


Figure 3.3: *Visual vs Auditory contrast.* Visualization of the voxels selected by thresholding the *p-values* for the Visual vs Auditory contrast at different thresholds (0.05, 0.01 and 0.05 corrected by *Bonferroni*). The *Bonferroni correction* is very severe and keeps very few voxels.

One of the main limitations behind Bonferroni corrections is that it does not take into account the spatial structure of the SPM. As such the number of independent test can smaller than the number of voxels. Other approaches besides Bonferroni correction include random field theory [Friston et al., 1994, Worsley et al., 1992] and resampling techniques [Friman and Westin, 2005, Holmes, 2003]. The review papers [Logan and Rowe, 2004, Nichols, 2012] provide an overview of the different methods that have been proposed to overcome this problem.

3.2 Machine learning in fMRI

While classical statistical modeling emphasizes statistical inference (confidence intervals, hypothesis test, optimal estimators), the field of *machine learning*, also known as statistical learning and pattern recognition, emphasizes model validation as measured by its performance on unseen samples. That is, in machine learning the validity of an estimated model will be judged based on its generalization performance.

The first applications of machine learning to neuroimaging focused on distinguishing patterns of neural activity associated with different stimuli or cognitive states, a problem commonly known as *decoding*, *reverse inference* or *brain reading* [Dehaene et al., 1998, Cox and Savoy, 2003, LaConte et al., 2005, Thirion et al., 2006, Song et al., 2011] uses a machine learning model to discriminate patterns of neural activity associated with different stimuli or cognitive states. In this thesis we will also describe the *encoding* problem [Thirion et al., 2006, Kay et al., 2008, Mitchell et al., 2008], in which the patterns of brain activity are predicted based on the stimuli features. Encoding and decoding can be seen as complementary operations: encoding uses stimuli to predict activity while decoding uses activity to predict

information about the stimuli. We will further describe these settings in Section 3.2.5 and 3.2.6, respectively.

3.2.1 Supervised Learning

Supervised learning is the task of learning a function from labeled training data. We will now give a formal definition of the task.

We consider two spaces $\mathcal{X}$ and $\mathcal{Y}$. We will refer to $\mathcal{X}$ as the *sample space* and to $\mathcal{Y}$ as the *target space*. We assume that the pair (X, Y) is a random variable taking values in $\mathcal{X} \times \mathcal{Y}$ and distributed according to an *unknown* probability distribution P. We observe a sequence of n i.i.d. pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ sampled according to P and the goal is to construct a function $h : \mathcal{X} \to \mathcal{Y}$ which *predicts* Y from X.

We need a criterion to choose this function h. For this we are given a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that measures the disagreement between a pair of elements in the target space. This way $\ell(y_i, f(x_i))$ quantifies the penalty of predicting the target $f(x_i)$ when the true label is y_i. The objective is to construct a function h such that its *risk* is as small as possible. The risk of a function h is defined as:

$$\mathcal{R}(h) = \mathbb{E}_{X \times Y}(\ell(Y, h(X)))$$

A function that achieves the minimum risk over all possible measurable functions is called the *Bayes predictor* and is denoted h^*:

$$h^* \in \arg\min_h \mathcal{R}(h)$$

Figure 3.4: Vapnik–Chervonenkis theory (also known as VC theory) was developed during 1960–1990 by Vladimir Vapnik (right) and Alexey Chervonenkis (left). The theory attempts to explain the learning process from a statistical point of view.

However, in general the risk cannot be computed because the distribution P is unknown. As an alternative we can use an approximation of the risk, called the *empirical risk*, by averaging the loss function over the pairs $\{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ drawn from P. The empirical risk is defined as:

$$\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) \quad . \tag{3.3}$$

The task is then to find the function f that minimizes the empirical risk, a setting known as *empirical risk minimization* [Vapnik and Chervonenkis, 1974]. Although the methods studied in this thesis can be seen within the framework of empirical risk minimization, several alternatives exist to this framework. A different setting for the estimation of learning models is *maximum likelihood estimation*, in which the model parameters are chosen as the maximizers of the likelihood function. When the loss function ℓ can be written as the negative log likelihood: $\ell(y, f(x)) = -\log P(f(x)|x)$, then empirical risk minimization is equivalent to maximum likelihood estimation.

Classification. If the target space $\mathcal{Y}$ is a finite set then the learning problem is known as *classification*. In the special case that this target space contains only two different values, then this problem is known as *binary classification*. The common loss ℓ in this setting is the zer-one loss, defined as $\ell(y, \hat{y}) = 0$ if $y = \hat{y}$ and 0 otherwise.

Regression. If on the other hand the target space is identified with an interval of $\mathbb{R}$ we speak about a *regression* problem. For example, the task of predicting the gender of a person would be a (binary) classification task since only two outcomes are possible. On the other hand, the task of predicting the height of a person is considered a regression task since the target space is an interval from the real line. The encoding and decoding problems in fMRI that we will consider in this chapter can be framed either using classification or regression models. The *pairwise ranking* and *ordinal regression* models that we will consider in Chapter 5 and 6 can be seen as a special case of classification problems in which the loss function depends on the distance between the respective labels. As we will see in Chapter 5, one of the contributions of this thesis is to show that certain decoding problems can be formulated using ranking and ordinal regression models rather than multiclass or regression.

For most practical applications, the sample space $\mathcal{X}$ is identified with $\mathbb{R}^p$, where p is referred to as the *dimensionality* or *number of features* of the learning problem and the target space $\mathcal{Y}$ is identified with $\mathbb{R}$.

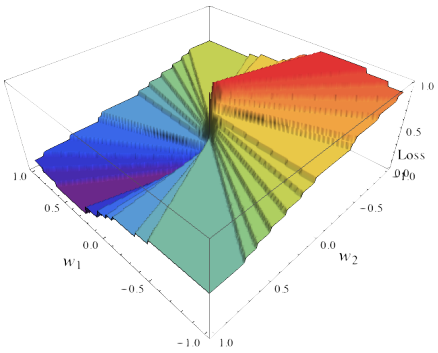


Figure 3.5: The direct minimization of the empirical risk for the 0-1 loss is a difficult computational problem due to the discontinuity of and non-convexity of the loss function. In the figure: plot of the surface $g(w_1, w_2) = \mathcal{R}(f)$, where f is the linear classification function $f(\mathbf{x}) = \text{sign}(\mathbf{x}^T\mathbf{w})$ with $\mathbf{X} \in \mathbb{R}^{10\times 2}$ a random normally distributed matrix. This surface is discontinuous with large, flat regions and is thus not amenable for optimization using gradient-based methods.

3.2.2 Surrogate loss functions.

The direct minimization of the empirical risk is often not a tractable optimization problem. For example, consider the binary classification 0-1 loss, defined as

$$\ell_{0-1}(y, \hat{y}) = \mathcal{H}(-y \cdot \hat{y}) \quad ,$$

where $\mathcal{H}$ is the Heaviside step function, defined as $\mathcal{H}(x) = 1$ if $x \geq 0$ and 0 otherwise. Minimization of the empirical risk associated with this loss is known to be NP-hard even for the class of functions as linear classifiers. See Figure 3.5 for an informal justification and [Feldman et al., 2012] and reference therein for a formal discussion of these properties.

For this reason it is common to consider instead a function $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$ which is an approximation to the true loss known as *surrogate loss* function. d is an integer that will be determined by the surrogate loss function. For binary classification, d is usually 1, while for multiclass classification d is usually equal to the number of classes. The goal in this setting is to minimize the empirical ψ-risk, defined as

$$\mathcal{R}_n^{\psi}(g) = \frac{1}{n}\sum_{i=1}^{n} \psi(y_i, g(x_i)) \quad .$$

For computational reasons, ψ is often a convex function in its second variable (the variable with respect to which we will minimize). Note that in this case the function g has as output space $\mathbb{R}$ and not $\mathcal{Y}$ as was the case before, thus the function g is not a prediction function. In binary-class classification, the prediction of two classes is given by the sign of this function. In this case, we will call g a decision function and $\text{sign}(g(X))$ will be the prediction function while in multiclass classification the prediction function is usually given by $\arg\max_{c\in\{1,\ldots,k\}} g_i(x)$ [Zhang, 2004].

Compared to the empirical risk minimization setting, we have replaced the original problem by one with better computational properties. It is natural to ask whether what have we lost in the process. In Chapter 6 we will

present results on the consistency of surrogate loss functions, that is, under which conditions minimizing the ψ-risk leads to the same solution as minimizing the risk. There, we will review existing results for binary classification and prove novel results for the case of ordinal regression.

The following is a list of surrogate loss function that are commonly used in the context of encoding and decoding models. As classification models we will consider Support Vector Machines (SVM) and Logistic Regression. For simplicity we will only describe binary classification models and assume the target space consists only of the labels $\mathcal{Y} = \{-1, 1\}$. Several techniques exist to convert a binary classification model into a multiclass classification model, such as one-vs-all and one-vs-rest [Bishop, 2006]. As regression models we will mention Support Vector Regression and Least Squares. Pairwise ranking and ordinal regression models will be described in Chapter 5 and 6.

Let $y \in \mathcal{Y}$ and $\alpha \in \mathbb{R}$, then the surrogate loss functions are defined as:

- *Support Vector Machines (SVM).* Since its first use in decoding models by Cox and Savoy [2003], Support Vector Machines [Boser et al., 1992, Cortes and Vapnik, 1995] have become the reference approach for classification decoding studies. Its success comes from its availability in popular software packages, its overall good performance under a wide array of circumstances [Bottou et al., 1994, King et al., 1995, Caruana and Niculescu-Mizil, 2006] and its ability to cope with high-dimensional data. The following surrogate is known as the *hinge loss*,

$$\psi(y, \alpha) = \max(1 - y\alpha, 0) \quad . \tag{3.4}$$

  Support Vector Machines can be extended to non-linear decision functions through the use of *kernels* [Shawe-Taylor and Cristianini, 2004].

- *Logistic Regression.* Logistic Regression is a classification model that models the posterior probability as a sigmoid, that is, $P(y|X) = (1 + e^{-yf(X)})^{-1}$. This allows to provide the probability estimates for class membership. The surrogate loss function in this case is given by the negative log-likelihood, that is, also known as the *logistic loss*

$$\psi(y, \alpha) = \log(1 + \exp(y\alpha)) \quad . \tag{3.5}$$

- *Support Vector Regression.* This is a variant of Support Vector Classification for the regression setting proposed by [Drucker et al., 1997]. The surrogate loss function in this case is known as the *ε-insensitive loss* and is given by

$$\psi(y, \alpha) = \max(|y - \alpha| - \varepsilon, 0) \quad , \tag{3.6}$$

  where $\varepsilon > 0$.

- *Least Squares* is a regression model that minimizes the square of the distance to the prediction. The loss function is given by

$$\psi(y, \alpha) = (y - \alpha)^2 \quad . \tag{3.7}$$

The most popular choice for prediction functions in encoding and decoding models are linear decision functions [Cox and Savoy, 2003, LaConte

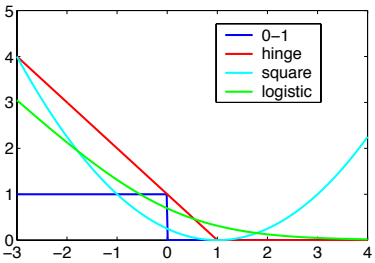


Figure 3.6: Different surrogate loss functions presented in the text (for y=1): hinge loss, logistic loss and squared loss

"In the terminology of statistics, this model is known as logistic regression, although it should be emphasized that this is a model for classification rather than regression.", Christopher M. Bishop (2006). *Pattern Recognition and Machine Learning*.p. 205.

et al., 2005, Song et al., 2011, Thirion et al., 2006, Naselaris et al., 2011], that is, functions of the form $f(\mathbf{x}) = \text{sgn}(\mathbf{x}^T\mathbf{w} + c)$ for a binary classification problem and $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + c$ for a regression problem, where $\mathbf{w} \in \mathbb{R}^p$ and $c \in \mathbb{R}$ are unknown parameters to be estimated.

3.2.3 Regularization

Regularization has long played a fundamental role in statistics and related mathematical fields. First introduced by Tikhonov and Arsenin [1977] in the context of solving ill-posed integral equations, it has since become a standard part of the statistical toolkit.

The purpose of regularization is to use prior knowledge of the problem to bias the estimated model. This can be desirable to solve an ill-posed problem or to avoid overfitting. In this setting, the model is estimated as a solution to an optimization problem of the form

$$\arg\min_{f \in \mathcal{F}} \mathcal{R}_n^{\psi}(f) + \lambda\Omega(f) \quad ,$$

where $\Omega(f)$ is the *regularization*, which biases solutions toward a desired kind of solutions and $\mathcal{F}$ is a family of functions (e.g. the family of linear or polynomial functions). In this setting the parameter λ controls the trade-off between data-fidelity and the regularization term.

We will present the following penalties due to their widespread use in fMRI analysis. These assume a linear decision function f, i.e., $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + c$ or $f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + c)$ and the penalty will be expressed in terms of the parameters $\mathbf{w}$.

- *squared* ℓ_2 $\left(\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2\right)$. Equivalent to Gaussian normal prior with zero mean [Bishop, 2006, Chapter 3]. When loss is linear least squares, it is referred to as Ridge regression and the estimated model $(\hat{\mathbf{w}}, \hat{c})$ has the closed form solution for $\lambda > 0$:

  $$[\hat{\mathbf{w}},\ \hat{c}] = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda n\tilde{\mathbf{I}})^{-1}\tilde{\mathbf{X}}^T\mathbf{y} \quad ,$$

  where $\tilde{\mathbf{X}}$ is the matrix formed by stacking a column of ones to the original design matrix $\mathbf{X}$ and $\tilde{\mathbf{I}}$ is the diagonal matrix with all ones except a zero in the last diagonal element, i.e. $\tilde{\mathbf{I}} = \mathbf{I} - \mathbf{e}_n\mathbf{e}_n^T$. This penalty is sometimes also used for computational reasons since it makes the optimization problem better conditioned.

- ℓ_1 *regularization* $(\Omega(\mathbf{w}) = \|\mathbf{w}\|_1)$. Promotes *sparse* solutions, i.e. solutions with a large fraction of zero coefficients. When combined with a least squares loss function, the model is known as *lasso* [Tibshirani, 1996] and *basis pursuit denoising* [Chen et al., 2001].

- *elastic-net regularization* $\left(\Omega(\mathbf{w}) = \alpha\|\mathbf{w}\|_1 + (1-\alpha)\|\mathbf{w}\|_2^2\right)$. Linearly combines ℓ_1 and squared ℓ_2 regularization. In the case of severely correlated variables, the ℓ_1 penalty tends to select one variable from the group of highly correlated variables and ignore the rest. To mitigate this problem, elastic-net penalty adds a quadratic ℓ_2 norm to the penalty [Zou and Hastie, 2005].

$$\mathcal{R}_n^{\psi} + \lambda\ \Omega(f)$$

Figure 3.7: The machine learning models that we will consider are estimated as the minimization of a trade-off between data fidelity and a regularization term. Regularization is used to bias the estimated model towards a set of desired solutions.

- *total variation (TV)* [Rudin et al., 1992, Chan et al., 1999, Michel et al., 2011]. Total-variation, defined as the ℓ_1 norm of the gradient promotes piecewise constant solutions. It can be combined with ℓ_1 [Baldassarre et al., 2012, Gramfort et al., 2013, Dohmatob et al., 2014] and with elastic-net [Dubois et al., 2014] penalties. Figure 3.8 compares the estimated coefficients by the use of elastic-net and TV+ℓ_1 regularization.

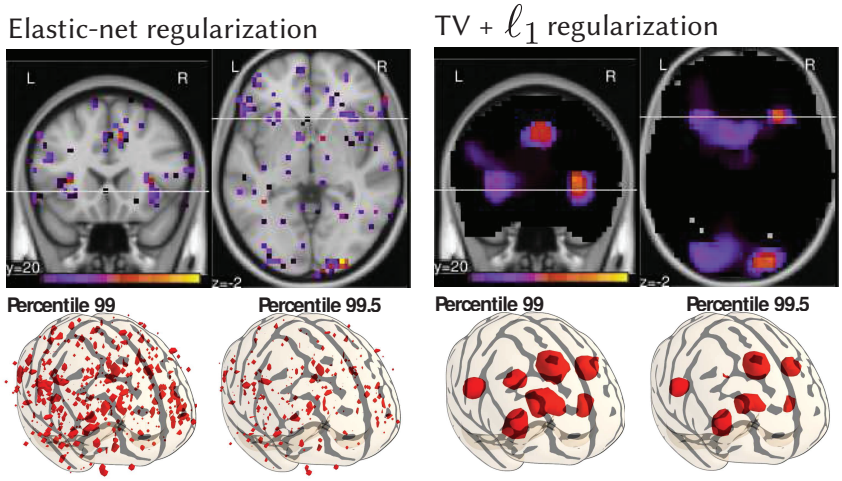


Figure 3.8: Regularization is an effective technique to inject prior knowledge to bias the estimated model. In this figure we show the estimated coefficients of a linear model with different regularizations. In this model the estimated coefficients correspond to voxels in a brain volume and are displayed over an anatomical image. In the left, elastic-net regularization yields sparse although very scattered coefficients. Moreover, this regularizer does not take into account the spatial structure of the image. The TV+ℓ_1 regularization in contrasts yields blobs of nonzero coefficients surrounded zero elements. This latter model has been proved to yield predictive regions which are meaningful from a cognitive point of view. Source: adapted from [Gramfort et al., 2013].

3.2.4 Model evaluation and cross-validation.

Since it is possible to construct a classifier that predicts perfectly on the train set but with very poor generalization performance (e.g. the classifier that returns the right label for a sample it has already seen and random otherwise), computing the empirical error on the training set yields a very poor estimator of the true risk of a model.

Cross-validation is a technique to iteratively partition the input dataset in order to obtain a more reliable estimator of the risk [Mosteller and Tukey, 1968, Stone, 1977, Geisser, 1975, Arlot and Celisse, 2010]. In this setting a subset of the data is used for training (the *training set*) and the rest of the data is used to compute the accuracy of the trained model (the *test set* or *validation set*). Repeating the process several times and averaging the accuracy of the predictions across the validation sets yields an estimator of the risk. One form of cross-validation leaves out a single observation at a time; this is known as leave-one-out. Another form, known as K-fold cross-validation, splits the data into K subsets; each is held out in turn as the validation set.

The cross-validation score is the average of the empirical risk across all the folds, which is itself an estimator of the risk. This can then be used to perform hypothesis relative to the risk of two predictors f and g. For example, consider the test in which we compare the risk of two estimators, that is, $H_0 : \mathcal{R}(f) = \mathcal{R}(g)$ and $H_1 : \mathcal{R}(f) \neq \mathcal{R}(g)$. This statistical test can be performed using the Wilcoxon signed-rank test presented in Section 3.1.2.

This test takes as input two sequences in which the samples are the empirical risk at the different cross-validation folds.

Cross-validation is an attractive estimator of the risk since it makes no assumptions on the model or the loss function. Alternatives exist for specific loss functions such as Stein's unbiased risk estimate [Stein, 1981, Donoho and Johnstone, 1995] which is an unbiased estimator of the mean-squared error.

Cross-validation can be used to select the regularization parameter in a setting known as *nested cross-validation*. In this setting, the train set is again split into the different cross-validation folds and the risk associated with the different parameters is computed using this inner cross-validation loop. The procedure can be repeated for the different training sets in the uppermost cross-validation loop. Through this thesis we will use this procedure to estimate the regularization parameter of the different models that we will consider, although it is not the only approach for this purpose. Other methods include Bayesian inference [Bishop, 2006, Chapter 8 & 10] and marginal likelihood maximization [Bock and Aitkin, 1981].

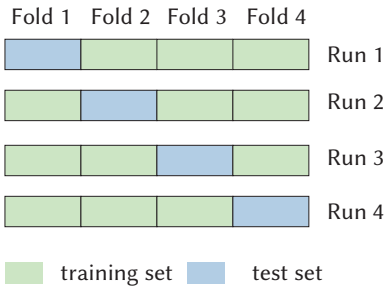


Figure 3.9: The technique of K-Fold cross-validation, illustrated here for the case $K = 4$, involves taking the available data and partitioning it into K groups. Then $K - 1$ groups are used (in green) to train a set of models that are then evaluated on the remaining group (in blue).

3.2.5 fMRI-based brain activity decoding

The paradigm of predicting the stimuli provided to the subject from the concurrent brain activity is known as *brain decoding* and accurate predictions support the hypothesis that the brain activity encodes those stimuli. Early studies [Dehaene et al., 1998] were able to predict right hand versus left hand movement based on fMRI images. In [Haxby et al., 2001, Cox and Savoy, 2003], the authors showed that different high-level visual stimulus categories (faces, animals and objects) were associated with distinct patterns of brain activity in visual areas. Subsequent work has shown that decoding can also distinguish many other brain states, for example low-level visual features in the early visual cortex [Haynes and Rees, 2005, Kamitani and Tong, 2005] and auditory stimuli in the auditory cortex [Formisano et al., 2008, Staeren et al., 2009], as well as more abstract brain states such as intentions [Haynes et al., 2007, Soon et al., 2008] and the contents of working memory [Harrison and Tong, 2009].

The neuroscientific questions that brain decoding is able to address are commonly shaped within the statistical hypothesis testing framework. The inference that we want to establish is whether the classifier designed on data from a given brain area of one subject is accurate enough to claim that the area encodes some information about the stimuli. In this setting, the null hypothesis is that a given brain area does not contain stimuli-related information. The ability of the classifier to correctly predict some information about the stimulus is considered a positive evidence in support of the alternate hypothesis of presence of stimuli-related information within the brain activity.

In [Borghesani et al., 2014], we have used decoding models to establish in which regions of the brain it is possible to decode different aspects of words representing real-world objects. One of the tasks was to decode the size of items from the words representing those objects. In this case, the different stimuli are ordered according to their relative size, so the target variable is

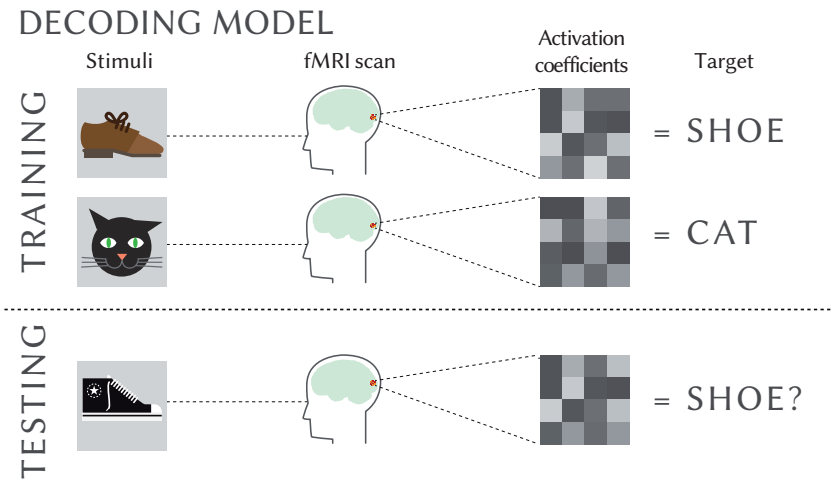


Figure 3.10: Decoding models use patterns of activity to discriminate between cognitive states. Different activation coefficients reflect different mental states; for example, those associated with different images viewed by the subject. In a training phase, the classifier will learn to discriminate between activity patterns measured under different cognitive states. In the testing phase the generalization performance of the trained model is quantified by evaluating the classifier on the testing set and comparing the output of the classifier with the true labels associated with the stimuli. Adapted from [Smith, 2013].

of ordinal nature. We predict the target variable from the brain activation on 6 anatomically defined regions of interest (ROI, which correspond to different Brodman areas). The metric is Kendall tau, which is a measure of the association between two measured quantities. This metric lies between -1 and 1. This metric and the used model will be presented in full detail in Chapter 5).

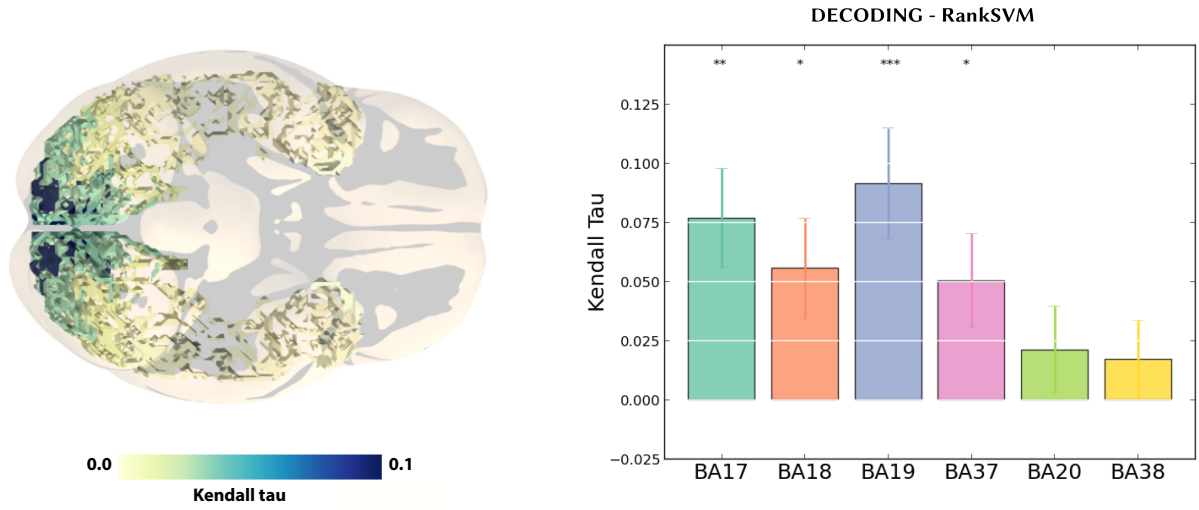


Figure 3.11: Cross-validation scores for the prediction of the length of words from [Borghesani et al., 2014]. The metric is Kendall tau (higher is better). In the left, the same scores are depicted for the different regions (Brodman areas).

As can be seen in Figure 3.11, this decoding model results in a higher decoding score in primary and secondary visual areas. In this case, a Wilcoxon signed-rank test was used to asses the statistical significance (p-value $<$ 0.05) of the scores. This is denoted by the $*$ symbol that reflects the significance of a Wilcoxon test that the mean is significantly different than zero, $*$ = p-val $<$ 0.05, $**$ = p-val $< 10^{-3}$, $***$ = p-val $< 10^{-6}$. This is achieved in Brodman areas BA17, BA18 and BA19. This experiments allows us to establish that the aforementioned areas encode some information related to the size of the stimuli.

3.2.6 fMRI-based brain activity encoding

fMRI-based *encoding* models (also known as *voxel-wise modelling*) [Thirion et al., 2006, Kay et al., 2008, Mitchell et al., 2008], seek to predict the patterns of brain activity from the the stimuli features. A machine learning model is learned from the stimuli features to the activation coefficients of a single voxel. The sample space in this case is the space of features derived from the stimuli, e.g. spatial Dirac functions in [Thirion et al., 2006] or Gabor filters in the case of natural images [Kay et al., 2008, Naselaris et al., 2014]. The predicted activation coefficient can then be compared to the true activation coefficient measured on left out data by using some distance metric such as Pearson or Spearman correlation coefficient.

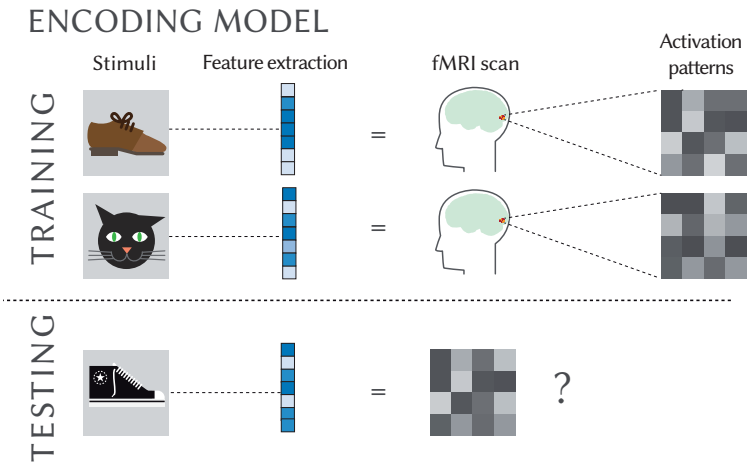


Figure 3.12: In an encoding model the patterns of brain activity are predicted by a machine learning model based on the stimuli features. The sample space in this case is the space of features derived from the stimuli, e.g. a set of Gabor filters in [Kay et al., 2008]. The predicted activation coefficient can then be compared to the true activation coefficient measured on left out data by using some distance metric such as Pearson's correlation coefficient. Adapted from [Smith, 2013].

To the best of my knowledge, all of the encoding models that have been published in the literature make assume a use of linear relationship features to the activation coefficients. That is, they assume that there is a mapping from the stimulus space to the feature space that, and a linear mapping between the feature space and the activity space. However, the success of an encoding model depends in great measure on deriving the right features from the stimuli, a transformation that might be nonlinear. For example, Naselaris et al. [2009] constructed two different models for each voxel: a model based on phase-invariant Gabor wavelets, and a semantic model that was based on a scene category label for each natural scene. The authors showed that the Gabor wavelet model provided good predictions of activity in early visual areas, while the semantic model predicted activity at higher stages of visual processing.

Encoding and decoding can be seen as complementary operations: while encoding uses stimuli to predict activity, decoding uses activity to predict information about the stimuli. Furthermore, encoding offers the advantage over decoding models that they can be used to predict information about an unseen stimuli. In this setting encoding models have been used to reconstruct stimuli from brain activity patterns in [Miyawaki et al., 2008, Naselaris et al., 2009, Nishimoto et al., 2011]. A similar setting was used in [Kay et al., 2008] to identify natural images. Here, the predicted activation coeffi-

cients were used to select the image that matched most closely the measured activation coefficients.

3.3 Conclusion

In this chapter we have presented the statistical methods that will be used for drawing conclusions from fMRI experiments in further chapters. The chapter is divided into two sections. In the first section we have introduced the framework of statistical hypothesis testing and presented several parametric and non-parametric tests. We have presented an application of voxel-wise hypothesis testing known as Statistical Parametric Maps (SPMs).

In the second part of this chapter we have presented the setting of supervised learning. We described different surrogate loss functions and penalties that have found applications in the context of fMRI analysis. The surrogate loss functions that we described are as Support Vector Machines, Logistic Regression, Support Vector Regression and Least Squares. The penalties that we have present are: squared ℓ_2, ℓ_1, elastic-net ($\ell_2^2 + \ell_1$) and total-variation (TV). Finally, we present two neuroscientific problems that can be model as a supervised learning problem: *encoding* and *decoding*.

Bibliography

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4: 40–79, 2010.

Luca Baldassarre, Janaina Mourao-Miranda, and Massimiliano Pontil. Structured sparsity models for brain decoding from fmri data. In *Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on*, pages 5–8. IEEE, 2012.

Christopher M. Bishop. *Pattern recognition and machine learning*, volume 1. Springer, 2006.

R Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981.

Valentina Borghesani, Fabian Pedregosa, Evelyn Eger, Marco Buiatti, and Manuela Piazza. A perceptual-to-conceptual gradient of word coding along the ventral path. In *Pattern Recognition in Neuroimaging*, Tubingen, Germany, June 2014. IEEE.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM.

Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Lawrence D Jackel, Yann LeCun, Urs A Muller, Edward Sackinger, Patrice Simard, et al. Comparison of classifier methods: a case study in handwritten digit recognition. In *International Conference on Pattern Recognition*, pages 77–77. IEEE Computer Society Press, 1994.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

George Casella and Roger L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Tony F. Chan, Gene H. Golub, and Pep Mulet. A nonlinear primal-dual method for total variation-based image restoration. *SIAM Journal on Scientific Computing*, 20(6):1964–1977, 1999.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

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.

Stanislas Dehaene, Gurvan Le Clec'H, Laurent Cohen, Jean-Baptiste Poline, Pierre-Francois van de Moortele, and Denis Le Bihan. Inferring behavior from functional brain images. *Nature Neuroscience*, 1(7):549–549, 11 1998.

Elvis Dohmatob, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Benchmarking solvers for tv-l1 least-squares and logistic regression in brain imaging. In *Pattern Recoginition in Neuroimaging (PRNI)*, Tübingen, Germany, June 2014. IEEE.

David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, 90(432):1200–1224, 1995.

Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

Mathieu Dubois, Fouad Hadj-Selem, Tommy Lofstedt, Matthieu Perrot, Clara Fischer, Vincent Frouin, and Edouard Duchesnay. Predictive support recovery with tv-elastic net penalty and logistic regression: an application to structural mri. In *Pattern Recognition in Neuroimaging, 2014 International Workshop on*, pages 1–4. IEEE, 2014.

Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

Elia Formisano, Federico De Martino, Milene Bonte, and Rainer Goebel. "who" is saying" what"? brain-based decoding of human voice and speech. *Science*, 322(5903):970–973, 2008.

Ola Friman and Carl-Fredrik Westin. Resampling fmri time series. *NeuroImage*, 25(3):859 – 867, 2005.

Karl J. Friston, Keith J Worsley, RSJ Frackowiak, John C Mazziotta, and Alan C Evans. Assessing the significance of focal activations using their spatial extent. *Human brain mapping*, 1(3):210–220, 1994.

Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975.

Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *Pattern Recognition in Neuroimaging (PRNI)*, Philadelphia, United States, June 2013. IEEE.

Stephenie A Harrison and Frank Tong. Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458(7238):632–635, 2009.

James V. Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

John-Dylan Haynes and Geraint Rees. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, 8(5):686–691, 2005.

John-Dylan Haynes, Katsuyuki Sakai, Geraint Rees, Sam Gilbert, Chris Frith, and Richard E Passingham. Reading hidden intentions in the human brain. *Current Biology*, 17(4):323–328, 2007.

Susan Holmes. Bootstrapping phylogenetic trees: theory and methods. *Statistical Science*, pages 241–255, 2003.

Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–685, 2005.

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.

Ross D. King, Cao Feng, and Alistair Sutherland. Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence an International Journal*, 9(3):289–333, 1995.

Stephen LaConte, Stephen Strother, Vladimir Cherkassky, Jon Anderson, and Xiaoping Hu. Support vector machines for temporal classification of block design fmri data. *NeuroImage*, 26(2):317–329, 2005.

Brent R. Logan and Daniel B. Rowe. An evaluation of thresholding techniques in fmri analysis. *NeuroImage*, 22(1): 95 – 108, 2004.

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(7):1328–1340, 2011.

Rupert G. Miller. *Simultaneous statistical inference*, volume 196. Springer, 1966.

Tom M. Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880): 1191–1195, 2008.

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):915–929, 2008.

Frederick Mosteller and John W. Tukey. *Data analysis, including statistics*. 1968.

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, 2009.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–10, May 2011.

Thomas Naselaris, Cheryl A. Olman, Dustin E. Stansbury, Kamil Ugurbil, and Jack L. Gallant. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage*, (0):–, 2014.

Thomas E. Nichols. Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, 62 (2):811 – 815, 2012.

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(19):1641–1646, 2011.

Philippe Pinel, Bertrand Thirion, Sébastien Meriaux, Antoinette Jobert, Julien Serres, Denis Le Bihan, Jean-Baptiste Poline, and Stanislas Dehaene. Fast reproducible identification and large-scale databasing of individual functional cognitive networks. *BMC neuroscience*, 8(1):91, 2007.

John Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. 60(1): 259–268, 1992.

John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Kerri Smith. Brain decoding: Reading minds. *Nature News*, 2013.

Sutao Song, Zhichao Zhan, Zhiying Long, Jiacai Zhang, and Li Yao. Comparative study of svm methods combined with voxel selection for object category classification on fmri data. *PLoS ONE*, 6(2), 02 2011.

Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. Unconscious determinants of free decisions in the human brain. *Nature neuroscience*, 11(5):543–545, 2008.

Noël Staeren, Hanna Renvall, Federico De Martino, Rainer Goebel, and Elia Formisano. Sound categories are represented as distributed patterns in the human auditory cortex. *Current Biology*, 19(6):498–502, 2009.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

Mervyn Stone. Asymptotics for and against cross-validation. *Biometrika*, pages 29–35, 1977.

Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Le Bihan, and Stanislas Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4):1104–16, December 2006.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Andrej Nikolaevich Tikhonov and Vasiliy Yakovlevich Arsenin. Solutions of ill-posed problems (translated from the russian). 1977.

Larry E. Toothaker. *Multiple comparison procedures.* Number 89. Sage, 1993.

Vladimir N. Vapnik and Alexey Y. Chervonenkis. Teoriya raspoznavaniya obrazov. statisticheskie problemy obucheniya (theory of pattern recognition. statistical problems of learning), 1974.

Peter H Westfall. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993.

Keith J. Worsley, Alan C. Evans, S. Marrett, and P. Neelin. A three-dimensional statistical analysis for rCBF activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism*, 12:900–918, 1992.

Tong Zhang. Statistical Analysis of Some Multi-Category Large Margin Classification Methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

4 Data-driven HRF estimation for encoding and decoding models

We have seen in Chapter 3 that encoding and decoding models take as input brain activation coefficients (also known as activation patterns or beta-maps). These are usually computed by means of the general linear model (GLM), which relies on a data-independent *canonical* form of the hemodynamic response function (HRF).

In this chapter we describe a novel method for the simultaneous estimation of HRF and activation coefficients based on low-rank modeling, forcing the estimated HRF to be equal across events or experimental conditions, yet permitting it to differ across voxels. The estimation of this model leads to an optimization problem that we propose to solve with using a quasi-Newton method, exploiting fast gradient computations. We compare 10 different HRF modeling methods in terms of encoding and decoding score on two different datasets. These 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.

The contributions developed in this chapter have been published in:

- F. Pedregosa, M. Eickenberg, P. Ciuciu, and B. Thirion, *"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. 3nd Int. Work. Pattern Recognit. NeuroImaging, 2013.

Contents

4.1 Increased sensitivity via HRF estimation

fMRI acquisitions consist of successive brain scans, given in intervals ranging from 1 to 4 seconds. The extraction of time-independent activation coefficient from the BOLD time course is commonly done with a model known as Linear General 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 *reference* form of the hemodynamic response function (HRF) to estimate the activation coefficient (also known as *canonical HRF*). However it is known [Handwerker et al., 2004, Badillo et al., 2013b] that the shape of this response function can vary substantially across subjects, age 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. to generalize poorly 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., 2011, Chapter 5].

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 time course). 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

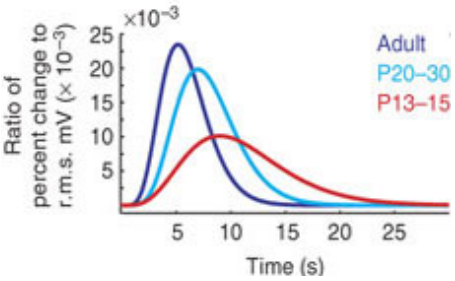


Figure 4.1: The HRF can vary substantially between subjects, brain regions and age. In Colonnese et al. [2007], the authors studied the evolution of the HRF across age in rats. By comparing fMRI measurements with electrophysiological recordings, they observed two significant trends as age increased: growing amplitude and decreasing time to peak. In the figure, estimated HRF for three groups of rats (with age P13-15 < P20-30< Adult). Source: [Colonnese et al., 2007]. A comparison of the HRF in human subjects was performed in [Badillo et al., 2014].

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 setting, the development of voxel-wise estimation procedures is complementary to the development of parcellation methods in that more robust estimation methods at the voxel level would naturally translate into more robust parcellation methods. In this thesis we focus on voxel-wise estimation methods.

Contribution In this chapter we have described a method for the simultaneous estimation of HRF and activation coefficients based on low-rank modeling. While the assumptions of this model are not novel (cf. [Makni et al., 2008, Vincent et al., 2010, Degras and Lindquist, 2014]), the formulation of this model as a least squares problem with a rank-one constraint is a novel contribution. This formulation allows to efficiently solve the problem using gradient-based methods. Finally, we evaluate the proposed model on three publicly available datasets.

4.2 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.

4.2.1 Basis-constrained GLM

The reference HRF models a general response function that has been proven successful under a wide range of circumstances. However, a number of studies have shown that the shape of the hemodynamic response differ substantially among subjects [Aguirre et al., 1998] and brain regions [Schacter et al., 1997]. One popular approach to model small offsets in the time to peak and dispersion is to consider that the HRF is modeled from a basis set consisting of the reference HRF plus its derivative with respect to time and dispersion (see Figure 4.2). The rationale for considering this basis set comes from the fact that it corresponds to the first-order approximation to the Taylor expansion of the reference HRF. Given the reference HRF, $h(t)$, a time-shifted version of the hemodynamic response can be described as $h(t + \delta)$. A Taylor series expansion of $h(t + \delta)$ with respect to δ gives the approximation $h(t) + \delta h'(t) + \dots$, implying that small offsets can be modeled by considering a linear combination of the reference HRF plus its time derivative. In similar fashion we can model small perturbations in dispersion (the width of the response) by considering the reference HRF plus its dispersion derivative.

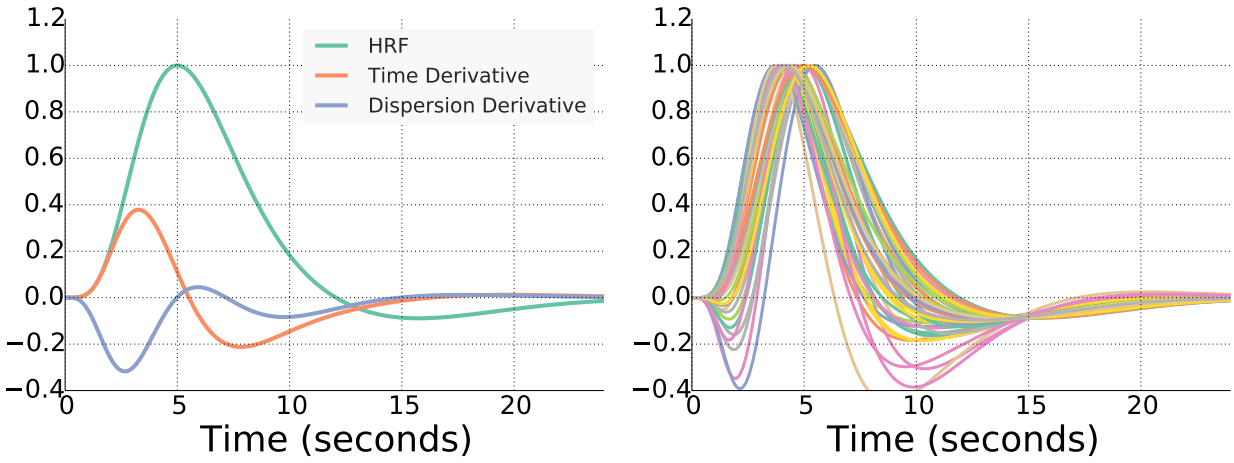


Figure 4.2: A popular basis set to generate a family of HRF functions is the "reference HRF plus derivatives". In the left plot, we show a reference HRF together with its time and dispersion derivatives. This basis set can model small variations in temporal shifts and dispersion with respect to the reference HRF. In the right plot we show a sample set of HRFs generated by this basis. The weights of these response functions are Gaussian random vectors centered around the reference HRF.

A popular example of basis set is presented in Figure 4.2 and consists of the reference HRF plus its time and dispersion (width) derivatives. While in the GLM with fixed HRF each regressor of the design matrix consisted of the convolution of the reference HRF with the stimulus function, in this case each regressor consist in the convolution of a basis element with the stimulus function. This results in a design matrix of size $n \times dk$ instead of $n \times k$, where d is the number of basis elements. If $d = 1$ and the basis element is the reference HRF, then this setting coincides with the standard GLM. A least squares estimate of the activation coefficients $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ will result in a vector of d elements for each condition.

The design matrix of a GLM using the basis set of the "refence HRF plus derivatives" is shown in Figure 2.10. The columns in this design matrix are each one of the basis elements convolved with the stimulus function for the different conditions.

4.2.2 Basis and rank-constrained GLM

In the basis-constrained GLM, 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 increased variance of the estimates.

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 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 $\boldsymbol{\beta}_{\mathbf{B}} = \text{vec}(\mathbf{h}\boldsymbol{\beta}^T)$. This can be seen as a constraint on the vector of coefficients to be the vectorization of a rank-one matrix, hence

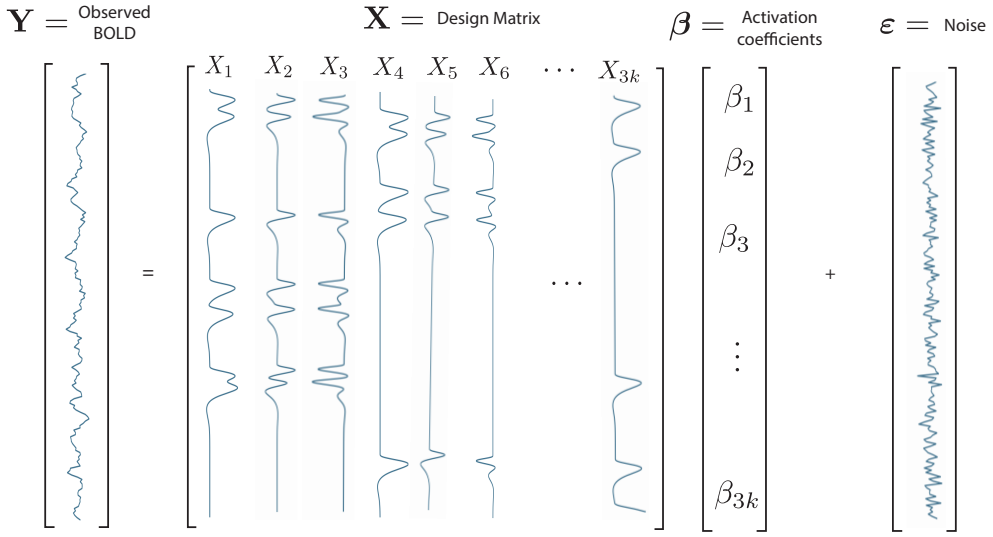


Figure 4.3: A basis-constrained GLM design matrix. The basis set consists of the reference HRF plus its time and dispersion derivative. As in the GLM introduced in Chapter 2 (Fig. 2.10), each column is the convolution of one basis function with the stimulus function. Here, the usage of 3 basis functions (instead of one) results in a design matrix with $3k$ regressors

the name *Rank-1 GLM (R1-GLM)*.

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

$$\begin{aligned} \hat{\mathbf{h}},\ \hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\omega}} \ = \ & \underset{\mathbf{h},\boldsymbol{\beta},\boldsymbol{\omega}}{\arg\min}\ F_{\mathrm{R1}}(\mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{X_B}, \mathbf{y}, \mathbf{Z}) \\ & \text{subject to } \|\mathbf{Bh}\|_\infty = 1 \text{ and } \langle \mathbf{Bh}, \mathbf{h}_{\mathrm{ref}} \rangle > 0 \ , \end{aligned} \tag{4.1}$$

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. Within its feasible set, 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. (4.1) 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 $h : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^d$ of some parameters $\boldsymbol{\alpha}$, we can formulate the analogous model of Eq. (4.1) as an optimization over the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with the design matrix $\mathbf{X}_{\text{FIR}}$ given by the convolution of the event matrix with the FIR basis:

$$\begin{aligned} \hat{\boldsymbol{\alpha}},\ \hat{\boldsymbol{\beta}},\ \hat{\omega} \ &= \ \underset{\boldsymbol{\alpha},\boldsymbol{\beta},\omega}{\arg\min}\, F_{\text{R1}}(h(\boldsymbol{\alpha}), \boldsymbol{\beta}, \omega, \mathbf{X}_{\text{FIR}}, \mathbf{y}, \mathbf{Z}) \\ &\text{subject to } \|h(\boldsymbol{\alpha})\|_\infty = 1 \text{ and } \langle h(\boldsymbol{\alpha}), \mathbf{h}_{\text{ref}} \rangle > 0 \end{aligned} \tag{4.2}$$

In section 4.2.4 we will discuss optimization strategies for both models.

4.2.3 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 ${\mathbf{X}_{\text{S}i}}^{\dagger}\mathbf{y}$, where $\mathbf{X}_{\text{S}i}$ 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].

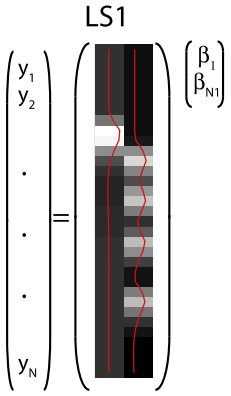


Figure 4.4: In the GLM with separate designs model of Mumford et al. [2012], the design matrix contains two regressors. The first one is the regressor associated with a given condition and the second one is the sum of all other regressors. Source: [Turner et al., 2012]

This approach was further extended in [Turner et al., 2012] to include 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 $F_{\text{R1-S}}(\mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{X}_{\mathbf{B}}, \mathbf{y}, \mathbf{Z}) = \frac{1}{2}\sum_i^k \|\mathbf{y} - \beta_i \mathbf{X}^0_{\text{BS}i}\mathbf{h} - r_i \mathbf{X}^1_{\text{BS}i}\mathbf{h} - \mathbf{Z}\boldsymbol{\omega}\|^2$, 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

$$\begin{aligned} \hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\omega}},\ \hat{\mathbf{h}},\ \hat{\mathbf{r}} \ &= \ \underset{\mathbf{h},\boldsymbol{\beta},\boldsymbol{\omega},\mathbf{r}}{\arg\min}\, F_{\text{R1-S}}(\mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{X}_{\mathbf{B}}, \mathbf{y}, \mathbf{Z}) \\ &\text{subject to } \|\mathbf{B}\mathbf{h}\|_\infty = 1 \text{ and } \langle \mathbf{B}\mathbf{h}, \mathbf{h}_{\text{ref}} \rangle > 0 \end{aligned} \tag{4.3}$$

4.2.4 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 (4.1) is to alternate the minimization with respect to the variables $\boldsymbol{\beta}$, $\mathbf{h}$ and $\boldsymbol{\omega}$. By recalling the Kronecker product identities [Horn and Johnson, 1991, Chapter 4.3], and using the identity $\mathrm{vec}(\mathbf{h}\boldsymbol{\beta}^T) = \boldsymbol{\beta} \otimes \mathbf{h}$ we can rewrite the objective function (4.1) to be minimized as:

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}_\mathbf{B}(\boldsymbol{\beta} \otimes \mathbf{h}) - \mathbf{Z}\boldsymbol{\omega}\|^2 = \tag{4.4}$$

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}_\mathbf{B}(\mathbf{I} \otimes \mathbf{h})\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\omega}\|^2 = \tag{4.5}$$

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}_\mathbf{B}(\boldsymbol{\beta} \otimes \mathbf{I})\mathbf{h} - \mathbf{Z}\boldsymbol{\omega}\|^2 \ . \tag{4.6}$$

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 $\mathbf{X}_\mathbf{B}(\boldsymbol{\beta} \otimes \mathbf{I})$ and $\mathbf{X}_\mathbf{B}(\mathbf{I} \otimes \mathbf{h})$ 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, $\mathbf{z} = \mathrm{vec}([\mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\omega}])$, 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 $\partial F_{\mathrm{R1}}/\partial \mathbf{z} = \mathrm{vec}([\partial F_{\mathrm{R1}}/\partial \mathbf{h}, \partial F_{\mathrm{R1}}/\partial \boldsymbol{\beta}, \partial F_{\mathrm{R1}}/\partial \boldsymbol{\omega}])$, whose expression can be easily derived using the aforementioned Kronecker product identities:

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

If instead a parametric model of the HRF is used as in Eq. (4.2), 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 $\mathbf{a} \in \mathbb{R}^k$ it computes $M(\mathbf{a}) = \mathbf{X}(\mathbf{a} \otimes \mathbf{h})$, avoiding thus the explicit computation of $\mathbf{I} \otimes \boldsymbol{\beta}$.

Similar equations can be derived for the rank-1 model with separate designs of Eq. (4.3) (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. 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 with $\mathbf{B}$ 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. (4.1) 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. 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

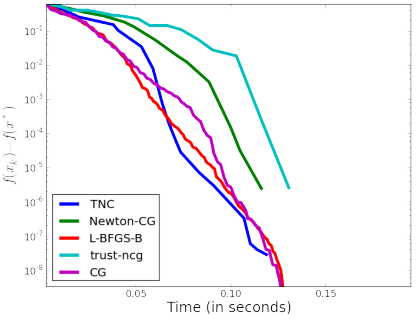


Figure 4.5: Convergence of different first-order and quasi-newton optimization algorithms for the R1-GLM model on a single voxel. "TNC" and "Newton-CG" are two different implementations of the truncated Newton [Nash, 1984] method (the first one in C and the second one in Python), "L-BFGS-B" is the Limited-memory BFGS algorithm with box constraints as implemented in [Zhu et al., 1997], "trust-ncg" is the Newton conjugate gradient trust-region algorithm and "CG" is the conjugate gradient algorithm, both of them described in [Nocedal and Wright, 2006]. We found that in general the L-BFGS-B gives the best performance among these methods.

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

Require: 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 $\mathbf{H}_0$.

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

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

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

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

4: Compute search direction. Set $\mathbf{p}_m \leftarrow -\mathbf{H}_m \nabla f(\mathbf{h}_m, \boldsymbol{\beta}_m, \boldsymbol{\omega}_m(, \mathbf{r}_m))$ by means of the L-BFGS algorithm.

5: 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 $\|\mathbf{h}_m\|_\infty \leq 1$.

6: $m \leftarrow m + 1$

7: **end while**

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

9: 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 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.

4.2.5 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 available at https://pypi.python.org/pypi/hrf_estimation . This software is further described in Section 7.4.1.

4.3 Data description

With the aim of making the results easily reproducible, we have chosen two freely available datasets to validate our approach and to compare different HRF modeling techniques.

4.3.1 Dataset 1: encoding of visual information

The first dataset we will consider is described in [Kay et al., 2008, Naselaris et al., 2009, Kay et al., 2011]. It contains BOLD fMRI responses in human subjects viewing natural images. As in [Kay et al., 2008], we performed prediction of BOLD signal following the visual presentation of natural images and compared it 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 10 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 2mm×2mm×2.5mm 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.

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 2 orientations and 4 scales. Ridge regression (regularization parameter chosen by Generalized Cross-Validation [Golub et al., 1979]) 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.

In a first step, we perform the image identification task from [Kay et al., 2008] (Fig. 4.6). 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

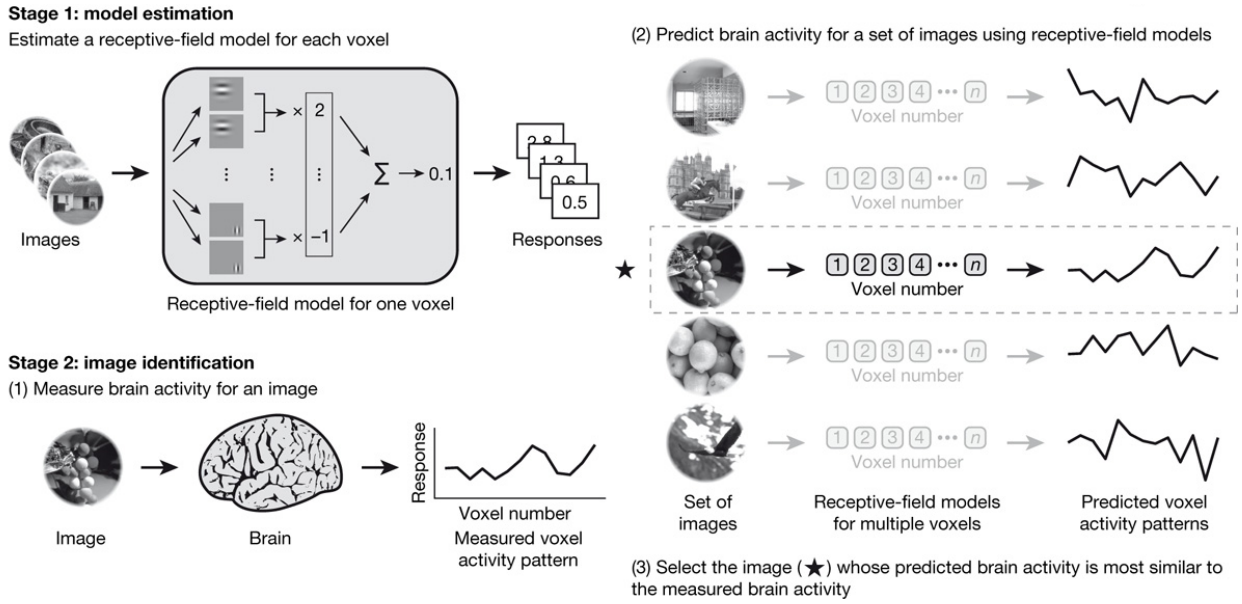


Figure 4.6: The original analysis performed in [Kay et al., 2008] allowed to identify natural images from human brain activity. The analysis consisted of two stages. In the first stage, model estimation, fMRI data were recorded while each subject viewed a large collection of natural images. These data were used to estimate an encoding model for each voxel. In the second stage, image identification, fMRI data were recorded while each subject viewed a collection of novel natural images. For each measurement of brain activity, they attempted to identify which specific image had been seen. This was accomplished by using the estimated encoding models to predict brain activity for a set of potential images and then selecting the image whose predicted activity correlates best with the measured activity. Source: Adapted from [Kay et al., 2008].

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.

4.3.2 Dataset 2: decoding of potential gain levels

The second dataset described in [Tom et al., 2007] 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 Pypreprocess[1] interface.

For all subjects three runs were recorded, each consisting of 240 images with a repetition time (TR) of 2 seconds and a stimulus presentation at ev-

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

ery 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 sensible to shrinkage of the prediction that might occur when penalizing 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 is equivalent to minimizing the number of the pairwise inversions, which was was previously proposed for fMRI decoding with ordered labels in [Pedregosa et al., 2012].

4.4 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.

4.4.1 Dataset 1: encoding of visual information

In the original study, 500 voxels were used to perform image identification. These voxels were selected as the voxels with the highest correlation with the true BOLD signal on left-out data using a (classical) GLM with the reference HRF. These voxels are therefore not the ones naturally benefiting the most from HRF estimation.

We first present the scores obtained in the image identification task for different variants of the GLM. This can be seen in Figure 4.7. The displayed score is the count of correctly identified images over the total number of im-

ages (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 acheive 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 being be the probability of recovering the correct image with method A and p_B being be 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 4.7: 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.

We will now use a different metric for evaluating the performance of the encoding model. This metric is the Pearson correlation between the

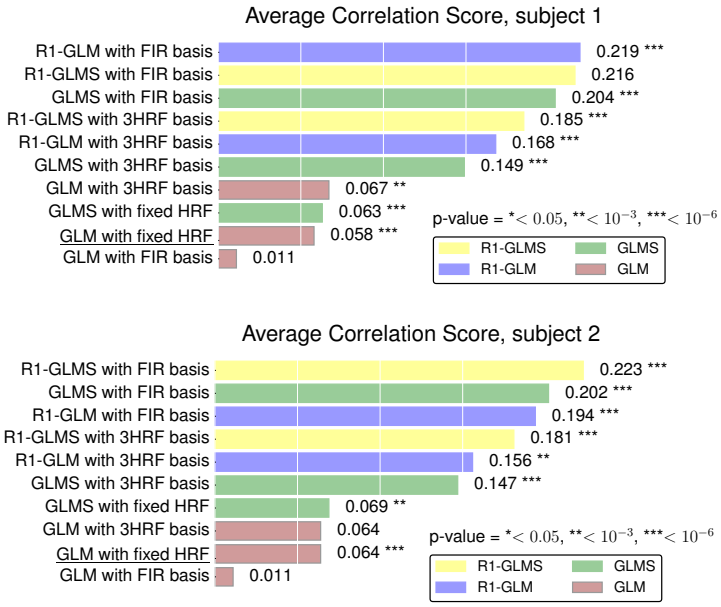


Figure 4.8: Average correlation score (higher is better) on two different subjects from the first dataset. The average correlation score is the Pearson correlation between the predicted BOLD and the true BOLD signal on left-out session, averaged across voxels and sessions. Methods that perform constrained 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.

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 4.8 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 4.8 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 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.

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.

To further inspect the results, we investigated the estimation and encoding scores at the voxel level. This provides some valuable information. For

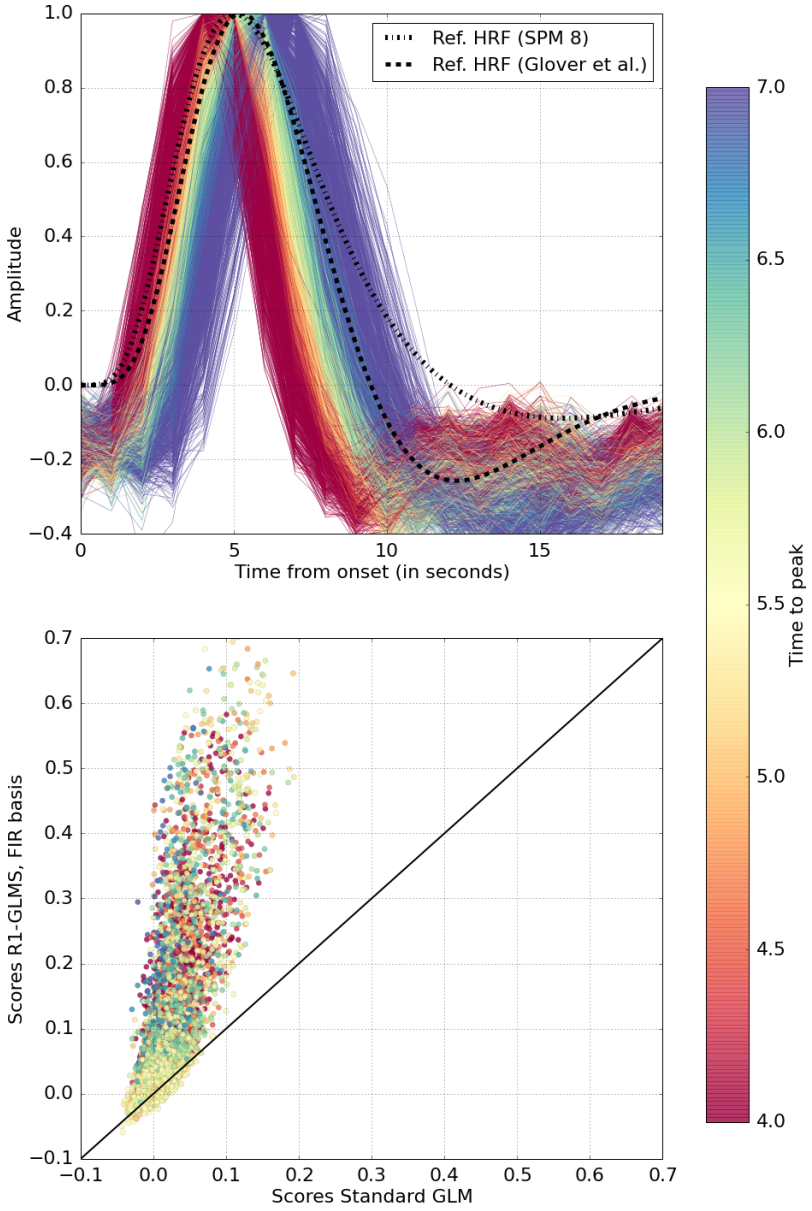


Figure 4.9: Top: HRF estimated by the R1-GLMS method on voxels for which the encoding score was above the mean encoding score (first dataset), color coded according to the time to peak of the estimated HRFs. The difference in the estimated HRFs suggests a substantial variability at the voxel level within 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.

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 4.9 we show the HRF estimated on 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, auditory study] 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 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.

[2] http://nipy.org

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

In the lower part of Figure 4.9 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 4.10 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

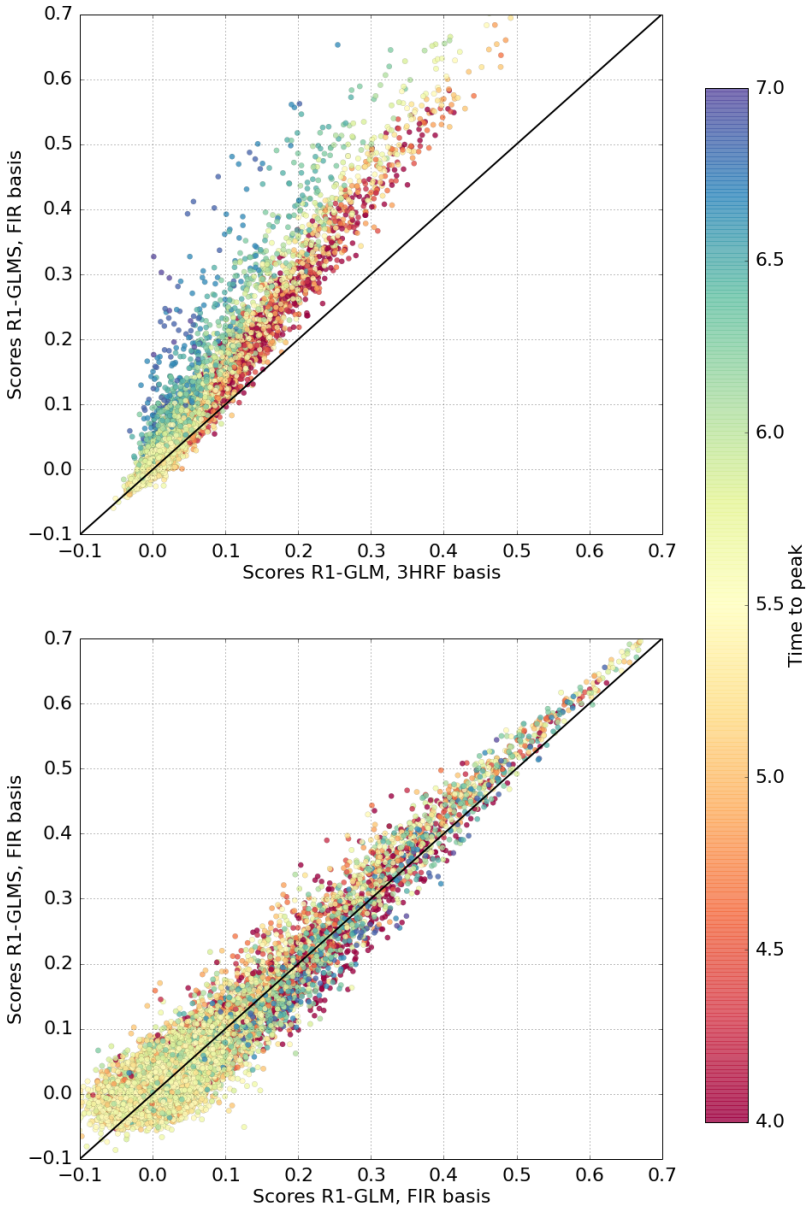


Figure 4.10: Voxel-wise encoding score for different models that perform HRF estimation (first dataset). As in figure 4.9, color codes for the time to peak of the estimated HRF at the given voxel. Top: two Rank-1 separate design models with different basis functions: FIR with 20 elements in the Y-axis 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 $< 10^{-3}$.

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 4.11 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 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

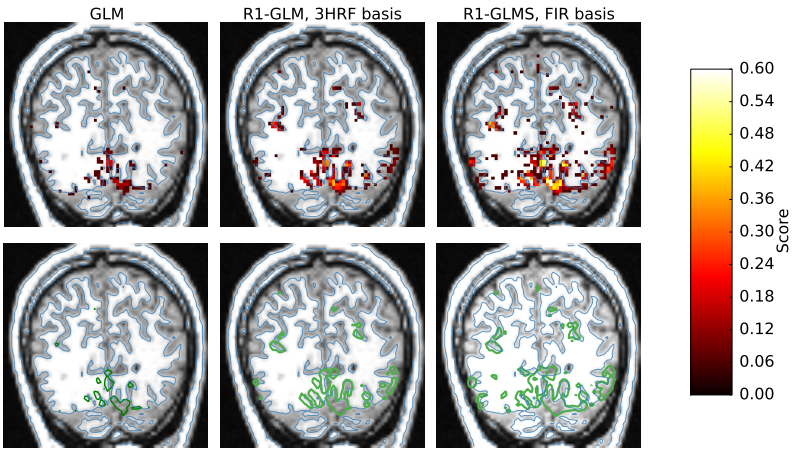


Figure 4.11: Voxel-wise encoding scores on a single acquisition slice for different estimation methods (first dataset). The metric is Pearson correlation. In the upper column, the voxel-wise score is thresholded at a value of 0.045 (p-value < 0.05), while in the bottom row the 0.055 contour (p-value < 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 sensitivity than the standard GLM method. In the bottom row it can be seen how the top performing voxels follow well the folding of the gray matter.

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.

4.4.2 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. 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 4.12. 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 4.12.

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.

4.5 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

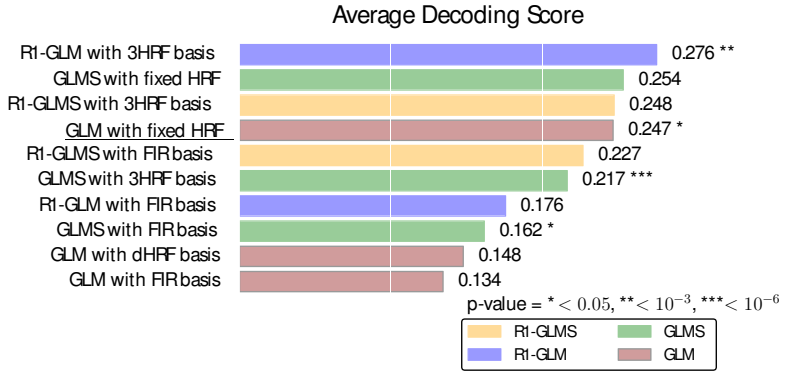


Figure 4.12: Averaged decoding score across subjects for the different method considered (higher is better) on the second dataset. The metric is Kendall tau. Methods that perform constrained HRF estimation significantly outperform methods that use a fixed (reference) HRF. In particular, the best performing 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 4.8, 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.

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 [Marrelec 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 was investigated using an encoding model 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 results were obtained for both subjects within the dataset, but since the results were similar for both subjects, we only show the results for the first subject. 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 is investigated using 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.

4.6 Conclusion

We have presented a method for the joint estimation of HRF and activation coefficients within the GLM framework. Based on ideas from previous literature [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 use a quasi-Newton method for its optimization. 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.

Bibliography

G.K. Aguirre, E. Zarahn, and M. D'Esposito. The variability of human, BOLD hemodynamic responses. *NeuroImage*, 8(4):360 – 369, 1998.

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.

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.

Solveig Badillo, Severine Desmidt, Chantal Ginisty, and Philippe Ciuciu. Multi-subject bayesian joint detection and estimation in fmri. In *Pattern Recognition in Neuroimaging, 2014 International Workshop on*, pages 1–4. IEEE, 2014.

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.

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.

Lofti Chaari, F. Forbes, T. Vincent, and Philippe 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.

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.

Matthew T. Colonnese, Marnie A. Phillips, Martha Constantine-Paton, Kai Kaila, and Alan Jasanoff. Development of hemodynamic responses and functional connectivity in rat somatosensory cortex. *Nature neuroscience*, 11(1): 72–79, 2007.

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

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

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.

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

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.

Cyril Goutte, Finn A. Nielsen, and Lars Kai 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.

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.

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

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.

Kendrick N. Kay, Naselaris, and Jack L. Gallant. fMRI of human visual areas in response to natural images. *CRCNS.org*, 2011.

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

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.

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.

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

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.

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.

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.

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.

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, 2009.

Stephen G Nash. Newton-type minimization via the lanczos method. *SIAM Journal on Numerical Analysis*, 21(4): 770–788, 1984.

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

Fabian Pedregosa, Elodie Cauvet, Gaël Varoquaux, Christophe Pallier, Bertrand Thirion, and Alexandre Gramfort. Learning to rank from medical imaging data. In *Third International Workshop on Machine Learning in Medical Imaging - MLMI 2012*, Nice, France, July 2012. INRIA.

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

Jean-Baptiste Poline and Matthew Brett. The general linear model and fMRI: does love last forever? *NeuroImage*, 62 (2):871–80, August 2012.

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.

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

Daniel L. Schacter, Randy L. Buckner, Wilma Koutstaal, Anders M. Dale, and Bruce R. Rosen. Late onset of anterior prefrontal activity during true and false recognition: An event-related fMRI study. *NeuroImage*, 6(4):259 – 269, 1997.

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.

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 2007. 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.

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.

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.

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.

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.

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.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

5 Decoding with Ordinal Labels

We have presented in Chapter 2 the decoding problem in fMRI. In this setting it is often the case that the target variable consists of discretely ordered values. This occurs for example when target values consists of human generated ratings, such as values on a Likert scale, the symptoms of a physical disease or a rating scale for clinical pain measurement.

In this chapter we propose the usage of two metrics to assess the performance of a decoding model when the target consists of discretely ordered values: the absolute error and pairwise disagreement. These two loss functions emphasize different aspects of the problem: while the absolute error gives a measure of the closeness of a predicted label to the true label, the pairwise disagreement gives a measure of correct ordering of the predicted labels. The choice of either metric will depend on the particular application at hand. For example, in clinical applications it is often desirable to predict a label as close as possible to the true label, in which case the absolute error is the appropriate metric. If however, the purpose of the decoding study is to perform a statistical hypothesis test to claim that the area encodes some information about the stimuli, then the pairwise disagreement can be considered.

We present three models based on different convex surrogates of the absolute error: least absolute error, ordinal logistic regression and cost-sensitive multiclass classification. We also consider a model that minimizes a surrogate of the pairwise disagreement: the RankLogistic model. We examine the generalization performance of the presented models on both synthetic data and three fMRI decoding problems from two datasets. We conclude that the best performing models is the last absolute error and ordinal logistic when considering the absolute error as metric and the RankLogistic model when considering the pairwise disagreement as metric.

The contributions relative to the use of the pairwise disagreement loss function have been published in:

- F. Pedregosa, E. Cauvet, G. Varoquaux, C. Pallier, B. Thirion, and A. Gramfort, *"Learning to rank from medical imaging data"*, in Proceedings of the 3rd International Workshop on Machine Learning in Medical Imaging, 2012.

Contents

5.1 Learning from ordinal labels

Let us motivate the problem of learning from ordinal labels by a decoding example. In the context of an fMRI acquisition, a subject is presented a set of words that represent real world objects: hammer, cow, sheep and whale. We are then interested to know whether it is possible to predict (i.e. *decode*) the implied real world size of the associated objects (i.e. the size of a goat rather than the size of the word "goat") based on the brain activation maps. How can we do this ?

Figure 5.1: In this experiment, the stimuli are words that represent real world objects. As in the Figure, these can be ordered according the the size of the associated concepts. The decoding problem will be then to predict the size of the associated concepts based on the brain activation maps.

In order to frame this problem as a decoding problem, we must choose a metric to evaluate the quality of our prediction (i.e. a *loss function*). Furthermore, as we have seen in Section 3.2.2, many models can be seen as the minimization of a convex surrogate of a given loss function. Thus, the chosen metric will determine which are the appropriate models to choose.

We have seen in previous chapters how the 0-1 loss can be applied to situations in which the target values consists of several categories. In this case, however, the 0-1 loss might give an overly pessimistic estimate of the performance of a classifier since it treats all misclassification errors alike. Suppose that a classifier predicts always the correct size $\pm$ 1, that is, never predicts the correct label but always predicts one of the adjacent elements in terms of size. This classifier will have the worst performance possible in terms of the 0-1 error, although we might still consider that this classifier is able predict with acceptable accuracy the size of an object.

It thus seems reasonable to choose a loss function that takes into account the distance among the labels. In this Chapter we present two metrics that fulfill this request and are adapted to the problem of supervised learning with ordinal labels. These loss functions are the *absolute error* and the *pairwise disagreement*. The use of the of the pairwise disagreement loss in the context of brain imaging is an original contribution first proposed in [Pedregosa et al., 2012].

We will describe in Section 5.4 three different surrogate loss functions of the absolute error and one surrogate of the pairwise disagreement. In section 5.5, we present the performance accuracy of these models in one synthetic dataset and three fMRI datasets. It is our intention for these results to provide guidelines on what methods are overall best suited in the context of decoding with ordered labels.

We have motivated the decoding problem with ordinal labels from a simple fMRI experiment in which the target variable is ordered according to the size of real world objects. However, the framework presented here can be applied to any situation in which the target variable consists of *discrete measures with some embedded order*. For example, this includes situations

in which the target variable consists of the symptoms of a physical disease such as Alzheimer's [Mueller et al., 2005], pain levels [Hartrick et al., 2003] or the syntactic complexity of a phrase [Pallier et al., 2011] to name a few.

5.2 Loss functions

A metric that arises naturally to evaluate the quality of an ordinal prediction the distance between the predicted label $\hat{y}$ and the true label y, which will promote classifiers that predict a label that is close to the correct label. This is known as the *absolute error* loss function, and is defined as:

$$\ell_{\mathcal{A}}(y, \hat{y}) = |y - \hat{y}|$$

This metric is very common when evaluating models with ordinal labels [Cardoso and Sousa, 2011]. A related loss worth mentioning is the squared error, often used in the regression setting. Compared to the squared error, the absolute error provides the advantage of being more robust to outliers [Bloomfield and Steiger, 1983].

The second loss function that we will present takes a different approach to measure the closeness of an ordinal response that does not take into account the value of the labels but rather only its relative ordering. Unlike the absolute error loss, this loss function acts on pairs of elements. Given two elements $y_1, y_2 \in \mathcal{Y}$ and the predicted values $\hat{y}_1, \hat{y}_2 \in \mathbb{R}$, this loss is defined as [Schapire and Singer, 1998, Herbrich et al., 2000]:

$$\ell_{\mathcal{P}}(y_1, y_2, \hat{y}_1, \hat{y}_1) = \mathcal{H}(-(y_1 - y_2)(\hat{y}_1 - \hat{y}_2)) \quad ,$$

where we recall that $\mathcal{H}$ is the Heaviside function, defined as $\mathcal{H}(x) = 1$ if $x \geq 0$ and 0 otherwise. That is, the loss $\ell_{\mathcal{P}}($ will be equal to one if $\hat{y}_1 - \hat{y}_2$ has not the same sign than $y_1 - y_2$ and zero otherwise. Since this loss is based on pairs of samples, the empirical risk will be evaluated on all possible pairwise combinations of elements in the training set (excluding those with same label). Given the training pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, the evaluation metric is defined as:

$$\hat{\mathcal{R}}_{\ell_{\mathcal{P}}}(f) = \frac{1}{m} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ y_i \neq y_j}}^{n} \ell_{\mathcal{P}}(y_i, y_j, f(x_i), f(x_j)) \quad ,$$

where m is the amount of pairwise combinations of samples with different labels.

This expression can be further simplified if we consider the symmetry of the loss function. Since all pairs of labels appear twice (once for $y_i > y_j$ and once for $y_i < y_j$), we can restrict ourselves to the set of elements which verify $y_i > y_j$, in which case we can write the empirical risk as

$$\hat{\mathcal{R}}_{\ell_{\mathcal{P}}}(f) = \frac{2}{m} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ y_i > y_j}}^{n} \mathcal{H}(f(x_j) - f(x_i)) \tag{5.1}$$

5.3 Ranking and ordinal regression

The different loss functions considered here lead to two different supervised learning problems. The problem in which we seek to predict an ordering as

close as possible to the true ordering of a sequence of labels is traditionally known as *ranking* while the problem of predicting a label as close as possible to the correct label is known as *ordinal regression.*

The *ordinal regression* setting was first studied by [McCullagh, 1980] and further developed in [Frank and Hall, 2001, Rennie and Srebro, 2005, Chu and Keerthi, 2007, 2005, Chu and Ghahramani, 2005] to name a few. The minimization of the absolute error can be seen as a special case of ordinal regression. In Chapter 5 we will study ordinal regression in a general setting that includes the minimization of other loss functions such as the squared loss.

Ranking models appear chronologically later than ordinal regression. The minimization of the pairwise disagreement was proposed in [Schapire and Singer, 1998] although the first attempt to minimize a convex surrogate of this loss is in [Herbrich et al., 2000][1]. There has been great interest in theory of ranking models in recent years with the application of these models to the field of information retrieval [Joachims, 2002, Burges et al., 2007, Sculley, 2009]. Analog theoretical results to the ones developed in Chapter 5 have been studied for the case of pairwise disagreement in [Duchi et al., 2010, Calauzènes et al., 2012]

[1] Although in that paper the authors refers to the proposed model (now known as RankSVM) as still as an regression models, later those same methods would be re-discovered as ranking models [Joachims, 2002].

5.4 Models

The models that we present here minimize a convex surrogate of the absolute error or the pairwise disagreement loss function. We present three different surrogate loss functions for the absolute error (least absolute error, ordinal logistic regression and cost-sensitive multiclass classification) and one for the pairwise disagreement (RankLogistic).

Because of the high dimensionality of the decoding problem and the associated risk of overfitting, the most popular choice for prediction functions in encoding and decoding models are linear decision functions [Cox and Savoy, 2003, LaConte et al., 2005, Song et al., 2011, Thirion et al., 2006, Naselaris et al., 2011], i.e., models in which the decision function is a linear mapping f from the sample space $\mathcal{X}$ onto $\mathbb{R}^d$. d is an integer that depends on the model: $d = 1$ in the case of least absolute error and RankLogistic, $d = k - 1$ in the case of ordinal logistic regression and $d = k$ in the case of multiclass support vector machines, where k is the number of classes. All models are estimated as a trade-off between a data-fitting term (the surrogate loss function) and a squared ℓ_2 penalty that controls for overfitting. The amount of penalty is chosen by nestted cross-validation.

The training set consists of n pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i$ is a p-dimensional vector and $y_i \in [k] = \{1, 2,' ld.$

5.4.1 Least absolute error

The first possibility that we will explore it is known as *least absolute deviations* [Bloomfield and Steiger, 1980, Narula and Wellington, 1982]. We consider the following surrogate loss function $\psi_{\mathcal{A}} : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$:

$$\psi_{\mathcal{A}}(y, \alpha) = |y - \alpha| \quad .$$

Note that this surrogate has the same expression as the absolute error

$\ell_{\mathcal{A}}$. The difference strives in that the surrogates are continuous functions in their second arguments while the loss functions take values in the discrete set $[k]$. The prediction function for these surrogates is given by rounding to the closest integer in $[k]$, i.e., $\text{pred}(\alpha) = \min_{i \in [k]} |i - \alpha|$. Although this prediction rule might seem somewhat ad-hoc for the moment, we will see in the next chapter (Section 6.3.2) that it is indeed the "optimal" prediction function for this surrogate (in some yet to be defined notion of optimality).

The model parameters are estimated by finding the minimizer of a trade-off between a data fidelity term (the $\psi_{\mathcal{A}}$-risk) and the penalty term:

$$\mathbf{w}^*, b^* \in \arg\min_{\mathbf{w},b} \frac{1}{n} \sum_{i=1}^{n} |y_i - b - \langle \mathbf{x}_i, \mathbf{w} \rangle| + \lambda \|\mathbf{w}\|^2 \quad ,$$

where $\mathbf{w} \in \mathbb{R}^p$ and $b \in \mathbb{R}$ is referred to as the *bias* or *intercept* term. This model can be seen as a particular instance of *support vector regression* with linear kernel and parameter ε in the ε-insensitive loss set to zero. This is a well studied model for which efficient implementations have been developed [Ho and Lin, 2012, Fan et al., 2008]. In the experiments section we will use the implementation provided in the LIBLINEAR library [Fan et al., 2008].

5.4.2 Ordinal logistic regression

The second approach that we consider is known as *ordinal logistic regression* [Rennie and Srebro, 2005] and can be seen in the larger family of *threshold-based ordinal regression* models [McCullagh, 1980, Rennie and Srebro, 2005, Chu and Keerthi, 2005, Lin and Li, 2006]. Let $\boldsymbol{\alpha} \in \mathbb{R}^{k-1}$ be the image of a decision function, that is, $\boldsymbol{\alpha} = f(x)$ for some $x \in \mathcal{X}$ and consider the following prediction function:

$$\text{pred}(\boldsymbol{\alpha}) = 1 + \sum_{i=1}^{k-1} \mathcal{H}(-\alpha_i) \quad . \tag{5.2}$$

In this case, we can express the absolute error loss function in the following form:

$$\begin{aligned} \ell_{\mathcal{A}}(y, \hat{y}) &= |y - \left(1 + \sum_{i=1}^{k-1} \mathcal{H}(-\alpha_i) \right)| \\ &= y - 1 - \sum_{i=1}^{y-1} \mathcal{H}(-\alpha_i) + \sum_{i=y}^{k-1} \mathcal{H}(-\alpha_i) \\ &= \sum_{i=1}^{y-1} \mathcal{H}(\alpha_i) + \sum_{i=y}^{k-1} \mathcal{H}(-\alpha_i) \quad , \end{aligned}$$

where we have used the following property of the Heaviside function: $\mathcal{H}(x) = 1 - \mathcal{H}(-x)$.

This last formula makes it clear that the absolute error can be seen as an addition of zero-one loss functions [2]. If we replace the Heaviside function by one of its convex surrogates such as the logistic loss, we obtain the following surrogate loss function:

$$\psi_{\mathcal{M}}(y, \boldsymbol{\alpha}) = \sum_{i=1}^{y-1} \varphi(-\alpha_i) + \sum_{i=y}^{k-1} \varphi(\alpha_i) \quad , \tag{5.3}$$

[2] the zero-one loss can be defined in terms of the Heaviside step function as $\ell_{0-1}(y, \hat{y}) = \mathcal{H}(-y \cdot \hat{y})$.

where $\varphi : \mathbb{R} \to \mathbb{R}$ is the logistic loss, defined as $\varphi(t) = \log(1 + e^{-t})$. Note that for $k = 2$, this coincides with the logistic regression model for binary 0-1 classification (Section 3.2.2).

When φ is the hinge loss, this approach has been proposed under the name of *Support Vector Ordinal Regression* (the implicit constraints variant) [Shashua and Levin, 2003, Chu and Keerthi, 2007]. For φ the exponential loss, this approach was proposed by [Lin and Li, 2006] as *Ordinal Regression Boosting (ORBoost)*. Rennie and Srebro [2005] formulated this model for an arbitrary surrogate loss function (as it is presented here) and considered a number of surrogates, including the hinge loss and logistic loss. We have chosen the logistic regression as a surrogate of the 0-1 loss for ease of implementation (the surrogate is a smooth function in this case, which allows us to use gradient-based methods) rather than the more popular *Support Vector Ordinal Regression* that arises when considering the hinge loss instead. However, due to the similarity between the hinge and logistic surrogates we expect both methods to yield similar results.

In this setting we consider $\boldsymbol{\alpha}$ to be of the form $\alpha_i = \theta_i - \mathbf{x}^T\mathbf{w}$, where $\mathbf{w} \in \mathbb{R}^p$ and $\boldsymbol{\theta} \in \mathbb{R}^k$ is a non-decreasing vector known as the *vector of thresholds*. Let us introduce the variable $s_{ij} = \operatorname{sign}(j - y_i + \frac{1}{2})$ for notational convenience. Then we can write the surrogate loss function from Eq. (5.3) as $\sum_{i=1}^{k-1} \varphi(s_{yi}\alpha_i)$. The coefficients $\mathbf{w} \in \mathbb{R}^p$ and the vector of thresholds $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k-1})$ will be estimated as the minimizers of the regularized empirical $\psi_{\mathcal{M}}$-risk, defined as

$$\mathbf{w}^*, \boldsymbol{\theta}^* \in \arg\min_{\mathbf{w},\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{k-1} \varphi(s_{ij}(\theta_j - \langle \mathbf{x}_i \mathbf{w} \rangle)) \right) + \lambda \|\mathbf{w}\|^2 \tag{5.4}$$

Unlike the other models presented in this section, the optimization of this model has not been extensively studied in the literature nor does it have a freely available implementation[3]. We will thus briefly discuss the optimization strategy that was employed to learn this model.

[3] The similar model Support Vector Ordinal Regression (the function φ is the hinge loss instead of the logistic loss) does have a freely available implementation. However, its optimization uses the dual form of the SVM while our optimization is based on the optimization of the primal formulation

For the problem sizes considered in this thesis, it is known that Newton and quasi-Newton methods yield excellent performance for ℓ_2-regularized logistic regression [Lin et al., 2008, Fan et al., 2008, Pedregosa, 2013]. Given the similarities with ordinal regression we decided to use the quasi-Newton L-BFGS-B algorithm for this problem. This algorithm requires to compute the objective function and its gradient.

The gradient of the objective function from (5.4) with respect to $\mathbf{w}$ and its partial derivatives with respect to θ_j can be computed as

$$\begin{aligned} \nabla_{\mathbf{w}} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \left(\sum_{j=1}^{k-1} \sigma(-s_{ij}\alpha_{ij}) \right) + 2\lambda_1 \mathbf{w} \\ \frac{\partial}{\partial \theta_j} &= \frac{1}{n} \sum_{i=1}^{n} s_{ij}(\sigma(s_i \alpha_{ij}) - 1) \quad . \end{aligned} \tag{5.5}$$

where σ is the sigmoid function, i.e. $\sigma(t) = 1/(1 + \exp(-t))$.

5.4.3 Multiclass classification

Since we aim at predicting a finite number of labels with a specific loss functions, it is also possible to use generic multiclass formulations such as the

one proposed in [Lee et al., 2004] which can take into account generic losses. As before, given ϕ the logistic loss function, this formulations considers the following surrogate

$$\psi_{\mathcal{L}}^{\ell}(y, \boldsymbol{\alpha}) = \sum_{i=1}^{k} \ell(y, i)\phi(-\alpha_i) \tag{5.6}$$

for $\boldsymbol{\alpha} \in \mathbb{R}^k$ such that $\sum_{i=1}^{k} \alpha_i = 0$. The prediction function in this case is given by $\text{pred}(\boldsymbol{\alpha}) = \arg\max_{i \in [k]} \alpha_i$. Note however that this method requires the estimation of k decision functions. For this reason, in practical settings threshold-based are often preferred as these only require the estimation of one decision function and $k-1$ thresholds.

In practice the matrix of coefficients $\mathbf{W} \in \mathbb{R}^{k \times p}$ is estimated as the minimizer of the following optimization problem

$$\mathbf{W}^*, \mathbf{b}^* \in \arg\min_{\mathbf{W},\mathbf{b}} \sum_{i=1}^{n} \sum_{j=1}^{k} |j - y_i| \phi(b_j - \langle \mathbf{x}_i, \mathbf{W}_j \rangle) + \lambda \|\mathbf{W}\|_{\mathcal{F}}$$

subject to the constraint $\mathbf{W}^T \mathbf{1}_k = \mathbf{0}$, $\mathbf{b}^T \mathbf{1}_k = 0$. Implementation details for this model can be found in [Zhang et al., 2008, Zhang and Jordan, 2006, Statnikov et al., 2005]

5.4.4 RankLogistic

This loss described in Eq. (5.1) function suggests as a natural choice for a surrogate loss is one of the form [Herbrich et al., 2000, Freund et al., 2003, Dekel et al., 2004] $\psi_{\mathcal{P}}(y_1, y_2, \alpha_1, \alpha_2) = \varphi((y_1 - y_2)(\hat{y}_1 - \hat{y}_2))$. This yields the following expression for the empirical $\psi_{\mathcal{P}}$-risk:

$$\hat{\mathcal{R}}_{\psi_{\mathcal{P}}}(f) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ y_i > y_j}}^{n} \varphi(f(x_i) - f(x_j)) \tag{5.7}$$

where as before $\varphi : \mathbb{R} \to \mathbb{R}$ is a surrogate of the zero-one loss such as the hinge or logistic loss. Here we will consider the case in which φ is the logistic loss. For the case in which φ is the hinge loss this model is sometimes referred to as *RankSVM* [Herbrich et al., 2000, Joachims, 2002].

In case the prediction function f is given by a linear function, this expression can be further simplified. In this case, we have that $f(x_i) - f(x_j) = f(x_i - x_j)$. That is, given two samples (x_i, x_j) and their associated labels (y_i, y_j) $(y_i \neq y_j)$ we form a new sample $x_i - x_j$ with label $\text{sign}(y_i - y_j)$. Due to the linearity of f, predicting the correct ordering of these two images, is equivalent to predicting the sign of $f(x_i) - f(x_j) = f(x_i - x_j)$ [Herbrich et al., 2000]. We can now write the model as the solution to the optimization problem

$$\mathbf{w}^*, b^* \in \arg\min_{\mathbf{w},b} \frac{2}{m} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ y_i > y_j}}^{n} \varphi((\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{w} + b)$$

This optimization problem can be viewed as a binary class classification problem on all pairwise combinations of $(\mathbf{x}_i - \mathbf{x}_j, \text{sign}(y_i - y_j))$ and thus can be solved using standard supervised classification algorithms. For consistency with previous sections, we will use the logistic loss instead and denote

this model *RankLogistic*. One of the possible drawbacks of this method with respect previous methods is that it requires to consider all possible pairs of images. This scales quadratically with the number of training samples, and the problem soon becomes intractable as the number of samples increases. However, specialized algorithms exist with better asymptotic properties [Joachims, 2006, Sculley, 2009]. For our study, we used the Support Vector Machine algorithms implemented in the LIBLINEAR library [Fan et al., 2008].

Figure 5.2: Maurice G. Kendall (6 September 1907 – 29 March 1983) was a British statistician, widely known for his contribution to statistics. The Kendall tau rank correlation is named after him.

Relationship with Kendall's τ. Some authors (e.g. [Joachims, 2002, Chen et al., 2009, Wauthier et al., 2013]) present the RankSVM model as the model that maximizes a surrogate of the *Kendall* τ correlation coefficient. We will show that maximizing Kedall's τ and minimizing the pairwise disagreement yield equivalent optimization problems.

Kendall's τ can be defined as

$$\tau = \frac{P - Q}{P + Q}$$

where P is the number of *concordant pairs*, that is, the number of elements $i > j$ such that $\alpha_i \geq \alpha_j$ and Q is the number of of *discordant pairs*, that is, the number of elements $i > j$ such that $\alpha_i < \alpha_j$. An equivalent formulation of Kendall's τ is [Joachims, 2002]

$$\tau = 1 - \frac{2Q}{\binom{n}{2}}$$

From here it is clear that maximizing the Kendall τ coefficient is equivalent to minimizing the number of discordant pairs. Since the pairwise disagreement counts the number of discordant pairs, both approaches are equivalent.

5.5 Experiments

5.5.1 Ordinal regression and dimensionality

The models that we have presented vary greatly in terms of parameters to estimate. Given that k is the number of classes and p is the dimensionality (number of features) of the dataset, the least absolute error and RankLogistic models estimate $p + 1$ parameters, the ordinal logistic model estimates $p + k$ parameters and the multiclass classification model estimates $k \times (p + 1)$ parameters. While methods with more parameters can express a richer set of decision functions, the increase in the number of parameters to estimate also induces a higher variance of the estimates which can result in poor generalization performance in settings such as decoding in which the number of samples is very limited and the dimensionality of the dataset is high.

To illustrate this problem we computed the generalization error of the different methods as we increase the dimensionality of a synthetic dataset. The setting is the following: the data is generated by applying a random linear regression model with 10% of informative nonzero regressors and Gaussian centered noise such that the signal-to-noise ratio is 10:1. The target variable is the discretized in 5 bins such that the number of samples is

equal for each class. All models have a squared ℓ_2 penalization term that has been set chosen among a grid of 10 log-spaced values between 10^{-3} and 10^6 by nested 5-fold cross-validation. The generalization score is computed by 10-fold cross-validation on an equally spaced grid of features between 100 and 600 features.

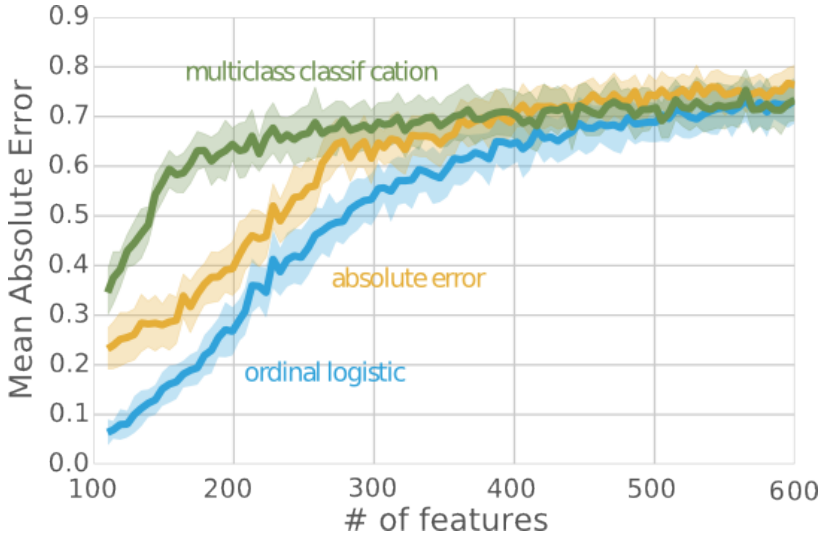


Figure 5.3: Generalization error as the number of dimensions increase on a synthetic dataset (lower is better). In the low sample regime, ordinal logistic regression outperforms the other methods, but as the number of dimension increases the gap between the methods vanishes. The poor performance of multiclass classification even in the low sample regime can be explained by the model we used to generate the data, which corresponds to the assumptions of ordinal logistic regression.

The generalization errors (lower is better) are displayed in Figure 5.3. It can be observed that in the regime with low number of features, ordinal logistic regression significantly outperforms the other methods, but as the number of dimension increases the gap between the methods vanishes. The data was generated as a discretized linear regression model, which corresponds very closely to the models assumed by the ordinal logistic model[4]. This might give an advantage to this model and explain the poor behavior of multiclass classification even in the regime with low number of features.

[4] in which the prediction function is of the form $\sum_i \mathcal{H}(\theta_i - \mathbf{x}^T\mathbf{w})$

5.5.2 Results on two fMRI datasets

To assess the performance of the different methods presented on the decoding problem, we investigate two fMRI datasets.

The first dataset that we will considered served as motivation for this chapter. It was presented in [Borghesani et al., 2014] and has already been mentioned in Section 3.2.5. The goal of this experiment is to predict different aspects of the words that subjects were seeing while undergoing an fMRI acquisition. We can consider two different decoding problems based this dataset. As a first step, we investigate the effect of the low level perceptual features characterizing the stimuli: the number of letters composing each word. We will call this decoding problem *length of word*. A second decoding problem that can be investigated on this dataset is to test the relationship between activation images and the real size of items. In this case, the different stimuli are ordered according to their relative size, i.e. hammer is smaller than cow which is smaller than a whale, etc., so the target variable is of ordinal nature. We will call this decoding problem *size of object*. In both cases we extracted the activation coefficients (beta-maps) using the R1-

GLM model with 3hrf basis described in Chapter 4. Since we are interested in predicting the target from low level visual features we restrict the decoding problem on an anatomically defined ROI for the primary visual cortex (V1) using the SPM toolbox PickAtlas (13940 voxels). 6 sessions were available for each subject. We trained the model on 5 sessions and evaluated the model on the left out session. We report the average generalization score across subjects.

The second dataset, described in [Cauvet, 2012], consists of 34 healthy volunteers scanned while listening to 16 words sentences with five different levels of complexity. These were 1 word constituent phrases (the simplest), 2 words, 4 words, 8 words and 16 words respectively, corresponding to 5 levels of complexity which was used as class label in our experiments. To clarify, a sentence with 16 words using 2 words constituents is formed by a series of 8 pairs of words. Words in each pair have a common meaning but there is meaning between each pair. A sentence has therefore the highest complexity when all the 16 words form a meaningful sentence. This dataset contains four manually labeled regions of interest that can be seen in Figure 5.4: Anterior Superior Temporal Sulcus, Temporal Pole, Inferior Frontal Gyrus Orbitalis and Inferior Frontal Gyrus triangularis. Further analysis will be limited to these regions of interest. In this dataset each subject only has two sessions, which is insufficient to compute the leave-one session out score. Because of this we instead train the model on 33 subjects and report the cross-validation score on a left-out subject.

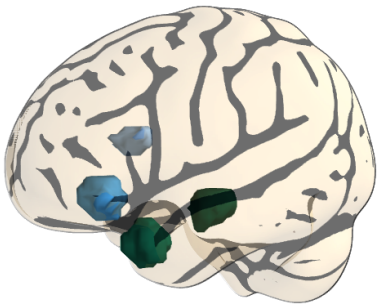


Figure 5.4: Manually labeled ROIs in the *language complexity* dataset [Cauvet, 2012].

The generalization errors (lower is better) for these three decoding problems (spanning two datasets) are displayed in Figure 5.5. We considered two different metrics, represented as rows in the figure: mean absolute error and mean pairwise disagreement. We ordered the models by performance and performed a Wilcoxon signed-rank test between each method and the next best performing method to assess whether the difference between both methods is statistically significant. This test is performed by considering the sequence of cross-validation scores obtained for each model. The p-value associated with this statistical test is denoted by one or two asterisks, with the convention that $* < 0.05, ** < 10^{-3}$.

When considering the mean absolute error, ordinal logistic regression and least absolute error are the best performing method. The difference between both methods is not significant in any of the three experiments. Multiclass classification is the worst performing method due to the high dimensionality of the problem and the high number of parameters to estimate ($p \times (k-1)$ versus $p + k - 1$ for ordinal logistic).

When considering the pairwise disagreement error, the best performing method is the RankLogistic model. RankLogistic is also the only model that minimizes a surrogate of the evaluation metric.

5.6 Discussion

From the experiments we have examined the relative performance of several classifiers and concluded that ordinal logistic and least absolute error are the best performing methods when evaluated using mean absolute error and RankLogistic is best model when evaluated using mean pairwise

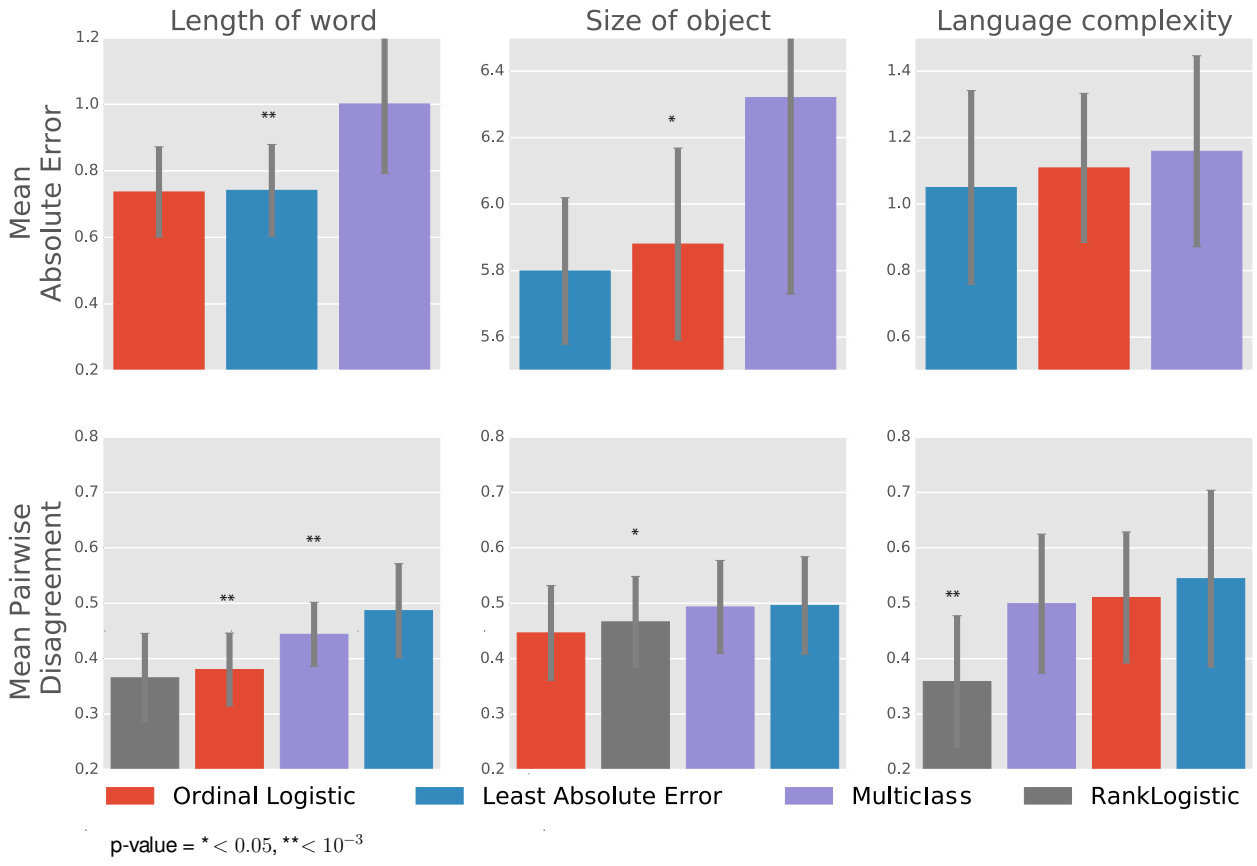


Figure 5.5: Generalization errors (lower is better) for three fMRI decoding problems. Two different metrics are used corresponding to the rows in the figure: mean absolute error and mean pairwise disagreement. The $*$ symbol represents the p-value associated with a Wilcoxon signed-rank test. This test is used to determine whether a given method outperforms significantly the next best-performing method.

disagreement. The superiority of RankLogistic highlights the importance of choosing a model that minimizes a surrogate of the evaluation metric.

A question that arises in practice is: when should the absolute error metric be used and when should the pairwise disagreement metric be used?. The use of one or the other will depend on the particular application in mind. For example, for clinical applications it is often necessary to predict the exact label. If the target variable consists of the different degrees of Alzheimer's disease it is natural to consider an evaluation metric that reflects how close to the true label the prediction is. In this case we would favor the mean absolute error. If however, we are only interested in performing a statistical hypothesis test to claim that the area encodes some information about the stimuli, then the pairwise disagreement can be considered.

In this study we have considered the absolute error, but we could have as well considered the squared error loss instead. The linear least squares model, which minimizes a surrogate of this loss, has advantageous computational properties when compared to its absolute error counterpart, the least absolute deviation model: strong convexity, smoothness and analyticity of solutions. However, the use of absolute error resulted in a higher significance when performing hypothesis testing, which is often the end goal of a decoding study. For example, when performing the omnibus test on the "lenght of word" decoding problem, we could reject the null hypothesis that the explained variance is not significantly greater than the unexplained variance with a p-value < 0.001 when considering the mean absolute error

metric and the least absolute error. The p-value when considering the mean squared error metric with a linear least squares model (both models have the same number of parameters) was only < 0.005. Similar effects were observed on the other decoding problems.

5.7 Conclusion

In this chapter, we have proposed the usage of two evaluation metrics in the context of brain decoding when the target variable consists of ordered values: the absolute error loss and the pairwise disagreement loss function. We have presented models that optimize a convex surrogate of these loss functions and discussed estimation strategies for these models based on convex optimization.

We examined the performance of these models on both synthetic and two real world fMRI datasets and identified the best methods for each evaluation metric. Our results show that when considering the absolute error as evaluation metric, the least absolute error and the logistic ordinal model are the best performing methods while when considering the mean pairwise disagreement the RankLogistic was the best performing methods. For neuroimaging studies, this contribution outlines the best strategies to choose when faced with a decoding problem in which the target variable has a meaningful order.

Bibliography

Peter Bloomfield and William Steiger. Least absolute deviations curve-fitting. *SIAM Journal on scientific and statistical computing*, 1(2):290–301, 1980.

Peter Bloomfield and William L. Steiger. Least absolute deviations: Theory, applications and algorithms. *Birkhiiuser, Boston*, 1983.

Valentina Borghesani, Fabian Pedregosa, Evelyn Eger, Marco Buiatti, and Manuela Piazza. A perceptual-to-conceptual gradient of word coding along the ventral path. In *4th International Workshop on Pattern Recognition in Neuroimaging*, pages 3–6, 2014.

Christopher J. C. Burges, Robert Ragno, and Quoc Viet Le. Learning to Rank with Nonsmooth Cost Functions. *Machine Learning*, 19(17):193–200, 2007. ISSN 10495258.

Clément Calauzènes, Nicolas Usunier, and Patrick Gallinari. On the (Non-) existence of Convex , Calibrated Surrogate Losses for Ranking. *Advances in Neural Information Processing Systems 2012*, pages 1–9, 2012.

Jaime S. Cardoso and Ricardo Sousa. Measuring the Performance of Ordinal Classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(08):1173–1195, December 2011. ISSN 0218-0014.

Elodie Cauvet. *Traitement des Structures Syntaxiques dans le langage et dans la musique*. PhD thesis, Ecole doctorale n°158, Cerveau - Cognition - Comportement, 2012.

Wei Chen, Y. Lan, T.Y. Liu, and Hang Li. A unified view on loss functions in learning to rank. Technical report, Technical Report, Microsoft Research, MSR-TR-2009-39, 2009.

Wei Chu and Zoubin Ghahramani. Gaussian Processes for Ordinal Regression. *Journal of Machine Learning Research*, 6:1–24, 2005.

Wei Chu and S. Sathiya Keerthi. New Approaches to Support Vector Ordinal Regression. In *Proceedings of the 22th International Conference on Machine Learning (ICML)*, 2005.

Wei Chu and S. Sathiya Keerthi. Support Vector Ordinal Regression. *Neural computation*, 815(2001):792–815, 2007.

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.

Ofer Dekel, Christopher D. Manning, and Yoram Singer. Log-linear models for label ranking. In S. Thrun, L.K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 497–504. MIT Press, 2004.

John C. Duchi, Lester W. Mackey, and Michael I. Jordan. On the Consistency of Ranking Algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

Eibe Frank and Mark Hall. A Simple Approach to Ordinal Classification. *ECML '01: Proceedings of the 12th European Conference on Machine Learning*, 2001.

Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *The journal of machine learning research*, 4:933–969, 2003.

Craig T. Hartrick, Juliann P. Kovan, and Sharon Shapiro. The numeric rating scale for clinical pain measurement: A ratio measure? *Pain Practice*, 3(4):310–316, 2003. ISSN 1533-2500.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. *Large margin rank boundaries for ordinal regression*, volume 88, pages 115–132. MIT Press, Cambridge, MA, 2000.

Chia-Hua Ho and Chih-Jen Lin. Large-scale linear support vector regression. *The Journal of Machine Learning Research*, 13(1):3323–3348, 2012.

Thorsten Joachims. Optimizing Search Engines using Clickthrough Data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002.

Thorsten Joachims. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 217–226, New York, NY, USA, 2006. ACM.

Stephen LaConte, Stephen Strother, Vladimir Cherkassky, Jon Anderson, and Xiaoping Hu. Support vector machines for temporal classification of block design fmri data. *NeuroImage*, 26(2):317–329, 2005.

Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99 (465):67–81, 2004.

Chih-Jen Lin, Ruby C. Weng, and S. Sathiya Keerthi. Trust region newton method for logistic regression. *The Journal of Machine Learning Research*, 9:627–650, 2008.

Hsuan-Tien Lin and Ling Li. Large-margin thresholded ensembles for ordinal regression: Theory and practice. In *Algorithmic Learning Theory*, pages 319–333. Springer, 2006.

Peter McCullagh. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society*, 42(2):109–142, 1980.

Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford R Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. Ways toward an early diagnosis in alzheimer's disease: The alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

Subhash C. Narula and John F. Wellington. The minimum sum of absolute errors regression: A state of the art survey. *International Statistical Review/Revue Internationale de Statistique*, pages 317–326, 1982.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–10, May 2011.

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527, 2011.

Fabian Pedregosa. Numerical optimizers for logistic regression. `http://fa.bianp.net/blog/2013/numerical-optimizers-for-logistic-regression/`, 2013. Accessed: 2014-11-30.

Fabian Pedregosa, Elodie Cauvet, Gaël Varoquaux, Christophe Pallier, Bertrand Thirion, and Alexandre Gramfort. Learning to rank from medical imaging data. In *Machine Learning in Medical Imaging*, pages 234–241. Springer Berlin Heidelberg, 2012.

Jason D. M. Rennie and Nathan Srebro. Loss Functions for Preference Levels : Regression with Discrete Ordered Labels. In *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, 2005.

William W. Cohen Robert E. Schapire and Yoram Singer. Learning to order things. In *Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference*, volume 10, page 451. MIT Press, 1998.

D. Sculley. Large Scale Learning to Rank. *NIPS 2009 Workshop on Advances in Ranking*, pages 1–6, 2009.

Amnon Shashua and Anat Levin. Ranking with large margin principle : Two approaches. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2003.

Sutao Song, Zhichao Zhan, Zhiying Long, Jiacai Zhang, and Li Yao. Comparative study of svm methods combined with voxel selection for object category classification on fmri data. *PLoS ONE*, 6(2), 02 2011.

Alexander Statnikov, Constantin F Aliferis, Ioannis Tsamardinos, Douglas Hardin, and Shawn Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005.

Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Le Bihan, and Stanislas Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4):1104–16, December 2006.

Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. Efficient ranking from pairwise comparisons. In *Proceedings of the 30th International Conference on Machine Learning*, pages 109–117, 2013.

Hao Helen Zhang, Yufeng Liu, Yichao Wu, and Ji Zhu. Variable selection for the multicategory svm via adaptive sup-norm regularization. *Electronic Journal of Statistics*, 2:149–167, 2008.

Zhihua Zhang and Michael I. Jordan. Bayesian multicategory support vector machines. In *In Uncertainty in Artificial Intelligence, 2006. Ji Zhu, Saharon Rosset, Trevor*, 2006.

6 Fisher Consistency of Ordinal Regression Methods

Ordinal regression is the supervised learning problem of learning a rule to predict labels from an ordinal scale. Ordinal regression models enjoy a wide applicability and some ordinal regression models have already been used in Chapter 4 to model the decoding problem when the target variable consists of ordered values.

Many of the ordinal regression models that have been proposed in the literature can be viewed as methods that minimize a convex surrogate of the zero-one, absolute (as the methods presented in Chapter 4), or squared errors. In this chapter we investigate some theoretical properties of ordinal regression methods. The property that we will investigate is known as *Fisher consistency* and relates the minimization of a given loss to the minimization of its surrogate.

We provide a theoretical analysis of the Fisher consistency properties of a rich family of surrogate loss functions, including proportional odds and support vector ordinal regression. For all the surrogates considered, we either prove consistency or provide sufficient conditions under which these approaches are consistent. Finally, we illustrate our findings on real-world datasets.

The contributions developed in this chapter are available in the submitted paper

- F. Pedregosa-Izquierdo, F. Bach, and A. Gramfort, *"On the Consistency of Ordinal Regression Methods"*.

Contents

6.1 Introduction

In ordinal regression the goal is to learn a rule to predict labels from an ordinal scale, i.e., labels from a discrete but ordered set. This arises often when the target variable consists of human generated ratings. Besides the examples of ordinal labels in the context fMRI-based brain decoding presented in Chapter 4, examples of ordinal scales include ("do-not-bother" $\prec$ "only-if-you-must" $\prec$ "good" $\prec$ "very-good" $\prec$ "run-to-see") in movie ratings [Crammer and Singer, 2001], ("absent" $\prec$ "mild" $\prec$ "severe") for the symptoms of a physical disease [Armstrong and Sloan, 1989] and the NRS-11 numeric rating scale for clinical pain measurement [Hartrick et al., 2003]. Ordinal regression models have been successfully applied to fields as diverse as econometrics [Greene, 1997], epidemiology [Ananth and Kleinbaum, 1997], fMRI-based brain decoding [Doyle et al., 2013] and collaborative filtering [Rennie and Srebro, 2005].

In this chapter we turn to study some theoretical properties of these methods. The aim is that a theoretical approach allows a better understanding the methods at hand. For example, Chu and Keerthi [2005] proposed two different models for the task of ordinal regression: SVOR with explicit constraints and SVOR with implicit constraints. In that work, the second approach obtained better generalization error in terms of the absolute error loss function. Similar results were obtained by [Lin and Li, 2006] replacing the hinge loss by an exponential loss. Yet again, [Rennie and Srebro, 2005] arrived to similar conclusions by considering the logistic loss instead. One of the motivations behind this chapter is to answer the question: is there a theoretical reason that can explain this behavior? By the end of the chapter we will give arguments to answer this and other relation questions.

Before introducing the general formulation of ordinal regression, we briefly recall the supervised learning setting described in Section 3.2.1. Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. Let (X, Y) be two random variables with joint probability distribution P, where X takes its values in $\mathcal{X}$ and Y is a random label taking values in a set of *ordered categories* that we will denote $\mathcal{Y} = \{1, 2, \ldots, k\}$. In the ordinal regression problem, we are given a set of n observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ drawn i.i.d. from $X \times Y$ and a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. The goal is to learn from the training examples a measurable mapping called a *classifier* $h : \mathcal{X} \to \mathcal{Y}$ so that the *risk* given below is as small as possible:

$$\mathcal{R}_\ell(h) = \mathbb{E}_{X \times Y}(\ell(Y, h(X))) \quad . \tag{6.1}$$

The setting above looks similar to that of a multiclass classification problem. However, a loss function used for multiclass classification such as the 0-1 loss is not sensitive to the distance among target values. On the other hand, in the ordinal regression setting, because of the order between labels, the loss function becomes lower as the distance among classes decreases. This has been formalized as the *V-shape* property [Li and Lin, 2007]. We will say that a loss function is V-shaped if its forward difference, $\Delta\ell(i, j) = \ell(i, j+1) - \ell(i, j)$, verifies $\Delta\ell(i, j) \leq 0$ for $j \leq i$ and $\Delta\ell(i, j) \geq 0$ for $j > i$.

The *absolute error* loss function ($\ell_{\mathcal{A}}(y, k) = |y - k|$) is an example of V-

shaped loss function, although this property includes other loss functions, such as the *squared error*, $\ell_S(y,k) = (y-k)^2$ and the $0-1$ loss.

Attempting to directly minimize Eq. (6.1) is not feasible in practice for two reasons. First, the probability distribution P is unknown and the risk must be minimized approximately based on the observations. Second, due to the non-convexity and discontinuity of ℓ, the risk is difficult to optimize and can lead to an NP-hard problem [Feldman et al., 2012, Ben-David et al., 2003] (note that binary classification can be seen as a particular case of ordinal regression). It is therefore common to approximate ℓ by a function $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$, called a *surrogate loss function*, which has better computational properties. Here d is an integer that depends on the surrogate. For the methods that we consider d will be equal to $1, k-1$ or k. The goal becomes to find the *decision function* f that minimizes instead the ψ-risk, defined as

$$\mathcal{R}_n^{\psi}(f) = \mathbb{E}_{X\times Y}(\psi(Y, f(X))) \ . \tag{6.2}$$

Fisher consistency is a desirable property for surrogate loss functions [Lin, 2004]. It implies that in the population setting, i.e., if the probability distribution P were available, then optimization of the ψ-risk would yield a function (not necessarily unique) with smallest possible risk, known as *Bayes predictor* and denoted by h^*. This implies that within the population setting, the minimization of the ψ-risk and the minimization of the risk both yield solutions with same risk. From a computational point of view, this implies that the minimization of the ψ-risk, which is usually a convex optimization problem and hence easier to solve than the minimization of the ℓ-risk, does not penalize the quality of the obtained solution.

The chapter is organized as follows. In Section 6.2 we present the ordinal regression models that we will consider for study. These can be broadly separated into *regression-based* and *threshold-based*. Section 6.3 is divided into several parts. In the first part, we extend results from Ramaswamy and Agarwal [2012] and prove consistency of regression-based surrogates. Because of its practical interest, the rest of this section is devoted to investigate the consistency of threshold-based surrogates. Here we present our main results, which gives sufficient conditions under which these surrogates are consistent. We finish with experiments and conclusions.

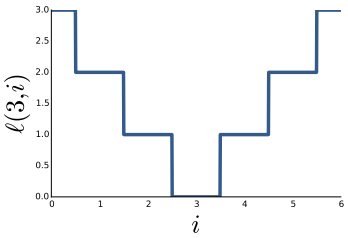


Figure 6.1: The absolute error, defined as $\ell(i,j) = |i-j|$, is a loss function that verifies the V-shape property. In the figure, a plot of the absolute error loss $\ell(i,j) = |i-j|$ with $j = 3$.

6.1.1 Related work

Fisher consistency of binary and multiclass classification for the zero-one loss has been studied for a variety of surrogate loss functions (see Bartlett et al. [2003], Tewari and Bartlett [2007] and references therein). Ramaswamy and Agarwal [2012] investigated the more general setting of multiclass classification with an arbitrary loss function, a setting that includes ordinal regression. The authors proved Fisher consistency of a surrogate loss function of the absolute error for the case of $k = 3$. However, this work did not prove consistency of this surrogate for $k > 3$, nor did it prove consistency for any squared error surrogate or for any of the threshold-based surrogates that represent the majority of traditional approaches for ordinal regression.

A related, yet different, notion of consistency is *asymptotic consistency*. A surrogate loss is said to be asymptotically consistent if the minimization of

the ψ-risk converges to the optimal risk as the number of samples tends to infinity. It has also been studied in the setting of supervised learning [Stone, 1977, Steinwart, 2002]. This chapter focuses solely on Fisher consistency, and for simplicity we will now use the term consistency to denote Fisher consistency.

Notation. As in the previous chapter we will denote the sequence of numbers from one to k as $[k] = \{1, 2, \ldots, k\}$. Throughout this chapter we will use letter k to denote the number of classes in the target space.

6.2 Ordinal regression models

Different methods have been proposed to learn an ordinal regression model. The *regression-based approach* treats the labels as real values. It uses a standard regression algorithm to learn a real-valued function, and then predicts by rounding to the closest label (see, e.g., Kramer et al. [2001] for a discussion of this method using regression trees). In this setting we will examine consistency of two different surrogate loss functions, the absolute error (that we will denote $\psi_{\mathcal{A}}$) and the squared error (denoted $\psi_{\mathcal{S}}$), which are convex surrogates of $\ell_{\mathcal{A}}$ and $\ell_{\mathcal{S}}$, respectively. Given $\alpha \in \mathbb{R}$, $y \in [k]$, these are defined as

$$\psi_{\mathcal{A}}(y, \alpha) = |y - \alpha|, \quad \psi_{\mathcal{S}}(y, \alpha) = (y - \alpha)^2 \quad . \tag{6.3}$$

Note that the loss functions $\ell_{\mathcal{A}}$ and $\ell_{\mathcal{S}}$ have the same expression as their surrogates, however the difference strives in that the surrogates are continuous functions in their second arguments while the loss functions take values in the discrete set $[k]$. The prediction function for these surrogates is given by rounding to the closest integer in $[k]$, i.e., $\text{pred}(\alpha) = \min_{i \in [k]} |i - \alpha|$. For half-integers, i.e., for number of the form integer + $\frac{1}{2}$, the rule is to round to the left, that is, $\text{pred}(1.5) = 1$, $\text{pred}(2.5) = 2$, etc.

While these approaches may lead to optimal predictors when no constraint is placed on the regressor function space as we will see in Section 6.3.2, in practice only simple function spaces are explored such as linear or polynomial functions. In these situations, the regression-based approach might lack flexibility. The *threshold-based approaches* [McCullagh, 1980, Rennie and Srebro, 2005, Chu and Keerthi, 2005, Lin and Li, 2006] provides greater flexibility by seeking for both a mapping $f : \mathcal{X} \to \mathbb{R}$ and a non-decreasing vector $\boldsymbol{\theta} \in \mathbb{R}^{k-1}$, often referred to as *thresholds*, that map the class labels into ordered real values.

The thresholds $\boldsymbol{\alpha}$ partition the real line into k segments, and the prediction is given by the segment into which the prediction $f(x)$ lies in (Fig 6.2). If we introduce the auxiliary variable $\alpha_i = \theta_i - f(x)$, an equivalent formulation of this prediction function is

$$\text{pred}(\boldsymbol{\alpha}) = 1 + \sum_{i=1}^{k-1} \mathcal{H}(-\alpha_i) \ , \tag{6.4}$$

where we recall that $\mathcal{H}$ is the Heaviside function, defined as $\mathcal{H}(x) = 1$ if $x \geq 0$ and 0 otherwise.

In the context of threshold-based functions we will consider two different families of surrogate loss functions. The first family of surrogate

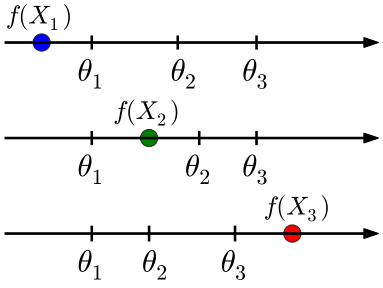


Figure 6.2: In the ordinal logistic model, the thresholds partition the real line into k segments and the prediction is given by the segment into which the decision function lies (assuming the segments are ordered by their relative order within the real line). Here, example for a 4-class problem. The prediction $f(X)$ for a given sample is denoted by a colored circle and $\theta_1, \theta_2, \theta_3$ are the estimated thresholds for that sample. Prediction in this example would be 1, 2, 4 respectively.

loss function that we will consider is the *cumulative link* surrogates of McCullagh [1980]. In such models the posterior probability is modeled as $P(Y \leq i|X = x) = \sigma(g_i(x))$, where σ is an appropriate link function. We will prove consistency for the case where σ is the sigmoid function, i.e., $\sigma(t) = 1/(1 + \exp(-t))$. In this case it is known as the *proportional odds* model or *cumulative logit* model. For $x \in \mathcal{X}, y \in [k]$ and $\alpha_i = g_i(x)$, the proportional odds surrogate (denoted ψ_C) is defined as

$$\psi_C(y, \boldsymbol{\alpha}) = \begin{cases} -\log(\sigma(\alpha_1)) & \text{if } y = 1 \\ -\log(\sigma(\alpha_y) - \sigma(\alpha_{y-1})) & \text{if } 1 < y < k \\ -\log(1 - \sigma(\alpha_{k-1})) & \text{if } y = k. \end{cases} \tag{6.5}$$

The second family of surrogate loss functions that we will consider are the *margin-based* surrogate loss functions of which the *ordinal logistic* model introduced in Chapter 4 is a particular example. For appropriate real-valued functions $\phi : \mathbb{R} \to \mathbb{R}$ such as the hinge loss or exponential loss, this surrogate separate target values by the largest margins centered around the thresholds [Lin and Li, 2006]. Given $x \in \mathcal{X}, y \in [k]$ and $\boldsymbol{\alpha} \in \mathbb{R}^{k-1}$, the margin-based surrogate (denoted $\psi^{\ell}_{\mathcal{M}}$) is given by

$$\psi^{\ell}_{\mathcal{M}}(y, \boldsymbol{\alpha}) = \sum_{i=1}^{y-1} \Delta\ell(y, i)\phi(\alpha_i) - \sum_{i=y}^{k-1} \Delta\ell(y, i)\phi(-\alpha_i) \ .$$

We recall that $\Delta\ell(y, i) = \ell(y, i + 1) - \ell(y, i)$. Note that the V-shape property implies $\Delta\ell(y, i) \geq 0$ for the elements in the first term and $\Delta\ell(y, i) \leq 0$ for elements in the second term, thus this surrogate is convex in its second argument if ϕ is a convex function.

This formulation parametrizes several popular approaches to ordinal regression. For example, let ϕ be the hinge loss and ℓ the zero-one loss, then ψ^T_{ℓ} coincides with the Support Vector Ordinal Regression ("explicit constraints" variant) of [Shashua and Levin, 2003, Chu and Keerthi, 2007]. If instead the mean absolute loss is considered, this approach coincides with the "implicit constraints" formulation of the same reference. For other values of ϕ and ℓ this loss includes the approaches proposed in [Shashua and Levin, 2003, Chu and Keerthi, 2005, Rennie and Srebro, 2005, Lin and Li, 2006]. In section 6.3.5 we will prove consistency results for arbitrary V-shaped loss function.

Since we aim at predicting a finite number of labels with a specific loss functions, it is also possible to use generic multiclass formulations such as the one proposed in [Lee et al., 2004] which can take into account generic losses. Given ϕ a real-valued function, this formulations considers the following surrogate

$$\psi^{\ell}_{\mathcal{L}}(y, \boldsymbol{\alpha}) = \sum_{i=1}^{k} \ell(y, i)\phi(-\alpha_i) \tag{6.6}$$

for $\boldsymbol{\alpha} \in \mathbb{R}^k$ such that $\sum_{i=1}^{k} \alpha_i = 0$. The prediction function in this case is given by $\text{pred}(\boldsymbol{\alpha}) = \arg\max_{i \in [k]} \alpha_i$. Note however that this method requires the estimation of $k - 1$ decision functions. For this reason, in practical settings threshold-based are often preferred as these only require the estimation of one decision function and $k - 1$ thresholds.

Consistency results of this surrogate was proven by Zhang [2004]. We will compare their results to our findings of consistency for threshold-based surrogates in Section 6.3.6.

Table 6.1 contains a list of the aforementioned surrogate loss functions, the (non-surrogate) loss function they target and their prediction function.

Loss	Surrogate	Prediction
Absolute error	$\lvert y-\alpha\rvert$	$\min_{i\in[k]} \lvert i-\alpha\rvert$
Squared error	$(y-\alpha)^2$	$\min_{i\in[k]} \lvert i-\alpha\rvert$
Absolute error	$\psi_C(y,\boldsymbol{\alpha})$	$1+\sum_{i=1}^{k-1}\mathcal{H}(-\alpha_i)$
Any V-shaped	$\psi^{\ell}_{\mathcal{M}}(y,\boldsymbol{\alpha})$	$1+\sum_{i=1}^{k-1}\mathcal{H}(-\alpha_i)$
Any	$\psi^{\ell}_{\mathcal{L}}(y,\boldsymbol{\alpha})$	$\arg\max_{i\in[k]}\alpha_i$

Table 6.1: Surrogate loss functions that we will examine in this paper. These include a number of popular approaches for ordinal regression, such as the support vector ordinal regression of Shashua and Levin [2003], Chu and Keerthi [2007] and the proportional odds model of McCullagh [1980].

6.3 Consistency of Surrogate Loss Functions

We will now give a precise definition for the (Fisher) consistency of a surrogate loss function. This notion originates from a classical parameter estimation setting. Suppose that an estimator T of some parameter θ is defined as a functional of the empirical distribution F_n, $T(F_n)$. The estimator is said to be Fisher consistent if its population analog, $T(F)$, coincides with the parameter θ. Adapting this notion to the context of risk minimization (in which the optimal risk is the parameter to estimate) yields the following definition, adapted from [Lin, 2004] to an arbitrary loss ℓ.

Definition 1. *(**Consistency**) Given a surrogate loss function $\psi : \mathcal{Y}\times\mathbb{R}^d \to \mathbb{R}$, a function space $\mathcal{F}$ and prediction rule pred : $\mathbb{R}^d \to [k]$, we will say that the pair $(\psi, pred)$ is consistent with respect to the loss ℓ if for every probability distribution over $X\times Y$ it is verified that every minimizer of the ψ-risk reaches Bayes optimal risk, that is,*

$$f^* \in \arg\min_{f\in\mathcal{F}} \mathcal{R}^{\psi}_n(f) \implies \mathcal{R}_\ell(\text{pred}\circ f^*) = \mathcal{R}_\ell(h^*) \quad .$$

By an abuse of notation we will refer to the consistency of a surrogate function ψ to designate the consistency of the pair (ψ, pred).

When the ψ-risk minimization is performed over all measurable functions, it is verified that

$$\begin{aligned}\inf_f \mathcal{R}^{\psi}_n(f) = \inf_f \mathbb{E}_{X\times Y}\left(\psi(Y, f(X))\right) = \\ \mathbb{E}_X\left[\inf_f \mathbb{E}_Y(\psi(Y, f(X)))\right] \quad .\end{aligned} \tag{6.7}$$

Hence in this case in order to compute the decision function with optimal risk it is sufficient to compute the decision function with minimal expected value (over Y) for every $x \in \mathcal{X}$.

6.3.1 Bayes predictor

In order to prove consistency of a surrogate loss we will find useful to have an explicit form for Bayes predictor. For example, in the case of binary

classification with the zero-one loss, Bayes predictor is known and is given by $\text{sign}(P(y=1|X=x)-1/2)$. In this section we will derive similar results for arbitrary V-shaped loss functions.

We first introduce the following notation. Let $\eta_i(x) = P(Y=i|X=x)$ denote the conditional probability at $X=x$. For $1 \le i < k$ we also define the functions $u_i, v_i : \mathcal{X} \to \mathbb{R}$ as

$$\begin{aligned} u_i(x) &= \sum_{j=1}^{i} \eta_j(x)\Delta\ell(j,i) \\ v_i(x) &= -\sum_{j=i+1}^{k} \eta_j(x)\Delta\ell(j,i) \quad . \end{aligned} \tag{6.8}$$

If ℓ is V-shape, then $\Delta\ell(j,i)$ is positive for $j \ge i$ and $(u_1(x), u_2(x), \ldots, u_k(x))$ is a non-decreasing positive sequence. Similarly, $\Delta\ell(j,i) \le 0$ for $i < j$ and $(v_1(x), v_2(x), \ldots, v_k(x))$ is a non-increasing positive sequence.

We now derive a formula for Bayes predictor of an arbitrary V-shaped loss function.

Theorem 1 (Bayes predictor for an ordinal regression loss)**.** *Let $\ell(i,j)$ be a V-shaped loss function. Then Bayes predictor is given by*

$$h^*(x) = 1 + \sum_{i=1}^{k-1} \mathcal{H}(v_i(x) - u_i(x)) \quad . \tag{6.9}$$

Proof. Let $x \in \mathcal{X}$ and $r = h^*(x)$. By the V-shape property we have that $(v_1(x)-u_1(x), v_2(x)-u_2(x), \ldots, v_k(x)-u_k(x))$ is a non-increasing sequence of i. Hence, $1 + \sum_{i=1}^{k-1} \mathcal{H}(v_i(x)-u_i(x)) = r$ implies that $(v_i - u_i) \ge 0$ for $1 \le i < r$ and $(v_i - u_i) < 0$ for $i \ge r$.

We will first prove $\mathbb{E}_Y(\ell(Y,r)) - \mathbb{E}_Y(\ell(Y,s)) \le 0$ for any $s \in [k]$. Suppose $s > r$, then we have

$$\begin{aligned} &\mathbb{E}_Y(\ell(Y,r)) - \mathbb{E}_Y(\ell(Y,s)) = \\ &\sum_{i=r}^{s-1} \mathbb{E}_Y(\ell(Y,i) - \ell(Y,i+1)) = \\ &\sum_{i=r}^{s-1} \left(-\sum_{j=1}^{k} \eta_j(x)\Delta\ell(j,i) \right) = \sum_{i=r}^{s-1} (v_i(x) - u_i(x)) \le 0 \end{aligned}$$

Similarly, for $s < r$

$$\begin{aligned} &\mathbb{E}_Y(\ell(Y,r)) - \mathbb{E}_Y(\ell(Y,s)) = \\ &\sum_{i=s}^{r-1} \mathbb{E}_Y(\ell(Y,i+1)) - \ell(Y,i) = \\ &\sum_{i=s}^{s-1} \left(\sum_{j=1}^{k} \eta_j(x)\Delta\ell(j,i) \right) = -\sum_{i=s}^{r-1} (v_i(x) - u_i(x)) < 0 \end{aligned}$$

We have proven that for any classifier h

$$\mathbb{E}_Y(\ell(Y,h^*(x))|X=x) - \mathbb{E}_Y(\ell(Y,h(x))|X=x) \le 0$$

Integrating both sides with respect to X yields

$$\mathcal{R}(h^*) \le \mathcal{R}(h) \quad ,$$

that is, h^* is Bayes predictor. □

An immediate consequence of this theorem is that Bayes predictor for the mean absolute error and the mean squared error admit the following simple form:

Corollary 1. . *Bayes predictor for the absolute error loss is given by*

$$h^*(x) = \min_{r\in[k]}\{r : \sum_{i=1}^{r} \eta_i(x) > \frac{1}{2}\} \quad . \tag{6.10}$$

Proof. By the V-shape property $(v_i(x) - u_i(x))$ is a non-increasing sequence of i. Hence if $h^*(x) = 1 + \sum_{i=1}^{k-1} \mathcal{H}(v_i(x) - u_i(x)) = r$ then it must be verified that $v_i(x) - u_i(x) < 0$ for $i \geq r$ and $v_i - u_i \geq 0$ for $i < r$. Because of this an alternative formulation of Bayes predictor (Eq. 6.9) is $h^*(x) = \min_{r\in[k]}\{r : u_r(x) > v_r(x)\}$.

For the absolute error loss, $\Delta\ell(i,j) = 1\ \forall i,j$. Thus, $v_i(x) = (1 - u_i(x)$ and from this $u_r(x) > v_r(x) \iff u_r > \frac{1}{2}$. Hence we can write $h^*(x) = \min_{r\in[k]}\{r : u_i(x) > \frac{1}{2}\} = \min_{r\in[k]}\{r : \sum_{i=1}^{r} \eta_i(x) > \frac{1}{2}\}$. □

Corollary 2. . *Bayes predictor for the squared error loss is given by*

$$h^*(x) = \min_{r\in[k]}\{r : \sum_{i=1}^{k} i\eta_i(x) > r - \frac{1}{2}\} \quad . \tag{6.11}$$

Proof. For the squared error loss, $\Delta\ell(i,j) = 1 - 2(i-j)$ and thus, $v_r(x) - u_r(x)$ = - $\sum_{j=1}^{k} \eta_j(x)(1 - 2(r-j)) = -1 + 2r - 2\sum_{j=1}^{k} j\eta_j(x)$. Hence $u_r(x) > v_r(x) \iff \sum_{j=1}^{k} j\eta_j(x) > r - \frac{1}{2}$. Using the alternate formulation of Bayes predictor given in Corollary 1 we can then write $h^*(x) = \min_{r\in[k]}\{r : \eta_j(x) > r - \frac{1}{2}\}$. □

Bayes predictor predicts a label from the conditional probability $(\eta_1(x), \eta_2(x), \ldots, \eta_k(x))$ and as such induces a partitioning of the probability simplex k regions. The probability simplex is the set $\{x \in \mathbb{R}^k : \sum_{i=1}^{n} x_i = 1, x_i \geq 0\}$ and is contained within an hyperplane of dimension $n-1$. In Figure 6.3, the probability simplex in $\mathbb{R}^3$ is colored according to the output of Bayes predictor. These sets have been previously studied for the 0-1 loss in [O'Brien et al., 2008] and for the absolute error in [Ramaswamy and Agarwal, 2012].

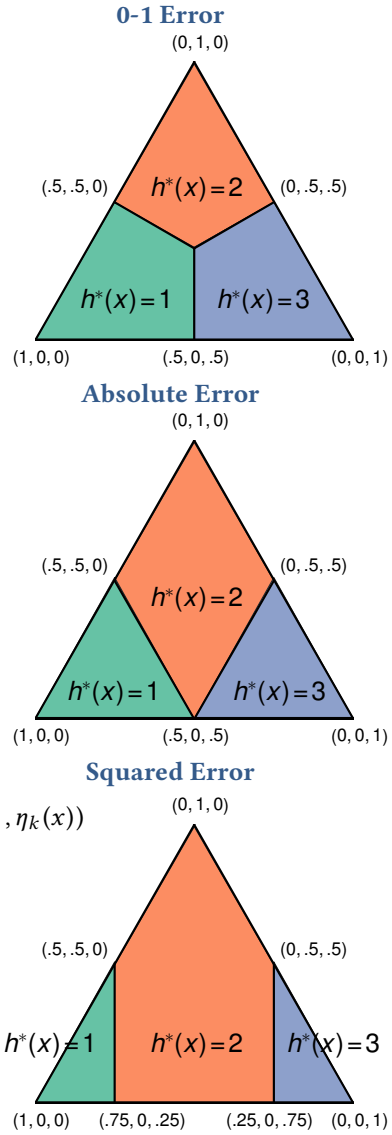


Figure 6.3: Bayes predictor on the probability simplex. Bayes predictor is a function of the conditional probability $\eta_i(x) = P(y = i|X = x)$. The vector $(\eta_1, \ldots, \eta_k)$ belongs to the probability simplex of R^n, which is contained within an hyperplane of dimension $k-1$. In the figure, probability simplex in $\mathbb{R}^3$ is colored according to the output of Bayes predictor.

6.3.2 Consistency of regression-based models

We will now examine the consistency of regression-based models. Consistency of the absolute error surrogate was proven by [Ramaswamy and Agarwal, 2012] for the case of 3 classes. Here we give an alternate simple proof that extends beyond $k > 3$.

Lemma 1. *The function with minimal $\psi_{\mathcal{A}}$-risk is $f^*(x) = \text{median}(Y|X=x)$, where* median *represents the median of a random variable (i.e. the value α such that $P(y \leq \alpha|X=x) \geq 1/2$ and $P(y \geq \alpha|X=x) \leq 1/2$). The function with minimal $\psi_{\mathcal{S}}$-risk is $f^*(x) = \mathbb{E}_Y(Y|X=x)$.*

Proof. By the application of optimality properties of the median and mean, the median and the mean are the scalar values that minimize $\mathbb{E}_Y(\psi_{\mathcal{A}}(Y,\alpha)|X=x) = \mathbb{E}_Y(|Y-\alpha||X=x)$ and $\mathbb{E}_Y(\psi_{\mathcal{S}}(Y,\alpha)|X=x) = \mathbb{E}_Y((Y-\alpha)^2|X=x)$, respectively. In light of Eq. (6.7) this is sufficient to obtain the minimal risk. □

Theorem 2. *The absolute error surrogate $\psi_{\mathcal{A}}$ is consistent with respect to $\ell_{\mathcal{A}}$.*

Proof. Let $x \in \mathcal{X}$, and $\alpha^* = \text{median}(Y|X = x)$. By definition of median, $P(y \leq \alpha^*|X=x) = \sum_{i=1}^{\alpha^*} \eta_i(x) \geq 1/2$ and $\sum_{\alpha^*}^{k} \eta_i(x) \leq 1/2$.

If $\sum_{i=1}^{\alpha^*} \eta_i(x) > 1/2$ and $\sum_{\alpha^*}^{k} \eta_i(x) \leq 1/2$ then $\alpha^* = \min_{r\in[k]}\{r : \sum_{i=1}^{r} \eta_i(x) > \frac{1}{2}\}$ and in light of Corollary 1 we predict the same label as Bayes predictor.

We have left out the case in which $\sum_{i=1}^{\alpha^*} \eta_i(x) = 1/2$. In this case the median would predict α^* but Bayes predictor would predict $\alpha^* + 1$. If we compute the conditional risk for these values we have

$$\mathbb{E}_Y(\ell(Y,\alpha^*)) - \mathbb{E}_Y(\ell(Y,\alpha^*)) =$$
$$\sum_{i=1}^{k} \eta_i(x)|i - \alpha^* + 1| - \sum_{i=1}^{k} \eta_i(x)|i - \alpha^*| =$$
$$\sum_{i=r}^{\alpha^*} \eta_i(x) - \sum_{i=\alpha^*+1}^{k} \eta_i(x) = 0$$

Hence in this case the risk associated with predicting α^* or $\alpha^* + 1$ is the same. We have shown thus that the risk associated with Bayes predictor is the same than the risk associated with the minimizer of $\psi_{\mathcal{A}}$ (the median), hence we have consistency of this surrogate.

□

Theorem 3. *The squared error surrogate $\psi_{\mathcal{S}}$ is consistent with respect to $\ell_{\mathcal{S}}$.*

Proof. Let $\alpha^* = \mathbb{E}_Y(Y|X=x) = \sum_{i=1}^{k} i\eta_i(x)$. Then $\text{pred}(\alpha^*) = \text{round}\left(\sum_{i=1}^{k} i\eta_i(x)\right)$ $= \min_{r\in[k]} \sum_{i=1}^{k} i\eta_i(x) > r - \frac{1}{2}$, which coincides with Bayes predictor from Eq. (6.11). □

6.3.3 Difficulty of consistency in the threshold-based setting

Although the threshold-based setting is of great practical importance, no consistency results exist for these surrogates to the best of our knowledge.

The difficulty of proving such results stems from the fact that within the space of allowed decision functions Eq. (6.7) is no longer valid. This implies that it is no longer possible to obtain the optimal decision function from the minimization at a fixed $x \in \mathcal{X}$, as we have done in the proof of Theorem 2 and 3.

In section 6.2, we have defined the decision function $\mathbf{g}(x) = (g_1(x), \ldots, g_{k-1}(x))$ to be of the form $g_i(x) = \theta_i - f(x)$, or equivalently to verify the condition that $g_{i+1}(x) - g_i(x)$ is a positive constant (i.e. does not depend on x) for all $1 \leq i < k - 1$. If $\mathbf{g}$ verifies this constraint, we will say that $\mathbf{g}$ is a *threshold-based decision function.*

In order to obtain sufficient conditions for the consistency of threshold-based methods, we will first consider the case in which the decision function $\mathbf{g}$ belongs to the space of all measurable functions. In this case we can construct the optimal decision function by considering each $x \in \mathcal{X}$ separately. Having an explicit form of the minimizer for the ψ-risk in this setting makes it possible to inspect under which conditions does this minimizer belong to the space of threshold-based decision functions.

An interesting relaxation of the threshold-based setting is given in [Peterson and Harrell, 1990] under the name of *partial thresholds.* In this setting, $\mathbf{g}(x) = (g_1(x), \ldots, g_k(x))$ is a non-decreasing vector for all $x \in \mathcal{X}$

which does not necessarily verify the constraints of a threshold-based decision function. In this setting, the decision function can represent any real-valued mapping that verifies the order constraints. We will call these decision functions *partial-threshold decision functions*. This setting is rarely used in practice because of the need to estimate $k-1$ functions.

6.3.4 Consistency of proportional odds

We begin by proving the strong convexity of proportional odds, whose proof can be found in the appendix. Through this section we will use ψ_C to denote the proportional odds surrogate as defined in Eq. (6.5).

Lemma 2. *The proportional odds surrogate ψ_C is a convex function of its arguments in the domain of definition.*

Proof. $\psi_C(1,\alpha)$ and $\psi_C(k,\alpha)$ (defined in Eq. (6.5)) are logistic loss functions, which are convex because they are log-sum-exp functions. We will prove that ψ_i is convex for $1 < i < K$. For convenience we will write this function as $f(a,b) = -\log\left(\frac{1}{1+\exp(a)} - \frac{1}{1+\exp(b)}\right)$, where $a > b$ is the domain of definition.

By factorizing the fraction inside f to a common denominator, f can equivalently be written as $-\log(\exp(a)-\exp(b)) + \log(1+\exp(a)) + \log(1+\exp(b))$. The last two terms are convex because they can be written as a log-sum-exp. The convexity of the first term, or equivalently the log-concavity of the function $f(a,b) = \exp(a)-\exp(b)$ can be settled by proving the positive-definiteness of the matrix $Q = \nabla f(a,b)\nabla f(a,b)^T - f(a,b)\nabla^2 f(a,b)$ for all (a,b) in the domain $\{b > a\}$ [Boyd and Vandenberghe, 2004]. In our case,

$$Q = \begin{pmatrix} \exp(a+b) & -\exp(a+b) \\ -\exp(a+b) & \exp(a+b) \end{pmatrix} = \exp(a+b)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Which is a positive semidefinite matrix with eigenvalues $2\exp(a+b)$ and 0. This proves that Q is positive semidefinite and thus the loss function ψ_i is convex. □

For the proportional odds surrogate ψ_C it is possible to find the explicit form of a function that minimizes the ψ_C-risk. We will use notation $\mathbf{g}$ to denote the vector-valued function $(g_1(x),\dots,g_{k-1}(x))$.

Theorem 4. *The function $\mathbf{g}: \mathcal{X} \to \mathbb{R}^{k-1}$ given by*

$$g_i^*(x) = \log\left(\frac{u_i(x)}{1-u_i(x)}\right) \quad ,$$

minimizes the ψ_C-risk.

Proof. Let $x \in \mathcal{X}$ and consider the optimization problem

$$\boldsymbol{\alpha}^* \in \underset{\boldsymbol{\alpha}\in\mathbb{R}^{k-1}}{\arg\min}\, \mathbb{E}_Y(\psi_C(Y,\boldsymbol{\alpha})|X=x)$$

The KKT conditions associated with this optimization problem are

$$
\begin{aligned}
&-\eta_1(x)\frac{1}{\sigma(\alpha_1)} + \eta_2(x)\frac{1}{\sigma(\alpha_2)-\sigma(\alpha_1)} = 0\\
&-\eta_i(x)\frac{1}{\sigma(\alpha_i)-\sigma(\alpha_{i-1})} + \eta_{i+1}(x)\frac{1}{\sigma(\alpha_{i+1})-\sigma(\alpha_i)} = 0\\
&-\eta_{k-1}(x)\frac{1}{\sigma(\alpha_{k-1})-\sigma(\alpha_{k-2})} + \eta_k(x)\frac{1}{1-\sigma(\alpha_{k-1})} = 0
\end{aligned}
$$

with $1 < i < k-1$. It is easy to verify that $\sigma(\alpha_i^*) = \sum_{j=1}^{i}\eta_j(x) = u_i(x)$ satisfy the optimality conditions.

Solving for α_i^* results in $\sigma(\alpha_i^*) = \sum_{j=1}^{i}\eta_j(x) \implies \alpha_i^* = \log\left(u_i(x)/(1-u_i(x))\right)$. By Eq. (6.7), the function that for all $x \in \mathcal{X}$ returns $\log\left(u_i(x)/(1-u_i(x))\right)$ is the function that minimizes the ψ-risk. □

Note that for $x \in \mathcal{X}$ fixed, the sequence $(g_1^*(x), \ldots, g_{k-1}^*(x))$, with g^* as defined in the previous theorem is non-decreasing since u_i is non-decreasing and due to the monotonicity of the logit function. This implies that $\mathbf{g}(x) = (g_1^*(x), \ldots, g_{k-1}^*(x))$ is a partial-threshold decision function. Consistency for this class of functions is now immediate since

$$
\begin{aligned}
\text{pred}(\mathbf{g}^*(x)) &= 1 + \sum_{i=1}^{k-1}\mathcal{H}\left(\log\left(\frac{u_i(x)}{1-u_i(x)}\right)\right) =\\
&= 1 + \sum_{i=1}^{k-1}\mathcal{H}\left(u_i(x) - 1/2\right)\\
&= \min_{r\in[k]}\{r : \sum_{i=1}^{r}\eta_i(x) \geq \frac{1}{2}\}
\end{aligned}
\tag{6.12}
$$

which coincides with Bayes predictor from Eq. (6.10). Thus, if the decision function belongs to the space of partial-threshold decision functions, the proportional odds is consistent. For threshold-based decision functions we have the following result:

Corollary 3. *Let P verify the property that the* odds-ratio *is constant, that is,*

$$
\frac{\eta_i(x)/(1-\eta_i(x))}{\eta_{i+1}(x)/(1-\eta_{i+1}(x))} \tag{6.13}
$$

is independent of $x \in \mathcal{X}$ for all $i \in [k-1]$. Then the proportional odds surrogate is consistent.

Proof. Let $g_i(x) = \log\left(u_i(x)/(1-u_i(x))\right)$ and $g_{i+1}(x) = \log\left(u_{i+1}(x)/(1-u_{i+1}(x))\right)$. Proving is that $g_i(x) - g_{i+1}(x)$ is constant is equivalent to proving that g is of the form $g_i(x) = \theta_i - f(x)$

Then

$$
\begin{aligned}
g_i(x) - g_{i+1}(x) &= \log\left(u_i(x)/(1-u_i(x))\right) -\\
&\quad \log\left(u_{i+1}(x)/(1-u_{i+1}(x))\right) =\\
&\quad \log\left(\frac{\eta_i(x)/(1-\eta_i(x))}{\eta_{i+1}(x)/(1-\eta_{i+1}(x))}\right)
\end{aligned}
$$

which is the log of a constant by assumption, hence constant. By Theorem 4 it follows that this function is the minimizer of the ψ_C-risk. Consistency is now a consequence of (6.12). □

6.3.5 Consistency of margin-based models

As done in the previous section, we will provide an explicit form of functions that minimize the $\psi^{\ell}_{\mathcal{M}}$-risk. This will allow to derive conditions under which threshold-based decision functions are consistent.

Theorem 5. *Let ℓ be V-shaped. Then the function $\mathrm{g} : \mathcal{X} \to \mathbb{R}^{k-1}$ minimizes the $\psi^{\ell}_{\mathcal{M}}$-risk for different values of ϕ:*

- *If ϕ is the hinge loss, i.e., $\phi(t) = \max(1-t, 0)$,*

$$g_i^*(x) = \operatorname{sign}(u_i(x) - v_i(x))$$

- *If ϕ is the logistic loss, i.e., $\phi(t) = 1/(1+\exp(-t))$,*

$$g_i^*(x) = \log(u_i(x)/v_i(x))$$

- *If ϕ is the exponential loss, i.e., $\phi(t) = \exp(-t)$*

$$g_i^*(x) = \frac{1}{2}\log(u_i(x)/v_i(x))$$

- *If ϕ is the squared loss, i.e., $\phi(t) = (1-t)^2$*

$$g_i^*(x) = \frac{u_i(x) + v_i(x)}{u_i(x) - v_i(x)}$$

Proof. Let u_i, v_i be as defined in Eq. (6.8), $x \in \mathcal{X}$ and $\boldsymbol{\alpha} = (g_1(x), \ldots, g_{k-1}(x))$. Then for any surrogate ψ we can write

$$\begin{aligned} &\mathbb{E}_Y(\psi(Y, \boldsymbol{\alpha})|X=x) = \\ &\sum_{j=1}^{k} \eta_j(x) \left(\sum_{i=1}^{j-1} \Delta\ell(y,i)\phi(\alpha_i) - \sum_{i=j}^{k-1} \Delta\ell(y,i)\phi(-\alpha_i) \right) = \\ &\sum_{i=1}^{k-1} \phi(\alpha_i)v_i(x) + \phi(-\alpha_i)u_i(x) \quad . \end{aligned} \tag{6.14}$$

If ϕ is the hinge loss, the values of α_i that minimize this expression verify $-1 \leq \alpha_i \leq 1$ for all $i \in [k-1]$, as otherwise truncation of these values at -1 or 1 gives a lower value of the surrogate loss. In this case we have

$$\begin{aligned} &\mathbb{E}_Y(\psi(Y, \boldsymbol{\alpha})|X=x) = \\ &\sum_{i=1}^{k-1} (1-\alpha_i)v_i(x) + (1+\alpha_i)u_i(x) = \\ &\sum_{i=1}^{k-1} \alpha(u_i(x) - v_i(x)) + C \end{aligned}$$

where C are terms that do not depend on α. Therefore, this expression minimized for $\alpha_i^* = \operatorname{sign}(v_i(x) - u_i(x))$.

If ϕ is the logistic loss, the expression from Eq. (6.8) is differentiable. The derivative with respect to α_i is $(1-\sigma(\alpha_i))v_i - \sigma(\alpha_i)u_i$, where $\sigma(\alpha_i) = 1/(1+\exp(-\alpha_i))$ is the sigmoid function. Equaling this expression to zero and solving for α_i yields the result.

The proof for ψ the rest of surrogates can be found in the appendix. □

In light of this result, it is possible to derive sufficient conditions under which margin-based decision functions are consistent.

Corollary 4. *Under the conditions of Theorem 5, if P is a probability distribution such that*

$$\alpha_i^*(x) - \alpha_{i+1}^*(x)$$

does not depend on x for all $1 \leq i < k$, then the surrogate $\psi_{\mathcal{M}}^{\ell}$ is consistent.

Proof. The optimal decision functions $\alpha_1^*, \ldots, \alpha_{k-1}^*$ are threshold-based decision functions by assumption. Furthermore, it is easy to verify that all the $\alpha_i^*(x)$ obtained in Theorem 5 verify $\mathcal{H}(\alpha_i^*(x)) = \mathcal{H}(u_i(x) - v_i(x))$, and thus prediction coincides with Bayes predictor of Eq. (6.9). □

As mentioned in Section 2, the surrogate $\psi_{\mathcal{M}}^{\ell}$ parametrizes several approaches that have appeared in the literature.

Corollary 5. *Under the assumptions of Corollary 4, the following surrogate loss functions are consistent with respect to the zero-one loss:*

- *Support Vector Ordinal Regression (SVOR), "explicit constraints" variant, from [Shashua and Levin, 2003, Chu and Keerthi, 2007] (ϕ = hinge loss),*
- *"Immediate threshold" from [Rennie and Srebro, 2005] (ϕ = logistic loss),*
- *"ORBoost with Left-Right margins" from [Lin and Li, 2006] (ϕ = exponential loss),*

Corollary 6. *Under the assumptions of Corollary 4, the following surrogate loss functions are consistent with respect to the mean absolute error:*

- *Support Vector Ordinal Regression (SVOR), "implicit constraints" variant, from [Shashua and Levin, 2003, Chu and Keerthi, 2007] (ϕ = hinge loss),*
- *"All threshold" from [Rennie and Srebro, 2005] (ϕ = logistic loss), used in Chapter 4 in the context of fMRI decoding models.*
- *"ORBoost with All margins" from [Lin and Li, 2006] (ϕ = exponential loss).*

We now have the necessary elements to answer the question that motivated our study at the beginning of this chapter: why does the SVOR with implicit thresholds often outperform the SVOR with explicit thresholds with respect to the mean absolute error metric? As the corollaries above outline, the SVOR with implicit constraints surrogate is consistent with respect to the absolute error loss while the SVOR with explicit constraints is not. Even more, SVOR with explicit constraints surrogate is instead consistent with respect to a different loss (the 0-1 loss). It is thus not surprising that the first approach performs better with respect to the absolute error. In the experimental section we discuss this issue further by comparing both methods with respect to different metrics.

The sufficient conditions of Corollary 4 translate into well-known conditions on the probability distribution for some values of ϕ. For example, let ϕ be the logistic loss and ℓ be the absolute error, the optimal decision function is given by $\alpha_i^*(x) = \log(u_i(x)/v_i(x)) = \log(u_i(x)/(1 - u_i(x)))$. Thus

we obtain the function that the optimal decision function for the proportional odds from Theorem 4. This implies (see Corollary 3) that if P verifies that the odds-ratio are constant as defined in (6.13), then the surrogate is consistent.

In light of these results, it is immediate to show that within the space of partial threshold decision functions, the aforementioned methods are consistent. Furthermore, in this case we can prove a slightly more general result. The following result states consistency while assuming only convexity and a condition of the differential at zero of the function ϕ.

Theorem 6. *Let ℓ be a V-shaped loss function. Given a convex function $\phi : \mathbb{R} \to \mathbb{R}_+$ such that ϕ is differentiable at zero and $\phi'(0) < 0$, then the surrogate loss function $\psi^\ell_{\mathcal{M}}$ is consistent with respect to ℓ if the decision functions are contains the space of partial-threshold decision functions .*

Proof. Let $x \in \mathcal{X}$ and $r = h^*(x)$ be the label predicted by Bayes predictor. As we did in the proof of Theorem 5, we can write $\mathbb{E}_Y(\psi(y, \boldsymbol{\alpha})|X = x) = \sum_{i=1}^{k-1} \phi(\alpha_i)v_i(x) + \phi(-\alpha_i)u_i(x)$. The KKT conditions for this optimization problem with respect to $\boldsymbol{\alpha}$ are

$$0 \in \partial F_i(\alpha_i) = \partial\phi(\alpha_i)u_i(x) - \partial\phi(-\alpha_i)v_i(x) \quad \forall i = 1, \ldots, k-1 \tag{6.15}$$

where ∂ denotes the subgradient operator.

By the V-shape property we have that $(v_1(x) - u_1(x), v_2(x) - u_2(x), \ldots, v_k(x) - u_k(x))$ is a non-increasing sequence of i. Hence, $1 + \sum_{i=1}^{k-1} \mathcal{H}(v_i(x) - u_i(x)) = r$ implies that $(v_i - u_i) \geq 0$ for $1 \leq i < r$ and $(v_i - u_i) < 0$ for $i \geq r$.

Let $\boldsymbol{\alpha}^*$ denote the vector that satisfies the KKT conditions and let $p \geq r$. Then have $u_p - v_p > 0$. Hence, $\partial F_p(0) = \phi'(0)(u_p - v_p) \leq 0$. The expression $\partial F_p(\alpha_p)$ is the subdifferential of a convex function and is thus a monotone operator [Rockafellar, 1970]. Hence $\partial F_p(0) < 0$ implies that the optimal value α_p^* will be located in the region $\{x : x > 0\}$. We have proved that $\alpha_p^* > 0$ for all $p \geq r$.

Suppose now $s < r$ and consider $\partial F_s(0) = \phi'(0)(u_s - v_s)$. Because of the V-shape property we have $u_s - v_s \leq 0$ and hence $\partial F_s(0) \geq 0$. The expression $\partial F_s(\alpha_s)$ is again a monotone operator and verifies $\partial F_s(0) \geq 0$ from where we can conclude that any zero of this expression will be located in the region $\{x : x \leq 0\}$. We have proved that $\alpha_s^* \leq 0$ for all $s < r$.

We have proved that $\alpha_p^* > 0$ for all $p \geq r$ and $\alpha_s^* \leq 0$ for all $s < r$. Hence, $1 + \sum_{i=1}^{k-1} \mathcal{H}(-\alpha_i) = 1 + (r-1) = r$ and the prediction coincides with Bayes predictor, hence the surrogate is consistent. □

6.3.6 Relationship with multiclass formulations

Let $\psi^\ell_{\mathcal{L}}$ the surrogate loss function defined in Eq. (6.6). For a given $x \in \mathcal{X}$, let $f_1^*(x), \ldots, f_k^*(x)$ be minimizers of $\mathbb{E}_Y(\psi^\ell_{\mathcal{L}}(Y, f(x)))$. Then it is verified

$$\sum_{i=1}^{k} \left(\sum_{j=1}^{k} \eta_j(x)\ell(j, i) \right) \psi(-f_i(x)) =$$

$$\sum_{i=1}^{k} (u_i(x) - v_i(x))\psi(-f_i(x))$$

For the hinge loss, it is shown in Lee et al. [2004] that given $x \in \mathcal{X}$, the optimal decision function is of the form $f_i^*(x) = 1$ for $i = \arg\min_i u_i(x) - v_i(x)$, and $-1/(k-1)$ otherwise. Thus, a sufficient condition for consistency is that the k functions above are in the class of functions we are considering for the decision function.

This is to be contrasted with the margin-based formulations, where, for the hinge surrogate, we need the $k-1$ functions $\text{sign}(u_i(x) - v_i(x)))$ to be in the class of functions of the decision function.

No requirement is stronger than the other. However, for the margin-based formulations, we have developed sufficient conditions under which we may use a single function and fixed thresholds.

6.4 Experiments

Although the focus of this chapter is a theoretical investigation of consistency, we have also conducted experiments that study empirical performance of some the methods outlined in this paper.

In this section we compare two approaches described earlier in terms of generalization accuracy. The different datasets used are described in [Chu and Ghahramani, 2004]. Following [Chu and Keerthi, 2005], we will consider two variants of the margin-based loss function ψ_C for $\ell = \ell_{0-1}$ and $\ell = \ell_{\mathcal{A}}$ with ϕ = hinge loss. Specifically, we compare the "explicit constraints on thresholds" formulation (denoted here ET) versus the "implicit constraints on thresholds" formulation (denoted IT). Corollary 4 states that under appropriate assumptions on the probability distribution P, ET is consistent with respect to the zero-one loss while AT is consistent with respect to the absolute error loss.

We show in Figure 6.4 the generalization scores of these two methods using as metric the zero-one loss and the absolute error on 8 different datasets. The generalization accuracy of both models has been computed using 5-fold cross validation. Although consistency results only apply under certain assumptions on the underlying probability distribution, we observe a correlation between consistent surrogates and the best performing model. Our findings provide a theoretical explanation of the poor performance of the ET surrogate compared with the IT surrogate when evaluated using the absolute error loss (since the IT surrogate is consistent w.r.t the absolute error). Similar results have been observed in the literature for different values of ϕ [Chu and Keerthi, 2005, Lin and Li, 2006, Shashua and Levin, 2003].

6.5 Conclusion

In this chapter we have characterized the consistency for a rich family of surrogate loss functions used for ordinal regression. In the regression-based setting we have extended work from Ramaswamy and Agarwal [2012] to prove consistency for the absolute error surrogate as well as the squared error surrogate.

In the threshold-based setting, we studied consistency of the proportional odds model and given sufficient conditions on the underlying probability distribution under which this surrogate is consistent. We also consid-

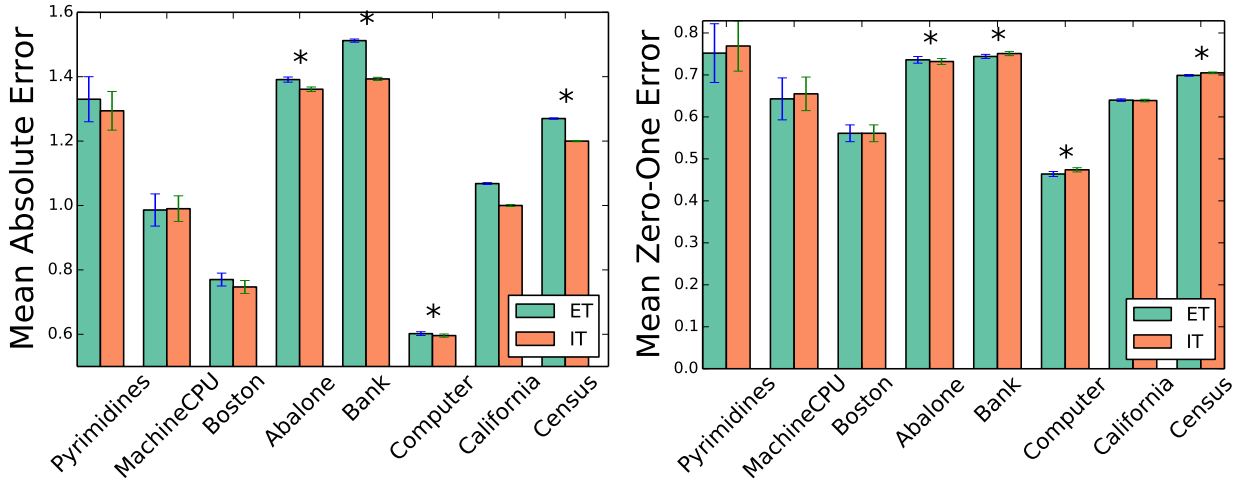


Figure 6.4: Performance of the "Explicit Threshold" (ET) and "Implicit Threshold" (IT) methods of Chu and Keerthi [2005] on 8 different datasets and for two different evaluation metrics. Top: the metric used is the mean absolute error. The IT method is consistent with respect to this loss and performs better on 7 out of 8 datasets. Bottom: the metric used is the mean zero-one loss. The ET method is consistent with respect to this metric and performs better on 6 out of 8 datasets. Datasets for which the difference of performance is significant (Wilcoxon signed-rank test with $p < 0.01$) are denoted with an asterisk ($*$).

ered formulations such as the Support Vector Ordinal Regression [Chu and Keerthi, 2005], the Ordinal Regression Boosting methods [Lin and Li, 2006] and the Logistic Regression formulation of [Rennie and Srebro, 2005]. We framed these methods under a common formulation that we call *margin-based surrogate*, and derived an explicit form of functions that minimize the ψ-risk. We gave sufficient conditions for the consistency of the aforementioned approaches. Thanks to these results, we could answer the question outlined in the introduction: the SVOR with implicit constraints surrogate is consistent with respect to the absolute error loss while the SVOR with explicit constraints is not. Even more, SVOR with explicit constraints surrogate is instead consistent with respect to a different loss (the 0-1 loss). This would explain why it has been repeatedly reported the superiority of the first approach when compared with the absolute error metric [Chu and Keerthi, 2005, Lin and Li, 2006, Rennie and Srebro, 2005]. In this respect, the importance of this work, more than to prove consistency, it to *identify the loss for which a given surrogate is consistent.*

Since consistency of the threshold-based approach is only proven subject to certain conditions on the underlying probability distribution P, we investigated under which conditions these surrogates are always consistent. Here we show that this is possible by considering an enlarged space for the decision functions that we called partial-threshold decision functions.

Finally, we illustrated our findings on by comparing the performance of two methods on 8 different datasets. Although the conditions for consistency that are required by the underlying probability distribution are not necessarily met, we observed that methods that are consistent w.r.t a given loss often outperform other methods that are not consistent with respect to that loss.

Bibliography

Cande V. Ananth and David G. Kleinbaum. Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology*, 26(6):1323–1333, 1997.

Ben G. Armstrong and Margaret Sloan. Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, 129(1):191–204, 1989.

Peter L. Bartlett, Michael I. Jordan, and Jon D. Mcauliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2003.

Shai Ben-David, Nadav Eiron, and Philip M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Wei Chu and Zoubin Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1–24, 2004.

Wei Chu and S Sathiya Keerthi. New Approaches to Support Vector Ordinal Regression. In *Proceedings of the 22th International Conference on Machine Learning (ICML)*, 2005.

Wei Chu and S Sathiya Keerthi. Support Vector Ordinal Regression. *Neural computation*, 815(2001):792–815, 2007.

Koby Crammer and Yoram Singer. Pranking with ranking. In *Advances in Neural Information Processing Systems 14*, 2001.

Orla M. Doyle, John Ashburner, F.O. Zelaya, Stephen C.R. Williams, Mitul A. Mehta, and Andre F. Marquand. Multivariate decoding of brain images using ordinal regression. *NeuroImage*, 81:347–357, 2013.

Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012.

William H. Greene. Econometric analysis, 1997, 1997.

Craig T. Hartrick, Juliann P. Kovan, and Sharon Shapiro. The numeric rating scale for clinical pain measurement: A ratio measure? *Pain Practice*, 3(4):310–316, 2003. ISSN 1533-2500.

Stefan Kramer, Gerhard Widmer, Bernhard Pfahringer, and Michael De Groeve. Prediction of ordinal classes using regression trees. *Fundamenta Informaticae*, 47(1):1–13, 2001.

Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99 (465):67–81, 2004.

Ling Li and Hsuan-tien Lin. Ordinal Regression by Extended Binary Classification. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2007.

Hsuan-Tien Lin and Ling Li. Large-margin thresholded ensembles for ordinal regression: Theory and practice. In *Algorithmic Learning Theory*, pages 319–333. Springer, 2006.

Yi Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68(1):73 – 82, 2004.

Peter McCullagh. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society*, 42(2):109–142, 1980.

Deirdre B. O'Brien, Maya R. Gupta, and Robert M. Gray. Cost-sensitive multi-class classification from probability estimates. In *Proceedings of the 25th international conference on Machine learning*, pages 712–719. ACM, 2008.

Bercedis Peterson and Frank E. Harrell. Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society*, 39:205–217, 1990.

Harish G. Ramaswamy and Shivani Agarwal. Classification Calibration Dimension for General Multiclass Losses. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2012.

Jason D M Rennie and Nathan Srebro. Loss Functions for Preference Levels : Regression with Discrete Ordered Labels. In *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, 2005.

Ralph Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33(1): 209–216, 1970.

Amnon Shashua and Anat Levin. Ranking with large margin principle : Two approaches. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2003.

Ingo Steinwart. Support Vector Machines are Universally Consistent. *Journal of Complexity*, 18(3):768–791, September 2002.

Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

Ambuj Tewari and Peter L. Bartlett. On the Consistency of Multiclass Classification Methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

Tong Zhang. Statistical Behavior and Consistency of Classification Methods based on Convex Risk Minimization. *The Annals of Statistics*, 32:56–85, 2004.

7 Conclusion and Perspectives

In this last chapter we detail the different contributions contained within this thesis and we point out possible extension that can be considered in the future. The proposed methods span all the different contributions of this thesis. We also enumerate the software packages that have been developed.

Contents

7.1 Contributions

In this thesis we have examined several aspects of the pipeline through which an fMRI datasets can be analyzed. We have made contributions at different stages of this pipeline.

In Chapter 4 we have studied a problem of *feature extraction*. The goal of this feature extraction step is to output time-independent activation maps from the BOLD time series. In this context we have introduced a new model for the joint estimation of hemodynamic response function (HRF) and brain activation coefficient. The novelty of our method stems from the observation that the formulation of the GLM model with a common (but unknown) HRF across conditions translates into a rank constraint on the vector of estimates. This allows to specify the model as a smooth optimization problem and to use gradient-based methods for its estimation.

A popular application of supervised learning to reveal cognitive mechanisms in fMRI studies is the problem of *brain decoding*, in which the goal is to predict some information about the stimuli given the activation coefficients. In Chapter 5 we examine the setting of practical importance in which the target variable consist of discretely ordered values. We identified two loss functions that are appropriate for the task: the absolute error and the pairwise disagreement. We presented several models based on the minimization of a convex surrogate of these loss functions. We examined their performance on both synthetic and two real world fMRI datasets.

Motivated by its applicability to decoding studies we turned in Chapter 6 to study some theoretical properties of *ordinal regression* models. We provided an analysis of the Fisher consistency properties of a rich family of surrogate loss functions, including proportional odds and support vector ordinal regression. For all the surrogates considered, we either proved consistency or provided sufficient conditions under which these approaches are consistent.

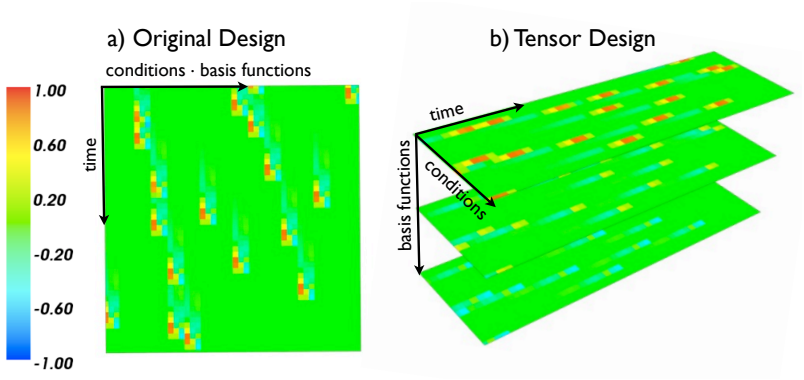
7.2 Research Perspectives

7.2.1 A tensor formulation of R1-GLM

Although the R1-GLM model presented in Chapter 4 has faster execution times than methods that implement similar assumptions [Makni et al., 2008, Vincent et al., 2010, Degras and Lindquist, 2014]), the algorithm still does not use all the structure within the problem.

For example, the algorithm fits independently a R1-GLM on every voxel (for around 5×10^4 voxels in a fMRI volume) without taking into account that *the design matrix is the same for all voxels*. In this context, a possible line of research is to use a *tensor-based formulation* of the R1-GLM model to incorporate this structure within the solver.

Let $\mathbf{X} \in \mathbb{R}^{n \times kd}$ be the design matrix of the GLM. This matrix can be naturally represented as the tensor $\mathcal{X} \in \mathbb{R}^{n \times k \times d}$ that verifies that its matrizialization along the last axis corresponds to the design matrix $\mathbf{X}$. In this case, using the n-mode tensor product [Kolda and Bader, 2009] we have the identity $\mathbf{X}\text{vec}(\mathbf{h}\boldsymbol{\beta}^T) = \mathcal{X} \times_2 \mathbf{h} \times_3 \boldsymbol{\beta}$, hence the R1-GLM can be seen as the



minimization of the objective function

$$\|\mathbf{y} - \mathcal{X} \times_2 \mathbf{h} \times_3 \boldsymbol{\beta}\|^2$$

subject to the usual constraints on $\mathbf{h}, \boldsymbol{\beta}$. Due to its analogies with a linear least squares problem, it seems reasonable to think that a tensor factorization of $\mathcal{X}$ (such as CP/PARAFAC or Tucker) might be able to solve or accelerate the optimization of this model.

7.2.2 Uniqueness of solution for R1-GLM

The R1-GLM model, being non-convex, comes with no guarantees of convergence to a global optimum for the algorithms considered.

However, some scarce theoretical results exist. For the case of infinite data (which would correspond in our model to an infinite number of fMRI scans), the uniqueness of solution of a similar model was proved by [Bai and Liu, 2006]. A possible area of research is to extend these results to the more practical setting of a limited number of samples.

7.2.3 Parcel R1-GLM model

This is a different extension of the R1-GLM model presented in Chapter 4 that aims at reducing the amount of HRFs estimated within the model.

Within the context of HRF estimation, some studies have proposed to perform the estimation of an HRF in a set of neighboring voxels called a *parcel*, thus taking advantage of the spatially dependent nature of fMRI [Wang et al., 2013, Chaari et al., 2012, Badillo et al., 2013].

The notion of parcel, i.e., a brain region which shares the same HRF, can be trivially incorporated into the R1-GLM model, and results in a modified R1-GLM model. Given m voxels in the parcel, let $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m]$ the concatenation of the BOLD signal for the voxels within the parcel and let $\mathbf{X}$ be the matrix formed by a block-diagonal matrix with m blocks in which every block is the design matrix for the current experiment. Then the R1-GLM model that assumes the HRF constant across the parcel can be written as

$$\begin{aligned} \hat{\mathbf{h}},\ \hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\omega}} \ &= \ \underset{\mathbf{h},\boldsymbol{\beta},\boldsymbol{\omega}}{\arg\min}\ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\,\mathrm{vec}(\mathbf{h}\boldsymbol{\beta}) - \mathbf{Z}\boldsymbol{\omega}\|^2 \\ &\text{subject to } \|\mathbf{B}\mathbf{h}\|_\infty = 1 \text{ and } \langle \mathbf{B}\mathbf{h}, \mathbf{h}_{\text{ref}} \rangle > 0 \ , \end{aligned} \tag{7.1}$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m]$ contains the activation coefficients for the different voxels within the region. Phrased differently, the estimation of a R1-GLM model within a parcel is itself a R1-GLM model with a modified design matrix.

However, these approaches must face the problem of choosing the right brain parcellation. An interesting approach, named hemodynamically-informed parcellations [Chaari et al., 2012, Badillo et al., 2013] relies on the computation of a large number of estimations at the voxel or sub-parcel level.

7.2.4 Weaker conditions for the consistency of threshold-based ordinal regression methods

In Chapter 6 we have presented consistency results for some ordinal regression methods. For threshold-based methods, in the practical setting in which the thresholds are constant across samples (a setting that we called model with *threshold-based decision function*), we have only been able to prove consistency under very restrictive conditions on the underlying probability distribution.

It is possible that similar consistency results can be obtained with weaker conditions on the probability distribution, which would result in conditions that are widely applicable. For example, in [Herbrich et al., 1999, Section 2], the authors present the cumulative models of [McCullagh, 1980] (described in Section 6.2) as a consequence of a stochastic ordering in the sample space.

The stochastic ordering assumption can be described as follows. Given the sample space $\mathcal{X}$ and a target space $\mathcal{Y}$, then for all different $x_1, x_2 \in \mathcal{X}$ either

$$P(y \leq r|X = x_1) \geq P(y \leq r|X = x_2) \text{ for all } r \in \mathcal{Y}$$

or

$$P(y \leq r|X = x_1) \leq P(y \leq r|X = x_2) \text{ for all } r \in \mathcal{Y}$$

The authors then conclude that stochastic ordering is satisfied by a model of the form

$$g^{-1}(P(y \leq r|X = x)) = \theta_r - f(x)$$

hence it seems reasonable to think that a stochastic ordering could be a sufficient condition in order to obtain consistency – at least for the cumulative logit model.

7.2.5 Application of ordinal regression methods to multiclass classification

Although ordinal regression methods have been initially developed for loss functions that minimize a distance between the labels, our theoretical results show that some ordinal regression methods are instead consistent to the 0-1 loss[1], i.e., with the usual loss used in multiclass classification. This suggests that some ordinal regression methods might be competitive in the context of multiclass classification. One of the advantages of ordinal regression models is that for linear decision functions, the learning only requires the estimation of $p + k - 1$ parameters versus e.g. $p \times (k - 1)$ in the case of one-vs-all multiclass classification, where p is the dimensionality of the dataset and k is the number of classes. It is possible that these methods have

[1] We recall that (although in a degenerate sense), the 0-1 loss does verify the V-shape property.

applications for the estimation of multiclass classification rules in very high-dimensional settings, although its usefulness still needs to be determined.

7.3 Publications by the author

CHAPTER III

- V. Borghesani, **F. Pedregosa**, E. Eger, M. Buiatti, and M. Piazza, *"A perceptual-to-conceptual gradient of word coding along the ventral path"* Proceedings of the 4th International Workshop on Pattern Recognition in Neuroimaging, 2014.

CHAPTER IV

- **F. Pedregosa**, M. Eickenberg, P. Ciuciu, B. Thirion, and 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. 3nd Int. Work. Pattern Recognit. NeuroImaging, 2013.

CHAPTER V

- **F. Pedregosa**, E. Cauvet, G. Varoquaux, C. Pallier, B. Thirion, and A. Gramfort, *"Learning to rank from medical imaging data"*, in Proceedings of the 3rd International Workshop on Machine Learning in Medical Imaging, 2012.
- **F. Pedregosa**, E. Cauvet, G. Varoquaux, C. Pallier, B. Thirion, and A. Gramfort, *"Improved brain pattern recovery through ranking approaches"*. 2nd International Workshop on Pattern Recognition in NeuroImaging, Jul 2012
- Y. Bekhti, N. Zilber, **F. Pedregosa**, P. Ciuciu, V. Van Wassenhove, and A. Gramfort, *"Decoding perceptual thresholds from MEG/EEG"*. Pattern Recoginition in Neuroimaging (PRNI) (2014)

CHAPTER VI

- **F. Pedregosa**, F. Bach, and A. Gramfort, *"On the Consistency of Ordinal Regression Methods"*.

OTHER PUBLICATIONS

- L. Buitinck, G. Louppe, M. Blondel, **F. Pedregosa**, A. Mueller, et al.. *"API design for machine learning software: experiences from the scikit-learn project"*. European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases, 2013.
- M. Eickenberg, **F. Pedregosa**, S. Mehdi, A. Gramfort, B. Thirion. *"Second order scattering descriptors predict fMRI activity due to visual textures"*. 3rd International Workshop on Pattern Recognition in NeuroImaging, 2013.
- A. Abraham, **F. Pedregosa**, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, B. Thirion and G. Varoquaux (2014). *"Machine learning for neuroimaging with scikit-learn"*. Frontiers in neuroinformatics, 8.
- F. Yepes-Calderon, **F. Pedregosa**, F., Thirion, B., Wang, Y., and Lepore, N. (2014, March). Automatic pathology classification using a single feature machine learning support-vector machines. In SPIE Medical Imaging (pp. 903524-903524). International Society for Optics and Photonics.

7.4 Software

A number of software distributions have been developed within the context of this thesis.

7.4.1 hrf_estimation

This package implements method for the joint estimation of hemodynamic response function (HRF) and activation coefficients (aka beta-maps) from fMRI data presented in Chapter 4. Full documentation for this package, including an example IPython notebook can be found at

`http://pythonhosted.org/hrf_estimation/`

7.4.2 mord

Ordinal Regression algorithms. Module that implements the ordinal regression models used in Chapter 5. The code can be found at the URL

`https://github.com/fabianp/mord`

7.4.3 pysofia

PySofia is a python wrapper around the methods present in the C++ sofia-ml library. These include Stochastic Gradient Descent implementations of some ranking algorithms, notably RankSVM [Sculley, 2009].

`https://pypi.python.org/pypi/pysofia/`

7.4.4 memory_profiler

This is a python module for monitoring memory consumption of a process as well as line-by-line analysis of memory consumption for python programs. It is a pure python module.

`https://pypi.python.org/pypi/memory_profiler`

The presentation of this module won the *Best Poster Award* at the conference EuroScipy 2012 (European Scientific Computing in Python)

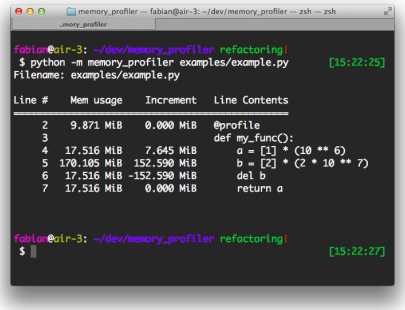


Figure 7.1: The memory_profiler module allows to quickly analyze the memory consumption of a program by using the line-by-line profiling (in the picture) or the time-based memory profiling.

Bibliography

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 2013.

Er-Wei Bai and Yun Liu. Least squares solutions of bilinear equations. *Systems & control letters*, 55(6):466–472, 2006.

Lofti Chaari, F. Forbes, T. Vincent, and Philippe 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.

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

Ralf Herbrich, Thore Graepel, Klaus Obermayer, and Fachbereich Informatik. Regression Models for Ordinal Data : A Machine Learning Approach. 1999.

Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, August 2009.

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.

Peter McCullagh. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society*, 42(2):109–142, 1980.

D. Sculley. Large scale learning to rank. In *NIPS 2009 Workshop on Advances in Ranking*, pages 1–6, 2009.

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.

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.

Glossary

activation coefficient amplitude for a single voxel associated with a stimuli in an fMRI study. 13, 15, 31, 52, 53, 63

BOLD fMRI contrast that measures oxygen change in blood flow. 30

conditions different stimuli in an fMRI study. 33

decoding distinguish patterns of neural activity associated with different stimuli or cognitive states. 45, 87

fMRI Functional Magnetic Resonance Imaging. 29

GLM General Linear Model. 17, 33, 44, 73, 74

Heaviside The real function that is zero for negative values and one otherwise. 47

hinge loss Loss function used by Support Vector Machines. 48

HRF Hemodynamic Response Function. 13, 31

Kendall τ Distance measure between two measurements. It is a measure of rank correlation, i.e., the similarity of the orderings of the data when ranked by each of the quantities. 93

LTI Linear Time Invariant assumption. 32

TR repetition time, sampling time in an fMRI scanner. 30

voxel unity of measure in a volumetric space. 30