Understanding the visual cortex by using classification techniques

Vincent Michel

▶ **To cite this version:**

Vincent Michel. Understanding the visual cortex by using classification techniques. Human-Computer Interaction [cs.HC]. Université Paris Sud - Paris XI, 2010. English. NNT : . tel-00550047

HAL Id: tel-00550047
https://theses.hal.science/tel-00550047

Submitted on 23 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

University Paris-Sud 11 - Faculty of Sciences
Graduate School of Informatics of Paris-Sud - EDIPS
INRIA Saclay - Parietal Team

PhD Thesis

submitted in partial fulfillment
of the requirements for the degree of
DOCTOR OF SCIENCE
Specialized in Computer Science

Understanding the visual cortex by using classification techniques

Vincent MICHEL

Advisors	Dr Bertrand Thirion	INRIA Saclay - Parietal team, Saclay, France
	Pr Gilles Celeux	INRIA Saclay - Select team, Saclay, France
	Dr Christine Keribin	Dept. of Mathematics , Uni. Paris-Sud 11, Orsay, France
Reviewers	Pr Polina Golland	MIT CSAIL, Boston, USA
	Pr Francis Bach	INRIA Rocquencourt - Willow team, Paris, France
Examiners	Dr Michèle Sebag	INRIA Saclay - TAO team, Saclay, France
	Dr Mathias Pessiglione	INSERM Unit 610, Paris, France

Université Paris-Sud 11 - Faculté de Sciences
Ecole Doctorale d'Informatique de Paris-Sud - EDIPS
INRIA Saclay - Equipe Parietal

Thèse de Doctorat

présentée le 15 décembre 2010 pour obtenir le grade de
Docteur en Sciences
Specialité Informatique

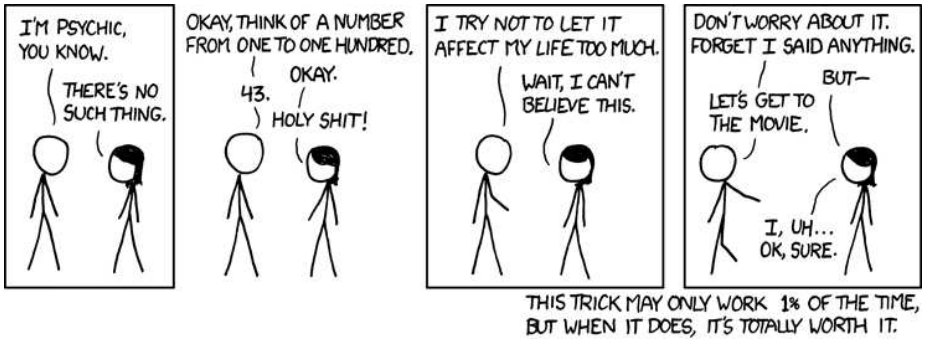
Améliorer la compréhension du cortex visuel à l'aide de techniques de classification

Vincent MICHEL

Directeurs	Dr Bertrand Thirion	INRIA Saclay - Equipe Parietal, Saclay, France
	Pr Gilles Celeux	INRIA Saclay - Equipe Select, Saclay, France
	Dr Christine Keribin	Dept. de Mathématiques , Uni. Paris-Sud 11, Orsay, France
Rapporteurs	Pr Polina Golland	MIT CSAIL, Boston, USA
	Pr Francis Bach	INRIA Rocquencourt - Equipe Willow, Paris, France
Examinateurs	Dr Michèle Sebag	INRIA Saclay - Equipe TAO, Saclay, France
	Dr Mathias Pessiglione	INSERM Unité 610, Paris, France

A mes parents.

Okay, brain. You don't like me, and I don't like you,
but let's get through this thing and then I can continue killing you with beer.
Homer Simpson.



http://xkcd.com/628/

Résumé

Dans ce mémoire, nous présentons différentes méthodes d'apprentissage statistique qui peuvent être utilisées pour comprendre le code neuronal des fonctions cognitives, en se basant sur des données d'Imagerie par Résonance Magnétique fonctionnelle du cerveau. Plus particulièrement, nous nous intéressons à l'étude de la localisation spatiale des entités impliquées dans le codage, et leur influence respective dans le processus cognitif. Dans cette étude, nous nous focalisons principalement sur l'étude du cortex visuel.

Dans la première partie de ce mémoire, nous introduisons les notions d'architecture fonctionnelle cérébrale, de codage neuronal et d'imagerie fonctionnelle. Nous étudions ensuite les limites de l'approche classique d'analyse des données d'IRMf pour l'étude du codage neuronal, et les différents avantages apportés par une méthode d'analyse récente, l'inférence inverse. Enfin, nous détaillons les méthodes d'apprentissage statistique utilisées dans le cadre de l'inférence inverse, et nous les évaluons sur un jeu de données réelles. Cette étude permet de mettre en évidence certaines limitations des approches classiquement utilisées, que cette thèse vise à résoudre. En particulier, nous nous intéressons à l'intégration de l'information sur la structure spatiale des données, au sein d'approches d'apprentissage statistique.

Dans la seconde partie de ce mémoire, nous décrivons les trois principales contributions de cette thèse. Tout d'abord, nous introduisons une approche Bayésienne pour la régularisation parcimonieuse, qui généralise au sein d'un même modèle plusieurs approches de références en régularisation Bayésienne. Ensuite nous proposons un algorithme de coalescence supervisé (*supervised clustering*) qui tient compte de l'information spatiale contenue dans les images fonctionnelles. Les cartes de poids résultantes sont facilement interprétables, et cette approche semble être bien adaptée au cas de l'inférence entre sujets. La dernière contribution de cette thèse vise à inclure l'information spatiale au sein d'un modèle de régularisation. Cette régularisation peut alors être utilisée dans un cadre de régression et de classification, et permet d'extraire des ensembles connexes de voxels prédictifs. Cette approche est particulièrement bien adaptée à l'étude de la localisation spatiale du codage neuronal, abordée durant cette thèse.

Mots clés :
codage neuronal, cortex visuel, neuroimagerie, Imagerie par Résonance Magnétique fonctionnelle (IRMf), inférence inverse, apprentissage statistique, information spatiale, clustering, méthode Bayésienne, régularisation parcimonieuse, Variation Totale.

Abstract

In this thesis, we present different approaches for statistical learning that can be used for studying the neural code of cognitive functions, based on brain functional Magnetic Resonance Imaging (fMRI) data. In particular, we study the spatial organization of the neural code, *i.e.* the spatial localization and the respective weights of the different entities implied in the neural coding. In this thesis, we focus on the visual cortex.

In the first part of this thesis, we introduce the notions of functional architecture, neural coding and functional imaging. Then, we study the limits of the classical approach for the characterization of the neural code from fMRI images, and the advantages of a recent method of analysis, namely inverse inference. Finally, we detail the statistical learning approaches used for inverse inference, and we evaluate them on real data. This study highlights the limitations of these approaches, that will be addressed during this thesis. In particular, we focus on the use of spatial information within statistical learning methods.

In a second part, we describe the three main contributions of this thesis. First, we introduce a Bayesian framework for sparse regularization, that generalizes two reference approaches. Then, we propose a supervised clustering method, that takes into account the spatial structure of the images. The resulting weighted maps are easily interpretable, and this approach seems particularly interesting in the case of inter-subjects inference. The last contribution of this thesis aims at including the spatial information into the regularization framework. This regularization is thus used in both regression and classification settings, and extracts clusters of predictive voxels. This approach is well suited for the decoding problem addressed in this thesis.

Keywords:
neural coding, visual cortex, neuroimaging, functional Magnetic Resonance Imaging (fMRI), inverse inference, statistical learning, spatial information, clustering, Bayesian approach, sparse regularization, Total Variation.

Acknowledgments

My first thought will be for Bertrand Thirion, who gave me this opportunity to take a look in the amazing world of science, and helping me during three years by a constant supervision. Also a great thank to Gilles Celeux for accepting to be the director of my thesis, and to Christine Keribin for co-supervision. I would also like to thank Polina Golland and Francis Bach for having spent their time on my thesis, and helped me to improve it with their comments. Also thank to Michèle Sebag and Mathias Pessiglione for accepting to examine my thesis.

A special thanks to my two "scientific big brothers", a.k.a the "Galex", Alexandre Gramfort and Gaël Varoquaux, who took the time to teach me a part of their huge knowledge, and decreased (I hope this is convex...) my ignorance in informatics, maths, and geek stuffs.

I also would like to acknowledge the "INRIOS Guapos" of the Parietal team for their support: Fabian "Consuela" Pedregosa, Alan "Graou" Tucholka, Pierre "Tonton" Fillard, Jean-Baptise Poline, Virgile Fritsch, Viviana Siless and Lise Favre. Thanks to Régine Bricquet, Marie Domingues and Stéphanie Druetta, for their help in the crual administrative world. Thanks to the guys from the Mimagen project, Yannick "As your sister" Schwartz, for showing me that I can find someone dumber than me, Benjamin Thyreau and Alexis Barbot. Thanks to Matthieu Perrot, Merlin Keller and Pamela Guevara, for exciting (scientific or not) conversations.

Thanks to my collaborators from the LNAO and LCOGN teams for their help: Evelyn Eger, Edouard Duchesnay, Philippe Ciuciu, Vincent Frouin, Alexis Roche, Thomas Vincent, Edith Le Floch, Anne-Laure Fouque, Pauline Roca, Eric Jouvent, Irina Kezele, Yann Cointepas, Denis Rivière, Linda Marrakchi, Valdis Gudmundsdottir, Dominique Geffroy, Soizic Laguitton, Grégory Operto, Julien Lefèvre, Laurent Risser, Nicolas Souedet, and Jean-François Mangin. Thanks to Mathias Pessiglione, Maël Lebreton, and Rodolphe Jenatton for their collaboration. Thanks to Tom Mitchell for giving my the opportunity to see how it works in the other side of the ocean, and for exciting discussions.

Thanks to Olivier, Mo, Manu, Sandrine, Marcus, Géraldine, Damdam, Anne-Carine and all the other folks, for being here and helping me to get off this thing. Also thanks to the PORC for making me suffer every week on the field, and making me understand that science will never help me to avoid being crushed by angry mountains of muscles.

Finally, I would like to thank my parents Evelyne and Jean-Louis, because I would not get here without them, and my brothers Christophe and Guillaume. A special thank to my beloved Valérie that had to cope with my stress and mood swings during two years, and (almost) never complained.

Résumé

Contexte de la thèse

Les *neurosciences cognitives* regroupent de nombreuses disciplines étudiant différentes composantes de la cognition humaine, telles que le comportement social, la mémoire ou les interactions sensorimotrices, en lien avec la structure et le fonctionnement du cerveau. Cette étude est particulièrement délicate, étant donné la structure multi-échelle du système nerveux (des synapses – connections entre neurones – 2 à 40 nm, au cerveau – 150 mm pour l'axe longitudinal), et son extrême complexité (jusqu'à 10^{15} connections). Les neurosciences cognitives sont donc basées sur un large éventail de disciplines, et en particulier la *neuro-imagerie*. Un des buts de la neuro-imagerie est de procurer une cartographie des régions fonctionnelles du cerveau et de leur interactions respectives. Cela inclut l'étude du *codage neuronal*, qui est la représentation interne d'informations dans le cerveau.

Le codage neuronal peut être effectué selon un grand nombre de schémas différents [Dayan 01], et sa caractérisation repose principalement sur les interactions entre les différentes entités impliquées dans ce codage, appelées *entités de codage*, et sur leur distribution spatiale. Les interactions entre entités de codage peuvent être expliquées par deux schémas de codage principaux : le *codage parcimonieux*, quand très peu d'entités sont impliquées (théoriquement une seule), et le *codage par populations*, quand le codage est effectué par un grand nombre d'entités de codage. Un autre aspect du codage neuronal est la distribution spatiale des entités de codage, qui peuvent être *groupées* dans une région bien localisée du cerveau, ou *distribuées*, *i.e.* éparpillées dans le volume entier.

La neuro-imagerie procure une opportunité unique d'étudier l'architecture fonctionnelle du cerveau (*imagerie fonctionnelle*) tout en étant minimalement invasive. Cette technique est donc bien adaptée à l'étude du codage neuronal. Différentes modalités existent, chacune ayant des résolutions spatiale et temporelle bien spécifiques. Parmi elles, l'*Imagerie par Résonance Magnétique fonctionnelle (IRMf)* [Ogawa 90b, Ogawa 90a] a émergé comme une modalité fondamentale pour l'imagerie fonctionnelle cérébrale. Depuis une vingtaine d'années, l'IRMf a été utilisée intensivement pour l'imagerie cérébrale, et est devenue une modalité de référence, grâce à sa bonne résolution spatiale. L'IRMf mesure par *Résonance Magnétique Nucléaire* une grandeur qui dépend, de manière indirecte et encore mal connue, de l'oxygénation du sang. La mesure est effectuée en utilisant un contraste appelé contraste *BOLD (Blood Oxygenation Level-Dependent)*. Quand certaines populations neuronales sont actives, l'augmentation du taux d'oxyhémoglobine augmente le contraste BOLD, et offre donc un accès indirect à des images de l'activité cérébrale.

Les images d'IMRf sont ensuite pré-traitées, puis modélisées au travers d'un *Modèle Linéaire Général (GLM)*, qui considère les différentes conditions expérimentales définies dans une *matrice de dessin*, ainsi que la dynamique de la

réponse hémodynamique. Les paramètres du modèle peuvent être représentés sous forme d'images, appelées *cartes d'activations*. Ces cartes représentent l'influence locale des différentes conditions expérimentales sur les signaux d'IRMf. L'approche classique, et très largement utilisée, pour l'analyse des cartes d'activations, est appelée *inférence classique*, et repose sur l'utilisation massive de tests univariés (un test par voxel). On crée ainsi des cartes de statistique paramétrique (*Statistical Parametric Maps (SPMs)*) [Friston 95], qui ont beaucoup d'intérêt en neurosciences, car elles permettent une localisation des voxels qui sont significativement actifs pour une condition expérimentale, et sont donc probablement impliqués dans le codage neuronal sous-jacent. Cependant cette inférence classique a une puissance limitée par le problème de comparaisons multiples, et ne tient pas compte de la structure multivariée des données d'IRMf.

Une approche récente, appelée *inférence inverse* (*"brain-reading"*) [Dehaene 98, Cox 03], a été proposée pour tenir compte des limitations de l'inférence classique. L'inférence inverse repose sur une approche de reconnaissance de formes (*pattern recognition*), et décode le codage neuronal en utilisant des méthodes *d'apprentissage statistique*. De manière concise, l'inférence inverse construit une fonction de prédiction, en utilisant les cartes d'activations, qui peut alors être utilisée pour prédire une variable comportementale liée à une nouvelle image d'activitation. La précision de la prédiction peut être vue comme une mesure de la quantité d'information présente dans les voxels utilisés dans la fonction de prédiction, en rapport avec la tâche cognitive. Cette approche est multivariée et permet de réaliser des analyses plus sensibles que la procédure basée sur des cartes *SPMs* [Kamitani 05, Haynes 06]. De nombreuses méthodes d'apprentissage statistique ont été testées pour la classification ou la régression (*e.g.* analyse linéaire discriminante, machines à vecteurs de support, régression elastic net). Cependant, dans le cas de l'analyse de données d'IRMf, la limitation majeure reste la localisation et l'extraction des régions prédictives au sein du volume cérébral. Nous avons de plus un problème de *malédiction de la dimension*, car le nombre *d'attributs* (voxels, régions) est beaucoup plus grand ($\sim 10^5$) que le nombre d'échantillons (images) ($\sim 10^2$), et la méthode de prédiction peut sur-apprendre les données d'apprentissage, et donc ne pas généraliser correctement à de nouvelles images.

L'objectif général de cette thèse est le développement de méthodes d'apprentissage statistique, utilisables en *inférence inverse*, qui tiennent compte des spécificités des données d'IRMf. D'un point de vue expérimental, nous nous focalisons particulièrement sur la compréhension du cortex visuel humain.

Organisation et contributions de cette thèse

Chapitre 1 - Accéder au codage neuronal

Dans ce premier chapitre, nous décrivons l'organisation fonctionnelle du cerveau humain, et détaillons la notion de codage neuronal. Nous nous focalisons sur les différentes distributions spatiales des entités impliquées dans le codage, et détaillons l'Imagerie par Résonance Magnétique fonctionnelle, une modalité d'imagerie bien adaptée à l'étude de ce codage.

Organisation fonctionnelle du cerveau

Le cerveau humain peut être décomposé en différentes régions qui correspondent à différentes étapes du traitement de l'information au sein du cerveau (voir Fig. *i*). Ces régions fonctionnelles correspondent grossièrement à des régions anatomiques, et peuvent être classées en trois catégories : les aires sensorielles (*e.g.* cortex visuel et cortex auditif) qui reçoivent et traitent les informations en provenance des organes sensoriels, les aires motrices (*e.g.* cortex moteur primaire, cortex pré-moteur) qui contrôlent les mouvements, et les aires associatives (*e.g.* aire de Broca, complexe latéro-occipital – *LOC* – sillon intra pariétal – *IPS*) qui traitent les informations relatives aux expériences perceptuelles. Les expériences détaillées dans cette thèse sont effectuées dans le cadre de l'étude de la reconnaissance des objets (cortex visuel et LOC), et du traitement des nombres (cortex parietal et IPS).

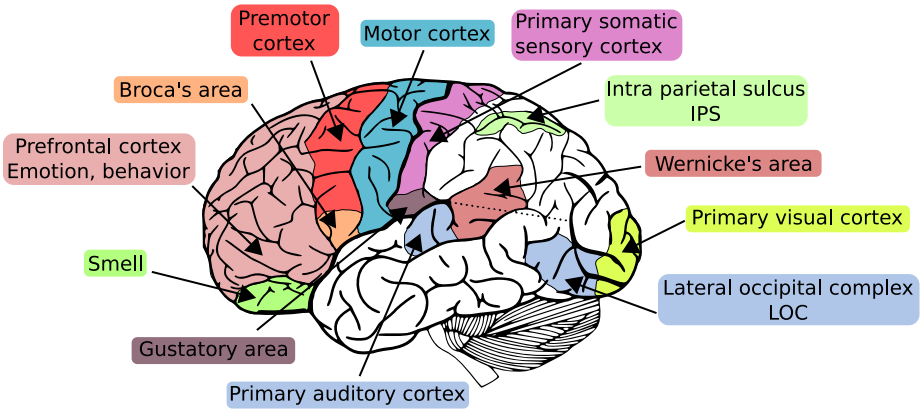


Fig. *i* : Les principales régions fonctionnelles du cerveau humain (hémisphère gauche), ainsi que les deux régions étudiées dans cette thèse (LOC et IPS). Adapté de http://agaudi.files.wordpress.com/.

Codage neuronal des processus mentaux

Le codage neuronal est la correspondance entre un stimulus et sa représentation par une unique réponse neuronale ou un ensemble de réponses neuronales. Accéder à l'organisation du codage neuronal est nécessaire pour comprendre les processus mentaux, et plus généralement la façon dont le cerveau traite l'information. L'étude du codage neuronal peut être effectuée à différentes échelles (du simple neurone aux grandes populations de neurones telles que les colonnes corticales avec 10^4 neurones), et nous appellerons désormais les structures impliquées dans le codage neuronal *entités* ou *populations neuronales*. Le codage neuronal peut être étudié par *décodage*, lorsque l'on reconstruit le stimulus (ou certains aspects de ce stimulus), à partir des signaux des entités de codage. Dans cette thèse, nous nous focalisons sur la notion d'organisation spatiale du codage neuronal, et nous abordons les deux problèmes suivants : la question de la sélectivité des populations neuronales impliquées dans une tâche cognitive (*codage parcimonieux* ou *codage par populations*), et la question de la distribution spatiale de ces populations neuronales (*codage groupé* ou *codage distribué*) au sein du cerveau (cortex cérébral, ganglions de la base, thalamus, ...). En général, le codage par populations semble une hypothèse plus plausible que le codage parcimonieux, mais par contre, la supériorité d'un modèle de distribution spatiale particulier reste plus sujet à controverse. Comprendre ces différents schémas de codage (voir Fig. *ii*) est crucial pour les études cognitives, et, dans cette thèse, nous proposons certains outils pour faciliter cette compréhension, basée sur la neuro-imagerie fonctionnelle.

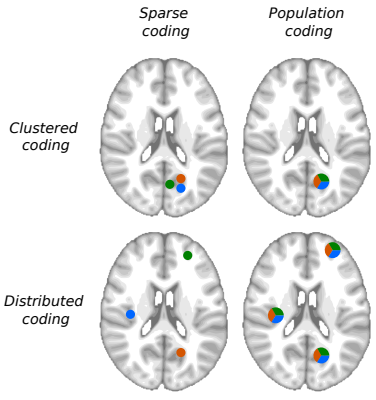


Fig. *ii* : Illustration des différents types d'entités impliquées dans le codage neuronal, et des différentes distributions spatiales de ces entités. Chaque couleur correspond à une condition, les neurones *gnostiques* (*i.e.* qui codent pour une seule condition) sont représentés par des disques de couleur uniforme, et les ensembles de neurones non spécifiques sont représentés par des disques de couleur mixte. Les deux notions d'entités et de distribution spatiale sont clairement distinctes. Dans le cas du codage groupé, les entités sont regroupées en petits ensembles, alors que dans le cas du codage distribué, les entités sont éparpillées à travers tout le cerveau. Le codage par population est basé sur des motifs d'activation qui doivent être analysés comme tels, c'est à dire décodés, alors que le codage parcimonieux repose sur très peu de neurones actifs.

Neuro-imagerie fonctionnelle et IRMf

La neuro-imagerie fonctionnelle vise à imager l'activité fonctionnelle du cerveau, afin d'étudier l'organisation spatiale du codage neuronal. Différentes approches peuvent être utilisées en neuro-imagerie, et nous donnons Fig. *iii*, leurs résolutions temporelles et spatiales.

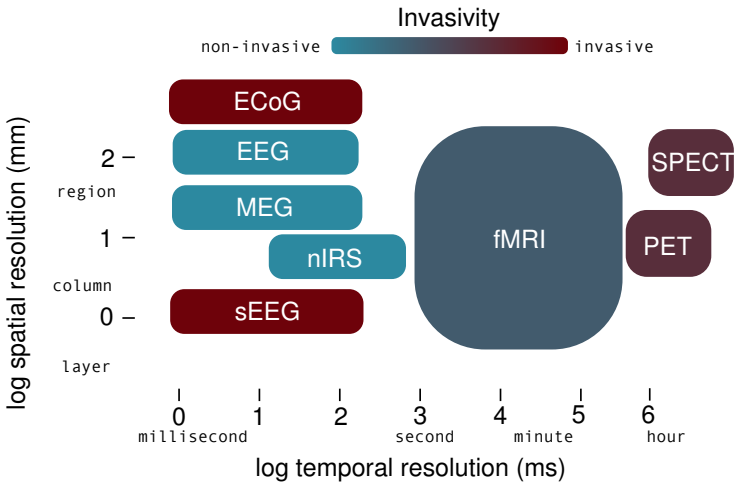


Fig. *iii* : Résolutions spatiale et temporelle des différentes modalités couramment utilisées pour l'imagerie fonctionnelle. Dans cette thèse, nous utilisons l'IRMf.

L'*Imagerie par Résonance Magnétique fonctionnelle (IRMf)* est une modalité couramment utilisée pour l'imagerie fonctionnelle cérébrale, car elle est non-invasive, a une bonne résolution spatiale (1-3mm), et permet d'avoir accès, même indirectement, à l'activité neuronale. De plus, dans le cadre standard d'acquisition, l'IRMf permet d'imager le cerveau dans son entier, ce qui permet de ne pas restreindre le décodage aux couches superficielles du cortex ou à des régions prédéfinies. L'IRMf est l'utilisation de l'IRM avec un contraste spécifique, appelé contraste *BOLD* [Ogawa 90b, Ogawa 90a]. Cette modalité mesure une fonction qui dépend du rapport entre les taux d'oxyhémoglobine et de déoxyhémoglobine dans le sang. Quand certaines populations de neurones sont actives, l'accroissement du taux de d'oxyhémoglobine dans le sang est observé par contraste BOLD. Cependant, cet effet n'est pas directement relié à l'activité neuronale et repose sur un chemin métabolique complexe et encore mal connu. Le signal d'IRMf reflète ainsi une activité qui peut se situer loin des neurones actifs. Cependant, malgré ces limitations, l'IRMf reste encore aujourd'hui la meilleure modalité pour l'étude de la distribution spatiale du codage neuronal.

Chapitre 2 - Des acquisitions *IRMf* au "brain-reading"

Dans le second chapitre, nous détaillons les pré-traitements requis pour l'analyse des données d'IRMf. Nous introduisons aussi le *Modèle Linéaire Général*, qui construit un ensemble de cartes d'activations depuis les données, en se basant sur la description du paradigme expérimental et sur les bases physiologiques du signal BOLD. Les cartes d'activations résultantes peuvent alors être utilisées pour une analyse statistique, afin d'étudier le codage neuronal spécifique à certaines taches cognitives. Nous détaillons dans ce chapitre les deux méthodes d'inférence, *l'inférence classique* et *l'inférence inverse*.

Le *Modèle Linéaire Général (GLM)* a été introduit pour l'analyse de données d'IRMf par Friston et al. [Friston 95]. Cette approche permet, au sein d'un unique modèle statistique, de tenir compte de tous les facteurs qui peuvent expliquer les décours temporels des signaux d'IRMf.

Inférence classique

L'inférence classique est une méthode couramment utilisée pour l'étude des données d'IRMf. Cette approche, intimement liée au GLM, repose sur des statistiques calculées au niveau des voxels, et créé des cartes statistiques (*Statistical Parametric Maps - SPMs*) pour les effets considérés (voir Fig. *iv*). Ces cartes permettent une bonne cartographie cérébrale. Cependant, en dépit de sa simplicité et de ses performances, cette méthode souffre de certaines limitations :

- l'inférence classique analyse chaque voxel séparément et ne tient donc rarement en compte des corrélations existantes entre les différentes régions du cerveau (localement, on peut néanmoins considérer la co-activation de voxels voisins par inférence *cluster-level*).
- la puissance statistique est limitée par un problème de comparaisons multiples ; nous effectuons un test statistique pour chacun des voxels, et nous devons donc corriger pour le grand nombre de tests effectués.

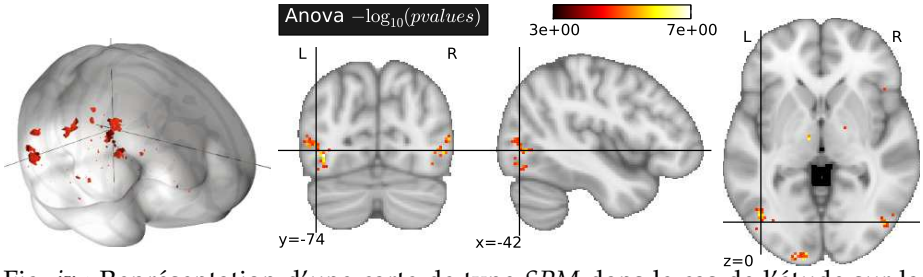


Fig. *iv* : Représentation d'une carte de type *SPM* dans le cas de l'étude sur la représentation mentale de la forme des objets. La carte est seuillée pour des p-valeurs inférieures à 10^{-3}. Nous pouvons remarquer que certaines régions du cerveau sont clairement délimitées, comme certaines régions du lobe occipital, connues pour être impliquées dans la reconnaissance visuelle.

Inférence inverse

Afin de tenir comptes des limitations de l'inférence classique, l'approche *d'inférence inverse* a été proposée [Dehaene 98, Cox 03]. Cette approche est basée sur des méthodes d'apprentissage statistique, et peut être utilisée pour vérifier l'implication de certaines zones cérébrales dans certains codes cognitifs. En évaluant la justesse de la prédiction d'une variable d'intérêt (la *cible*), basée sur les activations mesurées dans ces régions, il est possible de vérifier la pertinence des régions du cerveau étudiées. Cette approche a certains avantages, comparée à l'inférence classique :

- Cette approche est multivariée [Cox 03, Norman 06], ce qui est consistant avec l'hypothèse de codage par populations. En effet, dans le codage par populations, plusieurs entités sont impliquées dans le codage, et il faut donc les considérer ensemble afin de décoder le codage neuronal.
- Cette approche permet d'éviter le problème de comparaisons multiples, car elle réalise un seul test statistique (sur la variable prédite). En ce sens, l'inférence inverse permet de réaliser des analyses plus sensibles que l'inférence classique [Kamitani 05].
- Cette approche permet de généraliser la prédiction à des stimuli pouvant être inconnus [Mitchell 08, Knops 09], et ouvre la voie à une compréhension plus poussée de l'organisation fonctionnelle du cerveau, ainsi qu'à une possible reconstruction des stimuli [Thirion 06a, Kay 08].

La Fig. *v* représente les différentes étapes de l'analyse par inférence inverse. L'inférence inverse a cependant quelques défauts, comme la nécessité de prendre en compte la grande dimension des données lors de l'apprentissage de la fonction de prédiction. Nous détaillons aussi dans ce chapître certaines questions éthiques qui peuvent être soulevées par cette "lecture de pensées" [Farah 04].

Finalement, nous nous sommes particulièrement intéressés durant cette thèse à la généralisation de la prédiction à de nouveaux sujets. Cependant cette prédiction entre sujets est très sensible à la variabilité inter-sujet, qui rend la localisation des régions fonctionnelles variables entre sujets [Tucholka 10]. Il est donc particulièrement difficile de trouver un support spatial du codage neuronal entre sujets. Nous proposons dans cette thèse certains algorithmes qui visent à résoudre ce problème.

Chapitre 3 - Apprentissage statistique pour l'inférence inverse en IRMf

Dans le troisième chapitre, nous présentons les différentes étapes de l'inférence inverse : l'apprentissage de la fonction de prédiction, la réduction de dimension, la sélection de modèle et la validation. Nous détaillons aussi les différentes méthodes d'apprentissage statistique qui ont été utilisées ces dernières années dans le cadre de l'inférence inverse en IRMf.

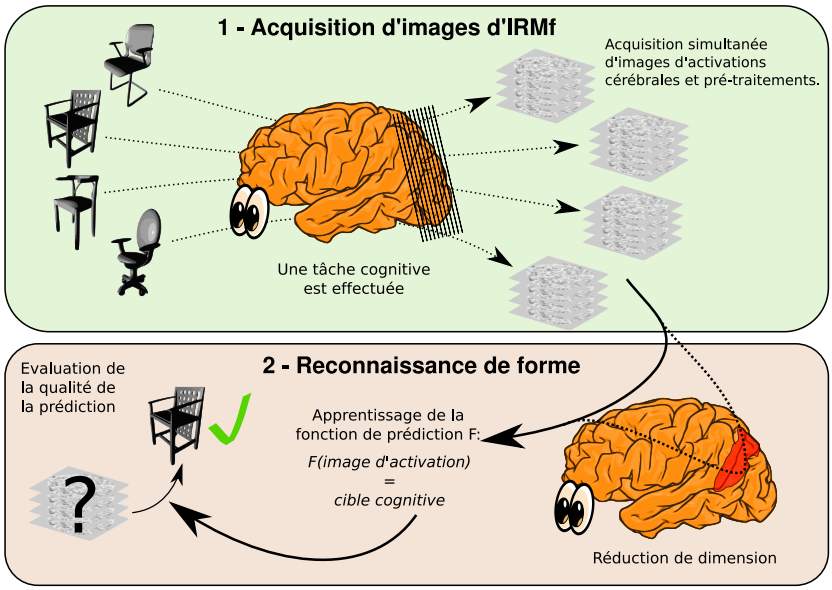


Fig. v : Illustration de l'inférence inverse. **Étape 1** : le sujet réalise une tâche cognitive, comme regarder des objets de différentes formes. Les images d'IRMf sont acquises simultanément et pré-traitées. **Étape 2** : un modèle prédictif est appris, et les prédictions correspondantes à des données test sont réalisées, puis comparées avec le vrai stimulus. Une *réduction de dimension* peut être réalisée, avant l'apprentissage du modèle prédictif, afin de sélectionner les zones du cerveau les plus pertinentes pour la prédiction ; cette étape peut être cruciale afin d'éviter le sur-apprentissage. La gestion du sur-apprentissage sera un des problèmes principaux abordés au cours de cette thèse.

L'inférence inverse cherche à décrypter le codage neuronal en trouvant des régions prédictives au sein des données d'IRMf. Il faut donc définir et entraîner une fonction de prédiction. Cependant, il y a un large choix de fonctions de prédiction, et nous détaillons dans ce chapitre les plus communément utilisées, en illustrant leur comportement sur des données réelles.

Modèle linéaire prédictif

La fonction de prédiction peut être non linéaire (*e.g. SVM* non linéaire), mais la supériorité d'une telle fonction non -linéaire dans le cadre de l'inférence inverse en IRMf n'a pas été montrée [Cox 03, LaConte 05]. Cependant, les raisons de cette supériorité ne sont pas encore complètement élucidées. Elle peut être expliquée par le fait que la sommation sur 10^5 neurones ayant une activité non-linéaire, peut être approximativement linéaire. Elle peut aussi simplement refléter le fait que les fonctions de prédictions non-linéaires utilisées sont incapables de capturer la non-linéarité réelle de cette relation. Dans cette thèse,

nous nous focalisons sur des fonctions de prédiction linéaires :

$$\mathbf{y} = f(\mathbf{X}, \mathbf{w}, b) = F(\mathbf{X}\mathbf{w} + b) \ , \tag{1}$$

où $(\mathbf{w}, b)$ sont les paramètres du modèle devant être estimés avec les données d'apprentissage ($b \in \mathbb{R}$ est appelé *l'ordonnée à l'origine*). $\mathbf{y}$ est la variable comportementale à prédire (la cible), et les données d'IRMf sont représentées par la matrice $\mathbf{X} \in \mathbf{R}^{n \times p}$, chaque ligne étant un échantillon p-dimensionnel (*i.e.* une carte d'activation). Nous avons n le nombre d'échantillons (images) et p le nombre d'attributs (voxels).

Dans le cas de la régression, nous avons $\mathbf{y} \in \mathbf{R}^n$, avec f :

$$f(\mathbf{X}, \mathbf{w}, b) = \mathbf{X}\,\mathbf{w} + b \tag{2}$$

Dans le cas de la classification, nous avons $\mathbf{y} \in [1, \ldots, K]^n$, avec f :

$$f(\mathbf{X}, \mathbf{w}, b) = \text{sign}(\mathbf{X}\,\mathbf{w} + b) \ , \tag{3}$$

où "sign" est la fonction signe. Nous détaillons aussi dans ce chapitre différentes heuristiques pour réaliser des classifications multi-classes.

Performances en prédiction

Les performances en prédictions peuvent être vues comme un test statistique sur les régions utilisées dans le modèle prédictif. Si une fonction de prédiction linéaire a des performances significativement supérieures au niveau de la chance, on peut considérer que le groupe de voxels utilisés dans la prédiction contient de l'information sur la variable cible (voir [Kamitani 05]).

Dans le cas d'une fonction de prédiction linéaire, il peut être intéressant de regarder les poids des voxels utilisés dans le modèle linéaire. Cependant, ces poids dépendent fortement de la fonction de prédiction, et nous n'avons aucune preuve que les voxels utilisés dans le modèle correspondent à la totalité des entités impliquées dans le codage neuronal étudié [Cox 03]. Les cartes obtenues ne peuvent donc pas être interprétées comme des cartes *SPMs* classiques. Cependant, il est toujours possible d'utiliser les cartes des poids du modèle pour interpréter certains aspects du codage neuronal. Nous nous attendons en effet à ce que l'organisation spatiale du codage neuronal soit parcimonieuse et possède une structure telle que les voxels de poids non-nuls soient groupés en composantes connexes. Les cartes de poids montrant de telles caractéristiques sont dites *interprétables*, car elles reflètent nos hypothèses sur l'organisation spatiale du codage neuronal. Dans le cas d'un modèle de prédiction non linéaire, la question de l'interprétation est plus complexe, et la non linéarité du modèle rend difficile l'accès à des cartes de poids.

Sélection de modèle et validation

Afin de valider le fait que les voxels utilisés par le modèle appartiennent effectivement au support du codage neuronal, nous devons évaluer la justesse

de la prédiction, *i.e.* tester si les prédictions effectuées par le modèle sont correctes. Cependant, le fait d'apprendre la fonction de prédiction et de la tester sur un même jeu de données, introduit un biais. Afin d'éviter ce biais de sur-apprentissage, nous devons définir deux jeux de données différents : un *jeu d'apprentissage* $(\mathbf{X^l}, \mathbf{y^l})$, qui est utilisé pour l'apprentissage de la fonction de prédiction, et un *jeu de test* $(\mathbf{X^t}, \mathbf{y^t})$, qui est utilisé pour tester la fonction de prédiction. Afin de ne pas dépendre d'un certain choix de jeux d'apprentissage et de test, et afin d'utiliser au mieux le nombre réduit d'échantillons disponibles, nous pouvons effectuer une *validation croisée*. Cette validation croisée sépare les données en un certain nombre de couples de jeux d'apprentissage et de test, le score final étant la moyenne des scores pour chacun des couples. Les différentes étapes d'apprentissage statistique de l'inférence inverse sont donc effectuées par validation croisée (voir Fig. *vi*). Une validation croisée (dite *interne*) peut aussi être utilisée sur les données d'apprentissage afin de sélectionner un modèle qui généralise optimalement.

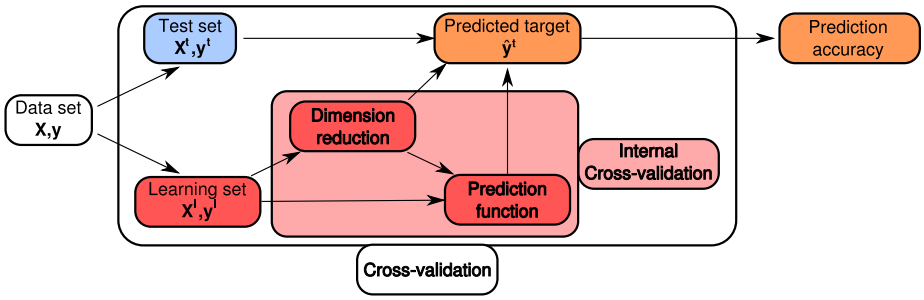


Fig. *vi* : Schéma global de l'approche d'apprentissage statistique pour l'inférence inverse, avec une sélection de modèle par validation croisée interne.

Réduction de dimension

Il a été montré [Hughes 68] qu'augmenter la complexité du modèle peut d'abord augmenter les performances de prédiction jusqu'à ce qu'une valeur optimale soit obtenue. Cependant, continuer à augmenter la complexité (*i.e.* la dimension) des données va réduire les performances en prédiction. Cet effet est appelé *malédiction de la dimension*, et est crucial dans l'analyse de données en IRMf. En effet, dans le cas où le nombre d'attributs p est très grand devant le nombre d'échantillons n (ce qui est le cas en IRMf, avec typiquement $p \sim 10^5$ et $n \sim 10^2$), il est toujours possible de trouver une fonction de prédiction qui donne une prédiction parfaite sur les données d'apprentissage. Cependant, une telle fonction ne peut pas généraliser (*i.e.* donner une prédiction correcte sur de nouveaux échantillons) car elle a appris des particularités non-informatives du jeu d'apprentissage, ou du bruit. On dit qu'une telle fonction *sur-apprend* les données d'apprentissage. Ce problème peut être évité en utilisant des méthodes de *réduction de dimension*, qui définissent un espace de petite

dimension qui contient l'information prédictive, tout en réduisant la dimension du problème. Pour l'inférence inverse en IRMf, la réduction de dimensions est menée avec deux objectifs différents, qui peuvent être remplis ou non : elle doit permettre d'obtenir un bonne performance en prédiction (*i.e.* extraire de l'information pertinente), et doit extraire des groupes de voxels interprétables (*e.g.* en construisant un espace de petite dimension qui correspond à un nombre réduit de régions cérébrales).

Pré-requis pour un algorithme d'apprentissage statistique en inférence inverse

Les différentes méthodes d'apprentissage statistique détaillées dans ce chapitre ont souvent été utilisées sans tenir compte des spécificités des données d'IRMf, et souffrent donc de certaines limitations. Les différentes études menées dans ce chapitre nous permettent de définir les pré-requis suivant pour un algorithme d'apprentissage statistique adapté à l'inférence inverse en IRMf :

1. **Modèle multivarié** : l'information d'intérêt peut être distribuée à travers des régions distantes du cerveau. L'algorithme d'apprentissage statistique doit donc tenir compte de la combinaison des signaux de ces différentes régions cérébrales, et, en ce sens, doit être *multivarié*.
2. **Tenir compte de la structure spatiale des données** : A cause de la structure spatiale particulière des données d'IRMf, il y a une redondance locale de l'information utilisable pour la prédiction, qui doit être considérée dans la fonction de prédiction, ou dans le processus de sélection d'attributs (par exemple en remplaçant les signaux des voxels par des moyennes locales).
3. **Approche multi-échelle** : Étant donné que les régions étudiées sont larges, et que les zones informatives peuvent être relativement petites, nous devons définir une approche qui peut se focaliser sur des petites sous-régions du volume considéré. Une méthode multi-échelle semble donc adaptée pour la recherche des régions prédictives. De plus, au contraire des approches purement géométriques, les procédures qui tiennent compte du signal et de la tâche de prédiction peuvent mieux respecter la structure multi-échelle sous-jacente des données.

Publications

Les méthodes présentées dans ce chapitre ont été utilisées pendant cette thèse dans le cadre des analyses neuroscientifiques suivantes :

- M. Lebreton, S. Jorge, V. Michel, B. Thirion and M. Pessiglione. *An automatic valuation system in the human brain : evidence from functional neuroimaging*. Neuron 64, 3, 2009.
- E. Eger, V. Michel, B. Thirion, A. Amadon, S. Dehaene and A. Kleinschmidt. *Deciphering Cortical Number Coding from Human Brain Activity Pattern*. Current Biology. 2009, 19 :1608.

– A. Knops, B. Thirion, E.M. Hubbard, V. Michel and S. Dehaene. *Recruitment of an area involved in eye movements during mental arithmetic*. Science. 2009 Jun 19 ;324(5934) :1583-5.
– A. Bachrach, A. Gramfort, V. Michel, E. Cauvet, B. Thirion and C. Pallier. *Decoding of syntactic trees*. In prep.

Des travaux méthodologiques ont aussi été présentés dans :

– V. Michel, C. Damon, and B. Thirion. Mutual information-based feature selection enhances *fMRI* brain activity classification. In 5th Proc. IEEE ISBI, pages 592-595, 2008.
– R. Genuer, V. Michel, E. Eger, and B. Thirion Random forests based feature selection for decoding *fMRI* data. In COMPSTAT 19th International Conference on Computational Statistics, pages 372 , 2010.

Chapitre 4 - Régression Bayésienne Multi-classe Parcimonieuse

Dans ce chapitre, nous proposons un modèle pour effectuer une régression adaptative, appelée *MCBR (Multi-Class Sparse Bayesian Regression)*. Nous groupons les attributs en Q classes différentes, et régularisons ces classes différemment, ce qui permet d'obtenir une régularisation stable et adaptative.

Les caractéristiques principales du modèle sont les suivantes :

– **Généralisation d'approches classiques** : la méthode proposée intègre dans un même modèle les approches *Bayesian Ridge Regression (BRR)* et *Automatic Relevance Determination (ARD)*, qui sont les deux principales approches de régression Bayésienne régularisée.
– **Régularisation adaptative** : en réalisant une régularisation différente pour les attributs pertinents et non pertinents, cette approche permet de réaliser conjointement une sélection d'attributs, et une estimation correcte des poids du modèle. La méthode proposée peut adapter, au sein d'un cadre Bayésien, le niveau de parcimonie aux données. La régularisation effectuée est ainsi aussi adaptative que l'ARD, mais est réalisée avec beaucoup moins d'hyper-paramètres, et est donc moins sensible au sur-apprentissage dans l'espace des hyper-paramètres.
– **Regroupement d'attributs** : l'approche proposée permet d'obtenir, au travers de la variable latente $\mathbf{z}$ d'appartenance aux classes, une information intéressante pour l'inférence inverse en IRMf. En effet, le regroupement d'attributs intrinsèque de *MCBR* permet d'extraire des groupes de voxels pertinents (voir Fig. *vii*).

Le modèle graphique de cette approche est donné Fig. *viii*. Ce modèle peut être estimé par une approche Bayésienne variationelle, et est alors appelé *VB-MCBR*, ou par une approche d'échantillonage de Gibbs, et est alors appelé *Gibbs-MCBR*.

Des expériences sur des données simulées et des données réelles montrent que notre approche est bien adaptée à la neuro-imagerie, car elle réalise des

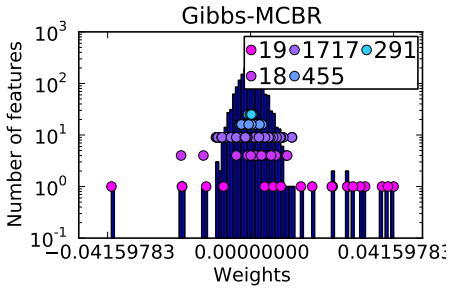
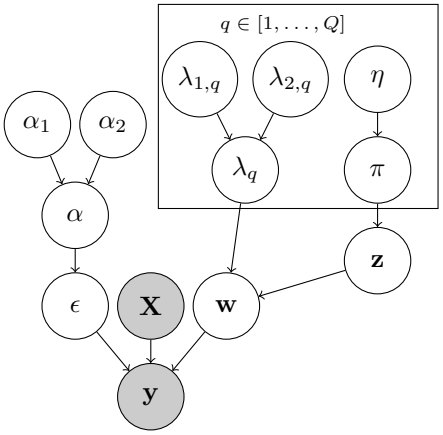


Fig. *vii* : Analyse inter-sujet de la représentation mentale de la taille des objets. Histogramme des poids obtenus par *Gibbs-MCBR*, et les classes correspondantes (chaque couleur représente une classe différente du modèle). Nous pouvons voir que l'approche *Gibbs-MCBR* crée des groupes d'attributs informatifs et non-informatifs, et que les différentes classes sont régularisées différemment, suivant la pertinence des attributs qui les composent.



$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\,\mathbf{w} + \epsilon + b \\
\epsilon &\sim \mathcal{N}(0, \alpha^{-\mathbf{1}}\mathbf{I_n}) \\
\alpha &\sim \Gamma(\alpha_1, \alpha_2) \\
\mathbf{z} &\sim \prod_{j=1}^{j=p}\prod_{q=1}^{q=Q} \pi_q^{\delta_{jq}} \\
\pi &\sim Dir(\eta) \\
\mathbf{w} &\sim \mathcal{N}(0, \mathbf{A}^{-1}) \\
&\text{with} \quad diag\,(\mathbf{A}) = \{\lambda_{z_1}, ..., \lambda_{z_p}\} \\
\lambda_q &\sim \Gamma(\lambda_{1,q}, \lambda_{2,q})
\end{aligned}
$$

Fig. *viii* : Modèle graphique de la régression Bayésienne multi-classes parcimonieuse.

prédictions correctes et stables, par rapport aux méthodes de l'état de l'art. Parmi différentes possibilités de recherche, il peut être intéressant d'ajouter un a-priori de Dirichlet au modèle MCBR. Cet a-priori peut permettre de régler automatiquement le nombre de classes Q, et peut donc adapter la parcimonie entre les deux extrêmes que sont *Bayesian Ridge Regression* (aucune parcimonie), et *Automatic Relevance Determination* (forte parcimonie). Une autre direction de recherche peut être d'implémenter un modèle spatial à l'approche proposée, de manière à extraire des groupes de voxels connectés, par exemple ajouter un a-priori Markovien sur les classes pour obtenir une consistance spatiale (mais une telle approche peut être coûteuse en temps de calcul).

Publications

Les contributions développées dans ce chapitre ont été publiées dans :

- V. Michel, E. Eger, C. Keribin and B. Thirion. *Adaptive multi-class bayesian sparse regression - an application to brain activity classification*. In MICCAI'09 Workshop on Analysis of Functional Medical Images, 2009.
- V. Michel, E. Eger, C. Keribin and B. Thirion. *Multi-Class Sparse Bayesian Regression for Neuroimaging data analysis*. Pages 50-57. In International Workshop on Machine Learning in Medical Imaging (MLMI) In conjunction with MICCAI, 2010.

Chapitre 5 - Coalescence supervisée - *"Supervised clustering"*

Un des défauts principaux des approches d'apprentissage statistique en inférence inverse, est qu'elles ne tiennent pas compte de *l'information spatiale*. En effet, à cause des processus métaboliques sous-jacents au signal IRMf, il y a un filtrage local de l'information qui doit être pris en compte. Ceci peut être fait par l'*agglomération d'attributs*, qui moyenne le signal de voxels voisins, afin de créer des structures intermédiaires appelées *parcelles*.

Dans ce chapitre, nous décrivons la seconde contribution de cette thèse, qui est *la coalescence supervisée ("supervised clustering")*, et qui est détaillée Fig. *ix*. Cette méthode introduit la structure des données par l'intermédiaire d'un algorithme de coalescence hiérarchique contraint spatialement, qui créé une parcellisation hierarchique ayant une structure d'arbre. Nous adaptons ensuite l'algorithme au problème de prédiction, en choisissant la meilleure coupure de l'arbre de parcellisation afin de maximiser la qualité de la prédiction. Une propriété particulièrement importante de cette approche est sa possibilité de se focaliser sur des régions relativement petites mais informatives, tout en laissant de larges zones non informatives non segmentées. De plus, cette approche n'est pas restreinte à une fonction de prédiction particulière, et peut être utilisée avec de nombreuses méthodes de régression ou de classification.

Les résultats expérimentaux démontrent que cet algorithme est efficace pour les analyses inter-sujet, car la moyenne spatiale du signal effectuée par la parcellisation est une manière efficace de résoudre le problème de la variabilité inter-sujet (voir Fig. *x*). Finalement, les cartes créées par la méthode proposée sont plus interprétables car elles présentent une structure spatiale simple, comparées aux approches basées sur les voxels (voir Fig. *xi*).

Publications

Les contributions développées dans ce chapitre ont été publiées dans :

- V. Michel, E. Eger, C. Keribin, J.-B. Poline and B. Thirion. *A supervised clustering approach for extracting predictive information from brain activation*

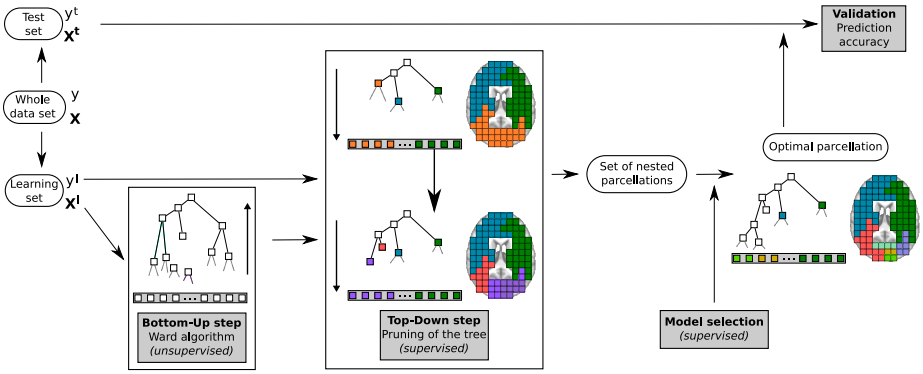


Fig. *ix* : Algorithme de *coalescence supervisée*. Étape ascendante de coalescence basée sur l'algorithme de Ward (*Bottom-Up step - Ward clustering*) : nous construisons un arbre de parcellisation $\mathcal{T}$ des feuilles jusqu'à la racine, en tenant compte de contraintes de connectivité. Étape descendance d'exploration de l'arbre (*Top-Down step - Pruning of the tree*) : l'arbre de Ward est divisé en choisissant la meilleure coupure en fonction d'un score de prédiction ζ. Étape de sélection de modèle (*Model selection*) : en utilisant les parcellisations créées lors de l'étape d'exploration, nous sélectionnons le sous-arbre $\widehat{\mathcal{T}}$, qui donne la meilleure valeur du score de prédiction ζ.

images. In IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA10) - IEEE Conference on Computer Vision and Pattern Recognition. 2010.

– V. Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin and B. Thirion. *A supervised clustering approach for fMRI-based inference of brain states*. Submitted to Pattern Recognition - Special Issue on 'Brain Decoding'. 2010

Chapitre 6 - Régularisation par Variation Totale (TV)

Dans ce dernier chapitre, nous introduisons la régularisation par Variation Totale, qui tient compte de l'information spatiale au sein de la régularisation. Cette régularisation repose sur des concepts d'optimisation convexe, et consiste à pénaliser l'estimation des poids du modèle de prédiction par la norme ℓ_1 de l'image du gradient des poids.

La régularisation par minimisation de la Variation Totale *TV*, *i.e.* de la norme ℓ_1 de l'image de gradient, a d'abord été utilisé pour le débruitage d'image [Rudin 92, Chambolle 04]. La motivation de l'utilisation de la régularisation *TV* pour l'imagerie cérébrale, vient du fait qu'elle permet de créer une carte de poids avec une structure par blocs (*i.e.* elle créé des composantes connexes avec des poids de valeurs identiques). Elle permet donc une bonne extraction des régions cérébrales impliquées dans la tâche cognitive. Dans ce chapitre, nous

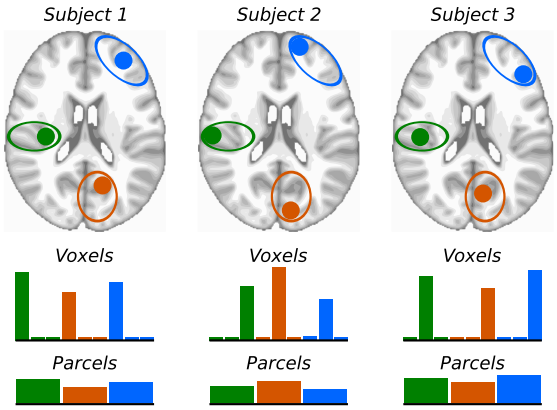


Fig. *x* : Illustration de l'agglomération d'attributs, comme méthode pour tenir compte de la variabilité inter-individuelle. Les régions impliquées dans la tâche cognitive sont représentées par des disques de différentes couleurs. Les populations de neurones actifs ne sont pas exactement à la même position à travers les sujets (haut), et le signal moyen à travers les sujets dans les voxels informatifs (milieu) ne porte pas beaucoup d'information. Il est donc clair, dans ce cas, que les approches de décodage utilisant les voxels vont avoir des performances faibles. Cependant, la moyenne des voxels informatifs au sein de chaque région à travers les sujets (bas) porte plus d'information, et permet d'améliorer la prédiction inter-sujet.

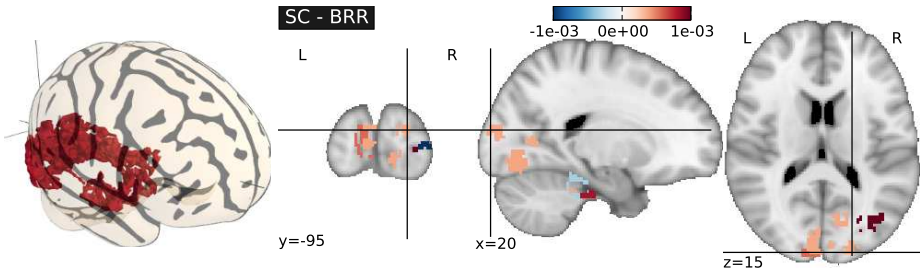


Fig. *xi* : Analyse inter-sujet de la représentation mentale de la taille des objets. La carte des poids obtenue par l'algorithme de coalescence supervisé montre des composantes connexes interprétables.

détaillons aussi la notion d'opérateur proximal qui permet le développement de procédures itératives telles que *ISTA* et *FISTA*, permettant de résoudre des problèmes d'optimisation convexe. Dans le cas spécifique de la régularisation par Variation Totale, l'optimisation est effectuée par une double boucle *ISTA* et *FISTA*. L'algorithme détaillé dans ce chapitre peut être utilisé pour la régression ou la classification.

La régularisation TV peut être utilisée pour extraire de l'information de

données d'IRMf. La sélection d'attributs et l'estimation du modèle sont réalisées conjointement. La méthode proposée capture l'information prédictive présente dans les données de manière plus précise que les méthodes de référence. Une propriété particulièrement importante de cette approche est sa tendance à créer des régions cohérentes spatialement, et ayant des poids similaires, réalisant ainsi des groupements d'attributs pertinents (voir Fig. *xii*). Les résultats expérimentaux montrent que cet algorithme est performant sur des données réelles, et qu'il est beaucoup plus précis que les méthodes de référence en analyse inter-sujet. Nous montrons aussi que les régions extraites sont robustes à la variabilité inter-individuelle. Ces observations démontrent que la régularisation TV est un outil puissant pour la compréhension de l'activité cérébrale et la cartographie des processus cognitifs. C'est de plus la première approche capable de réaliser des cartes de poids similaires à l'approche standard *SPM*, au sein d'un cadre d'inférence inverse.

Publications

Les contributions développées dans ce chapitre ont été publiées dans :

- V. Michel, A. Gramfort, G. Varoquaux and B. Thirion. *Total Variation regularization enhances regression-based brain activity prediction*. In 1st ICPR Workshop on Brain Decoding - Pattern recognition challenges in neuroimaging - 20th International Conference on Pattern Recognition. 2010.
- V. Michel, A. Gramfort, G. Varoquaux, E. Eger and B. Thirion. *Total variation regularization for fMRI-based prediction of behavior*. Submitted to IEEE Transactions on Medical Imaging. 2010.

Appendices

Appendice A - Une courte introduction à l'Imagerie par Résonance Magnétique

Dans cet appendice, nous détaillons brièvement les bases physiques de *l'Imagerie par Résonance Magnétique (IRM)*.

Appendice B - Description des jeux de données

Dans cet appendice, nous décrivons les jeux de données simulées et réelles qui ont été utilisés pendant cette thèse.

Appendice C - Scikit-learn pour l'inférence inverse en IRMf

Dans cet appendice, nous présentons l'utilisation du *Scikit-learn* pour l'inférence inverse en IRMf et nous donnons les principales fonctions qui peuvent être utilisées pour une telle analyse.

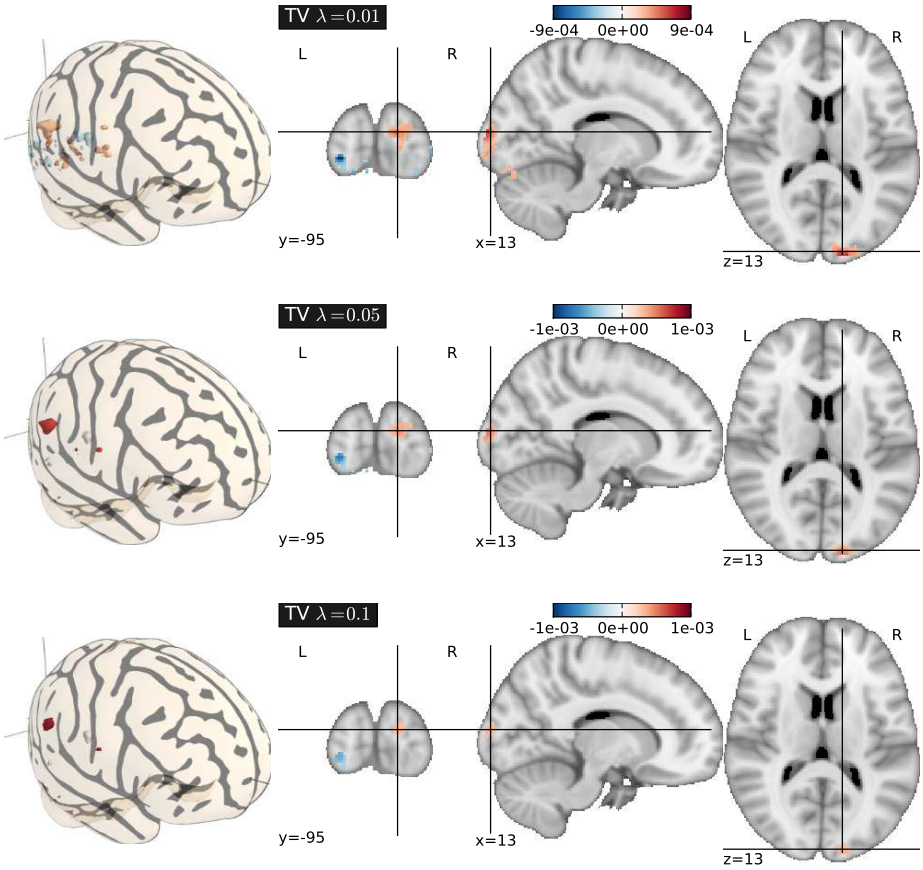


Fig. *xii* : Analyse inter-sujet de la représentation mentale de la taille des objets. Cartes de poids trouvées par la régression TV pour différentes valeurs du paramètre de régularisation. Quand la valeur de ce paramètre augmente, la régression TV créé différents groupes de poids avec des valeurs constantes. Ces groupes sont facilement interprétables.

Contributions logicielles

Au cours de cette thèse, nous avons contribué au *Scikit-learn* , une librairie logicielle en Python.

http://scikit-learn.sourceforge.net/

Conclusion

Dans cette thèse, nous avons présenté différentes contributions pour l'inférence inverse en IRMf. Cette approche, qui repose sur le concept de reconnaissance de formes, peut être utilisée pour décoder l'activité cérébrale, et plus précisément retrouver l'organisation spatiale du codage neuronal, à partir d'images du cerveau.

Contributions expérimentales - De nombreux algorithmes d'apprentissage statistique peuvent être utilisés pour la prédiction et la réduction de dimensions. Dans ce manuscrit, nous avons détaillé les algorithmes de l'état de l'art, et nous les avons implémentés et testés sur des données réelles. Nous avons de plus systématiquement étudié leurs performances dans le cas de leur utilisation sur des données d'IRMf. Cette étude rigoureuse nous a permis d'établir les pré-requis pour des algorithmes d'apprentissage statistique efficaces dans le cas d'utilisation qui nous intéresse.

Des études sur des données expérimentales ont aussi été réalisées en collaboration avec des neuroscientifiques, et nous avons obtenu des résultats de prédiction significatifs, dans des domaines d'applications variés, tels que la représentation mentale des quantités ou des préférences, ainsi que dans le cas plus complexe du recyclage cortical des fonctions cognitives de bas niveaux.

Contributions méthodologiques - Nos recherches se sont focalisées sur des méthodes améliorant l'interprétation des résultats d'inférence inverse :

- Une première contribution est une approche Bayésienne de régularisation parcimonieuse, appelée *Multi-Class Sparse Bayesian Regression – MCBR*. Cette approche est une généralisation des deux approches principales que sont *Bayesian Ridge Regression* et *Automatic Relevance Determination*.
- Un second axe de recherche a été motivé par le fait que les données d'IRMf ont une structure spatiale qui est rarement prise en compte dans les différentes méthodes de l'état de l'art. Nous avons donc proposé une approche, appelée *coalescence supervisée (supervised clustering)*, qui inclut l'information spatiale dans le modèle de prédiction, et permet d'obtenir des cartes de poids ayant une structure en composantes connexes. Cette méthode peut être utilisée avec n'importe quelle fonction de prédiction, et pour des données de très grande dimension.
- Notre dernière contribution a visé à implémenter la parcimonie et l'information spatiale au sein d'un même modèle. Nous avons proposé l'utilisation de la régularisation par Variation Totale pour des tâches de prédiction, et nous avons montré les bonnes performances de cette approche pour l'analyse de données d'IRMf.

Ces différentes approches ont été testées sur des données réelles, en rapport avec l'étude de la représentation mentale des tailles et des formes des objets, et permettent d'obtenir des cartes utilisables pour décoder certaines parties du système visuel.

Contributions logicielles - En plus des directions expérimentales et méthodologiques que nous avons décrites dans cette thèse, nous nous sommes aussi intéressé à l'implémentation des algorithmes étudiés et détaillés dans ce manuscrit. Une implémentation de bonne qualité est critique, à cause de la grande dimension des données d'IRMf. Nous avons aussi contribué au *Scikit-learn*, une librairie d'apprentissage statistique libre. Dans ces développements, nous nous sommes plus spécifiquement impliqués dans les modèles génératifs (*GNB*), les méthodes de *réduction de dimension* (*sélection d'attributs univariée, RFE*), les méthodes de *sélection de modèle*, et les *régularisations Bayésiennes*.

Perspectives de recherche

Analyses intra-sujet et inter-sujet : le point de vue de l'apprentissage statistique

Les performances en prédiction des méthodes de référence, ainsi que des méthodes proposées dans cette thèse, sont données dans le tableau *i* pour une analyse intra-sujet, et dans le tableau *ii* pour une analyse inter-sujet. Nous pouvons remarquer que les méthodes ont des niveaux de performance différents pour les analyses intra ou inter sujet. Cette variabilité peut être expliquée par la différence d'organisation spatiale du codage neuronal entre les deux expériences. En effet, l'organisation spatiale peut être, au niveau d'un seul sujet, très parcimonieuse et avec une organisation très fine, mais peut être très étendue dans le cas d'analyses inter sujet, à cause de la variabilité entre individus. Ainsi, les approches qui favorisent la parcimonie semblent être mieux adaptées à l'analyse intra-sujet, alors que les approches qui se basent sur des groupes de voxels semblent être mieux adaptées aux analyses inter-sujet. En conclusion, il y a beaucoup d'intérêt pour les méthodes qui peuvent adapter leur niveau de parcimonie aux données.

Extensions Une extension possible est l'addition d'une régularisation par norme ℓ_1 à la régularisation par Variation Totale, ce qui amène au problème de minimisation suivant :

$$\hat{\mathbf{w}}^l = \underset{\mathbf{w},b}{\operatorname{argmin}} \ \ell(\mathbf{w}) + \lambda_1 \|\mathbf{w}\|^1 + \lambda_2 TV(\mathbf{w}) \ \ , \ \ \lambda_1 \geq 0 \ , \ \lambda_2 \geq 0$$

En optimisant les deux paramètres λ_1 et λ_2 par validation croisée interne, nous pouvons adapter le modèle entre une régularisation parcimonieuse (ℓ_1) ou une régularisation générant plutôt des groupes d'attributs connexes (TV) Le

problème définie dans l'équation précédente est très similaire à celui du *smooth Lasso* [Hebiri 10] qui est basé sur la norme ℓ_2 du gradient. Cependant, la régularisation *TV* est plus adaptée pour l'extraction de groupes d'attributs, car elle pénalise la norme ℓ_1 du gradient. Une autre perspective intéressante serait de considérer au sein d'une même approche l'information intra et inter sujet, en utilisant des régularisations structurées se basant sur des normes mixtes (*e.g. group Lasso* [Yuan 06, Bach 08]).

Méthodes	Moyenne ζ	Dév. strd. ζ	max ζ	min ζ	p-val/VB-MCBR
SVR	0.82	0.07	0.9	0.67	0.0003 **
Elastic net	0.9	0.02	0.93	0.85	0.0002 **
BRR	0.92	0.02	0.96	0.88	0.0011 ***
ARD	0.89	0.03	0.95	0.85	0.0003 **
Gibbs-MCBR	0.93	0.01	0.95	0.92	0.0099 **
VB-MCBR	0.94	0.01	0.96	0.92	-
TV $\lambda = 0.05$	0.92	0.02	0.95	0.88	0.0002 **

Tab.*i* : Représentation mentale de la taille - Analyse **intra-sujet**. Variance expliquée ζ pour les différentes méthodes utilisées dans cette thèse. Les p-valeurs sont calculées en utilisant un test t apparié.

Méthodes	Moyenne ζ	Dév. strd. ζ	max ζ	min ζ	p-val/TV
SVR	0.77	0.11	0.97	0.58	0.0277 *
Elastic net	0.78	0.1	0.97	0.65	0.0405 *
BRR	0.72	0.1	0.94	0.6	0.0008 **
ARD	0.52	0.33	0.93	-0.28	0.0085 **
Gibbs-MCBR	0.79	0.1	0.97	0.62	0.0289 *
VB-MCBR	0.78	0.1	0.97	0.65	0.0151 *
SC - BRR	0.82	0.08	0.93	0.7	0.5816
TV $\lambda = 0.05$	0.84	0.07	0.97	0.72	-

Tab.*ii* : Représentation mentale de la taille - Analyse **inter-sujet**. Variance expliquée ζ pour les différentes méthodes utilisées dans cette thèse. Les p-valeurs sont calculées en utilisant un test t apparié.

Approches Bayésiennes et approches discriminantes classiques

Les méthodes présentées dans cette thèse peuvent être grossièrement classifiées en deux groupes : les approches Bayésiennes (*e.g. Bayesian Ridge Regression*, *Automatic Relevance Determination*, *Multi-Class Sparse Bayesian Regression*) et les approches discriminantes (*e.g. Lasso*, *Elastic net*, *SVC*, la régularisation par *Variation Totale*). En terme de temps de calcul, les approches Bayésiennes ne sont pas particulièrement efficaces, comparées aux approches discriminantes. Bien que le cadre Bayésien permette d'adapter automatiquement les paramètres

du modèle aux données, cela à un coût en termes de temps de calcul. Ce coût est souvent plus élevé que celui d'une validation croisée interne permettant d'optimiser les paramètres des approches discriminantes.

Une exception est la coalescence supervisée, où *Bayesian Ridge Regression* est bien adaptée pour adapter la régularisation à la complexité variable du problème lorsque l'on parcourt l'arbre. En effet, dans ce cas bien spécifique, le niveau de parcimonie peut varier entre les différents niveaux de l'arbre, et *Bayesian Ridge Regression* adapte finement la régularisation au niveau de parcimonie spécifique lors de chaque coupe de l'arbre.

En terme de performance en prédiction, les deux types d'approches sont similaires, avec un léger avantage pour les approches Bayésiennes dans le cas des analyses intra-sujet. En effet, de telles approches peuvent plus précisément adapter la régularisation à l'organisation spatiale très fine du codage neuronal spécifique à un sujet. Dans le cas de l'analyse inter-sujet, les approches discriminantes sont légèrement meilleures, car le choix de leurs paramètres par validation croisée interne est moins sensible au sur-apprentissage d'un ensemble particulier de sujets.

En conclusion, les approches discriminantes cherchent seulement à réaliser une prédiction correcte, alors que les méthodes Bayésiennes peuvent être utilisées pour construire des modèles plus interprétables qui tiennent compte de différentes hypothèses sur les données d'IRMf. Les modèles Bayésiens qui sont couramment utilisés ne sont pas spécifiques à l'analyse de données d'IRMf, et, en ce sens, il peut être intéressant de tenir compte au sein du modèle de certaines hypothèses sur le codage neuronal. Par exemple, dans le modèle MCBR, nous faisons l'hypothèse d'un codage par populations (i.e. différents groupes de voxels sont impliqués dans le codage) (voir aussi [Friston 08]). Donc, les approches Bayésiennes semblent prometteuses car elles peuvent plus facilement tenir compte d'information a-priori sur les données d'IRMf, par rapport aux approches discriminantes.

Extensions Les *processus Gaussiens (GPs)* [Rasmussen 05] ont été utilisés avec succès en IRMf [Marquand 10], et permettent de tenir compte d'a-priori complexes sur la moyenne et la covariance des poids. En particulier, ils peuvent être utilisés pour introduire l'information spatiale dans les modèles Bayésiens, d'une façon similaire à [Friston 08]. Une autre extension peut être de considérer la construction des cartes d'activations et le modèle prédictif, au sein d'une même approche. La méthode de *détection-estimation conjointe* développée dans [Vincent 07] ou le modèle hiérarchique proposé dans [Lashkari 10] sont deux alternatives intéressantes pour combiner l'identification des motifs fonctionnels et leurs utilisation pour la prédiction. Afin de tenir compte de l'information spatiale dans une approche purement Bayésienne, les *Champs de Markov aléatoires (MRFs)* sont aussi une approche prometteuse (voir [Ou 10] pour un exemple d'utilisation de l'information anatomique).

Information spatiale et *agglomération d'attributs*

Dans cette thèse, nous développons la notion d'agglomération d'attributs, et montrons qu'inclure l'information spatiale au sein d'une analyse au niveau des voxels, comme avec la régularisation par Variation Totale, ou en créant des structures intermédiaires telles que les parcelles, permet d'obtenir des résultats à la fois précis et interprétables en inférence inverse. Il y a donc un grand intérêt à utiliser l'information spatiale dans l'inférence inverse, et nous pensons que c'est une piste intéressante pour le développement de l'apprentissage statistique en neuro-imagerie.

Nous avons introduit la coalescence supervisée, qui est une approche particulièrement efficace dans les analyses inter-sujet, et qui est plus performante que les approches de l'état de l'art. Plus généralement, cette méthode n'est pas restreinte aux images cérébrales, et peut être utilisée avec n'importe quel jeux de données où la structure spatiale multi-échelle est considérée comme pertinente (*e.g.* images médicales ou satellitaires). De plus, un tel algorithme est bien adapté à la construction d'atlas anatomo-fonctionnel, car il considère de manière conjointe l'information spatiale et l'information fonctionnelle.

Extensions Parmi plusieurs extensions possibles, il peut être intéressant de développer une approche similaire aux *Forets aléatoires* [Breiman 01], en agglomérant différents arbres de parcellisations créés par *bootstrap* sur l'ensemble d'apprentissage. Les parcellisations résultantes sont combinées en moyennant leurs poids ou leurs prédictions. Des études préliminaires ont montrées un accroissement des performances en prédiction, mais certains travaux semblent nécessaires afin de préserver la structure spatiale des parcelles. De plus, une limitation majeure de l'algorithme de coalescence supervisée, est qu'il repose sur une exploration gloutonne de l'arbre, et l'optimalité n'est pas assurée. Afin d'améliorer cet aspect, il est possible d'introduire la structure hiérarchique de l'arbre au sein d'un problème d'optimisation convexe, par exemple en s'inspirant des travaux détaillés dans [Jenatton 10].

Contents

List of Figures

List of Tables

Introduction

Context

Many research fields aim at understanding some components of the human cognition, such as linguistic, social behavior, memory or sensory-motor interactions, and are referred to as *cognitive neurosciences*. The multi-scale nature of the nervous system (from the synapse – connection between neurons – 2 to 40 nm, to the brain – 150 mm for the longitudinal axis –), its extremely complex structure (up to 10^{15} connections) and the difficulty of in-vivo imaging, render this study very challenging. Cognitive neurosciences are thus based on a wide range of tools such as psychophysics, computer modeling and *neuroimaging*. One of the aims of *neuroimaging* is to provide a cartography of the functional regions of the brain and their respective relationships. This includes the study of the *neural code*, which is the internal representation of any given cognitive parameter within the brain, and thus, understanding this coding, *i.e. decoding* it, is particularly important in cognitive neurosciences.

The neural code is extremely rich and complex [Dayan 01], and its characterization principally relies on interactions between the different entities implied in this coding, called *coding entities*, and the spatial distribution of these entities. Two major schemes of coding have been proposed for neural coding: *sparse coding*, where very few coding entities are implied (theoretically only one), and *population coding*, where coding is performed by many coding entities. Another aspect of neural coding is the spatial distribution of the coding entities, that can be *clustered* when these coding entities are grouped together in compact regions of the brain, or *distributed*, when the coding entities are spread without any emerging structure.

Functional brain imaging (or *Neuroimaging*) provides a unique opportunity to study brain functional architecture, while being minimally invasive, and is thus well-suited for the challenging study of the spatial layout of neural coding. Different modalities exist, each one having specific spatial and temporal resolutions; among them *Functional Magnetic Resonance Imaging (fMRI)* has emerged as a fundamental modality for brain imaging. Over the last two decades, *fMRI* [Ogawa 90b, Ogawa 90a] has been widely used for brain imaging, and has become a reference method for neuroscientific studies, due to its good spatial resolution. *fMRI* consists in measuring the oxygenation of the blood by *Magnetic Resonance Imaging*, using a specific contrast called *Blood Oxygenation Level-Dependent (BOLD)* contrast. When some neural populations are active, the increase of deoxyhemoglobin ratio in the blood increases the *BOLD* contrast, and thus provides access to images of brain activity.

fMRI images are pre-processed, and modeled through a *General Linear Model*, that takes into account the different experimental conditions and the dynamics of the hemodynamic response in the *design matrix*. The resulting model parameters, *a.k.a. activation maps*, represent the influence of the different experimental conditions on local *fMRI* signals. The classical and widely used

approach for analyzing these activation maps is called *classical inference*, and relies on a mass-univariate statistical tests (one for each voxel), yielding the so-called *Statistical Parametric Maps (SPMs)* [Friston 95]. Such maps are of particular interest in neurosciences, as they open the door to localizing the voxels that are significantly active for any combination of experimental conditions, and thus are probably implied in the underlying neural code of the cognitive processes. However, this classical inference suffers from multiple comparisons issues, and does not take into account the multivariate structure of the *fMRI* data.

A recent approach, called *inverse inference* (or *"brain-reading"*) [Dehaene 98, Cox 03], has been proposed in order to cope with the limitations of the classical inference. Inverse inference relies on a *pattern recognition* framework, and aims at decoding the neural code by using *statistical learning* methods. Based on a set of activation maps, inverse inference builds a prediction function that can be used for predicting a behavioral target for a new set of images. The resulting prediction accuracy is a measure of the quantity of information about the cognitive task shared by the voxels. This approach is multivariate, and can provide more sensitive analysis than standard *statistical parametric mapping* procedure [Kamitani 05, Haynes 06]. Many methods have been tested for classification or regression of activation images (*Linear Discriminant Analysis, Support Vector Machines, Lasso, Elastic net regression, and many others*), but, in this problem, the major bottleneck remains the localization of predictive regions within the brain volume. Additionally, we have to deal with the curse of dimensionality, as the number of *features* (voxels, regions) is much larger ($\sim 10^5$) than the numbers of samples (images) ($\sim 10^2$), and thus the prediction method may overfit the training set and thus not generalize well to new samples.

The overall aim of this thesis is the development of statistical learning methods that take into account the characteristics of *fMRI* data, and that can be used for *inverse inference*. From an experimental point of view, we particularly focus on the understanding of the human visual cortex, but the presented framework can be used to study any brain system.

Organization and contributions of this thesis

Chapter 1 - Accessing the neural code

In the first chapter, we describe the functional organization of the human brain, and detail the notion of *neural code*. We focus on the different spatial distributions of the entities implied in coding, and we detail the *functional Magnetic Resonance Imaging* modality, that is well-suited for retrieving some large-scale features of neural coding.

Chapter 2 - From *fMRI* acquisitions to "Brain-Reading"

In the second chapter, we detail the different pre-processing steps required for *fMRI* data analysis. We introduce the *General Linear Model* that constructs a

set of activations maps from the data, given a description of the experimental paradigm and some physiological priors in the *BOLD* signal. Then, we develop the classical approach for exploiting these activations maps, called classical inference. This method relies on *mass-univariate statistics* within the whole brain, and is thus fast and easy to implement, but suffers from important drawbacks for *decoding*.

In a second part, we introduce the *inverse inference* framework, that is based on a *pattern recognition* approach. We present some specific uses of *inverse inference* in neurosciences, and how this approach can be used for *decoding*.

Chapter 3 - Statistical learning for *fMRI inverse inference*

In this chapter, we detail the key concepts of *statistical learning* that are used in the inverse inference framework, and we explain the different bottlenecks related to the characteristics of *fMRI* data, in particular the high dimensionality of the data.

In a second section, we introduce "historical" solutions based on *Support Vector Machine* and *Discriminant Analysis*, and we explain why such methods are not necessarily well-suited for the inverse inference framework. Then, we detail another approach for dealing with the high dimensionality of *fMRI* data, called *regularization*. We show that regularization-based methods perform well in our case, and we detail *Bayesian* frameworks that tune automatically the amount of regularization, based on the data. Finally, we discuss the role of *dimension reduction*, that deals with the curse of dimensionality issue associated with decoding problems by reducing the number of features before learning a prediction function.

We conclude by giving the requirements of a good statistical learning approach for *fMRI* inverse inference: it should be *multivariate*, *multi-scale*, and should take into account the *spatial structure* of the data.

Chapter 4 - Multi-Class Sparse Bayesian Regression

Bayesian regularizations are attractive for inverse inference but the classical approaches have to be adapted to the characteristics of *fMRI* data. In this chapter, we propose the first contribution of this thesis, a Bayesian framework for sparse regularization, that generalizes previous Bayesian regularizations. Based on a multi-class model, it creates a clustering of the features based on their relevance in the prediction, yields an adaptive regularization, and effectively copes with the limitations of other Bayesian regularization techniques.

Chapter 5 - Supervised clustering

A major drawback of state-of-the-art approaches in *inverse inference* is that they do not use *spatial information*. Due to the metabolic processes underlying the *fMRI* signal and the intrinsic structure of at least some neural codes, there is a local filtering of information that should be taken into account. This can be

done by *feature agglomeration*, that averages the signal in neighboring voxels to create intermediate structure called *parcels*.

In this chapter we describe the second contribution of this thesis, namely the *supervised clustering* approach. This method uses the image structure of the data through a spatially constrained hierarchical clustering. Additionally, we adapt this clustering to the predictive task. Thus the proposed approach yields interpretable maps and outperforms reference methods in inter-subjects analysis.

Chapter 6 - *Total variation* regularization

Based on previous studies and on the results obtained in this thesis, we conclude that both *regularization* and *spatial information* are important for improving both interpretability of the resulting maps and prediction accuracy. Based on this information, one can extract a correct model of the neural code.

In this chapter, we describe the last contribution of this thesis, that introduce spatial information into a generic regularization framework. We implement the *Total Variation* regularization (previously developed for image denoising) in a predictive framework, both in regression and classification settings. This approach outperforms reference methods, while extracting few interpretable clusters from the data.

Appendices

Appendix A - A short introduction to Magnetic Resonance Imaging

In this appendix, we briefly explain the physical basis of *Magnetic Resonance Imaging (MRI)*.

Appendix B - Description of the data sets

In this appendix, we describe the simulated and real data sets that are used in this thesis.

Appendix C - Scikit-learn for fMRI inverse inference

In this appendix, we describe how one can use the *Scikit-learn* for *fMRI inverse inference* and we give the principle functions that can be used for such analysis.

Software contributions

During this thesis, we have contributed to the *Scikit-learn*, a Python module integrating mainstream machine learning algorithms with a uniform API. It is open-source and aims at providing simple and efficient solutions to learning problems.

http://scikit-learn.sourceforge.net/

1

Accessing the neural code

In this thesis, we focus on retrieving the *spatial layout* of the *neural code*, *i.e.* the specific tuning of the brain tissues implied in this coding, and their spatial localization. In this chapter, we first give an overview of the functional structure of the brain. Then, we give some definitions about neural coding, and we describe the different possible distributions of the coding entities. More specifically, we focus on the notions of *Sparse coding* and *Population coding*. Finally, we briefly describe the principles and use of *fMRI* acquisitions, and how this modality can be used for functional imaging.

Contents

1.1 Brain functional architecture

In this section, we briefly introduce the neural basis of cognition, and the brain functional architecture.

1.1.1 Overview of the human nervous system

Neural cells

The human brain has a volume of around $1200\ cm^3$, and is roughly constituted by two types of cells: neurons (about 10^{11}) (see Fig. 1.1) that are responsible for information processing, and glia cells (4 times more than neurons) that are responsible for the structural and metabolic support of neurons. 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. Cells are connected through junctions called *synapses* (up to 10^4 by neuron), which transmit information using a chemical pathway (the release of *neurotransmitter*).

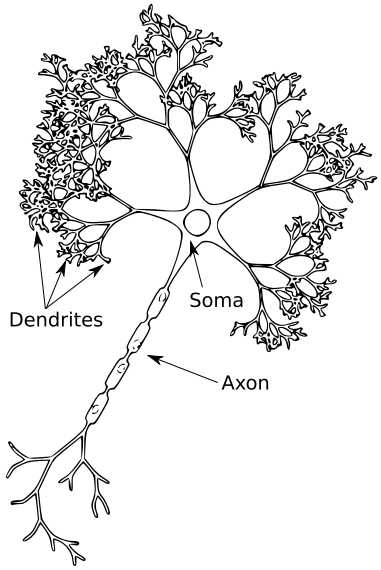


Figure 1.1: A neuron has a cell body (called the *soma*), many regions for receiving information from other neural cells (called *dendrites*), and often an *axon* (*nerve fiber*) for transmitting information to other cells (an *axon* can be longer than 1 meter in humans) The information in the *axon* is transmitted through an electrical signal called *action potential*, which is based on the electrical properties of the neuronal membrane. Adapted from http://commons.wikimedia.org/.

Human brain

The human brain is the center of the human nervous system, and is located in the cranium, protected by three membranes called *meninges*. It is constituted of different regions, in particular the two *cerebral hemispheres* (also called *telencephalon*) which are widely studied in neuroimaging (see Fig. 1.2), as they are responsible of the performance of a majority of the cognitive tasks. There are

also two important circulatory systems in the brain: one for the *cerebrospinal fluid – CSF –* (support and protection of neural cells), and one for the blood (supply in oxygen).

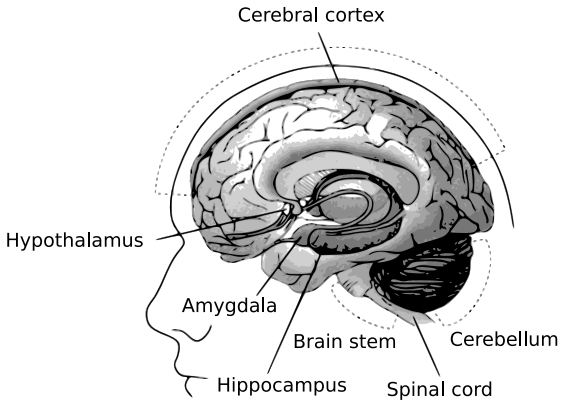


Figure 1.2: Different parts of the human brain.

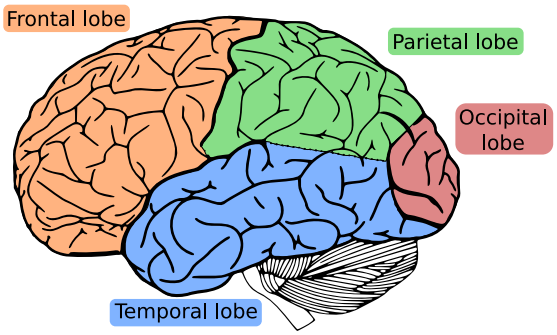


Figure 1.3: The four different lobes of the *cerebral cortex*. Adapted from http://agaudi.files.wordpress.com/.

The two *cerebral hemispheres* are the largest part of the human brain which 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 hemispheres is a highly circonvoluted 6-layered structure called *neocortex* (or more simply *cerebral cortex*). A cortical fold is called *sulcus*, and the area between two *sulci* is called a *gyrus*. The *cerebral cortex* can be decomposed in the left and right hemispheres, which can themselves be decomposed in four different lobes: *frontal lobe, parietal lobe, occipital lobe and temporal lobe* (see Fig. 1.3).

1.1.2 Functional regions of the human brain

The human brain can be decomposed in different functional regions which correspond to different part of the information processing within the brain (see Fig. 1.4). These functional regions roughly correspond to anatomical regions, and can be categorized in three general categories: 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*).

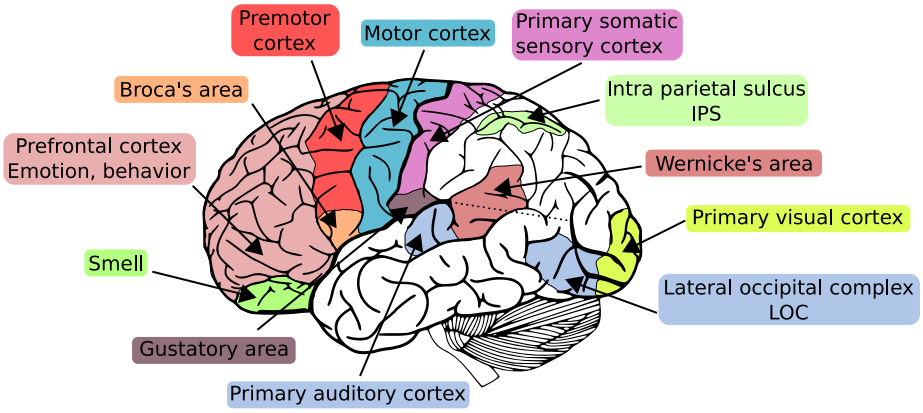


Figure 1.4: The main functional regions of the human brain (left hemisphere), and the two regions which are studied in this thesis (*LOC* and *IPS*). Adapted from http://agaudi.files.wordpress.com/.

1.2 Neural coding of mental processes

The *neural coding* is the correspondence between a stimulus and its representation by individual or ensemble of neuronal responses. Accessing this neural coding may be helpful for understanding the mental processes, and more generally, the way in which the brain processes information. As stated by Perkel [Perkel 68], the problem of neural coding is to elucidate:

> "[...] the representation and transformation of information in the nervous system. "

The study of neural coding can be carried on different scales or structures (from single neuron to large population of neurons such as cortical columns

with about 10^4 neurons), and we refer to these structures as *coding entities* or *neuronal populations*. Additionally, this study can be done by *decoding*. Decoding refers to the reconstruction of a stimulus, or certain aspects of that stimulus, from the signal it evokes in different coding entities.

In this thesis, we focus on the notion of *spatial layout* of neural coding, and we address two fundamental questions; the question of the specific tuning of the neuronal populations involved in a cognitive task (*Sparse coding* or *Population coding*); and the question of the spatial distribution (*Distributed coding* or *Clustered coding*) [Dayan 01] of these neural populations within the whole brain (cerebral cortex, basal ganglia, thalamus, ...). Understanding these different coding schemes is crucial for cognitive studies, and is addressed in this thesis using functional neuroimaging.

In this section, we introduce the different notions required for studying some aspects of neural coding. After describing the difference between *Sparse coding* and *Population coding*, we detail both *Distributed coding* and *Clustered coding*.

1.2.1 *Sparse coding* and *Population coding*

The first component of neural coding is the definition of the entities that are involved in the coding. Two types of organization are usually observed in neural activity: *Sparse coding* and *Population coding*.

Sparse coding

Sparse coding refers to the idea that the cognitive information is coded by sparsely distributed neural populations, the majority of the remaining populations being inactive or having very low activity when the corresponding information is presented. The extreme case is the well-known "grandmother cell" (also called *gnostic neuron* [Konorski 67]), where a single neuron is believed to code for a very specific information (e.g. the notion of grandmother). Those gnostic neurons are organized in some specific areas of the cortex called *gnostic fields* (e.g. the extra-striate visual cortex for specific visual processings). An identical, but more biologically plausible, model has been proposed by Barlow [Barlow 72], with the notion of *cardinal cells*. Such cells carry all the relevant information on a cognitive task, the remaining populations of neurons adding little additional information. More details on the concept of gnostic cell can be found in the review of Gross [Gross 02].

The justifications of the gnostic neurons model and sparse coding, as exposed by Konorski [Konorski 67], are the following :

1. the increasing specificity of neurons towards V2 and V3 (regions of the visual cortex) until some expected shape-specific neurons [Hubel 62].

2. preliminary results on the visual impairments in visual cognition created by some lesions in monkeys [Ettlinger 68].

3. the agnosias following cortical lesions in humans, e.g. prosopagnosia linked with ventral temporal lesions.

4. the fact that the firing rate of a neuron coding for some information is limited by the available energy. The limitation of resources within the brain implies that only a few fractions of the whole neural population will have a high firing rate, yielding sparsity in the coding.

Such sparse coding has been observed in some very specific cognitive systems, *e.g.* : the olfactory system of the insects, where each odor is individually represented by very few neurons which responded by only two action potentials [Perez-Orive 02]; a more debated result is the existence of individual specific recognition cells, which implies that some neurons respond specifically to some representations of individuals [Quiroga 05].

Limitations of *Sparse coding*

The sparse coding yields a simple view of neural coding. It is easy to understand, easy to link a stimuli with a neuronal response, and easy to model. However, more recent works have argued again the notion of *gnostic cell*. Gross et al. [Gross 92] show that complex visual stimuli are encoded within a pattern of responses over a population of neurons within the inferior temporal cortex, rather than encoded in specific gnostic cells (see also [Desimone 91]). The cells known to be gnostic respond, even weakly, to a variety of individual faces. They are not narrowly selective for one and only one face (independently of size, orientation and color). Thus, they violate the strict definition of sparse coding as exposed by Barlow [Barlow 72]. Instead, each stimulus is encoded in weakly selective neurons : *e.g* the notion of grandmother is "encoded" in a specialized population of neighboring cells.

One of the main argument against the assumption of sparse coding, is the combinatorial cost of such a model. Indeed, there is a clear limit of sparse coding: if each concept is coded by a single neuron, it is easy to see the combinatorial cost raised by the coding of all the stimuli encountered by a human being during is life. This limitation is well illustrated by the concept of *yellow Volkswagen cell* [Harris 80]. If the notion of *yellow* and the notion of *Volkswagen* are encoded within two different gnostic neurons, is the notion of *yellow Volkswagen cell* encoded within a third neuron ? Such a highly combinatorial a approach is biologically very unlikely, and shows the limit of the concept of *gnostic cell*.

Population coding

New explanations for neural coding have thus been developed following the old theory of *Ensemble Coding* [Young 02] and is known as *Population coding*. The main idea behind population coding is that information is encoded within a large population of weakly selective neurons (*i.e.* which do not respond to

only one stimulus), forming some *patterns of activation*. All the neurons involved in the pattern have to be taken into account to decode the corresponding neural information. This coding is based on a *many-to-many* relationship between two types of representations (concepts and neurons) :

- each concept is represented by many neurons.
- each neuron participates in the representation of many concepts.

This coding is particularly robust to biological noise such as cellular death and variability of the neuronal response (see illustration Fig.1.6), as explained in [Pouget 00]. However, such type of functional organization implies that one given neuron will activate for different stimuli [Treisman 96], and thus the decoding of a stimulus from the neuronal coding requires more complex approaches. Different cognitive processes exhibit a population coding, as the regions specific to the recognition of faces in monkeys [Tsao 06], or the motor cortex, where movement direction is encoded by a population of neurons [Georgopoulos 86].

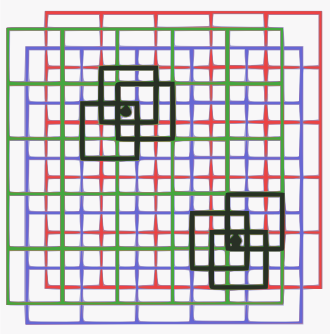


Figure 1.5: Illustration of *Sparse coding* and *Population coding* in a simple spatial discrimination experiment. The spatial position of the two dots can be encoded in a single 15×15 array (sparse coding). Each dot has two coordinates, and if one coordinate is wrong, the spatial position of the dot cannot be well approximated (error along one axis). The spatial combination can also be coded by three different 5×5 arrays (population coding). In this case, each dot has 2×3 coordinates, but, if a coordinate is wrong, the spatial position of one dot can still be well approximated (the dot is still at the intersection of the two others black arrays). Adapted from http://www.cs.toronto.edu/~hinton/.

1.2.2 *Clustered coding* and *distributed coding*.

Besides the notion of entities involved in the coding, it is useful to know how those entities are spatially distributed across the cerebral cortex (see Fig.1.6).

Spatial distribution

Two different types of spatial distribution of the coding entities have been defined :

- *Clustered coding*: The coding entities can be grouped into small clusters (e.g. cortical columns of the primary visual cortex). This is supported by some considerations on the minimization of the cost of connexions (such as axons, dendrites), which predicts that strongly interacting neurons should be close to each other (see [Chklovskii 04]).

- *Distributed coding*: The coding entities can also be widely distributed across the whole cortex. An example is given in [Haxby 01]: the shape of an object is coded within V4 (visual cortex) and the posterior inferotemporal cortex, by a large population of neurons that only code for simple shape features.

At the scale of a cortical region, clustered coding can be viewed as a sparse coding, as the information is encoded within a specific region of the cortex. In that case, the region involved in information coding is called a *gnostic region*, by analogy with the *gnostic neuron*. Similarly, at the same scale, distributed coding can be viewed as a population coding: the stimulus is encoded by a *pattern of active regions*.

Example of the visual system

As one of the cognitive paradigm that has been used in this thesis deals with visual object recognition, we briefly detail some notions of the coding process in the visual system. This system has been widely studied for several decades, but some aspects of its coding process still remain controversial. In particular, the existence of *gnostic regions* is still highly debated. The main regions that have been found within the brain (at the fMRI resolution) across several studies are the following :

- *Anterior Inferotemporal cortex* : specificity to body parts (a.k.a *Extrastriate Body Area - EBA*) [Downing 01]. The coordinates in the *MNI* space are $x = \pm 51\ mm$, $y = -71\ mm$ and $z = -4\ mm$.

- *Superior temporal sulcus* : specificity to faces (a.k.a *fusiform face area - FFA*) [Wada 01]. The coordinates in the *MNI* space are $x = \pm 44\ mm$, $y = -50\ mm$ and $z = -20\ mm$.

- *Human hippocampus* : specificity to particular places (a.k.a *Parahippocampal Place Area - PPA*) [Kreiman 00]. The coordinates in the *MNI* space are $x = \pm 30\ mm$, $y = -44\ mm$ and $z = -14\ mm$.

- *Left inferior temporal cortex* : specificity to processing of letter strings (a.k.a *Visual Word Form Area - VWFA*) [Cohen 00]. The coordinates in the *MNI* space are $x = -48\ mm$, $y = -60\ mm$ and $z = -16\ mm$.

The coordinates are given in the *MNI* space, that is described in section 2.1. The existence of the four gnostic regions can be explained by the familiarity of the presented stimulus in the sensory world (see [Op de Beeck 05]). Yet the existence of such specific regions involved in objects recognition is highly

debated. It seems that even if there exists some discriminative information for other visual categories outside the previously defined regions, this information is not sufficient for a normal perceptual prediction. In particular, see [Downing 06, Reddy 06] for studies showing the existence of gnostic regions, and [Haxby 01] for a study arguing for distributed coding.

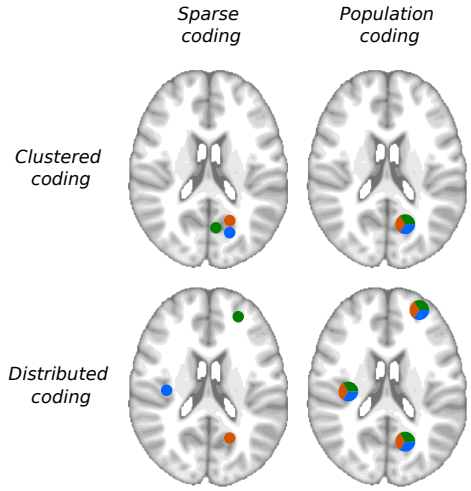


Figure 1.6: Illustration of the different types of entities involved in the neural coding, and the different spatial distributions of these entities. Each color corresponds to a condition, gnostic neurons are represented as disks with uniform color, and *patterns* of non-specific neurons are represented by disks with mixed colors. The two notions of entities and spatial distribution are clearly distinct. In the clustered coding case, coding entities are grouped into small clusters. In the distributed coding case, coding entities are widely spread across the whole cerebral cortex. Population coding defines a *pattern* of activation which has to be decoded, while sparse coding relies on very few active neurons.

1.3 Functional neuroimaging and *fMRI*

Functional neuroimaging (*a.k.a* functional brain imaging) aims at revealing brain physiological activity and its spatial distribution, and thus allow to study the spatial layout of the neural code. Different approaches (see Fig.1.7) can been used for functional neuroimaging. The aim of this section is to present the common modalities for functional imaging, and their characteristics in terms of spatial and temporal resolutions. Then, we detail *Functional Magnetic Resonance Imaging- fMRI*, which is the modality used in this thesis. Our focus on *fMRI* is driven by the good spatial resolution of this approach, as we aim at retrieving the spatial organization of the neural code.

1.3.1 Functional neuroimaging modalities

Electroencephalography - EEG

Electroencephalography – EEG – is a widely used modality for *in vivo* functional brain imaging. *EEG* measures the electrical activity of neurons, that can be recorded on the scalp. *EEG* is particularly used for the diagnosis of epilepsy. *EEG* signals are compared to the stimulus timing, and one can access fine temporal patterns of activation. However, 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

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* epilepsy) 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.

Magnetoencephalography - MEG

Magnetoencephalography – MEG – measures the magnetic field induced by neural electrical activity. The synchronized currents in neurons create magnetic fields (very weak, few hundreds of fT) that can be detected using specific devices (*SQUIDs*). As with *EEG*, the main challenge is to localize the sources of electric activity in the brain, and *MEG* has spatial resolution of a few millimeters to a few centimeters (a better precision can be obtained using *fMRI*). With a temporal resolution on the order of milliseconds, *MEG* is well-suited for recording the timing of the activity within the brain.

Positron emission tomography - PET

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* can 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. *PET* has two major limitations: the tracers required for *PET* are produced by cyclotrons (a type of particle accelerator), which implies an heavy logistic, and the use of radio-tracers is not harmless for the health of the subjects so that *PET* is now used for medical purpose only.

Single photon emission computed tomography - SPECT

Single photon emission computed tomography – SPECT – is similar to *PET*. 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 resolution is slightly worst 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

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 surface of the brain.

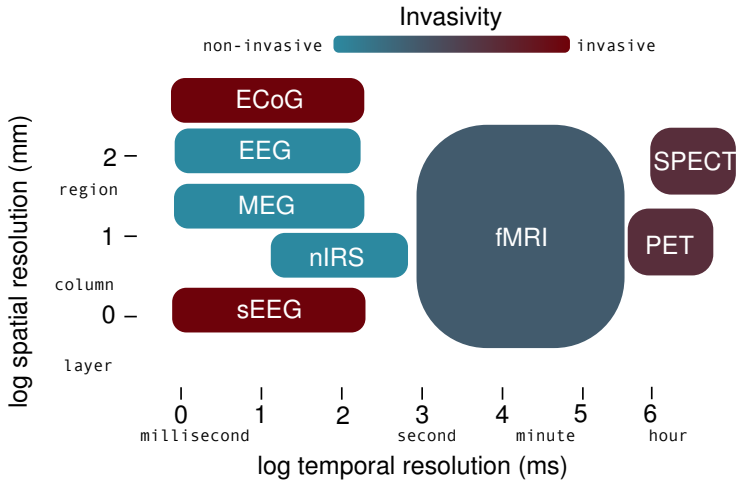


Figure 1.7: Spatial and temporal resolutions of the different modalities commonly used for functional imaging. In this thesis, we use *fMRI*.

1.3.2 *Functional Magnetic Resonance Imaging*

Functional MRI – fMRI – is a widely used method for functional brain imaging, because it is non-invasive, has a good spatial resolution (*1mm*), and provides access, albeit indirectly, to the neural activity. Moreover, in standard acquisitions, *fMRI* yields a full-brain coverage, which is useful for decoding, as it

does not restrict the study to superficial layers or predefined regions of the cortex and includes deep structures (*e.g.* basal ganglia) in the decoding. A short description of *Magnetic Resonance Imaging – or MRI –* working principles can be found in Appendix A. In this section, we recall the principle of *fMRI*, and we address the following points:

- How *MRI* is used for functional imaging.
- The link between *MRI* measurements and neural activity.
- The spatial and temporal characteristics of the functional signal.

Blood oxygenation level-dependent – BOLD – contrast

Ogawa et al. performed in 1990 [Ogawa 90b] the seminal experiment that would introduce the use of *MRI* as a functional imaging tool. The researchers were studying magnetic resonance of the protons in the brain of living rats, and noticed the existence of vertical rows, that correspond to some cerebral veins. The contrast is more accentuated in the case of an anoxic brain (lack of oxygen). This contrast is called *blood oxygenation level-dependent* (or *BOLD*), because it depends of the level of oxygenation of the blood. The *BOLD* contrast is observed through a gradient-echo *EPI* (*EchoPlanar Imaging*) sequence [Ogawa 90b, Ogawa 90a, Turner 91, Bandettini 92].

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. *Oxyhemoglobin* is diamagnetic (all its electrons are paired), and thus has no magnetic property. This has been explained by the particular distribution of electrons between oxygen and iron oxides (see [Pauling 36, Thulborn 82]). When the *oxyhemoglobin* loses its oxygen atoms and becomes the *deoxyhemoglobin*, it becomes paramagnetic (due to the iron oxides). The presence of *deoxyhemoglobin* in the blood modifies the *RMN* signal of the protons of the water molecules surrounding the blood vessels. Indeed, the difference of magnetic susceptibility between the blood vessel and the surrounding tissues, due to the paramagnetism of the blood, creates microscopic inhomogeneities in the magnetic field [Ogawa 90b, Thulborn 82]. Thus, the *BOLD* signal increases with the ratio oxyhemoglobin over deoxyhemoglobin.

BOLD contrast imaging

BOLD contrast relies on some physiological factors: it depends on the local equilibrium between the oxygen provided by the blood vessels, and the consumption of oxygen in brain tissue due to neural activity. In normal conditions,

the arterial blood is fully oxygenated and does not contribute to the *BOLD* contrast, but veins create low contrast regions, that explain the results found by Ogawa et al.

Based on these results, one can create an image of changes in *BOLD* contrast due to some given physiological events, and thus visualize, with a non-invasive method, the changes in blood oxygenation. The variations of *BOLD* contrast in the brain of a living rat during the inhalation of a gas that increases the *cerebral blood flow* (*CBF*), and thus blood oxygenation, have been studied in [Ogawa 90a] (see Fig. 1.8). These results show that *BOLD* contrast can be used to image blood oxygenation.

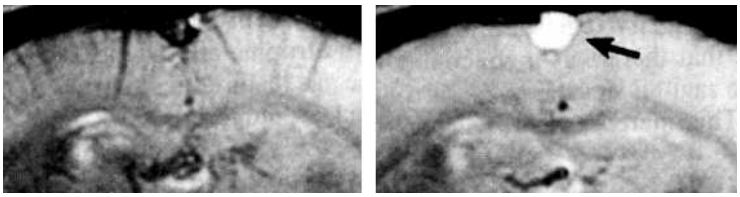


Figure 1.8: Illustration of the effect of the CO_2 on the *BOLD* contrast.
Left - Coronal slice showing the *BOLD* contrast of an anesthetized rat which has breathed pure O_2.
Right - Coronal slice of the same rat, showing the *BOLD* contrast after respiration of a mixture of 90% of O_2 and 10% of CO_2 (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. We can see a strong increase of intensity in this vein, that illustrates that the variation of blood oxygenation is visible in *BOLD* contrast. Adapted from [Ogawa 90a].

Functional MRI and hemodynamics

Some work has been done [Cooper 75, Frostig 90] to understand the correlation between changes in venous blood and increases in neuronal activity. In particular, the changes in blood flow in response to an increase in electrical activity has been studied (see [Frostig 90] among others). Using high resolution optical imaging, the authors were able to measure a coupling between neuronal activity and micro-circulatory responses. Moreover, they showed that the observed signal is probably linked to the supply in oxygen by the capillaries in response to the stimulus. Additionally, in [Bandettini 92], the authors show that, during a sensory stimulation, there is a local decoupling between the cerebral blood flow, and the cerebral metabolic rate of oxygen. An oversupply of oxygenated blood is delivered to the active region, which decreases the oxygen extraction fraction. This decrease implies a local increase in the average blood oxygen partial pressure, that can be observed using *BOLD* contrast.

Thus, *BOLD* contrast allows to visualize changes in hemodynamics that are

related to neural activity, and it is possible to use *BOLD* contrast for functional imaging. Indeed, it has been shown [Ogawa 92] that a visual stimulation creates an increase (easily detectable, $5 - 20\%$) in the intensity of the signal observed by *BOLD* contrast in *MRI* (see Fig. 1.9). Thus, by using changes in the oxyhemoglobin/deoxyhemoglobin ratio seen by *BOLD* contrast, it is possible to indirectly observe neural activity. This approach is called *fMRI* (*functional MRI*).

Some other approaches have been developed, using correlations between hemodynamics and neuronal activity. They are based on the direct mapping of changes in *Cerebral Blood Flow – CBF* [Kwong 92] (see [Fox 86] for correlation between neuronal activity and *CBF*), or changes in *Cerebral Blood Volume – CBV* [Mandeville 98]. However, approaches based on *CBV* require contrast agents that remain in vascularization during a long time, and approaches based on *CBF* have a limited covered volume. *BOLD* contrast, that relies on a complex combinations of *CBV* and *CBF*, does not suffer from these drawbacks.

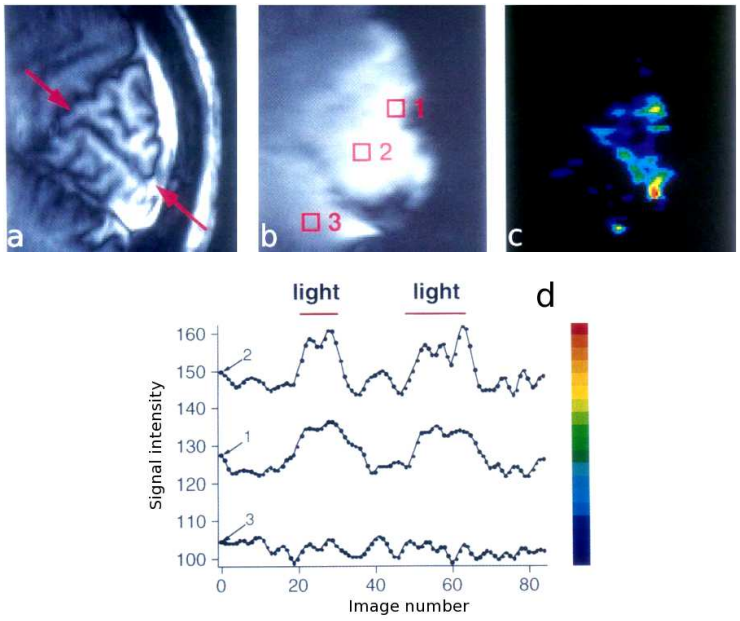


Figure 1.9: Illustration of the use of *BOLD* contrast for functional imaging. (a) Sagittal slice of an anatomical image, showing the occipital cortex. (b) Gradient echo image (i.e. *BOLD* contrast) for the same location. (c) Pseudo-color map of the difference in signal intensity between the mean of eights images acquired during visual (photic) stimulation, and eight images acquired in the dark. (d) Time course of signal-intensity changes (arbitrary units) for regions indicated by the three boxes in (b). Adapted from [Ogawa 92].

1.3.3 *fMRI* and neural coding

Based on *BOLD* contrast, *fMRI* allows to image neural activity related to a cognitive task, and thus, can be used to retrieve the spatial organization of a particular neural code. However, *BOLD* contrast is linked to neural activity through a complex, and still not fully understood, metabolic pathway, which strongly affects the conclusions that can be made from *fMRI* data. We detail here the different aspects of *fMRI* which are related to our study of neural coding.

Link to neuronal activity

In some experiments in monkeys and humans, *BOLD* contrast has been found to be more correlated with the input of neurons than with the output of neurons [Logothetis 01]. In that sense, *BOLD* contrast seems to reflect the input to a neuronal population as well as its intrinsic processing, not the outputs from that population. Moreover, it is important to note that the correlation between *BOLD* contrast and neuronal activity is not fully understood. Among other issues, an increase in *BOLD* signal can be due to a large activity of few neurons, or a small activity of a large population of neurons.

Moreover, one can not expect a better resolution than the millimeter (which is the scale of the cortical columns – about 10^4 neurons), by using *fMRI*, even if this modality has one of the highest spatial resolutions among imaging modalities. This points to a strong hypothesis, which is that neurons coding for the same cognitive task have to be grouped in the same region of the brain, with sufficient density to yield an *fMRI* signal. However, the study of neural coding can still be carried on neuronal populations rather that on single neurons, and *fMRI* is still the most well-suited functional neuroimaging modality for decoding various neural codes. From now, the smallest coding entity to be considered is the voxel (volumetric pixel, *i.e.* a $\sim 1.5mm \times 1.5mm \times 1.5mm$ cube).

Spatial resolution of *fMRI*

fMRI images can be used to extract information on the localization of the neural activity. However, due to the complex link between *BOLD* contrast and neural activity, *fMRI* suffers from some stringent limitations. In particular, it is necessary to assess the spatial correspondence between regions with a high *BOLD* contrast and regions of neural activity.

The spatial specificity of *fMRI* has been demonstrated by K. Ugurbil et al. [Ugurbil 03]. The authors have shown that *fMRI* images can not be considered as an accurate depiction of neural activity, and the precision of such maps depends on the spatial extension of the metabolic changes, and in particular:

- Blood flow and Vascularization: An increase in blood flow which is due to a strong neuronal activity, may exceed the region that contains active neurons. This create an intrinsic smoothing of the signal. The lack of spatial specificity can also be due to the vascularization, that is detected

by *MRI*. Indeed, large vessels have a larger contribution to *MRI* signal than small vessels, as the relative decrease in deoxyhemoglobin is higher in large vessels. As shown in [Iadecola 97], the dilatation of blood vessels by neural activity can propagate to distant blood vessels.

- Neuronal communication: The communication (*a.k.a.* synaptic activity) between neurons can be inhibitory, *i.e.* it can tend to shutdown neuronal activity. Thus, even if there is communication between neurons, the fact that this activity is inhibitory does not allow to see this neuronal processing through *BOLD* contrast.

These limitations can be illustrated by a simple experiment (see Fig. 1.10). *fMRI* images have been acquired in the visual area of a cat, using two different stimuli: parallel lines of two different orientations. *fMRI* fails to depict the expected columnar organization, as the spatial resolution of the map is between $3-5mm$. This can be explained by the fact that the changes in deoxyhemoglobin that coincide spatially with neuronal activity, do not remain confined to active regions, but instead, propagate to the larger vessels. They will incorrectly be observed as activation, even far from the initial places of neuronal activation. In this thesis, we study a specific framework for *fMRI* data analysis, called *inverse inference*, that can retrieve information about neural coding, even if the pattern of activation is smoothed by the spatial extent of *fMRI* signal, as in the described experiment.

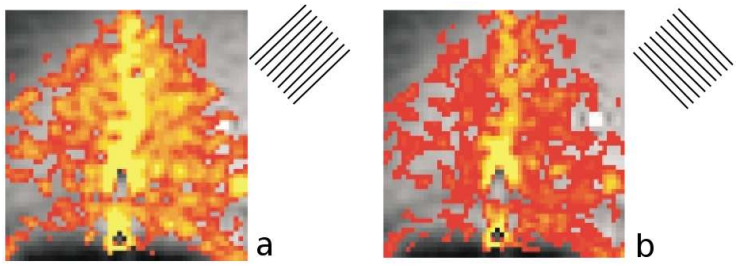


Figure 1.10: Illustration of the limited spatial resolution of *fMRI* images. (a) and (b): Functional images acquired during two different stimuli corresponding to two different orientations of parallel lines, on the visual area of the cat. *fMRI* fails to depict the expected columnar organization. Adapted from [Ugurbil 03].

Temporal resolution of *fMRI* and *Hemodynamic Response Function - HRF*

The function representing *fMRI* signal across time, due to a temporal increase of neural activity, is called *Hemodynamic Response Function* (or *HRF*). This function can be decomposed in different steps for a total duration of $20-25$ seconds:

- $1-6$ seconds: increase of *BOLD* signal, due to a huge increase in deoxygenated blood (consumption by active neurons). It reaches a maximum after 6s.

- $6-12$ seconds: increase of oxygenated blood will be larger than increase in deoxygenated blood. The decrease of the ratio of *deoxyhemoglobin* to *oxyhemoglobin* in the blood induces a decrease of *MRI* signal intensity.
- $12-20$ seconds: slow return to the baseline, after a small undershoot.

As we can see, the dynamic of the *HRF* is much slower than the dynamic of neural activity (on the order of few ms). Thus, *fMRI* has a poor temporal resolution [Horwitz 00]. This limitation can not be overcome, except by imaging a more direct measure of the neural activity than the *BOLD* contrast which is based on hemodynamics. However, in this thesis, we consider experiments for which temporal information is not crucial, and thus, we do not focus on the temporal precision of *fMRI*.

A model for *HRF* has been proposed [Glover 99], that has been successfully used in many experiments. The *HRF* h can be approximated by the following model :

$$h(t) = I_{t>0}\left\{\left(\frac{t_1}{a_1}\right)^{a_1}\exp\left(a_1-\frac{t}{b_1}\right)-c\left(\frac{t}{a_2 b_2}\right)^{a_2}\exp\left(a_2-\frac{t_2}{b_2}\right)\right\} \quad (1.1)$$

with $a_1 = 6$, $a_2 = 12$, $b_1 = 0.9$ $b_2 = 0.9$ and $c = 0.35$. A simulation is given Fig.1.11. It is interesting to notice that the proposed model for *HRF* can be seen as a low-pass filter. It keeps the low-frequency components of the underlying metabolic activity. Some more complicated models have been proposed to better deal with the different underlying processes of the *BOLD* effect; *Balloon model* [Buxton 98], and *Joint detection estimation* [Ciuciu 03]. In the last model, the authors take into account the fact that the *HRF* can vary spatially across different regions of the brain, and introduce an adaptive model of *HRF*.

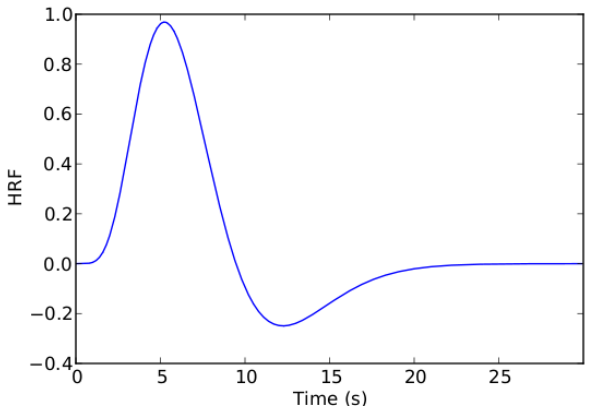


Figure 1.11: Model of the *HRF*, according to (1.1). The proposed model depicts the three important steps of the *HRF*: $1-6$s, increase of signal, $6-12$s, decrease of signal, $12-20$s return to baseline with undershoot.

1.4 Conclusion - Accessing the neural code

This chapter has introduced the complex functional architecture of the brain. In order to better assess cognitive hypothesis, it is necessary to extract reliable

support of the functional areas of the brain. Thus, accessing the spatial organization of neural coding is necessary for understanding how the brain processes information. This spatial organization can be characterized by two main factors: the entities involved in the code (*Sparse coding* or *Population coding*), and the spatial distribution of these entities (*Clustered coding* and *Distributed coding*). In general, population coding seems a more plausible hypothesis than sparse coding, but the superiority of one of the two spatial models is more controversial.

Among many different neuroimaging modalities, *MRI* has a good spatial resolution which is crucial in many cognitive neuroscience experiments reported in this thesis. *MRI* can be used with a specific contrast, called *BOLD contrast*, for functional neuroimaging (*a.k.a. fMRI*), and this modality measures a function that depends on the rate of oxyhemoglobin versus deoxyhemoglobin in the blood. When some neural populations are active, the increase of deoxyhemoglobin ratio in the blood increases the *BOLD contrast*. However, the observed effect is not directly created by the neural activity (it relies on a complex metabolic pathway), and thus, the *fMRI* signal may reflect activity far from the activated neurons. Despite these limitations, *fMRI* is currently the best modality for accurate inference and for studying the neural coding.

2

From *fMRI* acquisitions to "Brain-Reading"

In the previous chapter, we have seen that *functional MRI* can be used for functional imaging, and has a good spatial resolution for a whole brain coverage modality. After preprocessings and modeling of *fMRI* data, *classical inference* is the reference method for studying functional images, but this approach suffers from some limitations that we will discuss in this chapter. In the early 2000's, a new methodology for studying functional images has been developed, called *inverse inference* [Dehaene 98, Haxby 01, Cox 03]. The main idea is to use the functional images to predict a behavioral variable, based on statistical learning methods. This approach can be used to *decode* the *neural population* involved in a given cognitive function.
In this chapter, we first present the preprocessings and *General Linear Model* used for the statistical modeling of *fMRI* data. Then, we introduce *classical inference*, and finally, we explain the notion of *inverse inference* and detail the corresponding framework and its benefits.

Contents

2.1 Preprocessings and *fMRI* data modeling

In the previous chapter, we have presented *functional MRI* as a *functional neuroimaging* modality, based on *BOLD* contrast. During an *fMRI* experiment, many successive scans (or volumes) are acquired and processed to retrieve the spatial localization of brain activity. Next, some preprocessing and modeling steps, detailed in this section, have to be performed in order to extract relevant information.

2.1.1 Preprocessing of fMRI data

After the acquisition of *fMRI* data from an *MRI* scanner, some preprocessings are required, as they allow to remove some variability of the signal that is not related to the cognitive function under investigation, or account for specificities of the acquisition. As these preprocessings are a crucial step in the analysis of *fMRI* data, we describe them briefly. However, they will not be further studied in this thesis.

Slice-timing correction

Each slice of the *fMRI* volume is acquired with a slightly different time than the previous or the following one. The purpose of slice-timing correction is to correct this temporal shift within one *fMRI* volume, using temporal interpolation. This is done by adding a shift to the phase of each component of the Fourier transform of the signal, yielding a temporal interpolation of the signal while preserving its spectrum. Slice-timing correction is particularly recommended for event-related designs. It can be done before (but can be problematic if the movements have too large amplitude) or after (but in this case, it can slide some voxels between different layers, and thus it breaks the temporal coherency) realigning volumes to a common space. Realignment and slice-timing should ideally be performed simultaneously.

Motion correction - Spatial realignment

During an *fMRI* acquisition, hundreds of volumes are acquired, and motion can decrease the sensitivity of a statistical analysis of the *fMRI* data. One of the main hypothesis in intra-subject *fMRI* analysis is the voxel-to-voxel correspondence: a voxel represents a specific region of the brain, and should represent the same region during the whole acquisition. *Spatial realignment* aims at correcting the potential motion that occurs during the acquisition (head movement, spatial drift due to the warming of the scanner gradient). Realignment estimates the rigid transformation of every volumes with respect to a reference volume using a mean-square metric.

Coregistration of *fMRI* and anatomical Images

The *coregistration* is used to place two images of the same subject, but acquired with different modalities, in the same space. This is most often used to overlay functional images (*e.g. fMRI,PET*) onto structural images (*e.g. MRI*). This is crucial to report the anatomical locations of brain regions found by functional imaging.

In theory, this realignment should be easy, as the images are related to the same subject and acquired in the same position. In practice, since the images do not have the same contrasts, the square difference is not meaningful anymore. Techniques for coregistration usually resort to mutual information and try to optimize the probability of intensity of the first image knowing intensity in the second image. The images are coregistrated when the joint entropy is minimized, that is done by minimizing the dispersion of the joint histogram.

Spatial normalization

In a study including different subjects, anatomical images are not acquired in the same position with respect to the scanner. Moreover, brains have different size and shape across individuals, so that there is no obvious one to one mapping between voxels of images from different subjects. A way to solve this issue is to warp each brain, so that its main structures (large sulci, ventricles) correspond with a reference brain, or *template*. This is called *spatial normalization*. The aim of this preprocessing is to align brain images, with possible changes in size, shape and orientation of the brain. This enables comparison between individuals and but also between studies, and create a common reference space. The most widely used *template* is the *MNI template (Montreal Neurological Institute)* [Evans 93]. It has been constructed using *MRI* scans from 305 right handed healthy subjects. Spatial normalization can be linear (*i.e.* defined by 12 transformation parameters) or non-linear (*e.g.* defined by a set of sinus basis functions) (see among many others [Ashburner 99, Klein 09]).

Spatial smoothing

Spatial smoothing helps removing the high-frequency spatial noise, by strengthening the low spatial frequencies. In addition, this filter allows a better inter-subject comparison by increasing the overlap of activations across the different subjects. We typically choose a size of two or three times the size of one voxel, that will optimize the detection of clusters that have this size during the statistical analysis. However the spatial smoothing limits the spatial resolution, and can bias the position of fine-grained activation foci.

2.1.2 Modeling *fMRI* data

One of the milestones of *fMRI* data analysis has been the introduction of the *general linear model* (*GLM*) by Friston and al. [Friston 95]. This approach takes into account, in one statistical model, all the factors (experimental paradigm,

physiological effects, noise) that can explain the *fMRI* signal time series. The weight of each of these factors is estimated for each voxel, yielding the so-called *activation maps* (or β*-maps*), which can be used for *statistical inference*.

Design matrix

The design of an experiment will be reflected in the *design matrix*. The design matrix $\mathbf{X} \in \mathbb{R}^{\mathbf{n}\times\mathbf{r}}$ represents the r temporal factors that we consider as relevant for the modeling of the *fMRI* time series (one column by factor), during the n scans. These factors can be in particular related to:

- the presence or not of an experimental factor (i.e. task-related, pharmaceutically induced).
- the modeling of nuisance events such as motion, session-dependent effects, physiological noise.
- low-frequency signals (sinusoids or low-order polynomials) that model the drifts of the signal.
- a constant regressor.
- the time derivative of some regressors to take into account some errors in the *hrf* response delay or shape.

The design matrix is finally obtained by the convolution of the relevant regressor with an *hemodynamic response function* model (see chapter 1). An example of *design matrix* is given Fig. 2.1.

The *General Linear Model - GLM*

Let us introduce some notations. We denote $\mathbf{Y} \in \mathbb{R}^{\mathbf{n}\times\mathbf{p}}$ the matrix of the scans of *fMRI data*, where n is the number of scans, and p the number of voxels. An image (*i.e.* a function of space, $\mathbb{R}^3 \rightarrow \mathbb{R}$) is thus viewed as a vector in $\mathbb{R}^p$. The matrix $\beta \in \mathbb{R}^{\mathbf{r}\times\mathbf{p}}$ is the parameters of the model to be estimated, with r the number of regressors. The rows of β can be represented as images, called *activation maps*, reflecting the weights of the voxels for each regressor. The columns of β represent the weights of a voxel for the different regressors.

The *General linear model (GLM)* can be written as:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E} \tag{2.1}$$

Following this model, the value y_{ij} of the j^{th} voxel in the i^{th} *fMRI* scan is given by :

$$y_{ij} = \sum_{k=1}^{r} x_{ik}\beta_{kj} + e_{ij} \tag{2.2}$$

The residual error $\mathbf{E}$ of the model represents the fraction of the data which is not explained by the design matrix $\mathbf{X}$. In the case of *fMRI* data, the *GLM*

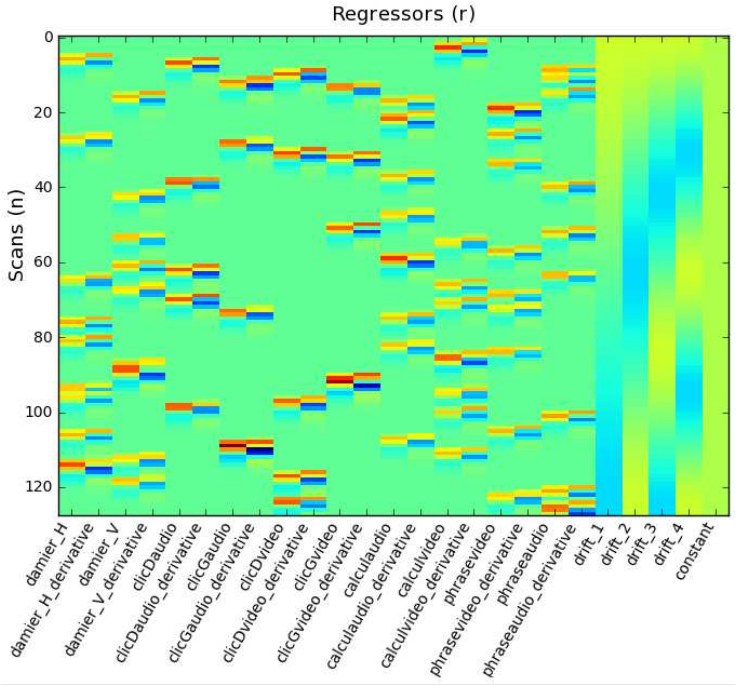


Figure 2.1: Example of *design matrix*. We can notice the different experimental conditions (*damierH*, *clicGaudio*, *calculaudio*, ...) and their derivatives. The last columns of the design matrix represent the low-frequency drifts at different frequencies and the constant regressor, that are added as confound regressors.

makes the hypothesis of a Gaussian auto-regressive process model for the noise $\mathbf{E}$. By hypothesis, at voxel j, we have $e_j \sim \mathcal{N}(0, \sigma_j^2 \mathbf{V_j})$, where σ_j^2 is the noise variance, and $\mathbf{V_j}$ is the normalized noise covariance matrix. The noise magnitude and covariance are voxel-dependent, and we assume the noise to be auto-regressive. We assume that the *design matrix* is full rank ($\text{rank}(\mathbf{X}) = \mathbf{r}$) and that the covariance matrix $\mathbf{V_j}$ is known and full rank. Thus we can derive the maximum likelihood estimation of the parameters β:

$$\hat{\beta}_\mathbf{j} = (\mathbf{X}^T \mathbf{V_j^{-1}} \mathbf{X})^{-1} \mathbf{X^T} \mathbf{V_j^{-1}} \mathbf{Y_j} \tag{2.3}$$

The values of β are the parameters of the *GLM* and can be seen as the effect of each regressor on each voxel time course. the

The noise variance is estimated as:

$$\hat{\sigma_j}^2 = \frac{1}{\hat{\nu}} (\mathbf{Y_j}^T \mathbf{V_j^{-1}} \mathbf{Y_j} - \mathbf{Y_j^T} \mathbf{V_j^{-1}} \mathbf{X} (\mathbf{X^T} \mathbf{V_j^{-1}} \mathbf{X})^{-1} \mathbf{X^T} \mathbf{V_j^{-1}} \mathbf{Y_j}) \tag{2.4}$$

where $\hat{\nu} = n - r$ is the degrees of freedom of the model.

2.2 Classical inference

In the previous chapter, we have seen that we can obtain brain functional images by using *fMRI*. After some preprocessings and fitting a *GLM*, we obtain a set of activation images. These images can be used to make an *inference* and to extract the spatial layout of the neural coding.

We describe here one approach for *inference*, called *classical inference*, which has been widely used during the last fifteen years [Friston 95]. This *inference* addresses the question of finding the regions of the brain which are more active for one condition compared to another conditions. This is the most used method in functional brain mapping and yields typical *Statistical parametric maps - SPMs*. This method relies on the classical hypothesis testing statistical framework. However, we will see that *classical inference* has important limitations, and is not necessarily adapted to the complexity of neural coding.

2.2.1 Statistical analysis and *SPMs*

In neuroimaging, the previously described *GLM* (see Eq. 2.1) allows to make statistical *inference*: one can test whether some variables of interest in the model fit a significantly part of the data. Let us introduce the notion of *contrast* c, as a linear combination of effects (*i.e.* experimental conditions) that are assumed to be of particular interest.

Mass univariate analysis

It is often easier to perform a *mass univariate analysis*, *i.e.* to test the significance of the effects of interest on all voxels considered separately. Apart from the fact that such analysis does not require a specific subset of voxels, the main interest of this method is that it gives a regional localization to the test. We have one test by voxel, and thus one score by voxel, which enables the creation of brain *maps*. This approach is often referred to as *classical inference*, the created maps being called *statistical parametric maps (SPMs)*. We can define the two following hypotheses:

- *Null hypothesis* H_0: the experimental conditions defined in the contrast c do not have an effect on the weights β: $H_0 \ : \ c^T\beta = 0$
- *Alternative hypothesis* (*Hypothesis of interest)* H_1: the experimental conditions defined in the contrast c have a significant positive effect on the weights β: $H_1 \ : \ c^T\beta > 0$

According to *Neyman-Pearson* lemma, the uniformly most powerful test to decide which hypothesis is true under standard assumptions, is the *likelihood ratio* test. We define the *likelihood ratio* as:

$$\Lambda = \frac{\mathcal{L}_{H_1}(Y|\beta)}{\mathcal{L}_{H_0}(Y|\beta)} \tag{2.5}$$

with $\mathcal{L}_{H_0}(Y|\beta)$ the likelihood of the data for the *null hypothesis*, and $\mathcal{L}_{H_1}(Y|\beta)$ the likelihood of the data for the *hypothesis of interest*.

Statistical tests

In order to perform a statistical test, we have to define the distribution P_{H_0} followed by the test statistics if the null hypothesis is true. This distribution allows us to compute a *p-value* p, which is the probability to observe under the null hypothesis, a value of the test statistics Λ as extreme as the one observed Λ_{obs}, *i.e.* $p = P_{H_0}(\Lambda > \Lambda_{obs})$. A *p-value* can be calculated using the *cumulative distribution function*, *i.e.* the integral of the *probability density function* of P_{H_0}. It can be seen as the percentage of the total are under the curve that is defined for a given statistical value.

Given the test statistics Λ, the two hypothesis H_0 and H_1 and the distribution under the null hypothesis P_{H_0}, we can define the *rejection region* as the set of values of the test statistic for which H_0 is rejected. For example, in a unilateral test, we can choose to not reject the *null hypothesis* H_0 if the observed value Λ_{obs} of test statistic is under a given threshold Λ_α, *i.e* $\Lambda_{obs} < \Lambda_\alpha$, and we can reject H_0 if $\Lambda_{obs} > \Lambda_\alpha$. α denotes the significance, *i.e.* the specificity of the test, and Λ_α is the corresponding statistics.

Classical tests for *fMRI* data: t test and F test

In the following analysis, we assume that the noise covariance matrix $\mathbf{V_j}$ at each voxel j is known. It follows that Λ is a monotonous function of the exhaustive statistic t (signed test) or F (unsigned and multi-dimensional test). When the noise covariance is unknown the equivalence between F-test and likelihood ratio test does not hold anymore.

In a first analysis, we can test if a voxel yields a similar signal for two different conditions, *i.e.* whether the means of the signal for the two different conditions are different or not. This can be done by using a *t*-test, which assumes that the distribution under the null hypothesis P_{H_0} is a *Student's t* distribution. We introduce the contrast c as a linear combination of two or more conditions to be studied (the coefficients of c sum to zero). For a given voxel j and a contrast of two conditions, we have the following *t-score*:

$$t_j = \frac{c^T \hat{\beta}_j}{\hat{\sigma_j}\sqrt{c^T(\mathbf{X_j^T V_j^{-1} X_j})^{-1}\mathbf{c}}} \tag{2.6}$$

with $\hat{\sigma_j}^2$ the noise variance and $\mathbf{V_j}$ is the covariance of the residuals. The variable t_j follows a Student's distribution with $\nu = n - r$ degrees of freedom.

In a second case, we can use a multi-dimensional contrast c, that yields a F-test. We assume that the test statistic has an F distribution under the null hypothesis. The value of the test statistic for a given voxel j (called *F-score* of

voxel j) is:

$$F_j = \frac{\text{Tr}(c^T \hat{\beta}_j \hat{\beta}_j^T c)}{\hat{\sigma_j}^2 \text{Tr}(c^T (\mathbf{X_j^T V_j^{-1} X_j})^{-1} \mathbf{c})} \tag{2.7}$$

The variable F_j follows a Fisher's distribution with $n - r, r - \text{rank}(c)$ degrees of freedom. For the two scores defined in Eq. 2.6 and Eq. 2.7, there is one value for each voxel, which can thus be mapped on the brain, creating the *statistical parametric maps – SPMs –* (see an example Fig.2.2).

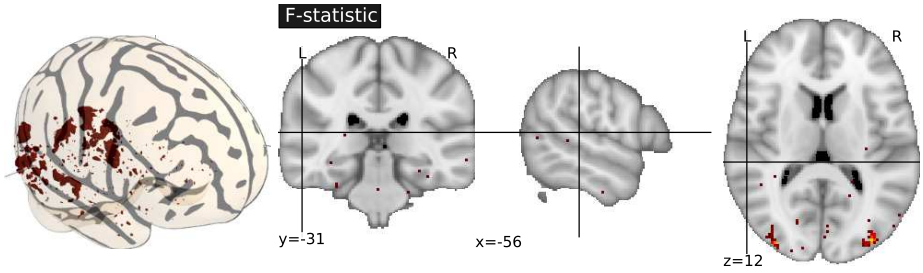


Figure 2.2: Effect of *mental representation of object shape* - subject 1 (see Appendix B.2). Representation of the *F-statistic* (i.e. mass univariate *F-test*) when we test if the shape of the object significantly activates brain sites. We can notice that some regions of the brain are outlined, such as the occipital lobe, which is a region known to be implied in visual recognition.

2.2.2 Multiple comparisons issues

The previously described tests allow to visualize very useful maps for brain mapping. However, classical inference suffers from a major drawback which is the multiple comparisons issue. Due to the huge number of voxels (60.10^3 at $(3mm)^3$ resolution), some tests can lead to a large amount of false positive results (*i.e.* some voxels found to be significant, but that are actually not). Thus, we have to take into account the fact that many tests have been performed.

Family-wise error correction can be performed instead, using the *Bonferroni correction*. This approach simply consists in dividing the threshold α by the number of tests p, which yields the new threshold $\alpha_b = \alpha/p$. This correction is very severe, and results in very strict significance values. However, this approach considers that all the tests performed are statistically independent. This does not take into account the spatial structure of the data, and is not appropriate for correlated data (as the effective number of tests may be reduced). To cope with these limitations, other approaches have been proposed (see [Nichols 03, Hayasaka 04] among others), such as *cluster-level thresholding*, *false detection-rate threshold*, or nonparametric methods can also be used, as permutation test.

In *fMRI*, multiple comparisons issues are crucial, as a high threshold yields a good specificity (*i.e.* we limit the reported voxels falsely activated) but a poor sensitivity (*i.e.* we miss truly activated voxels), and a low threshold yields a good sensitivity, but poor specificity. The maps of voxels selected by thresholding the *p-values* for the object recognition task (subject 1), are given in Fig.2.3, 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. In particular, we can observe that the voxels selected using the 0.05 corrected threshold are found in an region (*LOC*) known to be involved in high-level visual processes.

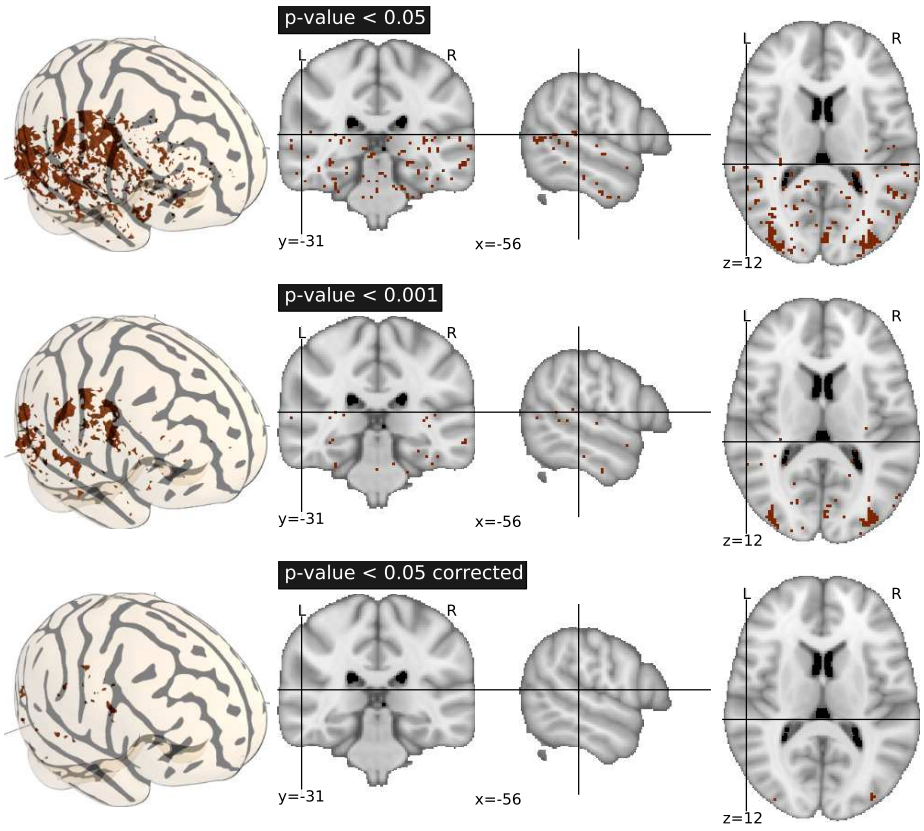


Figure 2.3: *Mental representation of shape* - subject 1 (see Appendix B.2). Visualization of the voxels selected by thresholding the *p-values* for the objects recognition task for different thresholds (0.05, 0.01 and 0.05 corrected by *Bonferroni*). The *Bonferroni correction* keeps very few voxels in the region of interest (*LOC*), and removes isolated voxels (in particular in the frontal lobe).

2.2.3 Other limitations of classical inference

The multivariate point of view

Another major drawback of classical inference is the fact that it is used to analyze each voxel separately. Thus, it does not take into account the correlations between different voxels (see Fig.2.4 for an illustration). Multivariate analysis is closely related to the notion of *population coding* explained previously. Indeed, the information related to the cognitive state can be scattered across several voxels. As explained in [O'Craven 00], it can be difficult to find a single region (or voxel) which can be used to predict the behavioral data. Thus, a classical univariate test may not be significant, while a multivariate decoding approach could detect information.

The crucial point is exposed by [Haynes 06]: by pooling together information from different locations within the whole brain, we can enhance the prediction. Some methods have been used to take into account the multivariate nature of the data. In the case of the classical inference approach and on a reduced set of voxels, multivariate tests can be made, such as *Manova* or *Mancova* [Friston 96]. However, those approaches require a small subset of voxels (less voxels than samples), and do not allow us to access the intrinsic organization of the information within a group of voxels. Indeed, multivariate tests do not extract the combination of voxels truly implied in the cognitive task, and do not clarify their respective involvement in the neural coding.

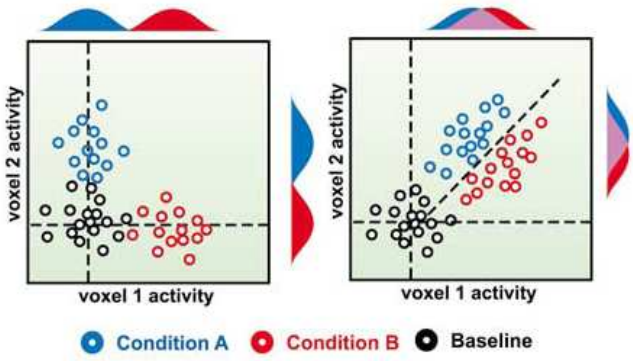


Figure 2.4: Illustration of univariate and multivariate decoding approaches. Left - This case can be handled with classical inference but might not be realistic. Right - This case is more realistic and requires to consider the two voxels together in a multivariate approach, to retrieve the different conditions. Adapted from [Cox 03].

The baseline issue

Classical inference is based on the hypothesis that the global activity of the brain is an additive combination of different activations (*a.k.a pure insertion hypothesis*). It compares the images where one condition is assumed to have some effect, with the ones where this condition is not present (which can be referred as the *baseline* for this condition). This method is detailed in [Fox 91], where each stimulus is compared to the control condition for each voxel separately. The *baseline* is empirically defined as the signal level in the absence of stimulation (control condition). The *pure insertion hypothesis* is questionable, as a control condition is rarely a steady state of the brain activity, even in the case of rest [Poeppel 08]. Neural populations activated in both the baseline and the task under investigation (even for different cognitive processes), may not be extracted by classical inference [Sidtis 03].

2.3 *Inverse inference*

In order to deal with the previously raised criticisms, a new approach, called *inverse inference* (or *brain-reading*), has been introduced [Dehaene 98, Cox 03]. This method relies on *statistical learning* tools, and more precisely on *pattern recognition* approaches. The main idea is to consider the *fMRI* analysis as a pattern recognition problem, *i.e.* using a *pattern* of voxels to predict a behavioral, perceptual or cognitive variable. In this approach, the accuracy of the prediction can be used to validate (or invalidate) that the pattern of voxels used in the predictive model is implied in the neural coding. In short, inverse inference is an approach for *decoding* the neural coding. Interestingly, this idea had been latent for a long time, as shown by this text of N. Tesla [Tesla 33] in 1933:

> "[...] I expect to photograph thoughts [...] I became convinced that a definite image formed in thought must, by reflex action, produce a corresponding image on the retina, which might possibly be read by suitable apparatus [...] "

In the following, we describe the characteristics and advantages of inverse inference, and how it copes with some limitations of classical inference. We also detail the framework used for this analysis, and contexts in which it was used, as well as the difficulties of decoding in inter-subject analysis.

Link with *Brain Computer Interfaces - BCI*

In *BCI* the predicted information is often extracted from functional imaging and is used for controlling a device. In that sense, the inverse inference approach is closely related to *BCI*. Using machine learning methods to decode the interaction between observed signals and a target has been done with EEG data (*e.g.* using correlation [Wang 04], ICA [Vallabhaneni 04] or neural network

[Phothisonothai 08]); or with neuronal recording data (*e.g.* using euclidean distance-based classification [Tsao 06], or linear SVM [Hung 05]).

However, even if the methodology is relatively similar between *BCI* and inverse inference, there is a major difference. *BCI* aims at a robust prediction and can use any kind of information that is extracted from the data, without always considering the neuroscientific meaning. On the other hand, inverse inference tries to extract information that can clearly be related to the neuronal population implied in the coding of the cognitive function, in order to confirm or infirm a neuroscientific hypothesis. One can refer to the review [Signe 09] for further applications of pattern recognition in neuroimaging.

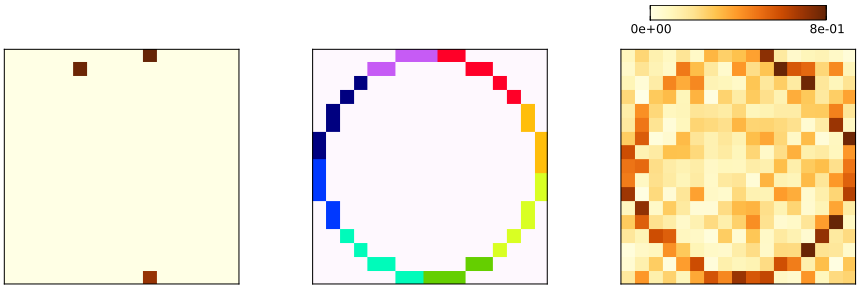
2.3.1 Multivariate pattern analysis – *MVPA*

Multivariate analysis

The link between inverse inference and pattern analysis/machine learning has been made early [Cox 03, Norman 06]. The conjunction of these two approaches is called *Multivariate Pattern Analysis* or *MVPA*. The use of machine learning techniques in fMRI data analysis has been justified as follows: while classical inference seeks the voxels which have a significant response to an experimental condition, a voxel with a non-significant response to a given condition can still carry some information about the presence or absence of this condition. This information can be detected only by testing if the voxel is useful for the prediction of the presence or absence of the condition by a pattern recognition approach. Moreover, as explained by Haynes and al. [Haynes 06], if a region responds to different cognitive processes with overlapping activities, it can be difficult (or even impossible) to use the signal of this region for discriminating between the different tasks. However, a multivariate pattern of different regions can be used to overcome this limitation.

This fact is illustrated by a simulation Fig. 2.5. A ring is divided in eight different sectors, and for each image, a sector is set to be active (this is the target to be decoded). Only voxels from the ring are relevant for prediction, and, when a sector is active, the neighboring sectors are also weakly active. We can see that, in the case where regions of interest have overlapping activities, classical inference (left, using *F-test*) is not able to retrieve the true spatial support of the neural code. However, a decoding technique as inverse inference can retrieve the true spatial organization by taking into account the multivariate structure of the signal.

These considerations about univariate and multivariate analyzes are crucial when seeking to extract the spatial layout of neural coding. In case of sparse coding, classical inference is able to find the relevant voxels, as the information is coded by only one (or few) voxel. However, in the more probable case of population coding, where the information is encoded by a set of voxels, we have to use inverse inference to access an accurate decoding.



Classical inference
F-score thresholded at $p < 0.01$

True substrate of the neural coding (labels)

Inverse inference
Weights of the model

Figure 2.5: We simulate two dimensional images with gaussian signal, and create a set of regions of interest by dividing a ring in 8 different sectors. For each image, a sector is defined to be active. This information (which will be the target to be predicted) is encoded by the corresponding sector of the ring, and the neighboring sectors (with decreasing weights). When there are overlapping activities, we can see (left image - thresholded at $p < 0.01$) that classical inference can not retrieve the true pattern, while a decoding technique (right image - weights of the predictive model) can retrieve the true substrate of the coding. The prediction accuracy is 28% (highly significant, with a p-value $< 10^{-3}$ under permutations), with a chance at 12.5% (8 classes).

Validation and multiple comparisons issue

More interestingly, inverse inference avoids the multiple comparison issue, as it performs only one statistical test (on the predicted behavioral variable). In that sense, it can provide more sensitive analyzes than standard statistical parametric mapping procedures [Kamitani 05]. In inverse inference, the validation relies on the comparison between the predicted label and the true target, and thus, is performed on the target and not in the feature space. The result of this comparison directly answers the question of the presence (or the absence) of stimulus-related information within the fMRI data. Moreover, the validation of the inverse inference procedure does not require a specific statistical structure, unlike classical inference (*e.g.* spatial correlation between voxels for the statistical test). However, the multiple comparisons issue manifests itself as curse of dimensionality in *MVPA*.

Spatial and local information

Despite its drawbacks, classical inference solves most of the questions in neuroscience, as it allows a functional brain mapping. The *SPMs* have become a reference in neuroscience studies. Smoothing is sometimes applied to *fMRI* data

in order to obtain a better signal-to-noise ratio (*SNR*) in univariate analysis. More importantly, smoothing improves the overlap of active regions in voxel-based inter-subjects studies. However, smoothing can erase the fine-grain information within a pattern of voxels, and thus hides some useful knowledge for brain mapping, as explained in [Norman 06] and [Haynes 06]. Using pattern recognition techniques, as in inverse inference, allows to access this information by comparing relative changes of signal within a set of voxels. In particular, this approach is well-suited for decoding some specific neural codes such as the *distributed codes* detailed previously, where information from different regions of the brain have to be considered simultaneously.

Few methods have been proposed to study the spatial distribution of the coding entities, within an inverse inference framework (see [Kriegeskorte 06]). Using spatial information for improving the inverse inference approach is a central subject of this thesis.

Overview of the experimental framework

Inverse inference framework is similar to the one commonly used in pattern recognition. This is illustrated in Fig.2.6, and relies on a few steps that are detailed here.

Step 1 - Acquisition and preprocessing of fMRI images In this step, fMRI images are acquired, during the performance of a cognitive task by a subject, and then processed to yield trial-by-trial activation maps. This initial step will not be studied in this thesis and a brief description of the operations performed as pre-processing can be found in the first section of this chapter.

Step 2 - Pattern recognition The important step of *inverse inference* framework is the step of *pattern recognition*, which relies on the following settings:

- *Prediction function:* In order to test the relevance of the patterns considered, and thus test whether or not they are part of the spatial layout of neural coding, we train a prediction function using these voxels. This function should be able to predict the target for a new *fMRI* image. The aim of this classifier is to give a prediction accuracy, which can be seen as a measure of the quantity of information within the considered patterns.

- *Dimension reduction:* This step can be mandatory due to the high dimension of the data. Indeed, keeping all the features (*i.e.* the whole image) can lead to *overfit*, which dramatically decreases the prediction accuracy: this will be made explicit in chapter 3.

- *Validation of the method:* The last step of the *pattern recognition* framework, is the computation of the prediction accuracy on a new set of images. One can apply the dimension reduction, to extract the putatively relevant features of the new images. The prediction function is used to predict the

target variables that correspond to behavior-related labels and/or stimulation parameters associated with new images, which are compared to the true target variables. This procedure makes the computation of the prediction accuracy possible, which is a measure of the quality of the prediction. A significantly accurate prediction, means that information about the different classes was indeed present within the considered features.

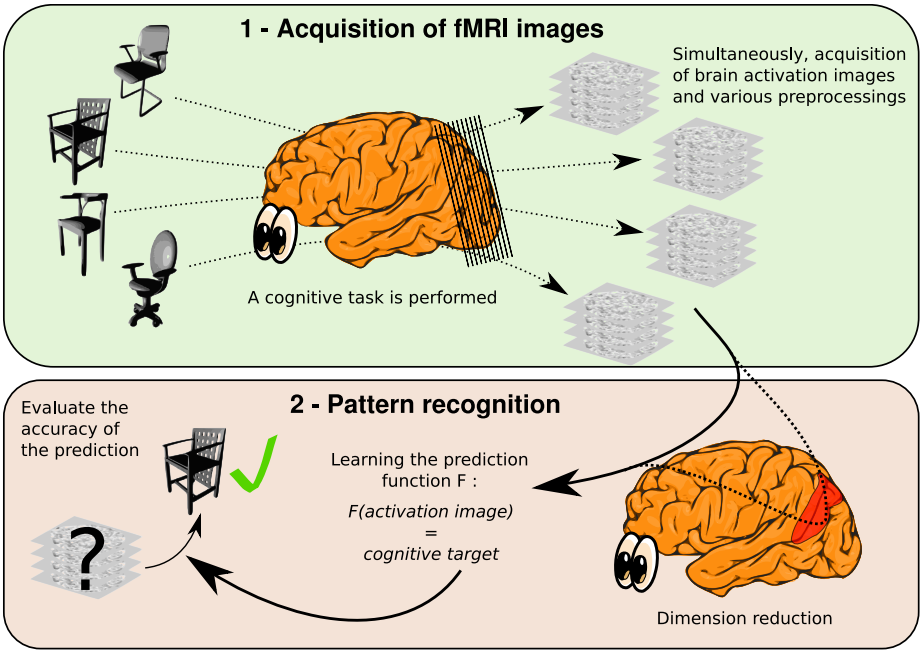


Figure 2.6: Illustration of the *inverse inference* scheme. **Step 1**: the subject is involved in a cognitive task, such as looking at objects of different shapes. *fMRI* images are simultaneously acquired. **Step 2**: the *prediction model* is learned, and the prediction corresponding to the *test set* is performed and compared to the true stimulus, in order to evaluate the accuracy of the prediction model. A *dimension reduction* can be applied, before or within the learning of the prediction function, in order to select the most relevant voxels; this step can be crucial to avoid *overfit*, and will be one of the main problems addressed in this thesis.

2.3.2 Inverse inference in cognitive neurosciences

Specific uses of inverse inference

Inverse inference is a powerful tool for accessing the representation of the neural code from *fMRI* data. However, this method also allows to perform specific

studies which were not possible with classical inference.

First, inverse inference can be used to make an inference on the cognitive state of the subject, which is in principle similar to building a ***Brain Computer Interface - BCI.*** Although *fMRI* is not well-suited for such a device, it promises powerful applications such as the lie detector detailed in [Davatzikos 05]. However, some ethical questions can be raised, as we will see later in this section.

Inverse inference can also be used to estimate the cognitive state in **a real-time framework** (more precisely, each image at a time). In [Mitchell 03], Mitchell et al. show that it is possible to retrieve the temporal changes in the behavior of a subject by using the prediction of a classifier. This is related with the previously described *BCI*, and such an approach can not be done using classical inference, which does not give a prediction. Inverse inference allows to help understanding the **representation of unconscious stimuli**: one can access to the prediction of the stimuli and their temporal characteristics, even if the subject is not conscious of them, in order to understand the conscious/unconscious decoding performed by the subject. An example of such an experiment is described in [Haynes 05a], on the prediction of unconscious visual stimuli.

Finally, inverse inference can provide **access to fine grained patterns** such as low level perceptual features (orientation, direction of motion), using low spatial resolution. In [Kamitani 05], Y. Kamitani et al. show that, by using a pattern of activations in the visual cortex, they could retrieve the orientation of visual stimuli. As the columns coding for orientation are too small for the *fMRI* resolution, they use the fact that each *fMRI* voxel has nevertheless a little bias for a specific orientation. Thus, by using redundancy of information across visual field within a multivariate analysis, the authors could predict low level sub-voxel visual features.

An ethical point of view

Inverse inference approach (based on *fMRI*, but also *EEG*, *MEG*) raises some very fundamental ethical questions. Relating to the conscience and free will, this method is, and must remain, subject to a significant ethical oversight. More specifically, the ethical problems posed by this approach are the following, and more details can be found in [Farah 04]:

- by revealing cognitive information in some given cerebral regions, and thus by allowing the decoding the cognitive or emotional states of a subject, it allows to bypass the "classical" lines of communication (voice and sign language) which are under the voluntary control of subject, and thus can access to the "privacy" of the mind (e.g. lie detection [Davatzikos 05], criminal justice [Farwell 01]). In 2008, an indian judge sentenced a suspect to life in prison, using *EEG*-based inverse inference. [1] Inverse inference can also be used for commercial purpose. This is called *neuromarketing*, and some details can be found in [Brammer 04], or in [McClure 04]

[1] http://www.nytimes.com/2008/09/15/world/asia/15brainscan.html?_r=1&scp=1&sq=Phansalkar-Joshi&st=cse.

.

- by being used for social discrimination and allowing to make prediction on the future behavior of a subject as it has already been done for criminality [Raine 98] and drugs [Childress 99].

The decoding approach outlined in this thesis can allow to access to a part of the neural coding of thoughts. Thus, it seems that this approach can be used to access unexpressed thoughts. However, inverse inference can at best be used to determine a very crude measure of personality, and the previous described approaches usually requires active participation/consent of the subject. The crucial point is that the subject should be aware of the use of the data and what it can reveal.

Generalization of inverse inference

Inverse inference seems to be a powerful approach for deciphering the neural coding, and the question of it generalization to new paradigms arises naturally [Cox 03]:

- One of the first possible generalization of inverse inference, it the generalization across time (*i.e.* using images acquired at different moments). An example of such generalization is given in [Cox 03], where the authors are able to predict a stimulus with a model learned a few days apart (see also [Kay 08]). However, this aspect of inverse inference is still weakly little developed, and should be considered as an additional validation rather than a fundamental dimension of neural coding.

- Another generalization of inverse inference is the prediction to new stimuli. The extrapolation to new stimuli is probably the most challenging problem. There is an infinity of possible cognitive states, and we have access to a limited collection of examples (*i.e. fMRI* images). However, this generalization is very interesting in order to find elements of neural coding that are robust to changes in stimuli, and thus are implied in the coding of abstract human thoughts. If activation patterns are ranked in a parametric space, we can extrapolate to new cognitive states.

  A more interesting solution is the construction of a prediction function that can identify new stimulus in a large dataset, based on already seen stimuli (as visual stimuli [Kay 08], or nouns associated to new images [Mitchell 08]). The more challenging generalization of the prediction to an unknown high level stimuli has been addressed in [Knops 09]. The authors show that, using inverse inference, they are able to generalize the prediction from ocular saccades to arithmetic tasks.

- Finally, the final goal of inverse inference can be the implicit reconstruction of the stimulus, as explained in [Thirion 06a] and [Miyawaki 08], in the case of reconstruction of visual stimuli.

2.3.3 Inter-subject inference

Among different possibilities of generalization of inverse inference, we focus during this thesis on the particular generalization across subjects. The main interest of this inter-subject prediction in the study of neural coding, is to possibly find predictive regions that are stable across subjects, and thus obtain a population-level validation of cognitive hypothesis.

However, inter-subject predictions are plagued by the inter-subject variability (lack of voxel-to-voxel correspondence) [Tahmasebi 10, Tucholka 10]. This variability in the location of activation can arise from variability in anatomical structure or/and in functional organization. *Spatial normalization* (or inter-subject registration), can be used to decrease anatomical variability, even if there is still no accurate voxel-to-voxel correspondence between subjects. However, some variability remains in the localization of activations, even with an accurate inter-subject registration, due to, among others effects, handedness [Kim 93] or genetics [Blokland 08]. Moreover, this variability can also be explained by different cognitive strategies that yield different spatial layouts of neural coding across subjects [Kirchhoff 06].

Thus, the localization of functional activity across subjects can vary and it is challenging to find a common spatial layout of neural coding across different subjects (see Fig. 2.7). Some approaches for inter-subject inverse inference that have been developed in this thesis are detailed in chapter 5 and chapter 6.

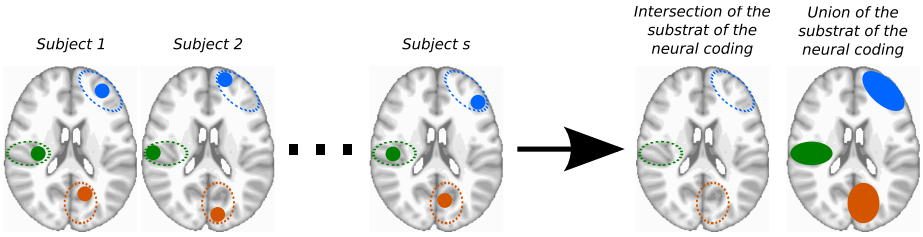


Figure 2.7: Illustration of the effect of inter-subject variability in the study of the neural coding. Regions implied in the neural coding are represented by disks of different colors. We can notice than the populations of active neurons are not at the same position across subjects (left), and thus, the intersection of the support of neural coding (middle right) is empty. However, by considering the union of the coding regions across subjects (right), we can notice that three regions of the brain are clearly outlined.

2.4 Conclusion - From *fMRI* acquisitions to "Brain-Reading"

After acquisition, raw *fMRI* data have to be preprocessed and modeled. The modeling of *fMRI* data relies on the *General Linear Model*, that takes into account the different parameters of the experiment defined in the *design matrix*. The resulting *activation maps* can then be used for statistical analysis, in order to study the neural code related to a specific cognitive parameter.

Classical inference is a widely used method to study *fMRI* data. This relies on voxel-based statistics, yielding significance maps (*a.k.a. Statistical Parametric Maps - SPMs*) for the effects under consideration. However, despite its simplicity and the accuracy of the *SPMs*, classical inference suffers from a major drawback: it analyzes each voxel separately and consequently cannot exploit the correlations existing between different brain regions to improve the inference. Moreover, statistical power in the case of classical inference is limited by the multiple comparison problem (one statistical test is performed for each voxel and the number of comparisons has to be corrected for).

In order to deal with these limitations, *inverse inference* has been proposed. Inverse inference is an approach for studying *fMRI* data based on machine learning methods. It can be used to assess the specificity of several brain regions for certain cognitive or perceptual functions, by evaluating the accuracy of the prediction of a behavioral variable of interest – the *target* – based on the activations measured in these regions. This inference relies on a prediction function, the accuracy of which depends on whether it uses the relevant variables, *i.e.,* the correct brain regions. The major advantages of this approach are:

- As multivariate approach, it is consistent with *population coding* models. Indeed, the neural information, which can be encoded by different populations of neurons, can be decoded using a *pattern* of voxels.
- It avoids the multiple comparison issue, as it performs only one statistical test (on the predicted behavioral variable). In that sense, it can provide more sensitive analyzes than standard statistical parametric mapping procedures.
- It addresses new challenges, in particular by allowing to identify a new stimulus in a large dataset, based on already seen stimuli. Moreover, it can be used for the more challenging generalization of prediction to unknown high level stimuli, which opens the way of a deeper understanding of brain functional organization.

3

Statistical learning for *fMRI inverse inference*

In this chapter, we describe the classical statistical learning framework used in *inverse inference* studies. Then, we detail the different methods of feature selection and prediction (classification or regression) that have been used in the literature for *fMRI* inverse inference. We illustrate each method on a real *fMRI* data set and explain its advantages and drawbacks. This study allows us to define the essential characteristics for an accurate *statistical learning* approach in the specific case of *fMRI* inverse inference.

Contents

3.1 *Inverse inference* framework

After the acquisition, we apply some preprocessings and fit a *GLM* to *fMRI* data. The resulting images of parameters (or *activation maps*), are then used to decipher the neural coding, within an inverse inference framework. From now and for the following chapters, $\mathbf{X}$ are the *activation maps*, and $\mathbf{y}$ is a behavioral target related to each of these maps.

In this section, we describe the statistical learning framework that has been used during this thesis for *fMRI*-based inverse inference. Especially, we detail the specificities of this framework related to the nature of *fMRI* data. In addition to the description of these concepts, the *Python* code used for implementing them is given in appendix C.

The notations used in this chapter are given in Fig. 3.1, and the state of the art of statistical learning in *fMRI* is summarized Tab. 3.1. We try to make this list as complete as possible, but more information can be found in the following review papers [Norman 06, Haynes 06, O'Toole 07, Spiers 07, Pereira 09].

Data			
$n \in \mathbb{N}$	number of samples	p	number of features (voxels)
$.^l$	reference to the learning set	$.^t$	reference to the test set
n^l	number of samples (learning set)	n^t	number of samples (test set)
$\mathbf{X} \in \mathbb{R}^{n \times p}$	data (*activation maps*)	$\mathbf{X^j}$	values of the j^{th} feature
$\mathbf{X_i}$	i^{th} image	x_i^j	j^{th} feature of the i^{th} image
$\mathbf{X^l} \in \mathbb{R}^{n^l \times p}$	data (learning set)	$\mathbf{X^t} \in \mathbb{R}^{n^t \times p}$	data (test set)
Model			
$b \in \mathbf{R}$	intercept	w_j	weight of the j^{th} feature
$\mathbf{w} \in \mathbf{R}^p$	true weights of the model	$\hat{\mathbf{w}} \in \mathbf{R}^p$	estimated weights
Classification settings			
$\mathbf{y} \in [1, \ldots, K]^n$	discrete target	y_i	target of the i^{th} image
$\hat{\mathbf{y}} \in [1, \ldots, K]^n$	predicted target	K	number of classes
$\mathbf{y^l} \in [1, \ldots, K]^{n^l}$	target (learning set)	$\mathbf{y^t} \in [1, \ldots, K]^{n^t}$	target (test set)
l_k^l	learning samples in the k^{th} class	l_k^t	test samples in the k^{th} class
Regression settings			
$\mathbf{y} \in \mathbb{R}^n$	continuous target	y_i	target of the i^{th} image
$\hat{\mathbf{y}} \in \mathbb{R}^{n^t}$	predicted target		
$\mathbf{y^l} \in \mathbb{R}^{n^l}$	target (learning set)	$\mathbf{y^t} \in \mathbb{R}^{n^t}$	target (test set)

Figure 3.1: Notations used in the following chapters.

3.1.1 Data representation

We note $\mathbf{y}$ the true behavioral variable (or *target*). This is the variable that will be inferred from the *activation maps*. In the case of a regression analysis, we have $\mathbf{y} \in \mathbf{R}^n$, and $\mathbf{y} \in [1, \ldots, K]^n$ for classification analysis.

The data can be represented by a matrix $\mathbf{X} \in \mathbf{R}^{n \times p}$, each row being a p-dimensional sample, *i.e.*, an activation map. We have n the number of samples, and p the number of features (voxels).

The data can be mapped into a high dimensional space, by using some

functions $\phi_l(\mathbf{X})$ called *basis functions* (with $1 \leq l \leq L$, L being the number of *basis functions*). For example, these functions can be *polynomial functions*, *wavelets* or even *identity* ($\phi(\mathbf{X}) = \mathbf{X}$). Such functions allow to extend the use of linear models, as a linear model applied to $\phi_i(\mathbf{X})$ can now be a non-linear model of the data $\mathbf{X}$. We call *feature space mapping* $\Phi(\mathbf{X})$, the transformation from the initial space of the data to the space defined by the *basis functions* (*a.k.a. feature space*). Thus, we can now define the *design matrix* $\mathbf{\Phi}$ as the matrix whose elements are given by $\mathbf{\Phi_{il}} = \phi_l(\mathbf{X_i})$. This matrix is the representation of the data in the *feature space*, and can be used indifferently in a linear model. In the experiments detail in this thesis, we use $\phi(\mathbf{X}) = \mathbf{X}$ (*i.e. identity*), as we already suffer from a high dimensionality issue, and other basis functions further increase the dimensionality of the *design matrix*. The *design matrix* is thus equal to $\mathbf{X}$, and, for more clarity, we now note $\mathbf{X}$ the *design matrix*.

Additional standard preprocessings can be applied to the data as *centering* (each sample has a zero-mean) and *variance normalization* (each sample has an unit-variance). Such preprocessings are required to have comparable activations maps. However, some preprocessings are more specific of *fMRI* data, and are often required. Some physical (*e.g.* increase of the temperature of the scanner), biological (*e.g.* fatigue and concentration of the subject) or behavioral (*habituation to the paradigm*) noise sources can exist and may not be taken into account in the *GLM*. Such noise sources have strong temporal correlations, and can introduce a bias in the prediction. Indeed, they create a drift (called *session effects*) that can be used by the prediction function to make an inference unrelated to the cognitive task. *Session effects* can be removed by centering the images within each different sessions of the experiment, and by using specific *cross-validation* schemes (namely *leave-one-session-out*).

3.1.2 Decoding and prediction model

Inverse inference aims at deciphering the neural code by finding predictive patterns within *fMRI* data. Thus, we have to define and train a prediction function, and different types of prediction function can be used. As we aim at extracting predictive patterns in order to decode the spatial layout of neural coding, we principally focus in this thesis on algorithms that extract relevant, but also *interpretable*, patterns of activation.

The prediction function can be non-linear (*e.g.* non-linear *SVM*), but the superiority of such non-linear prediction function has not been shown in the context of neuroimaging [Cox 03, LaConte 05]. In [Cox 03], a linear prediction function (*linear-SVM*) performs better than a non-linear prediction function, and similar results are reported in [Chu 10]: linear kernels performed better than non-linear kernels in most of the experiments. However this superiority is not yet fully understood. Indeed, it can be due to an intrinsic linear relationship between the signal within the support of neural coding and the cognitive target. Alternatively, it can simply reflect the fact that the non-linear prediction functions used cannot capture the true non-linearity of this relationship. In this thesis, we focus on linear prediction functions.

Predictive linear model

Let us introduce the following predictive linear model:

$$\mathbf{y} = f(\mathbf{X}, \mathbf{w}, b) = F(\mathbf{X}\mathbf{w} + b) \ , \tag{3.1}$$

where $(\mathbf{w}, b)$ are the parameters to be estimated on a learning set, and $b \in \mathbb{R}$ is called the *intercept*. Depending on whether the variable to be predicted takes scalar or discrete values, the learning problem is either a regression or a classification problem.

In a linear regression setting, f reads:

$$f(\mathbf{X}, \mathbf{w}, b) = \mathbf{X}\,\mathbf{w} + b \ , \tag{3.2}$$

In the case of a two-classes classification problem with a linear model, f is defined as:

$$f(\mathbf{X}, \mathbf{w}, b) = \text{sign}(\mathbf{X}\,\mathbf{w} + b) \ , \tag{3.3}$$

where "sign" denotes the sign function and $\mathbf{y} \in \{-1, 1\}^n$. The use of the intercept is fundamental in practice as it allows the separating hyperplane to be offseted from 0. However for the sake of simplicity in the presentation of the method, we will from now on consider b as an added coefficient in the vector $\mathbf{w}$. This is classically done by concatenating a column filled with 1 to the matrix $\mathbf{X}$. Multi-class classification problems are addressed using specific heuristics that are detailed in the case of *SVC* in section 3.2.

Different methods have been used for *prediction* in *fMRI*-based inverse inference. Besides the approaches detailed in this chapter, other methods have been tested as simple *linear regression* [Sidtis 03] or sign comparison [Dehaene 98], *hidden process model (HPM)* [Hutchinson 09], *projection pursuit* [Demirci 08], *Neural networks* [Mørch 97, Haxby 01, Onut 04, Polyn 05, Rissman 10], *CART / Random Forests* [Langlebe 05, Genuer 10, Langs 10], *Adaboost / Boosting* [Koltchinskii 04, Martinez-Ramon 06], *K-Nearest Neighbor (KNN)* and *Similarity-based classifier* [Mitchell 04, Sayres 06, Shinkareva 06, Williams 07, Mitchell 08]. These approaches are not further detailed in this thesis, and we focus on methods based on the model given in Eq. 3.1.

Generative and discriminative models

In classification settings (i.e. $\mathbf{y} \in \{1, \ldots, K\}$), in order to predict the label corresponding to new data, one has to access to the probability $p(\mathbf{y}|\mathbf{X})$. This probability allows us to perform prediction, using the *maximum a posteriori* estimate $\hat{\mathbf{y}} = \arg\max_k p(\hat{\mathbf{y}} = k|\mathbf{X})$. Generally, two different approaches, *discriminative* and *generative*, are possible for estimating the probability $p(\mathbf{y}|\mathbf{X})$:

- Discriminative approaches directly compute $p(\mathbf{y}|\mathbf{X})$, *i.e.* directly aim at solving the classification problem $\mathbf{y}$, using some assumptions on this probability (*e.g. Logistic Regression*).

- Generative approaches rely on the estimation of $p(\mathbf{X}|\mathbf{y} = k)$, and create a model which can be used to *generate* new data $\mathbf{X}$, knowing the target $\mathbf{y}$. In a predictive purpose, these approaches use the *Bayes' theorem* $p(\mathbf{y} = k|\mathbf{X}) \propto p(\mathbf{y} = k)p(\mathbf{X}|\mathbf{y} = k)$ to obtain the probability $p(\mathbf{y}|\mathbf{X})$, by estimating $p(\mathbf{y} = k)$ and $p(\mathbf{X}|\mathbf{y} = k)$ from the learning data. The definition of a model can be avoided by using only some local estimation of this probability, *e.g. KNN*.

Discriminative approaches yield predictive patterns that are not always easy to interpret, as there is no underlying model of the data. However, as we aim at deciphering some neural code by finding predictive regions, and as we measure the quality of the regions by computing a prediction accuracy, discriminative approaches seem still well-suited for inverse inference. Indeed, it has been shown [Ng 02] on many real data sets that discriminative methods have a lower asymptotic error than generative methods.

Generative approaches yield more interpretable results, and can more easily integrate neuroscientific prior through different hypothesis on $p(\mathbf{X}|\mathbf{y} = k)$, as this probability represents the hypothesis on how the features encode the stimuli. However, the estimation of the likelihood $p(\mathbf{X}|\mathbf{y} = k)$ is not trivial and require some assumptions about $p(\mathbf{X}|\mathbf{y})$ to simplify the computation. Some very simple assumptions can be made, yielding classical statistical learning approaches such as *Linear Discriminant Analysis*. A contrario, more complex assumptions about $p(\mathbf{X}|\mathbf{y})$ reflecting our knowledge on *fMRI* data are extremely difficult to make. Indeed, both the relationship between the stimuli and the neural signal, and the underlying metabolic pathway between the neural signal and the observed signal, are complex and poorly understood. Thus, "truly" *generative model* are rarely used in practice [Schmah 08], and as they rely on strong hypotheses, their use is limited to well-known systems, such as the visual system [Thirion 06a, van Gerven 10]. Such approaches will not be further developed in this document.

In conclusion, we will focus in this thesis on discriminative approaches, even if some classical machine learning generative methods will be detailed in this chapter. These approaches are generative from a statistical point of view, but rely on very general assumptions about $p(\mathbf{X}|\mathbf{y})$, and do not aim at modeling *fMRI* data.

3.1.3 Evaluation of the decoding

Different metrics can be used to assess the quality of a prediction, that is related to the information contained in the patterns used in the predictive model. In the case of regression analysis, the performance of the different models is evaluated using ζ, the ratio of explained variance (or R^2 coefficient):

$$\zeta(\mathbf{y^t}, \hat{\mathbf{y}}) = \frac{\text{var}(\mathbf{y^t}) - \text{var}\left(\mathbf{y^t} - \hat{\mathbf{y}}\right)}{\text{var}(\mathbf{y^t})} \tag{3.4}$$

This is the amount of variability in the response that can be explained by the model. A perfect prediction yields $\zeta = 1$, a constant prediction yields $\zeta = 0$, while $\zeta < 0$ if prediction is random and not correlated to $\mathbf{y}$. For classification analysis, the performance of the different models is evaluated using the *classification score* denoted κ , classically defined as:

$$\kappa(\mathbf{y^t}, \hat{\mathbf{y}}^\mathbf{t}) = \frac{\sum_{i=1}^{n^t} \delta(y_i^t, \hat{y}_i^t)}{n^t} \tag{3.5}$$

where n^t is the number of samples in the test set.

Prediction accuracy Prediction accuracy can be seen as a statistical test on the regions used in the predictive model. If a linear classifier has a prediction accuracy significantly above chance level, one can consider that the pattern of voxels shares information with the target, as explained in [Kamitani 05].

Interpretability of the resulting maps In the case of a linear prediction function, it can be interesting to directly look at the voxels weights used in the linear model. However, these weights depend strongly on the prediction function, and we have no proof that the features used in the model correspond to the whole support of the neural code under investigation [Cox 03]. The resulting maps cannot be interpreted as classical *SPMs*. However, one can still use these weighted maps to interpret some aspect of the neural coding. We are expecting (see chapter 2) that the spatial layout of neural coding is sparse and spatially structured in the sense that non-zero weights are grouped into connected clusters. Weighted maps showing such characteristics will be called *interpretable*, as they reflect our hypothesis on the spatial layout of neural coding.

In the case of non-linear classifier, the question of the interpretation is more difficult as we cannot access meaningful weighted maps due to the non-linearity.

3.1.4 Model selection and validation

In order to validate that the extracted patterns are part of the support of neural coding, we have to evaluate the accuracy of the prediction, *i.e.* test whether the predictions are correct. However, learning a prediction function and testing it on the same data yields a methodological bias. To avoid over-fitting, we have to define two different sets: a *learning set* $(\mathbf{X^l}, \mathbf{y^l})$ which is used for learning the prediction function (also called *training set*), and a *test set* $(\mathbf{X^t}, \mathbf{y^t})$ which is used for testing the prediction function.

However, by defining these two sets, we reduce the number of samples that can be used for learning the model, which is a crucial issue in *fMRI* data analysis. Moreover, the results can depend on a particular couple of learning set and test set.

A solution is to split the whole data into different learning and test sets, and to return the averaged value of the prediction scores obtained with the different sets. Such a procedure is called *cross-validation*. This approach can

be computationally expensive, but does not waste too much data (as it is the case when fixing an arbitrary test set), which is a major advantage in problem such as inverse inference, where the number of samples is very small. Among others, we use the following *cross-validation* schemes:

- *Leave-one-out: Leave-one-out* (or *LOO*) is one of the simplest cross-validation schemes. Each learning set is created by taking all the samples except one, the test set being the left out sample. Thus, for n samples, we have n different learning and test sets. This cross-validation procedure does not waste much data as only one sample is removed from the learning set.

- *K-fold: K-fold* cross-validation divides the sample set into K disjoint groups of samples, called *folds* (if $K = n$, we retrieve the *LOO*), of equal sizes (if possible). The prediction function is learned using $K - 1$ folds, and the left out fold is used for test.

- *Leave-one-subject-out: Leave-one-subject-out* (or *LOSO*) is a cross-validation scheme that is more specific to the problem of fMRI inverse inference. In inter-subject studies, it allows to test whether the prediction function learned in a given cohort can be generalized to other subjects. Moreover, it removes some subject-specific effects that can bias the prediction function, as there are some images from each subject in the learning set. It simply removes one subject from the data as test set, all the other subjects being the learning set.

It is important to note that both learning the prediction function and *dimension reduction* are performed within the cross-validation loop (see Fig. 3.2), in order to avoid *overfit* (*a.k.a. circular analysis* [Kriegeskorte 09]). In the case of extremely small sample size, an additional step can be the use of permutation tests for estimating the significance of the resulting prediction accuracy [Golland 03].

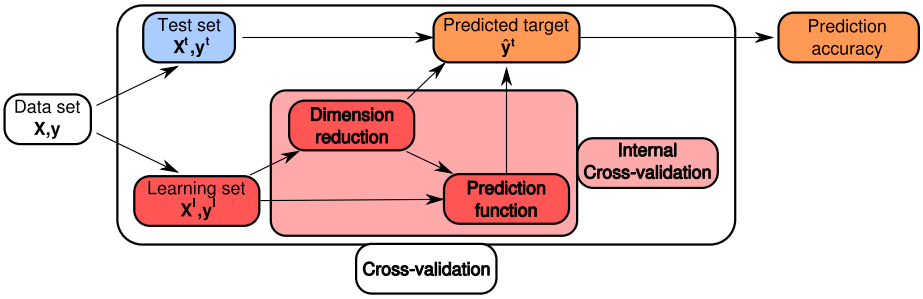


Figure 3.2: Global machine learning framework for inverse inference, with a model selection by cross-validation.

Internal cross-validation

The predictive model relies frequently on internal parameters, and thus different models can be used. We have to select the best model within a set of different models. This is often done by selecting the model that yields the best prediction accuracy. Thus, some cross-validation are also performed within the previously described global cross-validation, in order to select the best set of parameters. Such cross-validation is said to be *internal* or *nested*.

However, the choice of the best cross-validation scheme is difficult. The optimal ratio of data in the learning set/test sets (*e.g.* the number of folds in *K-folds*) is not the same if we want to have a consistent selection of the model, or if we want to test the generalization ability of a model [Larsen 99]. Moreover, the choice of cross-validation dramatically influences the computation time of the inverse inference framework, and thus, has to be done wisely.

3.1.5 Dimension reduction

It has been shown [Hughes 68] that increasing the complexity of the data can first increase the prediction accuracy until an optimal value is reached. After that, further increasing in the dimensionality of the data will reduce the prediction accuracy. This effect is called *curse of dimensionality*, and is crucial in *fMRI* data analysis. Indeed, in the case $p \gg n$ (with *fMRI*, we have typically $p \sim 10^5$ and $n \sim 10^2$) it is always possible to find a prediction function that yields a perfect prediction on the learning data. However, such function can not generalize (*i.e.* provide accurate predictions on new samples) as it has learned non-informative specificities of the learning set, or noise. Such a function is said to *overfit* the learning data. This problem can be overcome by using *dimension reduction* methods, that define a low-dimensional space that contains the predictive information while drastically reducing the dimensionality of the problem.

In *fMRI* inverse inference, *dimension reduction* is carried out with two different objectives, that may or may not be fulfilled. It has to yield good prediction accuracy (*i.e.* extract relevant information), and it has to extract interpretable patterns (*e.g.* constructing a low-dimensional space that corresponds to a reduced number of brain regions).

Prediction function / Dimension Reduction scheme	*Prior ROIs Atlas selection*	*Univariate feature selection*	*Feature Agglomeration*	*PCA*	*RFE*
Kernel Machines (Linear kernels)	[Mitchell 04, Kamitani 05, Ni 08, Knops 09, Chu 10, Rissman 10]	[Cox 03, Mitchell 04, Thirion 06a, Grazia 08]	[Mitchell 04, Fan 06, Grazia 08]	[LaConte 05, Mourao-Miranda 05, Mourao-Miranda 07, Wang 07, Sato 09, Wang 09]	[Martino 08, Hanson 08, Ryali 10]
Kernel Machines (Other kernels)	[Chu 10]	[Cox 03, Grazia 08]	[Davatzikos 05, Grazia 08, Koutsouleris 09]		
Regularized prediction	[Yamashita 08, Rissman 10]	[Carroll 09] [Ryali 10]			
Bayesian models	[Friston 08, Ganesh 08, Ni 08, Yamashita 08, Chu 10]				
Naive Bayes	[Mitchell 04, Palatucci 07, Shinkareva 08]	[Mitchell 04, Rustandi 06]	[Mitchell 04]		
KNN Similarity	[Mitchell 04, Williams 07]	[Sayres 06, Shinkareva 06, Mitchell 04, Mitchell 08]		[Mitchell 04]	
Discriminant analysis	[Haynes 05b]	[Cox 03, Haynes 05a]	[Davatzikos 05]	[Strother 02, Kjems 02, LaConte 03, Carlson 03, Ford 03, Jiang 04, Strother 04, Mourao-Miranda 05, Sato 09]	
Other methods	[Dehaene 98, Haxby 01, Sidtis 03, Koltchinskii 04, Polyn 05, Martinez-Ramon 06, Rissman 10]	[Mørch 97, Onut 04, Langlebe 05, Hutchinson 09, Langs 10]	[Genuer 10]	[Demirci 08]	[Langs 10]

Table 3.1: State of the art of statistical learning in *fMRI inverse inference*.

3.2 Some historical approaches

In this section, we detail the two historical approaches that have been first used in *fMRI-based* inverse inference, namely *Support Vector Machine* and *Discriminant Analysis*. These approaches have been first developed for other applications and are of course by no means limited to *fMRI* inverse problems.

3.2.1 *Support Vector Classification – SVC*

The first prediction function used in inverse inference [Cox 03] has been *Support Vector Machine (SVM)* [Cortes 95], and this approach, that is widely used, has became the reference approach for *fMRI* inverse inference. Its success comes from the fact that it can cope with relatively highly dimensional data, and that it yields a good prediction accuracy with many datasets.

It has been used for *fMRI* inverse inference with linear kernels [Cox 03, Mitchell 04, LaConte 05, Kamitani 05, Mourao-Miranda 05, Thirion 06a, Fan 06, Mourao-Miranda 07, Wang 07, Grazia 08, Martino 08, Hanson 08, Sato 09, Wang 09, Knops 09, Ryali 10, Rissman 10] or non-linear kernels (*RBF*, polynomial kernels) [Cox 03, Davatzikos 05, Grazia 08, Koutsouleris 09]. We describe here the *SVM* and illustrate it on real *fMRI* data. More details about the theory of *kernels machine* can be found in [Shawe-Taylor 04], and the use of others kernel approaches, such as *Kernel ridge*, can be found in [Ni 08, Chu 10].

Kernel trick

Some statistical learning algorithms can be solved using the *dual formulation* that is based on the *Lagrange dual problem*. Interestingly, in the estimation of this *dual problem*, the data enter only in the form of a scalar product $\mathbf{K} = \mathbf{X}^T\mathbf{X}$, called *Gram matrix*. The algorithm is said to work in the *dual space*. As the *Gram matrix* is a $n \times n$ matrix, this approach works in a very low-dimensional representation of the data (when $n \ll p$) compared to the *feature space*, and is thus well-suited for dealing with the high dimensionality issue of *fMRI* data.

More generally, one can use a *feature space mapping* $\Phi(\mathbf{X})$ that can take into account non-linear interactions between samples. The scalar product in the *Gram matrix* is replaced by the *kernel matrix* $\mathbf{K} = \Phi(\mathbf{X})^T\Phi(\mathbf{X})$, where $K(\mathbf{X_i}, \mathbf{X_j})$ is a *kernel*. The mapping is often not explicit as we directly work on the computed inner product. This is known as the *kernel trick* [Aizerman 64]. A *kernel* must satisfy the *Mercer's condition*, *i.e.* must be positive semi-definite, and the most common *kernels* are:

$$\begin{cases} \text{Linear kernel} & K(\mathbf{X_i}, \mathbf{X_j}) = {\mathbf{X_i}}^T\mathbf{X_j} \\ \text{Gaussian kernel} & K(\mathbf{X_i}, \mathbf{X_j}) = \exp^{-\gamma\|\mathbf{X_i}-\mathbf{X_j}\|^2} \end{cases} \tag{3.6}$$

Dual quadratic optimization problem

For binary classification, in the case of a predictive linear model, we introduce the following separating hyperplane:

$$\mathbf{w}\mathbf{X_i} + b = 0 \tag{3.7}$$

which corresponds to the decision function $\mathbf{y_i} = sgn(\mathbf{w}\mathbf{X_i} + b)$. We define the size of the *margin* $m = 1/\|\mathbf{w}\|$ as the distance of a sample to the separating hyperplane. We can rescale the parameters $\mathbf{w}$ and b such that the points closest to the hyperplane satisfy $|\mathbf{w}\mathbf{X_i} + b| = 1$. For the sake of simplicity in the presentation of the method, we will from now on consider $\mathbf{w}$ and b as the rescaled parameters. The hyperplane is thus chosen to separate the set of positive samples from the set of negative samples with maximum margin.

The conditions for classification without training error, *a.k.a. hard-margin* constraints, are:

$$\forall\ 1 \leq i \leq n^l \quad , \quad \mathbf{y_i^l}(\mathbf{w}\mathbf{X_i^l} + b) \geq 1 \tag{3.8}$$

i.e. the predicted target and the true target have the same sign. Minimizing the bound on the empirical risk and the complexity term can be done by minimizing $\|\mathbf{w}\|^2$, which reads:

$$\hat{\mathbf{w}} \qquad = \qquad \frac{1}{2}\min_{\mathbf{w},b} \|\mathbf{w}\|^2 \tag{3.9}$$

$$\text{subject to} \quad y_i^l(\mathbf{w}\mathbf{X_i^l} + b) \geq 1 \quad , \quad \text{for } 1 \leq i \leq n^l \tag{3.10}$$

This constrained optimization problem is solved using Lagrange multipliers α:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n^l} \alpha_i \left(y_i^l(\mathbf{w}\mathbf{X_i^l} + b) - 1\right) \tag{3.11}$$

which yields the following equations at the optimum:

$$\sum_{i=1}^{n^l} \alpha_i y_i^l = 0 \quad \text{and} \quad \hat{\mathbf{w}} = \sum_{i=1}^{n^l} \alpha_i y_i^l \mathbf{X_i^l} \tag{3.12}$$

By using the *kernel trick*, and replacing Eq. 3.12 in Eq. 3.11, we obtain the *dual quadratic optimization problem*:

$$\max_{\alpha} \qquad \sum_{i=1}^{n^l} \alpha_i - \frac{1}{2}\sum_{i=1}^{n^l}\sum_{j=1}^{n^l} \alpha_i \alpha_j y_i^l y_j^l K(\mathbf{X}_i, \mathbf{X}_j) \tag{3.13}$$

$$\text{subject to} \qquad \alpha_i \geq 0,\ i = 1, \ldots, n^l \ \text{ and } \ \sum_{i=1}^{n^l} \alpha_i y_i^l = 0 \tag{3.14}$$

The couples $\{\mathbf{X_i}, y_i\}$ with $\alpha_i \neq 0$ are called *Support Vectors*, and they define the margin. They are the samples that determine the decision function:

$$\hat{y} = \text{sgn}\left(\sum_{i=1}^{n^l} \alpha_i y_i K(\mathbf{X}, \mathbf{X_i}) + b\right) \tag{3.15}$$

The problem in Eq. 3.14 can be solved by different ways. A first approach is the *chunking* algorithm [Vapnik 82] and is based on the fact that removing rows and columns of the matrix with zero Lagrange multipliers do not change the quadratic form. This allows to solve Eq. 3.14 by solving a series of small *QP* problems. Another solution is given in [Osuna 97], and can be seen as a generalization of the *chunking* algorithm. It consists in solving smaller *QP* problems, with a constant size matrix. Finally, *Sequential Minimal Optimization* [Platt 99], is based on the previous approach. It solves at each step the smallest possible optimization problem, which involves only two Lagrange multipliers in the standard *SVM* problem. This can be done analytically, and allows a quick resolution of Eq. 3.14.

Slack variables and C-*SVC*

In order to relax the *hard-margin* constraints (null training error) if the data are not linearly separable, one can introduce the *slack variables* [Cortes 95], which yields the following conditions on the training set:

$$\forall\, 1 \leq i \leq n^l \;\;,\;\; \mathbf{y_i^l}(\mathbf{wX_i^l} + b) \geq 1 - \xi_i \;\;,\; \text{with } \xi_i \geq 0 \tag{3.16}$$

These *slack variables* allow for some classification errors, and yields the following minimization problem:

$$\hat{\mathbf{w}} \quad = \quad \frac{1}{2}\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C\sum_{i=1}^{n^l} \xi_i \tag{3.17}$$

$$\text{subject to} \quad y_i^l(\mathbf{wX_i^l} + b) \geq 1 - \xi_i \;\;,\;\; \text{for } 1 \leq i \leq n^l \tag{3.18}$$

where $C > 0$ is a regularization parameter. One can see that the *slack variables* allow some classification errors on the training set, the value ξ_i denotes the amount by which the corresponding sample is misclassified. The minimization problem defined in Eq.3.18 aims at minimizing the number of misclassified samples $C\sum_{i=1}^{n^l} \xi_i$, with a ℓ^2 norm regularization $\|\mathbf{w}\|^2$. This problem can be also viewed as minimizing an *Hinge loss* with a ℓ^2 norm regularization. The *Hinge loss* is defined as:

$$\ell(y_i^l(\mathbf{wX_i^l} + b)) = (1 - y_i^l(\mathbf{wX_i^l} + b))^+ = \max(y_i^l(\mathbf{wX_i^l} + b), 0) \tag{3.19}$$

This approach is called C-SVC, and is usually solved through the *dual quadratic*

optimization problem:

$$\max_{\alpha} \quad \sum_{i=1}^{n^l} \alpha_i - \frac{1}{2} \sum_{i=1}^{n^l} \sum_{j=1}^{n^l} \alpha_i \alpha_j y_i^l y_j^l K(\mathbf{X}_i, \mathbf{X}_j) \tag{3.20}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C,\ i = 1, \ldots, n^l \ \text{ and } \ \sum_{i=1}^{n^l} \alpha_i y_i^l = 0 \tag{3.21}$$

ν-SVC

However, as increasing C decreases the regularization, it can be useful to replace C by another constant ν in order to have a more intuitive regularization:

$$\max_{\alpha} \quad -\frac{1}{2} \sum_{i=1}^{n^l} \sum_{j=1}^{n^l} \alpha_i \alpha_j y_i^l y_j^l K(\mathbf{X}_i, \mathbf{X}_j) \tag{3.22}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq 1/n^l,\ i = 1, \ldots, n^l \ \text{ and } \ \sum_{i=1}^{n^l} \alpha_i y_i^l = 0 \tag{3.23}$$

$$\sum_{i=1}^{n^l} \alpha_i \geq \nu \tag{3.24}$$

This approach is called *ν-SVC* [Schölkopf 00], and the parameter ν can be seen as an lower bound on the number of support vectors, and an upper bound on the number of samples that lie on the wrong side of the hyperplane. Interestingly, the linear term $\sum_{i=1}^{n^l} \alpha_i$ does not appear in Eq. 3.22, compared to Eq. 3.20, and the *dual quadratic optimization problem* is now quadratically homogeneous in α.

Multi-class *SVC*

For *multi-class SVC*, we use some voting heuristics combined with the previously described binary *SVC* [Hsu 02]. These heuristics are combined with *voting strategy*, in order to make a decision from the different *multi-class* heuristics. Two *multi-class* heuristics are commonly used:

- *one-against-all* [Bottou 94]: binary classifiers are constructed by pooling together the data from all the classes except one, and the decision boundary of this new data set is computed. The *voting strategy* takes the prediction that yields the highest value of $\sum_{i=1}^{n^l} \alpha_i y_i K(\mathbf{X}, \mathbf{X_i}) + b$.
- *one-against-one* [Knerr 90, Friedman 96]: we construct a set of binary classifiers by considering all the possible pairs of classes. This approach yields a total of $K(K-1)/2$ classifiers, for a K classes problem. The training time is shorter than for the *one-against-all* heuristic, and this is

the heuristic used by *LibSVM* [Chang 01]. This heuristic is often combined with a *Max Wins voting strategy*, that predict the class that yields the highest number of prediction within all the different binary classifiers.

These *multi-class* heuristics are not limited to *SVC*, and will be used in this thesis for multi-class prediction (see chapter 6).

3.2.2 *Support Vector Regression – SVR*

Support Vector Machine can also be used for regression, yielding *Support Vector Regression – SVR* [Smola 98, Gunn 98, Smola 04]. We introduce the ϵ-insensitive loss function ℓ_ϵ, that enables sparsity in the *Support Vectors*:

$$\ell_\epsilon = 0 \text{ if } \|\hat{\mathbf{y}} - \mathbf{y}\| < \epsilon \tag{3.25}$$
$$= \|\hat{\mathbf{y}} - \mathbf{y}\| - \epsilon \text{ otherwise} \tag{3.26}$$

where $\hat{\mathbf{y}} = \hat{\mathbf{w}}\mathbf{X}_\mathbf{i}^\mathbf{l} + \hat{b}$ is the prediction. Adding *slack variables* ξ_i, ξ_i^* as in Eq.3.18, the minimization problem of *SVR* is:

$$\hat{\mathbf{w}} = \frac{1}{2}\min_{\mathbf{w},b}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n^l}(\xi_i + \xi_i^*) \tag{3.27}$$

subject to

$$y_i^l - (\mathbf{w}\mathbf{X}_\mathbf{i}^\mathbf{l} + b) \leq \epsilon + \xi_i \ , \ \text{for } 1 \leq i \leq n^l \tag{3.28}$$
$$\mathbf{w}\mathbf{X}_\mathbf{i}^\mathbf{l} + b - y_i^l \leq \epsilon + \xi_i^* \ , \ \text{for } 1 \leq i \leq n^l \tag{3.29}$$

This yields the *dual problem*:

$$\max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i=1}^{n^l}\sum_{j=1}^{n^l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(\mathbf{X}_i, \mathbf{X}_j) \tag{3.30}$$
$$-\epsilon\sum_{i=1}^{n^l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n^l} y_i^l(\alpha_i - \alpha_i^*) \tag{3.31}$$

subject to

$$0 \leq \alpha_i, \alpha_i^* \leq C, \ i = 1, \ldots, n^l \ \text{ and } \ \sum_{i=1}^{n^l}(\alpha_i - \alpha_i^*) = 0 \tag{3.32}$$

The estimate $\hat{\mathbf{w}}$ of the weights is thus:

$$\hat{\mathbf{w}} = \sum_{i=1}^{n^l}(\alpha_i - \alpha_i^*)\mathbf{X}_i \tag{3.33}$$

The choice of ϵ can be addressed using ν-SVR, where ν (in a similar way as ν-SVC) is the upper bound on the fraction of error samples or the lower bound on the fraction of samples inside the ϵ-insensitive tube [Schölkopf 00]. ϵ becomes a variable in the optimization process.

3.2.3 *SVM* for *fMRI* inverse inference

The large success of *SVM* in inverse inference is that it can handle high dimensional data. We evaluate the performance of *SVM* on the ten subjects of the *mental representation of shape* data set (see details in appendix B.2) The average cross-validated classification score (4-fold) obtained for different numbers of voxels and different values of C, is given Fig. 3.3 (left), the chance level being at 25%. We can notice that, when increasing the number of selected voxels (and thus increasing the proportion of possibly irrelevant features in the selected voxels), the prediction remains relatively accurate. The prediction accuracy is also stable for different number of features and different values of C.

Additionally, when dealing with *fMRI* datasets, it has been found that a linear *kernel* often gives better generalization performance than a polynomial *kernel* [Cox 03, Chu 10]. In Fig. 3.3, we give the average cross-validated regression score (4-fold) obtained for different numbers of voxels and different values of C, the parameter γ of the *RBF* kernel being optimized within an *internal cross-validation*. The *RBF SVM* is not stable and does not yield higher accuracy than the linear *SVM*. Thus, we now only use a linear *kernel* for *SVM*.

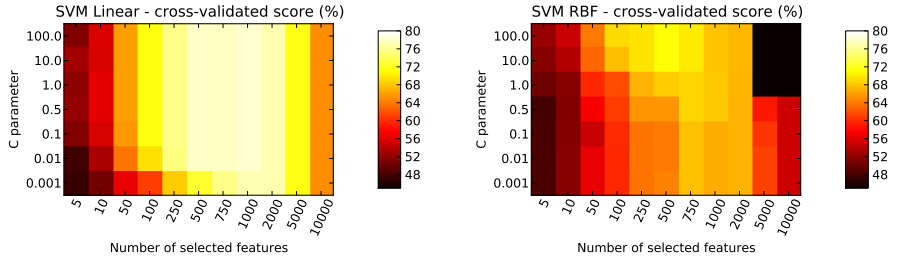


Figure 3.3: *SVC* - Average cross-validated classification score (4-fold) obtained for different numbers of voxels and different values of C for a *linear* kernel (left) and an *RBF* kernel (right).

SVM yields sparsity in the *dual space*, not in the *features space*, as can be seen in Fig. 3.4, for 1000 voxels selected using *F-score*-based *univariate feature selection*. It is thus more difficult to extract the spatial support of the neural code, as all the weighs are non-zero. Different methods have been proposed to have a more usable representation in the primal space: sensitivity maps (i.e. relative changes in class prediction when a given voxel is modified) [Kjems 02]; correlation between each voxel with the paradigm while excluding the images corresponding used to defined the margin (*i.e. support vectors*) [LaConte 05]. However, more generally, *kernel machines* do not provide interpretable maps.

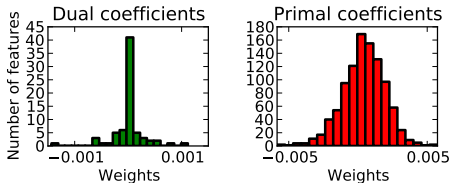


Figure 3.4: *Dual* and *primal* weights found by linear *SVM* ($C = 0.5$) for 1000 voxels. *SVM* yields sparsity in the *dual space*, which is not the case in the *features space*.

3.2.4 Generative models

Classical *generative* models have been among the first approaches used for *inverse inference*, due to their simplicity. The assumptions of the models are very general, and we detail here the models and the corresponding assumptions on $p(\mathbf{X}|\mathbf{y})$.

Gaussian Naive Bayes – GNB

The *Naive Bayes – NB –* algorithm is based on an hypothesis of *conditional independence* on $\mathbf{X}$, *i.e.* an hypothesis of independence between the different features $\mathbf{X_j}$:

$$p(\mathbf{X}|\mathbf{y}) = \prod_{j=1}^{p} p(\mathbf{X^j}|\mathbf{y}) \tag{3.34}$$

and the *Maximum a Posteriori* estimates is:

$$\hat{\mathbf{y}} = \arg\max_k \frac{p(\mathbf{y}=k)\prod_{j=1}^{p} p(\mathbf{X^j}|\mathbf{y}=k)}{\sum_{l=1}^{K} p(\mathbf{y}=l)\prod_{j=1}^{p} p(\mathbf{X^j}|\mathbf{y}=l)} \tag{3.35}$$

$$= \arg\max_k p(\mathbf{y}=k)\prod_{j=1}^{p} p(\mathbf{X^j}|\mathbf{y}=k) \tag{3.36}$$

The simplest hypothesis that can be made for $p(\mathbf{X^j}|\mathbf{y}=k)$ is a gaussian hypothesis, i.e. :

$$p(\mathbf{X^j}|\mathbf{y}=k) = \mathcal{N}(\mu_{j,k}, \sigma^2_{j,k}) \tag{3.37}$$

This model is called *Gaussian Naive Bayes – GNB*, and the *variance-covariance* matrix $\mathbf{\Sigma_k}$ ($\sigma^2_{j,k}$ is the j^{th} element of the diagonal of $\mathbf{\Sigma_k}$) is reduced to a simple diagonal matrix. The different parameters are estimated by *maximum likelihood* on the learning set, and we thus obtain $\forall j \in [1..p]$:

$$\begin{cases} \hat{\mu}_{j,k} = \frac{\sum_{i|y_i^l=k}(x_i^j)^l}{l_k^l} \\ \hat{\sigma}^2_{j,k} = \frac{\sum_{i|y_i^l=k}((x_i^j)^l-\hat{\mu}_{j,k})((x_i^j)^l-\hat{\mu}_{j,k})^T}{l_k^l} \end{cases} \tag{3.38}$$

where l_k^l is the number of samples in the k^{th} class.

GNB has been used for inverse inference [Mitchell 04, Rustandi 06, Palatucci 07, Shinkareva 08]. We evaluate the performance of *GNB* on the ten subjects of the *mental representation of shape* data set (see details in appendix B.2). The average cross-validated classification score (4-fold) obtained for different numbers of voxels is given Fig. 3.5, the chance level being at 25%. There is an optimal number of voxels (250), and the prediction accuracy dramatically decreases with larger numbers, due to the overfit issue of *GNB*. Thus *GNB* is often only used on small *regions of interest*. Moreover, *GNB* does not take into account the covariance between features, and thus may not be well-suited for complex classification task. Consequently, the optimal prediction accuracy (64%) is lower than the one found by *linear SVC* (78%).

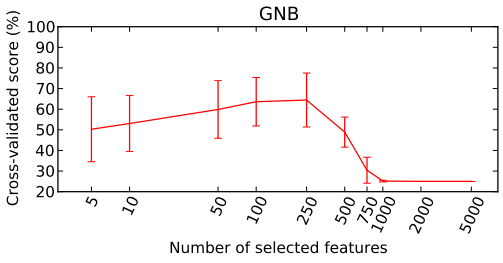


Figure 3.5: *Gaussian Naive Bayes* - Average cross-validated classification score (4-fold) obtained for different numbers of voxels. There is an optimal number of voxels (250 voxels), and the prediction accuracy dramatically decreases for larger number of voxels (overfit issue).

Linear Discriminant Analysis – LDA

In order to better take into account the multivariate structure of the data, one can remove the *conditional independence* hypothesis of the *GNB*. In conjunction with a Gaussian assumption, *Linear Discriminant Analysis – LDA* makes an *homoscedastic assumption*, *i.e.* that the covariance matrices are identical for all classes and have full rank:

$$p(\mathbf{X}|\mathbf{y}=k) = \mathcal{N}(\mu_\mathbf{k}, \mathbf{\Sigma_k}) \quad \text{with} \quad \forall k \in K \ : \ \mathbf{\Sigma_k} = \mathbf{\Sigma} \tag{3.39}$$

This *homoscedastic hypothesis* is equivalent to assume that each class is shifted from the others (same *variance-covariance* matrix, different means). For new samples $\mathbf{X}^t$ in the *test set*, the decision rule is:

$$\hat{\mathbf{y}} = \arg\max_k d(\mathbf{X}^t, k) \tag{3.40}$$

where:

$$d(\mathbf{X}^t, k) = 2\log p(\mathbf{y}^l = k) - (\mathbf{X}^t - \mu_\mathbf{k})^T \, \mathbf{\Sigma}^{-1} \, (\mathbf{X}^t - \mu_\mathbf{k}) \tag{3.41}$$

is the *discriminant function*. Thus, the decision rule is simply based on a linear combination of the samples:

$$d(\mathbf{X}^t, k) - d(\mathbf{X}^t, l) = 2\log\frac{p(\mathbf{y}^l = k)}{p(\mathbf{y}^l = l)} - \mathbf{X}^t\boldsymbol{\Sigma}^{-1}(\mu_\mathbf{k} - \mu_\mathbf{l}) - \frac{1}{2}{\mu_\mathbf{k}}^T\boldsymbol{\Sigma}^{-1}\mu_\mathbf{k} + \frac{1}{2}{\mu_\mathbf{l}}^T\boldsymbol{\Sigma}^{-1}\mu_\mathbf{l} \tag{3.42}$$

and the decision boundary between to classes is linear in $\mathbf{X}$:

$$2\log\frac{p(\mathbf{y}^l = k)}{p(\mathbf{y}^l = l)} - \mathbf{X}^t\boldsymbol{\Sigma}^{-1}(\mu_\mathbf{k} - \mu_\mathbf{l}) - \frac{1}{2}{\mu_\mathbf{k}}^T\boldsymbol{\Sigma}^{-1}\mu_\mathbf{k} + \frac{1}{2}{\mu_\mathbf{l}}^T\boldsymbol{\Sigma}^{-1}\mu_\mathbf{l} = 0 \tag{3.43}$$

The different parameters are estimated by *maximum likelihood* on the learning set:

$$\begin{cases} \hat{\mu}_k = \frac{\sum_{i|y_i^l=k}\mathbf{X_i^l}}{l_k^l} \\ \hat{\boldsymbol{\Sigma}} = \sum_{k=1}^{K}\frac{\sum_{i|y_i^l=k}(\mathbf{X_i^l}-\hat{\mu}_k)(\mathbf{X_i^l}-\hat{\mu}_k)^T}{n^l-K} \end{cases} \tag{3.44}$$

LDA has been used for *inverse inference* [Strother 02, Kjems 02, LaConte 03, Ford 03, Cox 03, Carlson 03, Jiang 04, Strother 04, Haynes 05a, Davatzikos 05, Mourao-Miranda 05, Haynes 05b, Sato 09]. We evaluate the performance of *LDA* on the ten subjects of the *mental representation of shape* data set (see details in appendix B.2). The average cross-validated classification score (4-fold) obtained for different numbers of voxels is given Fig. 3.6, the chance level being at 25%. This scores raises an optimum (for 500 voxels), and then decreases when the number of selected voxels increases. Compared to *GNB* (see Fig. 3.5), *LDA* yields higher prediction accuracy, but is still subject to overfit for high number of voxels.

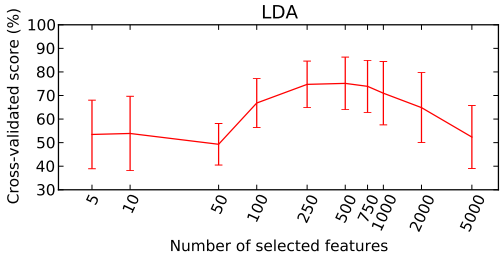


Figure 3.6: *Linear Discriminant Analysis* - Average cross-validated classification score (4-fold) obtained for different numbers of voxels. *LDA* yields higher prediction accuracy and is less subject to overfit for high number of voxels, than *GNB*.

Quadratic Discriminant Analysis – QDA

One can generalize the *LDA* model by removing the *homoscedastic assumption* in the *LDA* model, *i.e.* by making a *heteroscedastic assumption*. The resulting model

is called *Quadratic Discriminant Analysis – QDA*, and the *variance-covariance* matrices $\mathbf{\Sigma^k}$ are different for each class:

$$p(\mathbf{X}|\mathbf{y}=k) = \mathcal{N}(\mu_\mathbf{k}, \mathbf{\Sigma_k}) \ , \ \forall k \in [1, \ldots, K] \tag{3.45}$$

In this case, the *maximum a posteriori* estimate yields:

$$\hat{\mathbf{y}} = \arg\max_k \left\{ 2 \log p(\mathbf{y}=k) - (\mathbf{X}-\mu_\mathbf{k})^T \, \mathbf{\Sigma_k}^{-1} \, (\mathbf{X}-\mu_\mathbf{k}) - \log |\mathbf{\Sigma_k}| \right\} \tag{3.46}$$

where $(\mathbf{X}-\mu_\mathbf{k})^T \, \mathbf{\Sigma_k}^{-1} \, (\mathbf{X}-\mu_\mathbf{k})$ is the *Mahalanobis distance*, and the *discriminant function* is now $d(\mathbf{X}) = 2 \log p(\mathbf{y}=k) - (\mathbf{X}-\mu_\mathbf{k})^T \, \mathbf{\Sigma_k}^{-1} \, (\mathbf{X}-\mu_\mathbf{k}) - \log |\mathbf{\Sigma_k}|$.

For a new sample $\mathbf{X}_i^t$ in the *test set*, the decision rule is the same as the *LDA* (see Eq.3.40), but now is based on a quadratic combination of the samples in the learning set.

The different parameters are thus obtain as :

$$\begin{cases} \hat{\mu}_k = \frac{\sum_{i|y_i^l=k} \mathbf{X_i^l}}{l_k^l} \\ \hat{\mathbf{\Sigma}}_k = \frac{\sum_{i|y_i^l=k} (\mathbf{X_i^l}-\hat{\mu}_k)(\mathbf{X_i^l}-\hat{\mu}_k)^T}{l_k^l - 1} \end{cases} \tag{3.47}$$

QDA suffers from the fact that $\mathbf{\Sigma}_k$ is estimated with very few data points. The estimate $\hat{\mathbf{\Sigma}}_k$ is thus very unstable, and dramatically depends on the training set. Thus, it has not been used for *inverse inference*.

Regularization of *variance-covariance* matrices for *Discriminant Analysis*

A solution is to *regularize* the covariance estimation. Such an approach has not been used yet for inverse inference, but has emerged as an interesting method in *resting-state fMRI* data analysis [Varoquaux 10], among other applications. We give here different ways for less variable estimation of the *variance-covariance* matrices [Chen 10] based on *shrinkage*, and thus, allowing to use *LDA* and *QDA* in ill-posed problem such as ours.

The classical estimator of $\mathbf{\Sigma}_k$ is defined as:

$$\hat{\mathbf{\Sigma}}_k = \frac{\sum_{i|y_i^l=k} \tau_{i,k} \tau_{i,k}^T}{l_k^l - 1} \tag{3.48}$$

with $\tau_{i,k} = \mathbf{X_i^l} - \hat{\mu}_k$. This estimator is unbiased, but has a high variance and is ill-posed for $p \gg n$, which is the case in *fMRI*-based *inverse inference*. A most well-conditioned estimate of $\mathbf{\Sigma}_k$ is given by:

$$\tilde{\mathbf{\Sigma}}_k = \frac{\mathrm{Tr}(\hat{\mathbf{\Sigma}}_k)}{p} \mathbf{I} \tag{3.49}$$

which has a lower variance but higher bias than $\hat{\mathbf{\Sigma}}_k$. A trade-off between these two estimates can be obtained using:

$$\bar{\mathbf{\Sigma}}_k = (1-\rho)\hat{\mathbf{\Sigma}}_k + \rho \tilde{\mathbf{\Sigma}}_k \tag{3.50}$$

where $0 \leq \rho \leq 1$ is the *shrinkage* coefficient. Thus, we can see that the estimate given by Eq.3.50 is the classical estimator of $\mathbf{\Sigma}_k$ where the diagonal is reduced, in order to have a lower variance. The *Ledoit-Wolf* estimation of ρ can be obtained using [Ledoit 03]:

$$\hat{\rho} = \frac{\sum_{i|y_i^l=k} \|\tau_{i,k}\tau_{i,k}^T - \hat{\mathbf{\Sigma}}_k\|2}{n^2 \left(\text{Tr}(\hat{\mathbf{\Sigma}}_k^2) - \text{Tr}^2(\hat{\mathbf{\Sigma}}_k)/p\right)} \tag{3.51}$$

By using this *regularized* estimate of the *variance-covariance* matrix, it seems possible to used both *LDA* and *QDA* for *inverse inference*. We have not focused on such approaches during this thesis, but we detail *regularization* approaches for linear model in the next section.

3.3 Regularization

As previously stated, one of the main problem in *inverse inference*, is the huge dimensionality of the *fMRI* data, as the problem of learning the prediction function is plagued by the *curse of dimensionality*. A commonly used solution is the *regularization* of the weights used in the parametric prediction function, *i.e.* the values of the weights are constrained by some parameters. Such approach can be used for estimating a *variance-covariance* matrix in the case of very few samples (see section 3.2), and we detail here how *regularization* can be used within a linear model for *fMRI inverse inference*. Additionally to the regularization performed in approaches such as *SVM*, regularization of linear model has recently been successfully used for *fMRI*-based prediction [Carroll 09, Ryali 10, Rissman 10], both in *regression* and *classification*.

3.3.1 General form of regularization

Let us recall the following predictive linear model:

$$\mathbf{y} = f(\mathbf{X}, \mathbf{w}, b) = F(\mathbf{X}\mathbf{w} + b) \tag{3.52}$$

This estimation can be done by minimizing the difference between the estimated target $\hat{\mathbf{y}}$ and the true target $\mathbf{y}$. This difference can be seen as a function of the weights $\mathbf{w}$, called *loss function* (or more simply *loss*) and noted $\ell(\mathbf{w})$.

Loss function

The *loss function* represents the cost associated with an error in the estimation of $\mathbf{y}$ and it is usually chosen to be easily computed and convex. In regression settings, we usually use the *quadratic loss (or ℓ_2 loss)*:

$$\ell(\mathbf{w}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \tag{3.53}$$

and the following *logistic loss* for classification settings:

$$\ell(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp^{-y_i(\mathbf{X_i}^T\mathbf{w})}\right) \tag{3.54}$$

Ordinary Least Squares

In regression settings, the easiest way to solve the linear problem defined in Eq. 3.52 is to use the *Ordinary Least Squares estimate* (*OLS*), that minimizes the ℓ_2 *loss* without any constraints, *i.e.* $\hat{\mathbf{w}} = \text{argmin}_{\mathbf{w}}\ \ell(\mathbf{w})$. Assuming that $\mathbf{X}^T\mathbf{X}$ is invertible, the resulting estimate of the weights is given by $\hat{\mathbf{w}}^{\mathbf{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, and $\hat{\mathbf{w}}^{\mathbf{ols}}$ is a bias free estimate of $\mathbf{w}$. This estimator is based on the computation of $(\mathbf{X}^T\mathbf{X})^{-1}$ and is thus very sensitive to the conditioning of $\mathbf{X}$. One small eigenvalue of $\mathbf{X}^T\mathbf{X}$ yields a very unstable *OLS* estimation. In the case of ill-posed problem such as ours ($n \ll p$), $\mathbf{X}^T\mathbf{X}$ is not invertible, *i.e.* some eigenvalues of $\mathbf{X}^T\mathbf{X}$ are 0.

Regularization ℓ_p

A common way to perform a better estimation $\hat{\mathbf{w}}$, called **regularization**, is to sacrifice some bias for reducing the estimator variance [Hoerl 70]. A standard approach to perform the estimation of $\mathbf{w}$ with regularization uses penalization of a maximum likelihood estimator. It leads to the following minimization problem:

$$\hat{\mathbf{w}} = \underset{\mathbf{w},b}{\text{argmin}}\ \ell(\mathbf{w}) + \lambda J(\mathbf{w})\ \ ,\ \ \lambda \geq 0 \tag{3.55}$$

where $\lambda J(\mathbf{w})$ is the regularization term and $\ell(\mathbf{w})$ is the loss function. The parameter λ balances between the loss function $\ell(\mathbf{w})$ and the penalty $J(\mathbf{w})$. Note that the intercept b is not included in the regularization term. In the case of regression where $\ell(\mathbf{w}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$, and with $J(\mathbf{w}) = \|\mathbf{w}\|_\gamma^\gamma$, the problem defined in Eq. 3.55 is known as *Bridge Regression* [Frank 93]. Let us recall that:

$$\|\mathbf{w}\|_\gamma = \left(\sum_{j=1}^{p} |x^j|^\gamma\right)^{1/\gamma} \tag{3.56}$$

is the γ-norm, with the particular case of *Euclidean norm* for $\gamma = 2$, and *infinity norm* (or *maximum norm*) for $\gamma = \infty$. An approach is *sparse* if some weights are null, and in the case of ℓ_p norm penalization, sparsity is obtained for $p \leq 1$ [Nikolova 00].

3.3.2 *Ridge Regression* - ℓ_2 regularization

Ridge Regression is the special case of a ℓ_2 norm regularization [Hoerl 70]. The main idea is to penalize the *OLS* estimate by using $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}$ instead of $(\mathbf{X}^T\mathbf{X})^{-1}$, with $\lambda \geq 0$. This is equivalent to the following minimization problem:

$$\hat{\mathbf{w}}^r = \underset{\mathbf{w},b}{\text{argmin}}\ \ell(\mathbf{w}) + \lambda\|\mathbf{w}\|_2\ \ ,\ \ \lambda \geq 0 \tag{3.57}$$

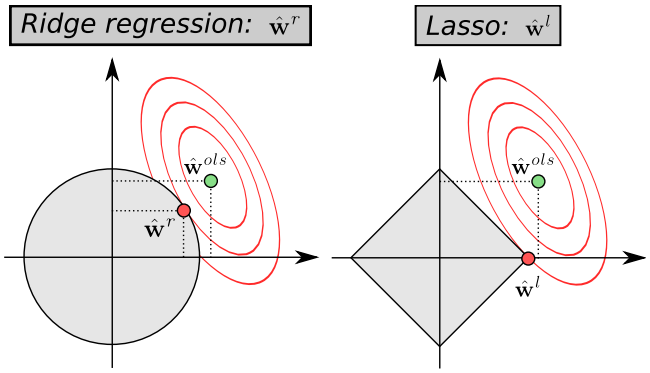


Figure 3.7: Illustration of the *Ridge regression* (ℓ_2 norm regularization) and *Lasso* (ℓ_1 norm regularization) estimations, in the case of two orthogonal regressors. The estimate $\hat{\mathbf{w}}$ is found at the intersection between the isocontours of the quadratic loss centered on the *OLS* estimate(red ellipsoids) and the constraint (gray region). With *Ridge regression*, the solution is not sparse, with two non-null components for $\hat{\mathbf{w}}^r$. With *Lasso*, the intersection is found on one axis, yielding a sparse solution (one non-null component for $\hat{\mathbf{w}}^l$).

The resulting estimation $\hat{\mathbf{w}}^r$ is more stable than $\hat{\mathbf{w}}^{ols}$, and we have the relationship to the *OLS* solution:

$$\hat{\mathbf{w}}^r = \left(\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^-\right)^{-1} \hat{\mathbf{w}}^{ols} \tag{3.58}$$

where $(.)^-$ is the *Moore–Penrose pseudoinverse*. The ℓ_2 norm regularization is equivalent to set a Gaussian prior on $\mathbf{w}$, $p_\lambda(\mathbf{w}) = C_\lambda \exp^{-\lambda\|\mathbf{w}\|_2^2}$, where C_λ is a constant depending on λ.

If we note $s(\hat{\mathbf{w}}^r)$ the residual sum of squares for the estimation $\hat{\mathbf{w}}^r$ of $\mathbf{w}$, we have $s(\hat{\mathbf{w}}^r) = (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}^r)^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}^r) = s(\hat{\mathbf{w}}^{ols}) + g(\hat{\mathbf{w}}^r)$ with $s(\hat{\mathbf{w}}^{ols})$ the residual sum of squares of the *OLS* estimation and $g(\hat{\mathbf{w}}^r)$ a quadratic form of $(\hat{\mathbf{w}}^r - \hat{\mathbf{w}}^{ols})$. The isocontours of $s(\hat{\mathbf{w}}^r)$ are hyper-ellipsoids centered on $\hat{\mathbf{w}}^{ols}$ (see an illustration Fig.3.7), $\hat{\mathbf{w}}^r$ being the intersection of those ellipsoids with the ℓ_2 constraint.

Properties of *Ridge Regression* for inverse inference

Ridge Regression exhibits groups of correlated features and works well when all the features are equally relevant. However, when correlations between variables increase, *Ridge Regression* tends to yield equal coefficients in order to minimize their ℓ_2 norm [Tibshirani 96]. This effect may be important in *fMRI* data analysis, as the underlying metabolic effects can extend among wide regions. Additionally, some consistency results for *Ridge Regression* in the case of n fixed and $p \to \infty$, which is the case for *fMRI* studies, can be found in [Luo 09].

Illustration of *Ridge Regression* on real data

We evaluate the performance of *Ridge Regression* on the ten subjects of the *mental representation of size* data set (see details in appendix B.2). The average cross-validated regression score (4-fold) obtained for different numbers of voxels and different values of the regularization parameter λ, is given Fig. 3.8 (left). We can see that the prediction accuracy is not too sensitive to the regularization parameter λ, and slightly decreases for high number of voxels (10^4). The *path* (*i.e.* the values of the weights in function of the regularization parameter) of *Ridge Regression* is given Fig. 3.8 (right) for 11 different features of different relevance (selected features have different ranks of *F-scores*). The weights are not sparse, even if some voxels have zero weights. Additionally, we can notice the grouping effect, *i.e.* the fact that similar features (in red) have a similar weights when the regularization increases. Moreover, all the relevant features (red and yellow) are extracted by *Ridge Regression*, even if they are correlated.

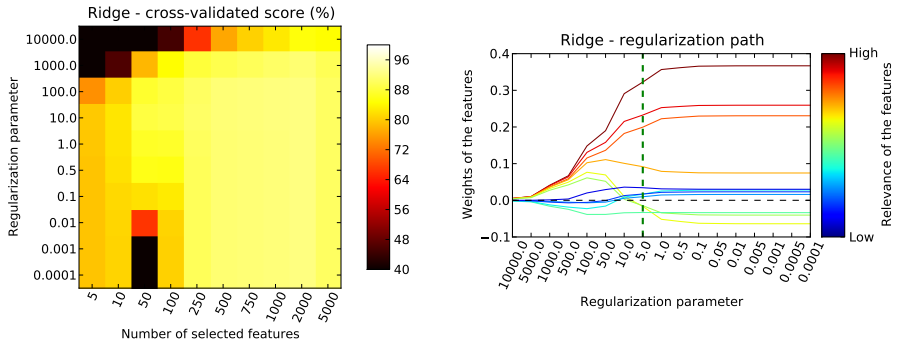


Figure 3.8: *Ridge Regression* - Average cross-validated regression score (4-fold) obtained for different numbers of voxels and different values of the regularization parameter λ (left). The path of *Ridge Regression* (right) for 11 features, shows the grouping effect (relevant features are in red, irrelevant ones in blue).

3.3.3 *Lasso* - ℓ_1 Regularization

We now study the case of the ℓ_1 regularization, also called *Lasso* for *Least Absolute Shrinkage and Selection Operator* [Tibshirani 96]. *Lasso* tries to deal with the weakness of *Ridge Regression* by forcing the uninformative features to have zero weights, and yields a sparse model. Moreover, *Lasso* can also be seen as a particular case of *subset selection*, but is less variable as it is not a discrete process as other subset selection methods. *Lasso* corresponds to the following minimization problem:

$$\hat{\mathbf{w}}^l = \underset{\mathbf{w},b}{\operatorname{argmin}} \ \ell(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 \ \ , \ \ \lambda \geq 0 \tag{3.59}$$

Lasso is equivalent to a Laplacian prior on $\mathbf{w}$, $p_\lambda(\mathbf{w}) = C_\lambda \exp^{-\lambda\|\mathbf{w}\|_1}$, where C_λ is a constant depending on λ.

In the same way as in *Ridge Regression*, isocontours of $s(\hat{\mathbf{w}}^l)$ are ellipsoids centered on $\hat{\mathbf{w}}^{ols}$. The sparsity of *Lasso* comes from the fact that, while the constraint of *Ridge regression* is a p-dimensional sphere, the constraint of *Lasso* has some singularities along the axes. The intersection is more likely to occur on these corners, yielding the nullity of many coefficients. This is easily seen on a simple case of two orthogonal regressors (Fig. 3.7). *Lasso* is also closely related to the *non-negative garrote* [Breiman 95], that stronger penalizes the small coefficients of *OLS*. By working on the initial *OLS* weights that are very sensitive to the conditioning of $\mathbf{X}$, the *non-negative garrote* is also sensitive to ill-posed problems.

The minimization problem defined in Eq. 3.59 can be solved using *Lars (Least Angle Regression)* [Trevor 02], *coordinate descent* [Friedman 07, Friedman 10] or iterative procedures based on *proximal operator* [Daubechies 04, Combettes 05].

Properties of *Lasso* for inverse inference

One of the most important property of *Lasso* is that when $n \ll p$, as in *fMRI* data, the solution yielded by *Lasso* as at most n non-zero coefficients [Osborne 99]. Indeed, it has been shown experimentally [Tibshirani 96] that *Lasso* often does not pick the correct model, and selects only one feature from a set of correlated voxels. Thus, the resulting model can be difficult to interpret, as the selection can be relatively unstable (*i.e.* the support of $\mathbf{w}$ can vary a lot). Moreover, *Lasso* does not yield consistent model when λ is chosen to minimize the prediction error [Leng 06], *i.e.* does not choose the right model when the number of samples tends to infinity. Yet, in the case of inverse inference, the best model is often selected by minimizing the prediction error within an *internal cross-validation*, and thus, the model yielded by *Lasso* should be interpreted carefully.

Illustration of *Lasso* on real data

We evaluate the performance of *Lasso* on the ten subjects of the *mental representation of size* data set (see details in appendix B.2). The average cross-validated regression score (4-fold) obtained for different numbers of voxels and different values of the regularization parameter λ, is given Fig. 3.9 (left). We can see that the prediction accuracy is far more sensitive to the regularization parameter λ than *Ridge Regression*. Moreover, the optimal value of λ is different for different number of features. The *path* of *Lasso* is given Fig. 3.9 (right) for 11 different features of different relevance. The weights are very sparse. We can notice than the relevant features (red) are not selected at the same point on the path. Contrariwise to *Ridge Regression*, slightly different regularization parameter can yield very different models.

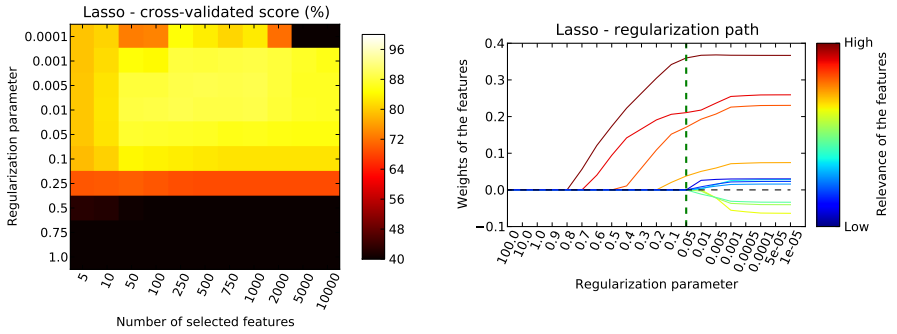


Figure 3.9: *Lasso* - Average cross-validated regression score (4-fold) obtained for different numbers of voxels and different values of the regularization parameter λ (left). The path of *Lasso* (right) for 11 features, shows that the weights are sparse, but *Lasso* chooses only one features from a set of correlated voxels (relevant features are in red, irrelevant ones in blue).

3.3.4 *Elastic net* and *Sparse Multinomial Logistic Regression* - $\ell_1 + \ell_2$ Regularization

More recently, a new approach, called *Elastic net*, has been proposed [Zou 05]. *Elastic net* deals with the limitation of the two previous approaches, by using a combined ℓ_1 and ℓ_2 penalization. Indeed, it is sparser than *Ridge Regression* and yields more interpretable models by setting many weights to zeros. *Elastic net* also allows to extract more features than samples and correlated features, contrariwise to *Lasso*. *Elastic net* yields the following minimization problem:

$$\hat{\mathbf{w}}^l = \underset{\mathbf{w},b}{\operatorname{argmin}} \ \ell(\mathbf{w}) + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2^2 \ , \ \ \lambda_1 \geq 0 \ , \ \lambda_2 \geq 0 \tag{3.60}$$

Trivially, *Elastic net* admits *Lasso* ($\lambda_2 = 0$) and *Ridge regression* ($\lambda_1 = 0$) as limit cases. Another parametrization of *Elastic net* can be used, where λ_2 is denoted λ and $\rho = \frac{\lambda_1}{\lambda_1+\lambda_2}$ is the fraction of the ℓ_1 norm in the total norm. As *Lasso*, *Elastic net* can be solved using *Lars* [Trevor 02] or *coordinate descent* [Friedman 07, Friedman 10]. An interesting result [De Mol 09] is the consistency of *Elastic net* for both prediction and variable selection, which is important when seeking for an interpretable model.

Illustration of *Elastic Net* on real data

We evaluate the performance of *Elastic Net* on the ten subjects of the *mental representation of size* data set (see details in appendix B.2). The average cross-validated regression score (4-fold) obtained for different values of the regularization parameter λ_1, and different values of the mixing parameter ρ, is given in Fig. 3.10 (left), for 5000 voxels selected using *F-score*-based *univariate feature*

selection (we fix the number of voxels in order to let both λ_1 and ρ vary). The prediction accuracy is high, and there is a correlation between the two parameters λ_1 and ρ. The *path* of *Elastic Net* is given Fig. 3.10 (right) for 11 different features of different relevance, and for $\rho = 0.2$. *Elastic Net* yields null weights for irrelevant features (blue), and selects relevant features (red), even if they are correlated. *Elastic Net* is thus a good compromise between *Lasso* and *Ridge regression*. It seems an attractive approach for inverse inference, as we expect to extract some groups of correlated features, while seeking for an interpretable model (*i.e.* few selected groups). In practice, the parameters of *Elastic Net* are selected by nested cross-validation.

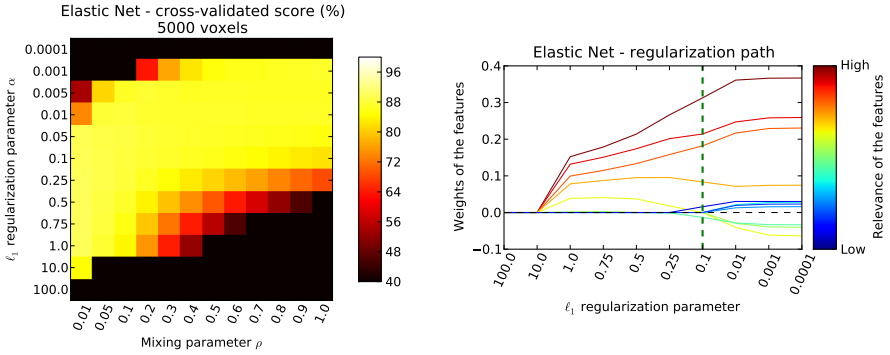


Figure 3.10: *Elastic Net* - Average cross-validated regression score (4-fold) obtained for different values of the regularization parameter λ_1, and different values of the mixing parameter ρ, for 5000 voxels (left). The path of *Elastic Net* (right, with $\rho = 0.2$) for 11 features, shares the properties of both *Ridge Regression* and *Lasso* (relevant features are in red, irrelevant ones in blue). Irrelevant features have null weights, and relevant features are selected together, even if they are correlated.

Sparse Multinomial Logistic Regression – SMLR

The previous combined ℓ_1 and ℓ_2 penalization can be used in classification settings, and is called *Sparse Multinomial Logistic Regression (SMLR)* [Krishnapuram 05]. This algorithm is based on a *logistic loss*, defined in Eq. 3.54 (see [Hastie 03] for more details). We now give the mathematical formulation for the binary case with $\mathbf{y} \in \{-1, 1\}^n$. The *logistic regression* model defines the conditional probability of y_i given the data $\mathbf{X}_i$ as:

$$p(y_i|\mathbf{X_i}, \mathbf{w}) = \frac{1}{1 + \exp^{-y_i(\mathbf{X_i}^T\mathbf{w})}} \tag{3.61}$$

The corresponding *loss* and the *loss* gradient read:

$$\begin{cases} \ell(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp^{-y_i(\mathbf{X_i}^T\mathbf{w})}\right) \\ \nabla\ell(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n} \frac{y_i\mathbf{X_i}}{1+\exp^{y_i(\mathbf{X_i}^T\mathbf{w})}} \end{cases} \tag{3.62}$$

This approach will not be further detailed in this thesis, but will be used for comparison purpose in our experiments.

3.4 Bayesian regularization

Regularization, presented in the previous section, is of great interest as it can deal with the high dimensionality within the prediction function. However, this regularization can be very sensitive to the regularization parameter (*e.g. Lasso*). *Bayesian* methods can be used to tune the regularization to the data, and thus avoid an optimization of the *regularization* parameter in a nested cross-validation. However, these approaches are often computationally expensive. Bayesian regularization have been used for inverse inference in [Friston 08, Ganesh 08, Ni 08, Yamashita 08, Chu 10].

3.4.1 Priors

We note $\mathcal{F}(\theta)$ a prediction function, parametrized by a set of parameters θ. Learning the prediction function is equivalent to estimating the set of parameters $\hat{\theta}$ that best fits the data. In the estimation of the model, one may want to introduce some prior knowledge on the parameters. We call **prior** this information introduced as a distribution over some parameters, *e.g.* $p(\theta_1)$ for the parameter θ_1. This distribution is set before processing the data. The parameters of a prior distribution are called **hyper-parameters**. This description is based on the following *Bayes' theorem*:

$$p(\theta|\{\mathbf{X},\mathbf{y}\}) = \frac{p(\{\mathbf{X},\mathbf{y}\}|\theta)p(\theta)}{p(\{\mathbf{X},\mathbf{y}\})} \tag{3.63}$$

With:

- $p(\{\mathbf{X},\mathbf{y}\}|\theta)$ the *likelihood*: it expresses how probable it is to observe $\{\mathbf{X},\mathbf{y}\}$ given θ.
- $p(\{\mathbf{X},\mathbf{y}\}) = \int p(\{\mathbf{X},\mathbf{y}\}|\theta)p(\theta)\mathrm{d}\theta$ the *marginal probability* of the data: it is used as a normalizing constant.
- $p(\theta)$ the *prior* over the parameters: it expresses the knowledge that we can have about θ before processing the data.
- $p(\theta|\{\mathbf{X},\mathbf{y}\})$ the *conditional probability* (or *posterior probability*): it expresses the uncertainty on θ after observing the data.

An estimate of the parameters, called *Maximum a Posteriori (MAP)*, can be found by maximizing the *posterior probability*:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, L_{\{\mathbf{X},\mathbf{y}\}}(\theta)p(\theta) = \underset{\theta}{\operatorname{argmax}} \left\{\log L_{\{\mathbf{X},\mathbf{y}\}}(\theta) + \log p(\theta)\right\} \quad (3.64)$$

In the case of an uninformative prior, we simply find the *maximum likelihood* estimate. The *maximum-likelihood* approach does not use any additional information and only needs the specification of a model $\mathcal{M}$. A contrario, the *maximum a posteriori* approach needs the definition of priors on the parameters. This choice of priors can yield a better estimation but bad priors can dramatically decrease the performance of the estimator.

Informative and non-informative priors

Distributions used as priors can be *informative* or *non-informative*. *Informative* priors add strong information on the parameters, and are difficult to set up, because they do not take into account the data and can strongly influence the estimation of the model. However, when we have little information or hypothesis about the data, it is often useful to introduce a prior that has only a small (or no) influence on the posterior estimate. Such a prior is called *non-informative*. Additionally, the choice of a prior is unfortunately often constrained by practical considerations. It is often interesting to choose some particular prior which gives a posterior distribution of the same form as the prior distribution. Such a prior is called a *conjugate prior*.

3.4.2 *Bayesian Ridge Regression – BRR*

Gaussian Bayesian regression is based on the following Gaussian assumption:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \alpha) = \prod_{i=1}^{n} \mathcal{N}(y_i|\mathbf{X_i}\mathbf{w}, \alpha^{-1}) \quad (3.65)$$

We assume that the noise ϵ is Gaussian with a precision (inverse of the variance) α, *i.e.* $p(\epsilon|\alpha) = \mathcal{N}(0, \alpha^{-1}\mathbf{I_n})$. For *regularization* purpose, one can add the following prior on $\mathbf{w}$:

$$p(\mathbf{w}|\lambda) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}} \exp^{-\frac{\lambda\|\mathbf{w}\|^2}{2}} = \mathcal{N}(\mathbf{w}|0, \lambda^{-1}\mathbf{I_p}) \quad (3.66)$$

The resulting model is called *Bayesian Ridge Regression –BRR*. As explained previously, λ is an **hyper-parameter** and the prior performs a *shrinkage* or *regularization*, by constraining the values of the weights to be small. Indeed, with a large value of λ, the Gaussian is narrowed around zero which does not allow large values of $\mathbf{w}$; with low value of λ, the Gaussian is flattened, which allows higher values for $\mathbf{w}$. We have:

$$\begin{aligned} p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha, \lambda) &\propto p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \alpha)p(\mathbf{w}|\lambda) & (3.67)\\ &= \mathcal{N}(\mathbf{w}|\mu, \Sigma) \ , & (3.68) \end{aligned}$$

where:

$$\begin{cases} \mu = \alpha \Sigma \mathbf{X}^T \mathbf{y} \\ \mathbf{\Sigma} = (\lambda \mathbf{I_p} + \alpha \mathbf{X}^T \mathbf{X})^{-1} \end{cases} \tag{3.69}$$

By choosing $\lambda = 0$, we have an uniform (uninformative) prior and we retrieve the *maximum-likelihood* estimate. We have the *log likelihood*:

$$\ln p(\mathbf{y}|\alpha, \lambda) = \frac{p}{2} \ln \lambda + \frac{n}{2} \ln \alpha - \frac{\alpha}{2} \sum_{i=1}^{n} (y_i - \mathbf{X_i}\mu)^2 - \frac{\lambda}{2} \mu^T \mu - \frac{1}{2} \ln |\Sigma^{-1}| - \frac{n}{2} \ln 2\pi \tag{3.70}$$

and the parameters α and λ can be estimated by maximizing Eq. 3.70:

$$\begin{cases} \hat{\lambda} = \frac{\gamma}{\mu^T \mu} \\ \hat{\alpha} = \frac{n - \gamma}{\sum_{i=1}^{n} (\mathbf{y_i} - \mathbf{X_i}\mu)^2} \end{cases} , \tag{3.71}$$

where:

$$\gamma = \sum_{i=1}^{p} \frac{\alpha s_i}{\lambda + \alpha s_i} , \tag{3.72}$$

where s_i are the eigenvalues of $\mathbf{X}^T \mathbf{X}$.

In order to have a full Bayesian framework and to avoid degenerate models, one can add classical Γ priors on α and λ:

$$\Gamma(\alpha; \alpha_1, \alpha_2) = \alpha_2^{\alpha_1} x^{\alpha_1 - 1} \frac{\exp^{-x\alpha_2}}{\Gamma(\alpha_1)} \text{ and } \Gamma(\lambda; \lambda_1, \lambda_2) = \lambda_2^{\lambda_1} x^{\lambda_1 - 1} \frac{\exp^{-x\lambda_2}}{\Gamma(\lambda_1)} \tag{3.73}$$

and the parameters update now reads:

$$\begin{cases} \hat{\lambda} = \frac{\gamma + 2\lambda_1}{\mu^T \mu + 2\lambda_2} \\ \hat{\alpha} = \frac{n - \gamma + 2\alpha_1}{\sum_{i=1}^{n} (\mathbf{y_i} - \mathbf{X_i}\mu)^2 + 2\alpha_2} \end{cases} , \tag{3.74}$$

In the experiments detailed in this thesis, we choose $\lambda_1 = \lambda_2 = \alpha_1 = \alpha_2 = 10^{-6}$, *i.e.* weakly informative priors. *Bayesian Ridge Regression* is solved using an iterative algorithm. Starting with $\alpha = \frac{1}{\text{var}(\mathbf{y}^t)}$ and $\lambda = 1$, we iteratively evaluate μ and $\mathbf{\Sigma}$ using Eq. 3.69, and use these values to estimate γ, $\hat{\lambda}$ and $\hat{\alpha}$, using Eq. 3.72 and Eq. 3.74. The convergence of the algorithm is monitored by the convergence of $\mathbf{w}$, and the algorithm is stopped if $\|\mathbf{w}_{s+1} - \mathbf{w}_s\|^1 < 10^{-3}$, where $\mathbf{w}_s$ and $\mathbf{w}_{s+1}$ are the values of $\mathbf{w}$ in two consecutive steps.

We evaluate the performance of *Bayesian Ridge Regression* on the ten subjects of the *mental representation of size* data set (see details in appendix B.2). The average cross-validated regression score (4-fold) obtained for different numbers of voxels is given Fig. 3.11. We can see that when increasing the number of selected voxels (and thus increase the proportion of possibly irrelevant features in the selected voxels), the prediction accuracy is stable, as the regularization allows to deal with high dimensional data.

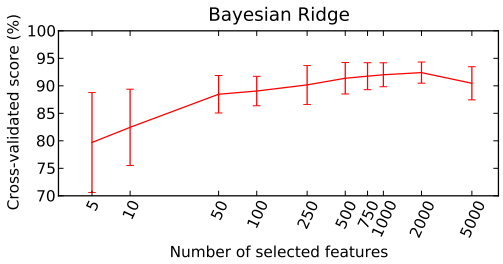


Figure 3.11: *Bayesian Ridge Regression* - Average cross-validated regression score (4-fold) obtained for different numbers of voxels.

Bayesian Ridge Regression is related to *Ridge Regression*, as it performs a regularization equals to *Ridge Regression* with a regularization parameter of λ/α. However, the *Bayesian* framework has the advantage to determine the parameters α and λ through the direct maximization of the *marginal likelihood*. By contrast, *Ridge Regression* tunes the parameters using a *cross-validation* on a grid, and thus requires to define a range and a step for the values of the parameters to be tested. We give in Fig. 3.12 the cross-validated regression score (4-fold) obtained for different numbers of voxels, for *Bayesian Ridge* (red), and the minimum (blue) and maximum (green) values obtained by the classical *Ridge* among the different values of the regularization parameter. We can notice that *Bayesian Ridge* is similar to the maximum value found by *Ridge*. Indeed, it automatically adapts the regularization parameter to the data, and similar results should be found by optimizing the regularization parameter of *Ridge* by *internal cross-validation*.

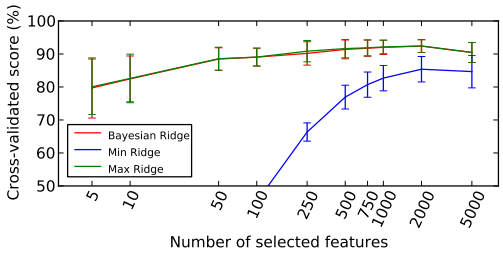


Figure 3.12: Cross-validated regression score (4-fold) obtained for different numbers of voxels, for *Bayesian Ridge* (red), and the minimum (blue) and maximum (green) values obtained by the classical *Ridge* among the different values of the regularization parameter (see Fig. 3.8).

3.4.3 *Automatic Relevance Determination – ARD*

We introduce here a more complex prior for $\mathbf{w}$, namely the *Automatic Relevance Determination – ARD* – [MacKay 92, Neal 96], where we assume that each weight w_i is drawn in a Gaussian distribution, centered on zero and with a precision λ_i ($\lambda_i \neq \lambda_j$ if $i \neq j$). *ARD* yields sparse model by this choice of hyper-parameters, a contrario to *Lasso* that deals with sparsity directly in the *features space*. We have:

$$p(\mathbf{w}|\lambda) = \mathcal{N}(0, \mathbf{A}^{-1}) \text{ with } diag\ (\mathbf{A}) = \Lambda = \{\lambda_1, ..., \lambda_p\} \tag{3.75}$$

As for *Bayesian Ridge Regression*, we find that:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha, \lambda) = \mathcal{N}(\mathbf{w}|\mu, \Sigma) \text{ with } \left\{ \begin{array}{l} \mu = \alpha \Sigma \mathbf{X}^T \mathbf{y} \\ \mathbf{\Sigma} = (\mathbf{A} + \alpha \mathbf{X}^T \mathbf{X})^{-1} \end{array} \right. \tag{3.76}$$

and the parameters α and λ are again estimated by maximizing $\ln p(\mathbf{y}|\alpha, \lambda)$:

$$\left\{ \begin{array}{l} \hat{\lambda} = \frac{\gamma}{\mu^T \mu} \\ \hat{\alpha} = \frac{n - \gamma}{\sum_{i=1}^{n} (\mathbf{y_i} - \mathbf{X_i} \mu)^2} \end{array} \right. \tag{3.77}$$

with $\gamma = 1 - \lambda_i \Sigma_{ii}$, and Σ_{ii} is the i^{th} diagonal component of Σ. One can add priors on α and γ as in *Bayesian Ridge Regression*:

$$\left\{ \begin{array}{l} \hat{\lambda} = \frac{\gamma + 2\lambda_1}{\mu^T \mu + 2\lambda_2} \\ \hat{\alpha} = \frac{n - \gamma + 2\alpha_1}{\sum_{i=1}^{n} (\mathbf{y_i} - \mathbf{X_i} \mu)^2 + 2\alpha_2} \end{array} \right. \tag{3.78}$$

and *ARD* is evaluated through a similar iterative procedure as *Bayesian Ridge Regression*.

ARD suffers from some convergence issues that are discussed in more details in chapter 4.

ARD for inverse inference

We evaluate the performance of *Automatic Relevance Determination* on the ten subjects of the *mental representation of size* data set (see details in appendix B.2). The average cross-validated regression score (4-fold) obtained for different numbers of voxels is given Fig. 3.13. The prediction accuracy increases while increasing the number of voxels in the model. This is due to the more adaptive regularization performed by *ARD*. The weights found by *ARD* for one subject, and 5000 selected voxels, are compared to the weights found by *Bayesian Ridge Regression* in Fig. 3.14. The weights are very sparse, and, due to the high adaptability of *ARD*, some voxels have higher weights than in *Bayesian Ridge Regression*, while more voxels have zero weights.

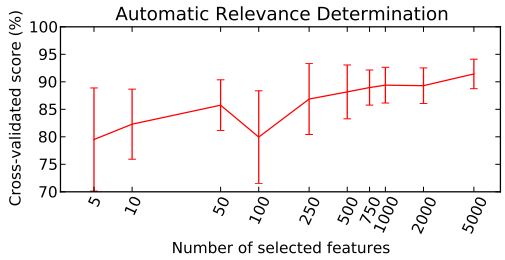


Figure 3.13: *Automatic Relevance Determination* - Average cross-validated regression score (4-fold) obtained for different number of voxels.

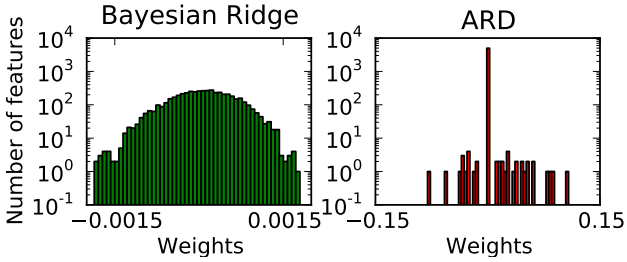


Figure 3.14: Weights found by *Bayesian Ridge Regression* and *ARD* for one subject, and 5000 selected voxels.

Relevance Vector Machine – RVM

Relevance Vector Machine – RVM [Tipping 00, Tipping 01] is an approach that combines a *Bayesian* framework with an *ARD* prior, and the *kernel trick*. By using the *ARD* prior, *RVM* allows to keep very few *Vector Machines*. For classification, *RVM* are combined with a *logistic function*. We do not give here example of use of *RVM* for *fMRI* inverse inference, as the approach is very similar to *ARD*.

3.5 Dimension reduction

One of the main difficulties in *fMRI* analysis is the high dimensionality of the data, especially in view of the low number of samples (a few tens). Among other solutions, *dimension reduction* has been widely used.

The most frequently used approach is *features selection*. This method simply consists in removing non-relevant features from the features space, based on different criteria. The features finally used in the predictive model are thus a subset of the initial features, there is no creation of new features. The *features selection* approaches can be based on *regions of interest* (they do not take into account the target to be predicted), *univariate* (they deal with each voxel

independently), or *multivariate* (they consider different voxels together).

3.5.1 Regions of interest

The most simple *feature selection* consists in selecting few *Regions Of Interest* (or *ROIs*) in the brain, and to keep only the voxels that belong to these *ROIs*. This selection is often performed using an anatomical atlas, or by performing a localizer scan. It allows to make explicit inference on the involvement of particular regions in information encoding. However, one major drawback of this approach is that it depends on a strong prior knowledge on the (supposed) relevant regions of the brain, and can thus miss unknown or unexpected relevant regions. This very simple and basic approach will not be further detailed, and has been used for *fMRI* inverse inference, in (among others) [Dehaene 98, Haxby 01, Sidtis 03, Hanson 04, Mitchell 04, Koltchinskii 04, Kamitani 05, Polyn 05, Haynes 05b, Martinez-Ramon 06, Williams 07, Palatucci 07, Friston 08, Ganesh 08, Yamashita 08, Knops 09, Rissman 10].

3.5.2 Univariate feature selection

In the case of *univariate feature selection*, the features are selected independently from each other, based on the computation of a score $g_i = g(x_i)$ for each feature (*e.g. F-score*, correlation-based score, activation-based score, significance values). The selection is performed by thresholding this score to a given value, by keeping the k best scores, or by using a p-value on this score. Among others, it has been used in [Mørch 97, Cox 03, Mitchell 04, Onut 04, Ji 04, Haynes 05a, Langlebe 05, Shinkareva 06, Rustandi 06, Thirion 06a, Sayres 06, Shinkareva 08, Grazia 08, Mitchell 08, Martino 08, Carroll 09, Hutchinson 09, Langs 10, Ryali 10].

These methods are quick and easy to implement. However, they suffer from two major drawbacks. First, they cannot avoid the redundancy of information. Indeed, two features with redundant information will have a similar score g and will be both selected. Additionally, such univariate approach tends to select voxels with high *SNR*, and can miss relevant voxels that are implied in a multivariate coding of the target and have a weaker *SNR*. A set of features selected jointly should be more informative than a group of features selected independently.

Classification settings

The standard score for *univariate feature selection* in *fMRI* relies on a *F-statistic*, which reads:

$$F_j = \frac{n-K}{K-1} \frac{\sum_{k=1}^{K} l_k^l (\mu_j^k - \mu_j)^2}{\sum_{k=1}^{K} \sum_{i|y_i=k} (x_i^j - \mu_j^k)} = \frac{n-K}{K-1} \frac{(\sigma^b)^2}{(\sigma^w)^2} \tag{3.79}$$

where μ_j^k is the average of the j^{th} feature for the samples in the class k, μ_j is the average of the j^{th} feature for all the images, and l_k is the number of samples

in the k^{th} class. This can also be written with the *between class variance* $(\sigma^b)^2 = \sum_{k=1}^{K} l_k^l (\mu_j^k - \mu_j)^2$ and the *within class variance* $(\sigma^b)^2 = \sum_{k=1}^{K} \sum_{xi|y_i=k} (x_i^j - \mu_j^k)$. This approach is also called *Analysis of Variance – Anova*.

Regression settings

For regression analysis, we introduce the following linear model:

$$\hat{y_i^j} = \hat{w_{1,j}} x_i^j + \hat{w_{0,j}} \tag{3.80}$$

The null hypothesis is that knowing $\mathbf{X^j}$ does not provide extra information about $\mathbf{y}$, *i.e.* $H_0 \ : \ w_{1,j} = 0$, and the alternative hypothesis is $H_1 \ : \ w_{1,j} \neq 0$. The resulting *F-statistic* reads:

$$F_j = n - 2 \frac{\sum_{i=1}^{n} (\hat{y_i^j} - \bar{\mathbf{y}})^2}{\sum_{i=1}^{n} (\hat{y_i^j} - \mathbf{y}_i)^2} \tag{3.81}$$

where $\bar{\mathbf{y}}$ is the mean value of $\mathbf{y}$, $\sum_{i=1}^{n} (\hat{y_i^j} - \bar{\mathbf{y}})^2$ is the variance of the regression model defined in Eq.3.80, and $\sum_{i=1}^{n} (\hat{y_i^j} - \mathbf{y}_i)^2$ is the residual variance. This analysis is closely related to the *GLM* used for creating the *activation maps* (see chapter 2), but is adapted here to the notations of this chapter. The *F-statistic* defined Eq.3.81 follows the F distribution $F_{1,n-2}$.

***Univariate feature selection* for inverse inference**

In *fMRI* inverse inference, *univariate feature selection* is widely used due to its speed and its simplicity. We evaluate the performance of *univariate feature selection* on the ten subjects of the *mental representation of shape* data set (see details in appendix B.2). In this section, we fix the prediction function (SVC, $C = 1$) and study the influence of the number of voxels. The average cross-validated classification score (4-fold) obtained for different numbers of voxels, is given Fig. 3.15 (bottom). There is an optimal number of selected voxels (500) which yields the highest prediction accuracy. For higher number of voxels (5000), we keep too much irrelevant features and the classifier (*SVM*) overfits the learning set. Thus, a crucial step is to choose the optimal number of features to be selected (this is often done by *internal cross-validation*).

We can see for one subject (top) that even a simple *univariate feature selection* can be used to retrieve the relevant regions of interest (occipital lobe). The voxels are selected by regions, due to the information redundancy in neighboring voxels. However such an approach is not able to take into account the *multivariate* structure of the data.

3.5.3 *Multivariate* feature selection

Alternatively, in the case of a *multivariate* feature selection, the features are selected by taking into account the fact that they can share information. In this

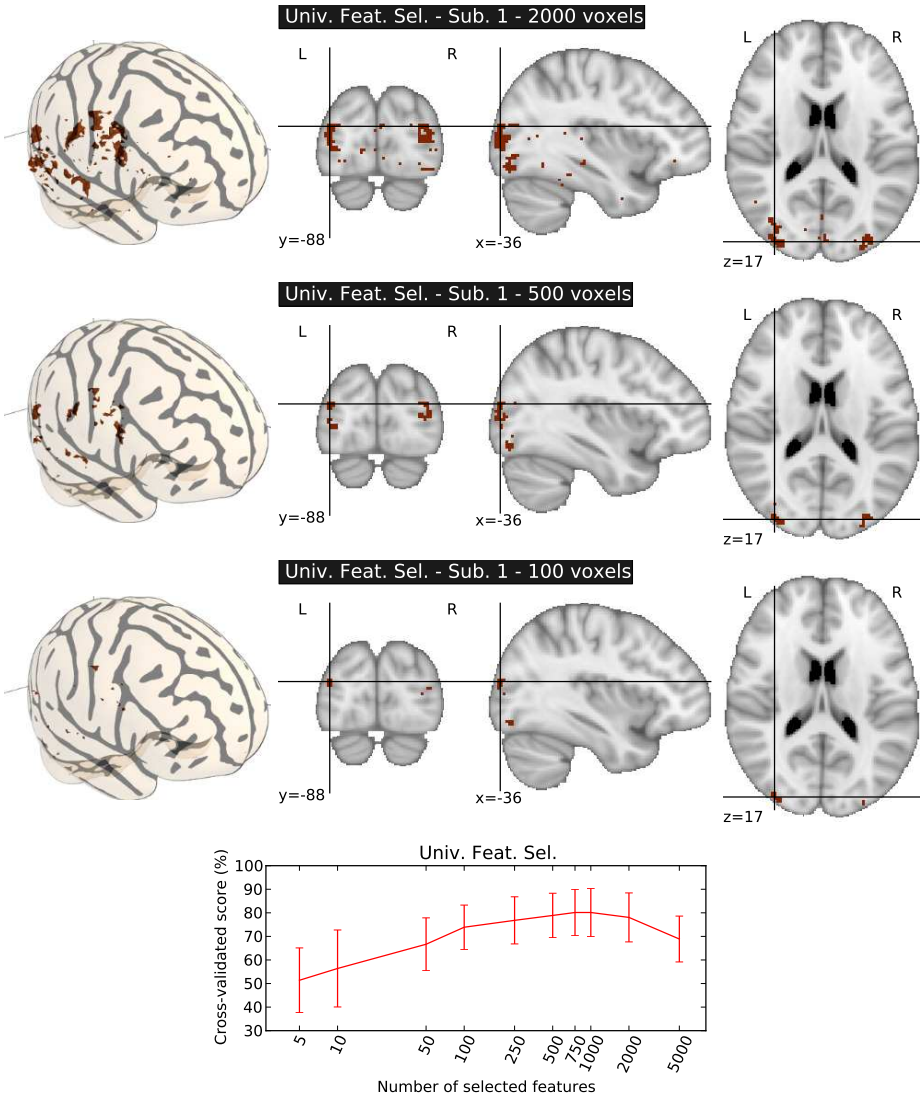


Figure 3.15: *Univariate feature selection - shapes recognition* data set. Top - Representation of the voxels selected by *univariate feature selection* using an *F-score* as ranking criterion, for 2000, 500 and 100 voxels. As it is univariate, the feature selection does not extract a fine predictive pattern, and selects possibly redundant information. Bottom - Average cross-validated regression score (4-fold) obtained for different number of voxels, with SVC ($C = 1$).

case, the score g is computed using a group of features $g(x_i, \ldots, x_j)$. This kind of selection is believed to perform better than most classical *univariate* methods,

as they better respect the hypothesis of population coding and extract features by taking into account the correlation between them. However, they suffer from a combinatorial complexity and are thus not often used in *fMRI* inverse inference. Indeed, for an exhaustive search, we have to consider all the possible subsets of features of all possible sizes, which can be impossible to do when p is large enough (which is the case in *fMRI* studies). Different strategies can be used (see [Kohavi 97]) to overcome this combinatorial limitation, such as *Forward* (starting from an empty subset of features, and then adding iteratively features to this subset), *Backward* (starting from a set containing all the features, then removing iteratively features from this subset) and *Stepwise – Forward-Backward* (starting from an empty subset of features, add iteratively features to this subset, while allowing to remove features from the subset at each step). An example of such *stepwise* strategy for *fMRI* inverse inference can be found in [Michel 08].

Recursive feature elimination – RFE

Recursive feature elimination – RFE – is a recent *Backward* strategy for *multivariate feature selection* that has been first introduced for genetic analysis [Guyon 02], and is related to backward feature elimination. It consists in iteratively removing irrelevant features based on some characteristics of a prediction function. *RFE* has been successfully used in *fMRI inverse inference* [Hanson 08, Martino 08, Ryali 10].

We define the *active subset* of features $\mathcal{S}$, as the set of features that are considered in the model. At each step, *RFE* removes from the *active subset* a given number of features (or a percentage γ of the *active subset*) with the smallest rank, until the number of voxels in $\mathcal{S}$ is smaller than a given number s. The removed features are kept ranked in a subset $\mathcal{R}$. The ranking is based on the weights of a classifier $\mathbf{w_i}$ (the weights of the prediction model). The output of *RFE* are nested subsets of features, and thus, an additional step of model selection is thereby required. The original approach, called *SVM-RFE*, used the weights obtained by *SVM*, but any other predictive model might be used.

We evaluate the performance of *SVM-RFE* on the ten subjects of the *mental representation of shape* data set (see details in appendix B.2). The average cross-validated classification score (4-fold) obtained for different numbers of voxels, with SVC ($C = 1$), is given Fig. 3.16 (bottom). The prediction accuracy globally increases as we remove irrelevant features, and an optimal pattern of about 700 voxels is found. If we remove more voxels, the prediction accuracy decreases. The prediction accuracy is very unstable across the different subjects, and a computationally expensive selection of the relevant model should be done within a cross-validation framework. Moreover the *SVM-RFE* heuristic may not be optimal.

The resulting predictive patterns for two subjects (top) show that *SVM-RFE* might extract very sparse patterns of voxels in the regions of interest (visual cortex). However, these resulting predictive patterns are difficult to interpret due to their extreme sparsity, and are unstable across subjects.

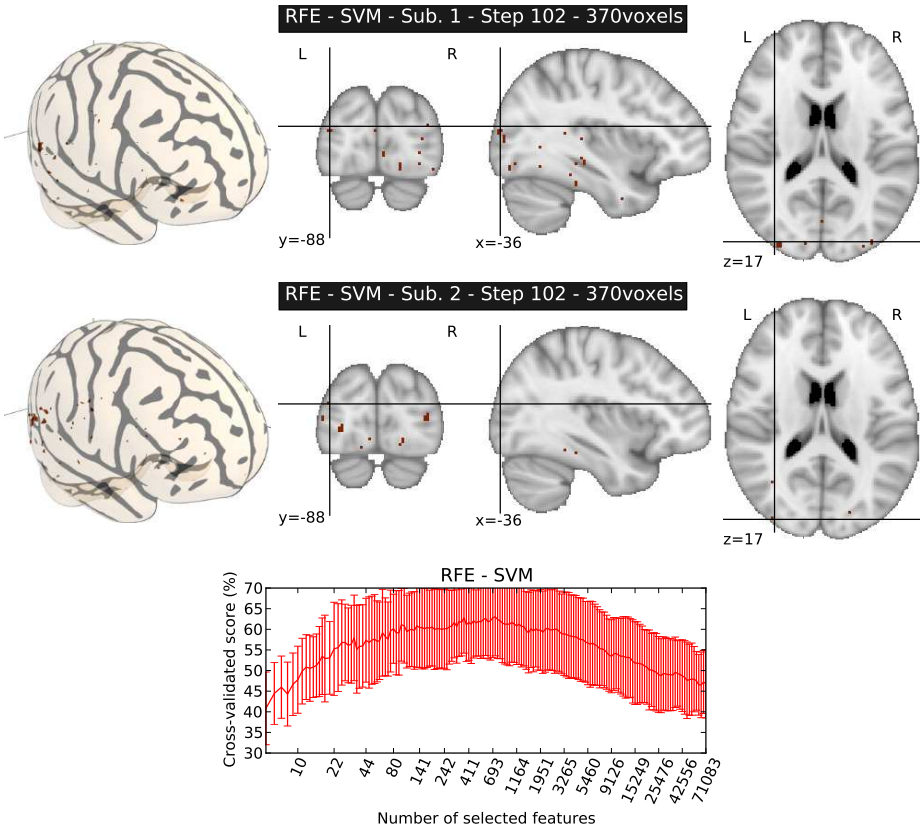


Figure 3.16: *Recursive feature elimination - shapes recognition* data set. Top - Representation of the voxels selected by *SVM-RFE* for two different subjects. The multivariate patterns are sparse but are unstable across subjects. Bottom - Average cross-validated regression score (4-fold) obtained for different number of voxels, with SVC ($C = 1$).

3.5.4 Features agglomeration

Feature agglomeration is a less common dimension reduction method, but it seems well-suited for *fMRI* data [Mitchell 04, Ji 04, Davatzikos 05, Fan 06, Grazia 08, Genuer 10]. The principle is to group features together, using a given criterion, and to create a single feature from all the features within a group (*e.g.* by averaging). Such methods can be used to take into account the spatial structure of the data as a prior. A contribution of this thesis has been to develop a method for introducing information about the target to be predicted in the clustering process (see chapter 5).

3.5.5 *Principal component analysis – PCA*

Principal Component Analysis – PCA is a method for extracting the components of higher variance in data. These components are called *principal components*, the r first components will usually be kept (in the case of *fMRI* data, each component can be viewed as an image). Thus, the decomposition of each sample of $\mathbf{X}$ on these r components can be used to reduce the dimensionality of the data in $\mathbf{X_r} \in \mathbb{R}^{n \times r}$. This method has been used within an *fMRI inverse inference* framework in [Strother 02, Kjems 02, Carlson 03, LaConte 03, Jiang 04, Strother 04, LaConte 05, Mourao-Miranda 05, Wang 07, Mourao-Miranda 07, Demirci 08, Wang 09, Sato 09, Koutsouleris 09].

PCA is an unsupervised dimension reduction approach, *i.e.* it does not take into account the target $\mathbf{y}$. Prior to the *PCA*, it is required to remove the mean of the data matrix $\mathbf{X}$. Then, we perform a *Singular Value Decomposition – SVD –* on the matrix $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{3.82}$$

The r first *principal components* $\mathbf{P_r}$ can be found by taking the r first rows of $\mathbf{V}^T$, and we have the reduced matrix $\mathbf{X_r} = \mathbf{X}{\mathbf{P_r}}^T$.

We evaluate the performance of *Principal Component Analysis* on the ten subjects of the *mental representation of shape* data set (see details in appendix B.2). The average cross-validated classification score (4-fold) obtained for different numbers of voxels, with SVC ($C = 1$), is given Fig. 3.17 (bottom). The prediction accuracy raises an optimum for a number of *principal components* of 50, and then remains relatively stable. However, the optimal prediction accuracy is lower than the one found using an *F-score*-based *univariate feature selection*. We represent the three first *principal components* (top) and we can see that these maps are not easy to interpret from a neuroscientific point of view, even if the third *principal component* shows an important amount of variability in the visual cortex.

3.5.6 Built-in *feature selection*

Let us mention here that sparsity inducing *regularizations* such as *Lasso* or *SMLR* include a built-in feature selection. By setting many features to have zero weights, such approaches extract predictive patterns from the data, while training the prediction function. This is illustrated in the chapter 4, based on a *Multi-class Sparse Bayesian Regression*.

3.6 Conclusion - Statistical learning for *fMRI inverse inference*

In this chapter we have seen that the prediction accuracy of predictive models quantifies the information carried by the features used in the predictive model, and thus allows to assess whether these features (*i.e.* voxel-based signal) belong to the spatial support of the neural coding. Many methods have been used

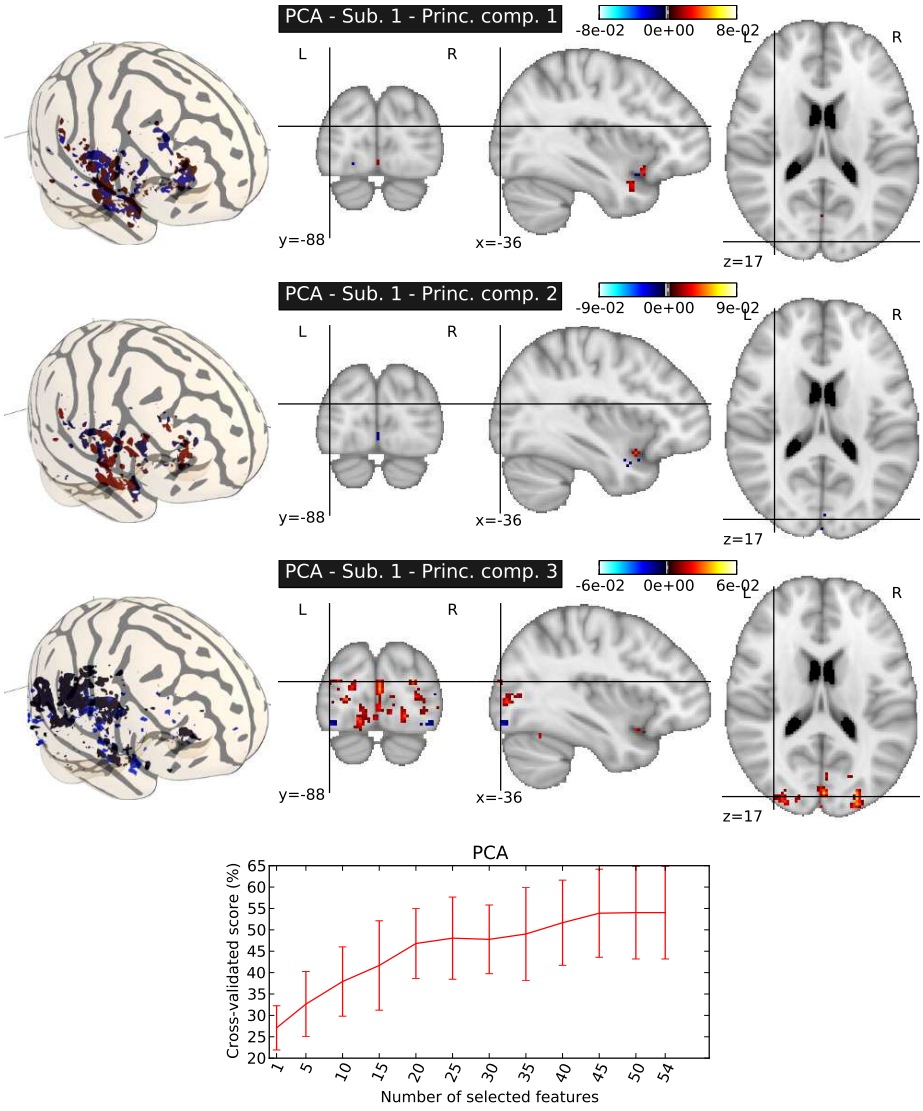


Figure 3.17: *Principal component analysis - shapes recognition* data set. Top - Representation of the three first *principal components*: the resulting maps are not easy to interpret from a neuroscientific point of view. Bottom - Average cross-validated regression score (4-fold) obtained for different number of voxels, with SVC ($C = 1$).

for few years, and most of them have a good predictive ability. However, these *machine learning* methods have most often been used without considering the

characteristics of *fMRI* data, and thus, they suffer from some major limitations. First, they do not take into account the image structure of *fMRI data*: considering the spatial structure of the data can increase the prediction accuracy and extract a more plausible spatial layout of a given neural code. Secondly, the resulting models are often difficult to interpret, as the weights can be too sparse (few voxels within the whole brain). Thus, the weighted maps do not allow a clear interpretation of the spatial organization of the neural coding, especially compared to classical *SPMs*. Finally, *dimension reduction* is often only used to increase the prediction accuracy without aiming at creating useful maps for neuroscientific studies.

Based on these considerations, we can define the following requirements for a good *statistical learning* algorithm for extracting information from *fMRI* data:

1. **A multivariate model**: The information of interest can be distributed over distant brain regions. The *statistical learning* algorithm should be able to account for combinations of signals over these different brain sites, hence it should be a multivariate approach. Indeed, *multivariate pattern analysis* is crucial to make accurate predictions.
2. **Taking into account the spatial structure of the data**: Due to the spatial structure of *fMRI* data, there is a local redundancy of the predictive information, which should be considered in the feature building procedure, *e.g.* by replacing voxel-based signals by local averages.
3. **A multi-scale approach**: Given that the investigated regions are wide if there is little prior information, while the truly informative regions can be relatively tiny, we need an approach that focuses on compact sub-regions of the search volume: a multi-scale approach might thus be useful to optimize the definition of predictive regions. Unlike purely geometrical clustering approaches, procedures that use the signal for clustering might better respect the underlying data structure.

In the following chapters, we present the three major contributions of this thesis, that try to implement these requirements. In the chapter 4, we detail a *Bayesian Regularization*, that regularizes groups of voxels differently, and thus yields an adaptive regularization. A *Supervised Clustering* approach, that creates clusters of voxels informed by the predictive task, is presented in chapter 5. Finally, in chapter 6, we expose the *Total Variation regularization*, that introduces the spatial structure of the data within a *regularization-based* approach.

Publications

The methods presented in this chapter have been used during this thesis in the following neuroscientific studies:

- M. Lebreton, S. Jorge, V. Michel, B. Thirion and M. Pessiglione. *An automatic valuation system in the human brain : evidence from functional neuroimaging*. Neuron 64, 3, 2009.

- E. Eger, V. Michel, B. Thirion, A. Amadon, S. Dehaene and A. Kleinschmidt. *Deciphering Cortical Number Coding from Human Brain Activity Pattern*. Current Biology. 2009, 19:1608.

- A. Knops, B. Thirion, E.M. Hubbard, V. Michel and S. Dehaene. *Recruitment of an area involved in eye movements during mental arithmetic*. Science. 2009 Jun 19;324(5934):1583-5.

- A. Bachrach, A. Gramfort, V. Michel, E. Cauvet, B. Thirion and C. Pallier. *Decoding of syntactic trees*. In prep.

Some methodological works have been presented in:

- V. Michel, C. Damon, and B. Thirion. Mutual information-based feature selection enhances *fMRI* brain activity classification. In 5th Proc. IEEE ISBI, pages 592-595, 2008.

- R. Genuer, V. Michel, E. Eger, and B. Thirion Random forests based feature selection for decoding *fMRI* data. In COMPSTAT 19th International Conference on Computational Statistics, pages 372 , 2010.

4

Multi-Class Sparse Bayesian Regression

In this chapter, we detail a novel *Bayesian regularization* approach based on a multi-class organization of the features, with the aim of adapting the amount of regularization to the informative content of each feature class. This approach is called *Multi-Class Sparse Bayesian Regression (MCBR)*. After detailing the priors of the *MCBR*, we propose two different estimation frameworks and we illustrate *MCBR* on both simulated and real data.

Contents

4.1 *Priors for Multi-Class Sparse Bayesian Regression*

In this section, we detail a model developed during this thesis for *Bayesian Regression*. We group the features into Q different classes, and regularize these classes differently. Regularization is performed in each class separately, leading to a stable and adaptive regularization. This approach, called *Multi-Class Sparse Bayesian Regression (MCBR)*, is thus an intermediate between *Bayesian Ridge Regression (BRR)* and *Automatic Relevance Determination (ARD)*. It reduces the number of parameters estimated by *ARD*, and is far more adaptive than *Bayesian Ridge Regression*. One another great asset of the proposed approach in *fMRI inverse inference* is that it creates a clustering of the features, and thus yields useful maps for brain mapping.

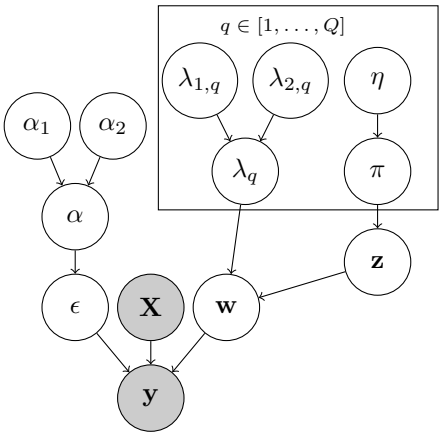
Sparse bayesian regularizations have been used in few studies for *fMRI* inverse inference. Indeed, these approaches can be computationally expensive on high dimensional data, and thus are often abandoned in favor of more easily optimized regularization methods such as *Lasso* or *Elastic net*. Most of the approaches used state of the art *sparse bayesian regularizations* and classical sparsity promoting priors such as *ARD* [Yamashita 08, Ni 08, Chu 10]. These studies used sparsity as built-in feature selection, and do not aim at introducing sparsity based on neuroscientific assumptions. A more interesting approach is the *Bayesian regression* detailed in [Friston 08], that is the closest work to our approach. The weights of the model are defined by $\mathbf{w} = U\eta$, where U is a matrix defining as set of spatial patterns (one pattern by column), and η is the decomposition of $\mathbf{w}$ in the basis defined by U. The sparsity is introduced within the covariance of η, that is assumed to be diagonal with only m possible different values $\text{cov}(\eta) = \exp(\lambda_1)\mathbf{I^{(1)}} + \ldots + \exp(\lambda_m)\mathbf{I^{(m)}}$. The matrices $\mathbf{I^{(i)}}$ are diagonal and defined subsets of columns of U sharing similar variance $\exp(\lambda_i)$. Due to is class-based model, this approach is similar to the one proposed in this chapter, but the construction of I relies on ad hoc voxel selection steps, so that there is no proof that the solution is correct.

4.1.1 Model and priors

We recall the linear model for regression:

$$\mathbf{y} = f(\mathbf{X}, \mathbf{w}, b) = \mathbf{X}\,\mathbf{w} + b \ , \tag{4.1}$$

and now detail the priors and parameters of the model. The complete generative model is summarized in Fig.4.1.



$$
\begin{array}{rcl}
\mathbf{y} & = & \mathbf{X}\,\mathbf{w} + \epsilon + b \\
\epsilon & \sim & \mathcal{N}(0, \alpha^{-\mathbf{1}}\mathbf{I_n}) \\
\alpha & \sim & \Gamma(\alpha; \alpha_1, \alpha_2) \\
\mathbf{z} & \sim & \prod_{j=1}^{p}\prod_{q=1}^{Q} \pi_q^{\delta_{jq}} \\
\pi & \sim & Dir(\eta) \\
\mathbf{w} & \sim & \mathcal{N}(0, \mathbf{A}^{-1}) \\
 & \text{with} & diag\,(\mathbf{A}) = \{\lambda_{z_1}, ..., \lambda_{z_p}\} \\
\lambda_q & \sim & \Gamma(\lambda_q; \lambda_{1,q}, \lambda_{2,q})
\end{array}
$$

Figure 4.1: Graphical model of *Multi-Class Sparse Bayesian Regression – MCBR*.

Priors on the noise

We use classical priors for regression, (see chapter 3), and we model the noise as an *i.i.d.* Gaussian variable:

$$\epsilon \sim \mathcal{N}(0, \alpha^{-\mathbf{1}}\mathbf{I_n}) \tag{4.2}$$
$$\alpha \sim \Gamma(\alpha; \alpha_1, \alpha_2) \tag{4.3}$$

where α is the precision parameter, and Γ stands for the *gamma density* with two hyper-parameters α_1, α_2:

$$\Gamma(x; \alpha_1, \alpha_2) = \alpha_2^{\alpha_1} x^{\alpha_1 - 1} \frac{\exp^{-x\alpha_2}}{\Gamma(\alpha_1)} \tag{4.4}$$

Priors on the class assignment

In order to combine the sparsity of *ARD* with the stability of *BRR*, we introduce an intermediate representation, in which each feature j belongs to one class among Q indexed by a discrete variable z_j ($\mathbf{z} = \{z_1, \ldots, z_m\}$). All the features within a class $q \in \{1, .., Q\}$ share the same precision parameter λ_q, and we use the following prior on $\mathbf{z}$:

$$\mathbf{z} \sim \prod_{j=1}^{p}\prod_{q=1}^{Q} \pi_q^{\delta_{jq}} \tag{4.5}$$

where δ is *Kronecker's* δ, defined as:

$$\begin{cases} \delta_{jq} = 0 \text{ if } j \neq q \\ \delta_{jq} = 1 \text{ if } j = q \end{cases} \tag{4.6}$$

We finally introduce an additional Dirichlet prior on π:

$$\pi \sim Dir(\eta) \tag{4.7}$$

with an hyper-parameter η. By updating at each step the probability π_k of each class, it is possible to prune classes. This model has no spatial constraints, and thus is not spatially regularized.

Priors on the weights

As in *ARD*, we make use of an independent Gaussian prior for the weights:

$$\mathbf{w} \sim \mathcal{N}(0, \mathbf{A}^{-1}) \text{ with } diag\,(\mathbf{A}) = \{\lambda_{z_1}, ..., \lambda_{z_p}\} \tag{4.8}$$

where λ_{z_j} is the precision parameter of the j^{th} feature, with $z_j \in \{1, \ldots, Q\}$. We introduce the following prior on λ_q:

$$\lambda_q \sim \Gamma(\lambda_q; \lambda_{1,q}, \lambda_{2,q}) \tag{4.9}$$

with hyper-parameters $\lambda_{1,q}, \lambda_{2,q}$.

4.1.2 Link with other *Bayesian regularization*

In the chapter 3, we have introduced the classical models of *Bayesian Ridge Regression* and *Automatic Relevance Determination* for *Bayesian regression*. The link between the proposed *MCBR* model and these other regularization methods is obvious:

- with $Q = 1$, *i.e.* $\lambda_{z_1} = \ldots = \lambda_{z_p}$, we retrieve the *BRR* model.
- with $Q = p$, *i.e.* $\lambda_{z_i} \neq \lambda_{z_j}$ if $i \neq j$, and assigning each feature to a singleton class (*i.e.* $z_j = j$), we retrieve the *ARD* model.

Moreover, the proposed approach is related to the one developed in [George 93]. In this paper, the authors proposed for the distribution of the weights of the features, a binary mixture of Gaussians with small and large precisions. This model is used for variable selection, and estimated by *Gibb's sampling*. Our work can be viewed as a generalization of this model to a number of classes $Q \geq 2$.

4.1.3 Issues of *ARD*

ARD suffers from some drawbacks that are explained here. We also detail how *MCBR* can cope with these different issues.

One of the main issues of *ARD* is convergence. The method detailed in chapter 3 for estimating the parameters, is not guaranteed to converge to a local maximum of the *marginal log likelihood* [Wipf 08]. In [Wipf 08], the authors proposed an estimation of *ARD* based on iterative *Lasso* procedures, that allows to avoid the convergence issue of *ARD*, even if $p \gg n$. There is no theoretical

guarantee on convergence, but Wipf's method performs well in practice. This issue is avoided in the estimation of *MCBR* by using *Gibb's sampling*, but the estimation based on another approach, called *Variational Bayes*, still suffers from convergence issue (see below).

Another issue of *ARD* comes from the fact that, among different models that classify the data equally well, *ARD* choose the simplest one (*i.e.* the most sparse one), that can be seen as a kind of *underfitting* [Qi 04]: as *ARD* is estimated by maximizing evidence, models with less selected features are preferred, as the integration is done on less dimensions, and thus the evidence is higher. This underfitting, which happens in the hyper-parameters space, does not directly fit noise (as "classical" overfitting), but merely corresponds to an underfitting in model selection (*i.e.* on the features to be pruned). A solution is to estimate *ARD* based on the optimization of the predictive performance within an internal cross-validation [Qi 04]. A contrario, *MCBR* requires far less hyper-parameters ($2 \times Q$, with $Q \ll p$), and thus the underfit in the hyper-parameters space might have less dramatic impact on feature selection.

Finally, a full Bayesian framework for estimating *ARD* requires to set some priors on the *hyper-parameters*. In the case of the widely used *Gamma* hyper-prior parametrized by α_1 and α_2, we have to define $2 \times p$ hyper-parameters. As these values α_1 and α_2 are feature-specific, *ARD* may be sensible to specific choice of these hyper-parameters. A solution is to use an *internal cross-validation* for optimizing these parameters, but this approach can be computationally expensive. In the case of *MCBR*, the distributions of the hyper-parameters are specific to a class and not to a specific feature, and thus, the proposed approach is less sensible to the choice of α_1 and α_2. Indeed, the choice of good hyper-parameters for the features are dealt with at the level of the class in which features belong to.

4.2 Model inference

For models with latent variables, such as *MCBR*, some singularities can exist. For instance in a mixture of components, a singularity is a component with one single sample and thus zero variance. In such cases, maximizing the *log likelihood* yields flawed solutions, and one can use the posterior distribution of the latent variables $p(\mathbf{z}|\mathbf{X}, \mathbf{y})$ for this maximization. However, the posterior distribution of the latent variables given the data has not always a closed-form expression, and some specific methods can be used as *Variational Bayes* or *Gibbs Sampling*.

We propose two different algorithms for inferring the parameters of the *MCBR* model. We first estimate the model by *Variational Bayes*, the resulting algorithm is thus called *VB-MCBR*. We also detail an algorithm, called *Gibbs-MCBR*, based on a *Gibbs Sampling* procedure.

4.2.1 Estimation by *Variational Bayes – VB-MCBR*

Variational Bayes

The *Variational Bayes* (or *VB*) approach provides an approximation $q(\Theta)$ of $p(\Theta|\mathbf{y})$, where $q(\Theta)$ is taken in a given family of distributions, and $\Theta = [\mathbf{w}, \lambda, \alpha, \mathbf{z}, \pi]$. Additionally, the *Variational Bayes* approach often uses the following *mean field approximation*, that allows the factorization between the approximate distribution of the latent variables and the approximate distributions of the parameters:

$$q(\Theta, \mathbf{X}) = q(\mathbf{w})q(\lambda)q(\alpha)q(\mathbf{z})q(\pi) \tag{4.10}$$

We introduce the *Kullback-Leibler* divergence $\mathcal{D}(q(\Theta, \mathbf{X}))$ that measures the similarity between the true posterior $p(\Theta, \mathbf{X}|\mathbf{y})$ and the variational approximation $q(\Theta, \mathbf{X})$. One can decompose the *marginal log-likelihood* $\log p(\mathbf{y})$ as:

$$\log p(\mathbf{y}|\Theta) = \int d\mathbf{X} d\Theta q(\Theta, \mathbf{X}) \log \frac{p(\mathbf{X}, \mathbf{y}, \Theta)}{q(\Theta, \mathbf{X})} \tag{4.11}$$

$$+ \int d\mathbf{X} d\Theta q(\Theta, \mathbf{X}) \log \frac{q(\Theta, \mathbf{X})}{p(\mathbf{X}, \Theta|\mathbf{y})} \tag{4.12}$$

$$\log p(\mathbf{y}|\Theta) = \mathcal{F}(q(\Theta, \mathbf{X})) + \mathcal{D}(q(\Theta, \mathbf{X})) \tag{4.13}$$

where $\mathcal{F}(q(\Theta, \mathbf{X}))$ is called *free energy*, and can be seen as measure of the quality of the model. As $\mathcal{D}(q(\Theta, \mathbf{X})) \geq 0$, the *free energy* is a lower bound on $\log p(\mathbf{y})$ with equality iff $q(\Theta, \mathbf{X}) = p(\Theta, \mathbf{X}|\mathbf{y})$. So, inferring the density $q(\Theta, \mathbf{X})$ of the parameters corresponds to maximizing $\mathcal{F}$, on all the free distribution $q(\Theta)$.

In practice, the *VB* approach consists in maximizing the *free energy* $\mathcal{F}$ iteratively with respect to the approximate distribution $q(\mathbf{z})$ of the latent variables, and with respect to the approximate distributions of the parameters of the model $q(\mathbf{w})$, $q(\lambda)$, $q(\alpha)$ and $q(\pi)$.

Update equations

The *Variational Bayes* approach yields the following variational distributions:

- $q(\mathbf{w}) \sim \mathcal{N}(\mathbf{w}|\mu, \mathbf{\Sigma})$ with:

$$\overline{\mathbf{A}} = \text{diag}(\overline{l_1}, \ldots, \overline{l_p}) \quad \text{with} \quad \overline{l_p} = \sum_{q=1}^{Q} q(z_j = q) \frac{l_{1,q}}{l_{2,q}} \tag{4.14}$$

$$\mathbf{\Sigma} = (\frac{a_1}{a_2}\mathbf{X}^T\mathbf{X} + \overline{\mathbf{A}})^{-1} \tag{4.15}$$

$$\mu = \frac{a_1}{a_2}\mathbf{\Sigma}\mathbf{X}^T\mathbf{y} \tag{4.16}$$

- $q(\lambda_q) \sim \Gamma(l_{1,q}, l_{2,q})$ with:

$$l_{1,q} = \lambda_{1,q} + \frac{1}{2}\sum_{j=1}^{p} q(z_j = q) \quad (4.17)$$

$$l_{2,q} = \lambda_{2,q} + \frac{1}{2}\sum_{j=1}^{p} (\mu_{jj}^2 + \Sigma_{jj}) q(z_j = q) \quad (4.18)$$

- $q(\alpha) \sim \Gamma(a_1, a_2)$ with:

$$a_1 = \alpha_1 + \frac{n}{2} \quad (4.19)$$

$$a_2 = \alpha_2 + \frac{1}{2}(\mathbf{y} - \mathbf{X}\mu)^T(\mathbf{y} - \mathbf{X}\mu) + \frac{1}{2}Tr(\mathbf{\Sigma}\mathbf{X}^T\mathbf{X}) \quad (4.20)$$

- $q(z_j = q) \sim \exp^{(\rho_{jq})}$ with:

$$\rho_{jq} = -\frac{1}{2}(\mu_j^2 + \Sigma_{jj})\frac{l_{1,q}}{l_{2,q}} + \ln(\pi_q) + \frac{1}{2}\left(\Psi(l_{1,q}) - \log(l_{2,q})\right) \quad (4.21)$$

$$\pi_q = \exp^{\left\{\Psi(d_q) - \Psi(\sum_{q=1}^{q=Q} d_q)\right\}} \quad (4.22)$$

$$d_q = \eta_q + \sum_{j=1}^{p} q(z_j = q) \quad (4.23)$$

where Ψ is the digamma function $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$.

The pseudo-code of the *VB-MCBR* algorithm is provided in Table 1. It maximizes the free energy $\mathcal{F}$. In practice, iterations are performed until convergence to a local maximum of $\mathcal{F}$. With an *ARD* prior (*i.e.* $Q = p$ and fixing $z_j = j$), we retrieve the same formulas than the ones found for *Variational ARD* [Tipping 00]. Moreover, the free energy $\mathcal{F}$ reads:

$$\mathcal{F} = \mathcal{L}_Y + \mathcal{L}_Z - \mathrm{D_{kl}}(\mathbf{w}) - \mathrm{D_{kl}}(\lambda) - \mathrm{D_{kl}}(\alpha) - \mathrm{D_{kl}}(\pi) \quad (4.24)$$

with the *likelihood* terms:

$$\mathcal{L}_Y = \frac{n}{2}\left(\Psi(a_1) - \log(a_2) - 2\pi\right) - \frac{1}{2}\frac{a_1}{a_2}\left\{\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mu + \mathrm{Tr}(\mathbf{\Sigma}\mathbf{X}^T\mathbf{X}) + \mu^T\mathbf{X}^T\mathbf{X}\mu\right\}$$

$$\mathcal{L}_Z = \sum_{j=1}^{j=p}\sum_{q=1}^{q=Q} q(z_j = q)\log\pi_q - \sum_{j=1}^{j=p}\sum_{q=1}^{q=Q} q(z_j = q)\log q(z_j = q)$$

and the *Kullback-Leibler* divergence terms:

$$\begin{aligned}
\mathrm{D_{kl}}(\mathbf{w}) &= -\frac{p}{2} - \frac{\log(|\mathbf{\Sigma}|)}{2} - \frac{1}{2}\sum_{j=1}^{j=p}\sum_{q=1}^{q=Q} q(z_j = q)\left\{\Psi(l_{1,q}) - \log(l_{2,q})\right\} \\
&\quad + \frac{1}{2}\sum_{j=1}^{j=p}\left\{(\mu_j^2 + \Sigma_{jj})\sum_{q=1}^{q=Q} q(z_j = q)\frac{l_{1,q}}{l_{2,q}}\right\} \\
\mathrm{D_{kl}}(\lambda) &= \sum_{q=1}^{q=Q} \log\Gamma(\lambda_{1,q}) - \log\Gamma(l_{1,q}) + l_{1,q}\log l_{2,q} - \lambda_{1,q}\log\lambda_{2,q} \\
&\quad -(\lambda_{1,q} - l_{1,q})(\Psi(l_{1,q}) - \log l_{2,q}) - l_{1,q}\frac{\lambda_{2,q} - l_{2,q}}{l_{2,q}}
\end{aligned}$$

and:

$$\begin{aligned}
\mathrm{D_{kl}}(\alpha) &= \log\Gamma(a_1) - \log\Gamma(\alpha_1) + a_1\log a_2 - \alpha_1\log\alpha_2 \\
&\quad -(\alpha_1 - a_1)(\Psi(a_1) - \log a_2) - a_1\frac{\alpha_2 - a_2}{a_2} \\
\mathrm{D_{kl}}(\pi) &= \ln\Gamma(\sum_{q=1}^{q=Q} d_q) - \ln\Gamma(\sum_{q=1}^{q=Q}\eta_q) \\
&\quad - \sum_{q=1}^{q=Q}(-\ln\Gamma(\eta_q) + \ln\Gamma(d_q) - (d_q - \eta_q)\Psi(d_q - \eta_q))
\end{aligned}$$

Algorithm 1: *VB-MCBR* algorithm

Initialize $a_1 = \alpha_1$, $a_2 = \alpha_2$, $l_1 = \lambda_1$, $l_2 = \lambda_2$ and $d_q = \eta_q$
Randomly initialize $q(z_j = q)$
Set a number of iterations *max steps*
repeat
  Compute A using Eq. 4.14, Σ using Eq. 4.15 and μ using Eq. 4.16.
  Compute l_1 using Eq. 4.17 and l_2 using Eq. 4.18.
  Compute a_1 using Eq. 4.19 and a_2 using Eq. 4.20.
  Compute ρ_{jq} using Eq. 4.21.
  Compute π_q using Eq. 4.22 and d_q using Eq. 4.23.
until *max steps* ;
return μ

4.2.2 Estimation by *Gibbs Sampling – Gibbs-MCBR*

We develop here an estimation of the model *MCBR* using *Gibbs Sampling*. The resulting algorithm is called *Gibbs-MCBR*, and the pseudo-code of the algorithm is provided in Table 2.

Gibbs Sampling

The *Gibbs Sampling* algorithm [Geman 87] is a particular case of the *Metropolis-Hastings* where the *acceptance probability* is always equal to 1. It is used for generating a sequence of samples from the joint distribution to approximate marginal distributions. Suppose we want to compute both marginals $p(\theta_1)$ and $p(\theta_2)$ from a complex distribution $p(\theta_1, \theta_2)$. The main idea is to use conditional distributions $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$, that should be known and possibly easy to sample from, instead of directly computing the marginals from the joint law by integration (the joint law may not be known or hard to sample from). The sampling is done iteratively between the different parameters.

The final estimation of the parameters is obtained by averaging the values of the different parameters across the different iterations. One may not consider the first iterations, this is called the *burn in*.

Candidate distributions

With $\Theta = [\mathbf{w}, \lambda, \alpha, \mathbf{z}, \pi]$, we have the following candidate distributions (*i.e.* the distributions used for the sampling of the different parameters):

- $p(\mathbf{w}|\Theta - \{\mathbf{w}\}) \propto \mathcal{N}(\mathbf{w}|\mu, \mathbf{\Sigma})$ with:

$$\mathbf{\Sigma} = (\mathbf{X}^T\mathbf{X}\alpha + \mathbf{A})^{-1} \quad \text{with} \quad \mathbf{A} = \text{diag}(\lambda_{z_1}, ..., \lambda_{z_p}) \tag{4.25}$$

$$\mu = \mathbf{\Sigma}\alpha\mathbf{X}^T\mathbf{y} \tag{4.26}$$

- $p(\lambda|\Theta - \{\lambda\}) \propto \Pi_{q=1}^{Q}\Gamma(\lambda_q|l_{1,q}, l_{2,q})$ with:

$$l_{1,q} = \lambda_{1,q} + \frac{1}{2}\sum_{j=1}^{p}\delta(z_j = q) \tag{4.27}$$

$$l_{2,q} = \lambda_{2,q} + \frac{1}{2}\sum_{j=1}^{p}\delta(z_j = q)w_j^2 \tag{4.28}$$

- $p(\alpha|\Theta - \{\alpha\}) \propto \Gamma(a_1, a_2)$ with:

$$a_1 = \alpha_1 + \frac{n}{2} \tag{4.29}$$

$$a_2 = \alpha_2 + \frac{1}{2}(\mathbf{y} - \mathbf{X}\mu)^T(\mathbf{y} - \mathbf{X}\mu) \tag{4.30}$$

- $p(z_j|\Theta - \{\mathbf{z}\}) \propto \text{mult}(\exp\rho_{j,1}, ..., \exp\rho_{j,Q})$ with:

$$\rho_{jq} = -\frac{1}{2}w_j^2\lambda_q + \ln(\pi_q) + \frac{1}{2}\log\lambda_q \tag{4.31}$$

- $p(\pi_q|\Theta - \{\pi\}) \propto \text{Dir}(d_q)$ with:

$$d_q = \eta_q + \sum_{j=1}^{p}\delta(z_j = q) \tag{4.32}$$

Algorithm 2: *Gibbs-MCBR* algorithm

Initialize α_1, α_2, λ_1, λ_2 and η_q
Randomly initialize z
Set a number of iterations *burn number* for *burn-in*
Set a number of iterations *max steps*
repeat
  Compute Σ using Eq. 4.25 and μ using Eq. 4.26.
  Sample $\mathbf{w}$ in $\mathcal{N}(\mathbf{w}|\mu, \mathbf{\Sigma})$.
  Compute l_1 using Eq. 4.27 and l_2 using Eq. 4.28.
  Sample λ in $\Pi_{q=1}^{q=Q}\Gamma(\lambda_q|l_{1,q}, l_{2,q})$.
  Compute a_1 using Eq. 4.29 and a_2 using Eq. 4.30.
  Sample α in $\Gamma(a_1, a_2)$.
  Compute ρ_{jq} using Eq. 4.31.
  Sample $\mathbf{z}$ in mult$(\exp \rho_{j,1}, ..., \exp \rho_{j,Q})$.
  Compute d_q using Eq. 4.32.
  Sample π_q in Dir(d_q).
until *max steps* ;
return Average value of $\mathbf{w}$ after *burn number* iterations.

4.2.3 Initialization and priors on the model parameters

Our model needs few hyper-parameters; we choose here to use slightly informative and class-specific hyper-parameters in order to reflect a wide range of possible behaviors for the weights distribution. This choice of priors is equivalent to setting heavy-tailed centered *Student* distributions with variance at different scales as priors on the weights parameters. We set $Q = 9$, with weakly informative priors $\lambda_{1,q} = 10^{q-4}, q \in [1, .., Q]$ and $\lambda_{2,q} = 10^{-2}$, $q \in [1, .., Q]$. Moreover, we set $\alpha_1 = \alpha_2 = 1$. Starting with a given number of classes and letting the model automatically prune the classes, can be seen as a means to avoid costly model selection procedures. The choice of class-specific priors is also useful to avoid label switching issues and thus speeds up convergence. Crucially, the priors used here can be used in any regression problem, provided that the target data is approximately scaled to the range of values used in our experiments. In that sense, the present choice of priors can be considered as *"universal"*. We also randomly initialize $q(\mathbf{z})$ (or $\mathbf{z}$ for *Gibbs-MCBR*).

4.3 Illustration on simulated data

We now evaluate and illustrate *MCBR* on two different sets of simulated data.

4.3.1 Simulated regression data

Details on simulated regression data

We first test *MCBR* on a simulated data set, designed for the study of ill-posed regression problem, *i.e.* $n \ll p$. Data are simulated as follows:

$$\begin{aligned} \mathbf{X} &\sim \mathcal{N}(0,1) \text{ with } \epsilon \sim \mathcal{N}(0,1) \\ \mathbf{y} &= 2\left(\mathbf{X_1} + \mathbf{X_2} - \mathbf{X_3} - \mathbf{X_4}\right) + 0.5\left(\mathbf{X_5} + \mathbf{X_6} - \mathbf{X_7} - \mathbf{X_8}\right) + \epsilon \end{aligned}$$

We have $p = 200$ features, $n^l = 50$ images for the training set and $n^t = 50$ images for the test set. We compare *MCBR* with *Elastic net*, *SVR*, *Bayesian Ridge Regression* and *Automatic Relevance Determination*. *Elastic net* is optimized with a 5-folds cross-validation within the training set, with $\lambda_1 \in \{0.2\tilde{\lambda}, 0.1\tilde{\lambda}, 0.05\tilde{\lambda}, 0.01\tilde{\lambda}\}$ ($\tilde{\lambda} = \|\mathbf{X}^T\mathbf{y}\|_\infty$), and $\lambda_2 \in \{0.1, 0.5, 1., 10., 100.\}$. *SVR* is used with a linear kernel and the C parameter is optimized by a 5-folds cross-validation in the range 10^{-3} to 10^1 in multiplicative steps of 10.

Results on simulated regression data

We average the results of 15 different trials, and the average explained variance is shown Tab.4.1. *Gibbs-MCBR* outperforms the other approaches, yielding higher prediction accuracy than the reference methods *Elastic net* and *ARD*. The prediction accuracy is also more stable than the other methods. *VB-MCBR* falls into local maximum of $\mathcal{F}$ and does not yield an accurate prediction.

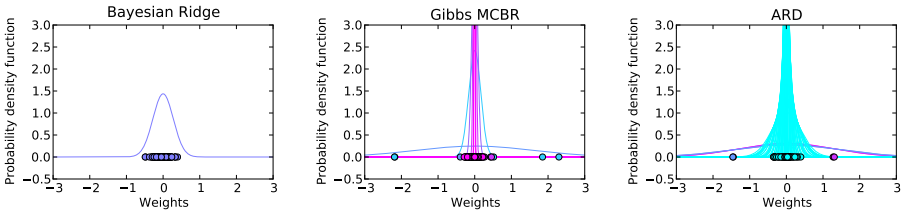
Methods	mean ζ	std ζ	max ζ	min ζ	p-value to Gibbs-MCBR
SVR	0.11	0.1	0.32	-0.03	0.0 **
Elastic net	0.77	0.11	0.92	0.49	0.0004 **
BRR	0.19	0.14	0.49	-0.04	0.0 **
ARD	0.79	0.06	0.89	0.65	0.0 **
Gibbs-MCBR	0.89	0.04	0.94	0.81	-
VB-MCBR	0.04	0.05	0.13	-0.02	0.0 **

Table 4.1: *Simulated regression data*. Explained variance ζ for different methods (average of 15 different trials). The p-values are computed using a paired t-test.

In Fig.4.2, we represent the probability density function of the distributions of the weights obtained with *BRR* (a), *Gibbs-MCBR* (b) and *ARD* (c). With *BRR*, the weights are grouped in a mono-modal density. *ARD* is far more adaptive and sets lots of weights to zero. The *Gibbs-MCBR* algorithm creates a multi-modal distribution, lots of weights being highly regularized (pink distribu-

tions), and the informative features are allowed to have higher weights (blue distributions).

With *MCBR*, the weights are clustered in different groups, depending on their predictive power, which is interesting in application such as *fMRI* inverse inference, as it can yields more interpretable models. Indeed, the class where the features with higher weights ($\{\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}\}$) belong to, is small (average size of 6 features) but has a high *purity* (percentage of relevant features in the class) of 74%.



(a) BRR *(b) Gibbs-MCBR* *(b) ARD*

Figure 4.2: Results on simulated regression data. Probability density function of the weights distributions obtained with *BRR* (a), *Gibbs-MCBR* (b) and *ARD* (c). Each color represents a different component of the mixture model.

Comparison *VB-MCBR* and *Gibbs-MCBR*

We now look at the values of w_1 and w_2 for the different steps of the two algorithms (see Fig.4.3). We can see that *VB-MCBR (b)* quickly falls into a local maximum, as *Gibbs-MCBR (a)* is able to visit the space in order to find the correct set of parameters (red dot). *VB-MCBR* is not optimal in this case.

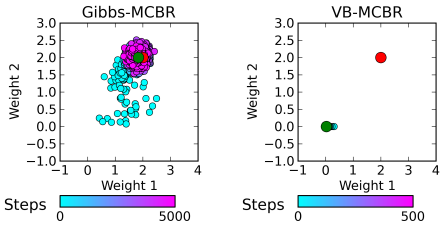
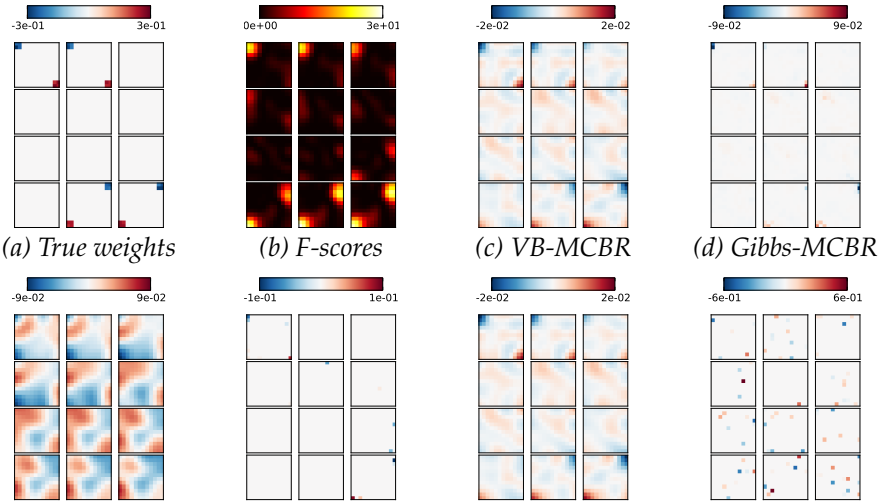


Figure 4.3: Results on simulated regression data. Weights of the first two features found for the different steps of *Gibbs-MCBR* (a) and *VB-MCBR* (b). The red dot represents the ground truth of both weights, and the green dot represents the final state found by the two algorithms. *VB-MCBR* is stuck in a local maximum, and *Gibbs-MCBR* finds the correct weights.

In order to avoid taking posterior means of the parameters and obtain a point estimate of the maximum a posteriori, we can combine the two estimation frameworks. In a first step, we launch *Gibbs-MCBR* and we then launch *VB-MCBR*, using the output of the first step as initialization. However, this approach does not yield higher prediction accuracy than *Gibbs-MCBR* alone, and thus has not been further developed in this work.

4.3.2 Simulated neuroimaging data

The simulated neuroimaging data are detailed in Appendix B.1. We compare *VB-MCBR* and *Gibbs-MCBR* with the different competing algorithms detailed in B.1, and with *Bayesian Ridge Regression* and *Automatic Relevance Determination*. The resulting images of weights are given Fig. 4.4, with the true weights (a) and resulting *Anova* F-scores (b). The reference methods can detect the truly informative regions (*ROIs*), but *Elastic net* (f) and *ARD* (h) only retrieve part of the support of the weights. Moreover, *Elastic net* yields an overly sparse solution. *BRR* (g) also retrieves the *ROIs*, but does not yield a sparse solution, as all the features are regularized in the same way. We note that the weights in the *feature space* estimated by *SVR* (e) are non-zero everywhere and do not outline the support of the ground truth. *VB-MCBR* (c) converges to a local maximum similar as *BRR* (g), *i.e.* creates only one non-empty class, and thus regularizes all the feature similarly. We can thus clearly see that, in this model, the *Variational Bayes* approach is very sensitive to the initialization, and can fall into non-optimal local maximum, for very sparse support of the weights. Finally, *Gibbs-MCBR* (d) retrieves the largest part of the whole support of the weights by performing an adapted regularization.



(a) True weights *(b) F-scores* *(c) VB-MCBR* *(d) Gibbs-MCBR*

(e) C.v. SVR *(f) C.v. Elastic net* *(g) BRR* *(h) ARD*

Figure 4.4: Two-dimensional slices of the three-dimensional volume of simulated data. Weights found by different methods, the true target (a), and *F-score* (b). The *Gibbs-MCBR* method (d) retrieves almost the whole support of weights. The sparsity-promoting reference methods *Elastic net* (f) and *ARD* (h) find an overly sparse support of the weights. *VB-MCBR* (c) converges to a local maximum similar to *BRR*(g) , and thus does not yield a sparse solution. *SVR* (e) yields smooth maps that are not similar to the ground truth.

4.4 *MCBR* for *fMRI-based inverse inference*

In this section, we assess the performance of *MCBR* in an experiment on the *mental representation of size*, where the aim is to predict the size of an object seen by the subject during the experiment (see Appendix B.2), in both intra-subject and inter-subject cases. We compare *MCBR* to *Elastic Net* and *SVR* (see B.2), as well as *Bayesian Ridge Regression – BRR* and *Automatic Relevance Determination – ARD*.

All these methods are used after an *Anova*-based feature selection, as this maximizes their performance. This selection is performed on the training set of each fold in an internal cross-validation loop, and the optimal number of voxels is selected within the range $\{50, 100, 250, 500\}$, for *SVR*, *Elastic net*, *BRR* and *ARD*. For *VB-MCBR* and *Gibbs-MCBR*, in order to avoid a costly *internal cross-validation*, we select 500 voxels, and this selection is performed on the training set. The number of iterations used is fixed to 5000 (*burn in* of 4000 iterations) for *Gibbs-MCBR* and 500 for *VB-MCBR*. As previously, we set $Q = 9$.

4.4.1 Intra-subject regression analysis

The results obtained by the different methods are given in Table. 4.2. The *p-values* are computed using a paired t-test across subjects. *VB-MCBR* outperforms the other methods. Compared to the results on simulated data, *VB-MCBR* still falls in a local maximum similar to *Bayesian Ridge Regression* that performs well in this experiment. Moreover, both *Gibbs-MCBR* and *VB-MCBR* are more stable than the reference methods.

Methods	mean ζ	std ζ	max ζ	min ζ	p-value to VB-MCBR
SVR	0.82	0.07	0.9	0.67	0.0003 **
Elastic net	0.9	0.02	0.93	0.85	0.0002 **
BRR	0.92	0.02	0.96	0.88	0.0011 **
ARD	0.89	0.03	0.95	0.85	0.0003 **
Gibbs-MCBR	0.93	0.01	0.95	0.92	0.0099 **
VB-MCBR	0.94	0.01	0.96	0.92	-

Table 4.2: *Regression - Mental representation of size - Intra-subject analysis*. Explained variance ζ for the three different methods. The p-values are computed using a paired t-test. *VB-MCBR* yields the best prediction accuracy, while being more stable than the reference methods.

4.4.2 Inter-subject regression analysis

The results obtained with the different methods are given in Table. 4.3. As in the intra-subject analysis, both *MCBR* approaches outperform the reference methods *SVR*, *Bayesian Ridge* and *ARD*. However, the prediction accuracy is

similar to *Elastic net*. In this case, *Gibbs-MCBR* performs slightly better than *VB-MCBR*, but the difference is not significant.

Methods	mean ζ	std ζ	max ζ	min ζ	p-value to Gibbs-MCBR
SVR	0.77	0.11	0.97	0.58	0.1356
Elastic net	0.78	0.1	0.97	0.65	0.7504
BRR	0.72	0.1	0.94	0.6	0.0094 **
ARD	0.52	0.33	0.93	−0.28	0.0189 *
Gibbs-MCBR	0.79	0.1	0.97	0.62	-
VB-MCBR	0.78	0.1	0.97	0.65	0.3845

Table 4.3: *Regression - Mental representation of size - Inter-subject analysis*. Explained variance ζ for the different methods. The p-values are computed using a paired t-test. *MCBR* yields highest prediction accuracy than the two other *Bayesian* framework *BRR* and *ARD*.

The maps of weights found by the different methods are detailed in Fig. 4.6. The methods are used combined with an *Anova*-based *univariate feature selection* (2500 voxels selected, in order to have a good support of the weights). As *Elastic net*, *Gibbs-MCBR* yields a sparse solution, but extracts a few more voxels. The map found by *Elastic net* is not easy to interpret, with very few informative voxels scattered in the whole occipital cortex. The map found by *SVR* is not sparse in the *feature space* and is thus difficult to interpret, as the spatial layout of the neural code is not clearly extracted. *VB-MCBR* does not yield a sparse map either, all the features having non-null weights.

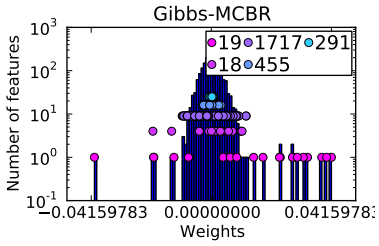


Figure 4.5: *Mental representation of size - Inter-subject analysis*. Histogram of the weights found by *Gibbs-MCBR*, and corresponding $\mathbf{z}$ values (each color of dots represents a different class), for the inter-subject analyzes on the *mental representation of size*. We can see that *Gibbs-MCBR* creates clusters of informative and non informative voxels, and that the different classes are regularized differently, according to the relevance of the features within them.

One major asset of *MCBR* (and more particularly *Gibbs-MCBR*, as *VB-MCBR* often falls into a one-class local maximum) is that it creates a clustering of the features, based on the relevance of the features in the predictive model. This clustering can be accessed using the variable $\mathbf{z}$, that is implied in the regularization performed on the different features. In Fig.4.5, we give the histogram of the weights of *Gibbs-MCBR* for the inter-subject analyzes on the *mental representation of size*. We keep the weights and the values of $\mathbf{z}$ of the last iteration, the

different classes are represented as dots of different colors, and are superimposed on the histogram. We can notice than the pink distribution represented at the bottom of the histogram corresponds to relevant features. This cluster is very small (19 voxels), compared to the two blue classes represented at the top of the histogram that contain many voxels (746 voxels) which are highly regularized, as they are non informative.

4.4.3 Discussion

Regularization of voxels loadings significantly increases the generalization ability of the predictive model. However, this regularization has to be adapted to each particular dataset. In place of costly cross-validation procedures, we cast regularization in a Bayesian framework and treat the regularization weights as hyper-parameters. The proposed approach yields an adaptive and efficient regularization, and can be seen as a compromise between a global regularization (*Bayesian Ridge Regression*) which does not take into account the sparse or focal distribution of the information, and *Automatic Relevance determination*, that may be subject to overfit in high-dimensional feature spaces.

On simulated data, our approach performs better than other classical methods such as *SVR*, *BRR*, *ARD* and *Elastic net* and yields a more stable prediction accuracy. Additionally, *MCBR* creates a clustering of the features based on their relevance, and thus explicitly extracts groups of informative features. Moreover, as seen on simulated neuroimaging data, by adapting the regularization to different groups of voxels, *MCBR* retrieves the true support of the weights, and recovers a sparse solution.

Results on real data show that *MCBR* yields more accurate predictions than other regularization methods. As it yields less sparse solution than *Elastic net*, it gives access to more plausible loading maps which are necessary for understanding the spatial organization of brain activity, *i.e.* retrieving the spatial layout of the neural coding. The explicit clustering of *Gibbs-MCBR* is also an interesting aspect of the model, as it can extract few groups of relevant features from many voxels.

In some experiments, the *Variational Bayes* algorithm yields less accurate predictions than the *Gibbs sampling* approach, which can be explained by the difficulty of initializing the different variables (especially $\mathbf{z}$) when the support of the weight is overly sparse.

The question of model selection (i.e. the number of classes Q) has not been addressed in this thesis. One can use the *free energy* in order to select the best model, but due to the instability of *VB-MCBR*, this approach does not seem promising. A more interesting method is the one detailed in [Chib 01], which can be used with *Gibbs sampling* algorithm. Here, model selection is performed implicitly by emptying classes that do not fit the data well. In that respect, the choice of heterogeneous priors for different classes is crucial: replacing our priors with class-independent priors (*i.e.* $\lambda_{1,q} = 10^{-2}\ , q \in [1, .., Q]$) in the inter-subject analysis on sizes prediction, leads *Gibbs-MCBR* to a local maximum similar to *VB-MCBR*.

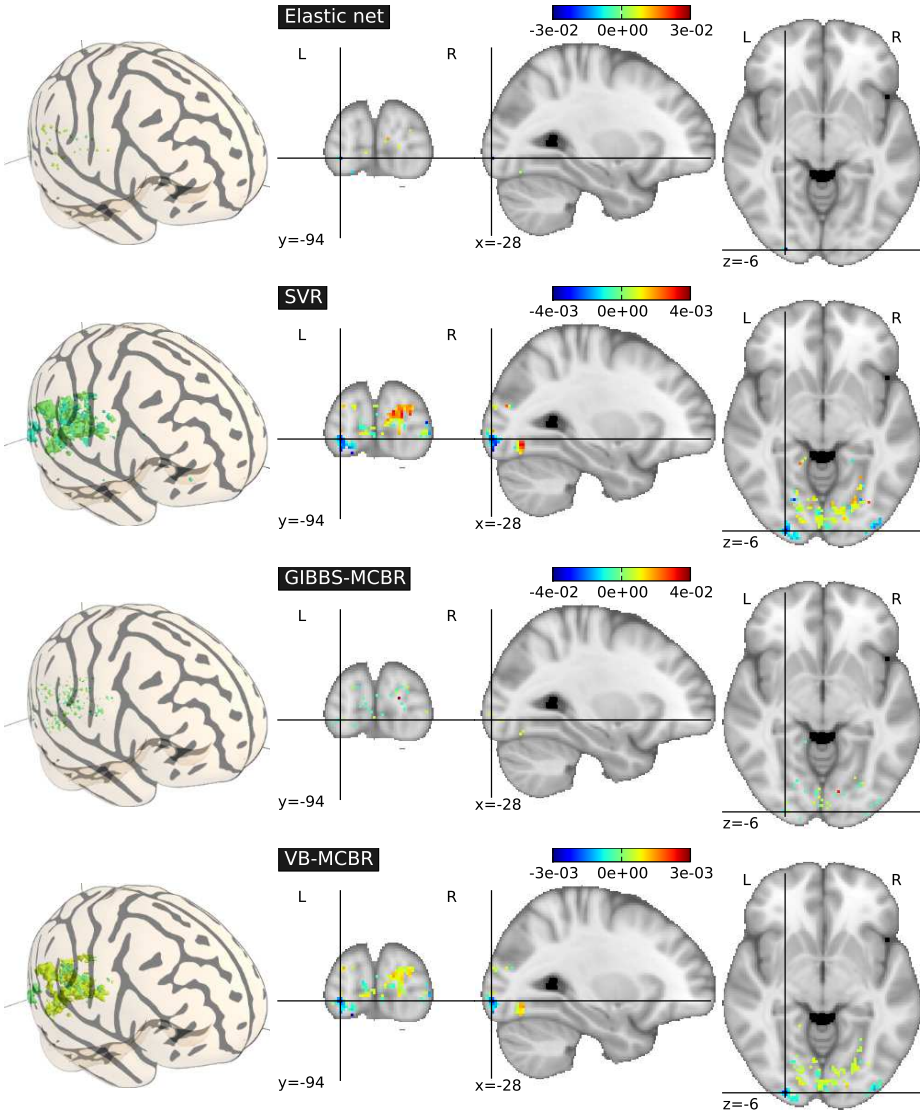


Figure 4.6: *Regression - Mental representation of size - Inter-subject analysis*. Maps of weights found by the different methods on the 2500 most relevant features by *Anova*. The map found by *Elastic net* is difficult to interpret as the very few relevant features are scattered within the whole brain. *SVR* and *VB-MCBR* do not yield a sparse solution. *Gibbs-MCBR*, by performing an adaptive regularization, draws a compromise between the other approaches, and yields a sparse solution, but also extract small groups of relevant features.

4.5 Conclusion - Multi-Class Sparse Bayesian Regression

In this chapter, we have proposed a model for adaptive regression, called *MCBR*. The main characteristics of the model are the following:

- **Generalization of classical approaches**: the proposed method integrates in the same framework *BRR* and *ARD*.
- **Adaptive regularization**: by performing a different regularization for relevant and irrelevant features, it enables both a subset selection, and an accurate estimation of the weights of the model. It can tune the regularization, within a bayesian framework, to the different level of sparsity of *fMRI* data. The regularization is as adaptive as *ARD*, but is performed will far less hyper-parameters, and thus is less subject to overfit in the hyper-parameters space. Indeed, *MCBR* does not have the convergence issue of *ARD*, and thus does not pick up the "simplest" model among different models with similar prediction accuracy.
- **Features clustering**: the proposed approach yields an interesting information for *fMRI inverse inference*, *i.e.* the $\mathbf{z}$ variable (latent class variable). Indeed, the intrinsic clustering of *MCBR* allows to extract clusters of relevant features.

Experiments on both simulated and real data show that our approach is well-suited for neuroimaging, as it yields accurate and stable predictions compared to state of the art methods. Among different research directions, it can be interesting to add Dirichlet prior to the *MCBR* method. This prior tunes the number of classes Q automatically, and thus can adapt the sparsity between the two extremal cases of *Bayesian Ridge Regression* (no sparsity), and *Automatic Relevance Determination* (high sparsity). Another direction can be to implement a spatial model in this framework, in order to extract groups of connected voxels. For instance, one can add a spatial Markovian prior on the labels for spatial consistency, but such an approach may be computationally prohibitive.

Publications

The contributions developed in this chapter have been published in:

- V. Michel, E. Eger, C. Keribin and B. Thirion. *Adaptive multi-class bayesian sparse regression - an application to brain activity classification*. In MICCAI'09 Workshop on Analysis of Functional Medical Images, 2009.
- V. Michel, E. Eger, C. Keribin and B. Thirion. *Multi-Class Sparse Bayesian Regression for Neuroimaging data analysis*. Pages 50-57. In International Workshop on Machine Learning in Medical Imaging (MLMI) In conjunction with MICCAI, 2010.

5

Supervised clustering

In this chapter, we propose an original contribution, called *supervised clustering* for *feature agglomeration* in *fMRI inverse inference*, that handles both the spatial structure of the images, and the multivariate nature of the signal.
We first detail how spatial information can be used in *inverse inference* and then, we introduce the *supervised clustering* framework. Finally, we illustrate this approach on simulated data, and compare it to reference methods on real data.

Contents

5.1 Spatial information and *Supervised Clustering*

We have seen in chapter 3 the different methods of *dimension reduction* and *regularization* that can be used to deal with the high dimensionality of the data. To date, the most widely used method for feature selection is voxel-based *Anova (Analysis of Variance)*, that evaluates each brain voxel independently. The selected features can be redundant, and are not constrained by spatial information, and thus can be spread in large regions within the whole brain. The resulting feature maps are difficult to interpret, especially compared to standard brain mapping techniques such as *SPMs* (see chapter 2). Constructing spatially-informed predictive features gives access to meaningful maps (*e.g.* by constructing informative and anatomically coherent regions [Cordes 02]) within the decoding framework of *inverse inference*.

In this section, we first detail how spatial information is classically used in *inverse inference*. Then, we introduce the *supervised clustering* approach.

5.1.1 Introducing the spatial information in *inverse inference*

Spatial information and voxel-based analysis

A first solution is to introduce the spatial information within a voxel-based analysis, *e.g.* by adding region-based priors [Palatucci 07] or by keeping only the neighboring voxels for the predictive model, such as in *searchlight* approach [Kriegeskorte 06] (but such an approach cannot handle long-range interactions in the information coding). Another contribution of this thesis is the use of a spatially-informed regularization, and is detailed in chapter 6.

Features agglomeration and *parcels*

A more natural way for using the spatial information is called *feature agglomeration*, and consist of replacing voxel-based signals by local averages (*a.k.a. parcels*) [Flandin 02, Flandin 04, Mitchell 04, Fan 06, Thirion 06b]. This is motivated by the fact that *fMRI* signal has a very strong spatial coherence due to the spatial extension of the underlying metabolic changes and of the neural code, and there is a local redundancy of the predictive information. Using these parcel-based averages of *fMRI* signals to fit the *target* naturally reduces the number of features (from $\sim 10^5$ voxels to $\sim 10^2$ *parcels*).

We define a *parcel* P as a group of connected voxels, a *parcellation* $\mathcal{P}$ being a partition of the whole set of features in a set of *parcels*:

$$\forall j \in [1, \dots, p] \ , \ \exists k \in [1, \dots, \delta] \mid v^j \in P^k \tag{5.1}$$

with

$$\forall k, k' \in [1, \dots, p] \ , \ P^k \cap P^{k'} = \emptyset \tag{5.2}$$

where δ is the number of parcels, P^k the k^{th} parcel and we note v^j the j^{th} voxel. The *parcel-based signal* $\mathbf{P}$ is the average of the voxels within each parcel (other

representation can be considered, *e.g.* median values of each parcel), and the k^{th} row of $\mathbf{P}$ is noted $\mathbf{P}^k$:

$$\mathbf{P}^k = \frac{\sum_{j|v^j \in P^k} \mathbf{X}^j}{\delta_k} \tag{5.3}$$

where δ_k is the number of voxels in the parcel P^k.

These parcels can be created using only spatial information, in purely geometrical approach [Kontos 04], or using atlases [Tzourio-Mazoyer 02, Keller 09]. In order to take into account both spatial information and functional data, clustering approaches have also been proposed, *e.g.* spectral clustering [Thirion 06b], Gaussian mixture models [Thyreau 06], K-means [Ghebreab 08] or fuzzy clustering [He 08]. The optimal number of clusters may be hard to find [Thyreau 06, Filzmoser 99], but probabilistic clustering provides a solution [Tucholka 08]. Moreover, such spatial averages can lose the fine-grained information, which is crucial for an accurate decoding of *fMRI* data [Cox 03, Haynes 06, Haynes 05a], and different resolutions of information should be considered [Golland 07].

Searchlight

The *searchlight* [Kriegeskorte 06] is a widely used approach for the study of the fine-grained patterns of information in *fMRI* analysis. Its principle is relatively simple: a small group of neighboring features is extracted from the data, and the prediction function is instantiated on these features only. The resulting prediction accuracy is thus associated with all the features within the group, or only with the feature on the center. This yields a map of local fine-grained information, that can be used for assessing hypothesis on the local spatial layout of the neural code under investigation.

The interest of such a method is to avoid the use of *feature selection*, and simply performs an extraction of the neighboring features. However, the *searchlight* has important drawbacks. First, it requires unsmoothed data to be fully optimal, and is very sensitive to voxel-to-voxel correspondence, across images or across sessions, because it relies on fine-grained patterns. It is thus difficult to use for inter-subject inference. Moreover, the *searchlight* does not perform a multivariate analysis within the whole brain, and thus can not extract long range interactions. Additionally, in a similar way as *Statistical Parametric Maps*, the *searchlight* only returns maps of local prediction score, and can not be directly used for prediction on a dataset. It can be used on the training set to extract relevant features, but this is difficult in practice, due to the high computational cost of this approach. Finally, the *searchlight* also suffers from multiple comparisons issue, because it performs as many statistical tests as classical inference approach.

Any prediction function can be used, but it is classically used jointly with *SVM*. Similarly, any kinds of neighborhood systems can be used, but a spherical spatial neighborhood is used in [Kriegeskorte 06].

5.1.2 Supervised clustering

We now present the *supervised clustering* algorithm, that *considers the target to be predicted as early as in the clustering procedure* and yields an adaptive segmentation into *both* large regions and fine-grained information, and can thus be considered as *multi-scale*. The proposed approach can be used with any type of prediction functions, in both classification and regression settings. The flowchart of the proposed approach is given in Fig. 5.1, and the corresponding pseudo-code in Algo. 3.

We first construct a hierarchical subdivision of the search domain using Ward hierarchical clustering algorithm [Ward 63]. The resulting nested parcels constructed from the functional data is isomorphic to a tree. By construction, there is a one-to-one mapping between cuts of this tree and parcellations of the domain. Given a parcellation, the signal can be represented by parcel-based averages, thus providing a low dimensional representation of the data (*i.e. feature agglomeration*). The method proposed in this contribution is a greedy approach that optimizes the cut in order to maximize the prediction accuracy, based on parcel-based averages. By doing so, a parcellation of the domain is estimated in a supervised learning setting, hence the name *supervised clustering*. We now detail the different steps of the procedure.

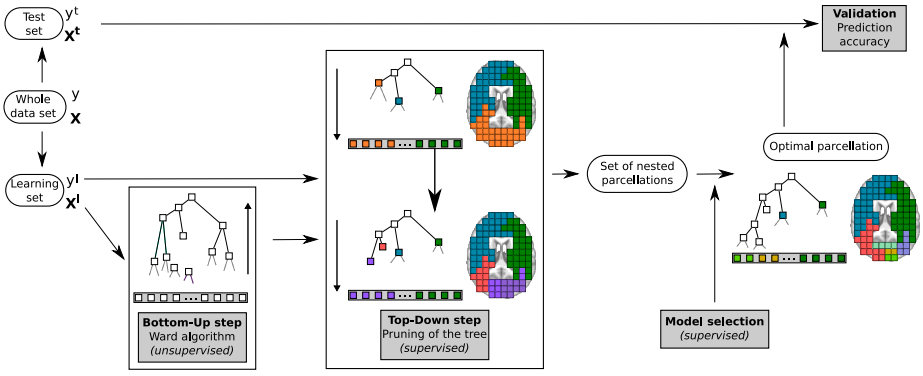


Figure 5.1: Flowchart of the *supervised clustering* approach. *Bottom-Up step (Ward clustering)* - step 1: the tree $\mathcal{T}$ is constructed from the leaves (the voxels in the gray box) to the unique root (*i.e.* the full brain volume), following spatial connectivity constraints. *Top-Down step (Pruning of the tree)* - step 2: the Ward's tree is cut recursively into smaller sub-trees, each one corresponding to a parcellation, in order to maximize a prediction accuracy ζ. *Model selection* - step 3: given the set of nested parcellations obtained by the pruning step, we select the optimal sub-tree $\widehat{\mathcal{T}}$, i.e. the one that yields the optimal value for ζ.

Bottom-Up step: hierarchical clustering

At this stage, we ignore the target information – *i.e.* the behavior variable to be predicted – and use a *hierarchical agglomerative clustering*. We add connectivity constraints to this algorithm (only adjacent clusters can be merged together) so that only spatially connected *clusters*, *i.e. parcels*, are created. This approach creates a hierarchy of *clusters* represented as a tree $\mathcal{T}$ (or dendrogram) [Johnson 67]. The root of the tree is the unique cluster that gathers all the voxels, the leaves being the clusters with only one voxel. As the resulting nested parcel sets is isomorphic to the tree $\mathcal{T}$, we identify any tree cut with a given parcellation of the domain. Any cut of the tree into δ sub-trees corresponds to a unique parcellation $\mathcal{P}_\delta$, through which the data can be reduced to δ parcels-based averages. Among different *hierarchical agglomerative clustering*, we use the variance-minimizing approach of Ward algorithm [Ward 63] in order to ensure that *parcel-based* averages provide a fair representation of the signal within each parcel. At each step, we merge together the two *parcels* so that the resulting parcellation minimizes the sum of squared differences within all *parcels* (*inertia criterion*).

Top-Down step: pruning of the tree $\mathcal{T}$

We now detail how the tree $\mathcal{T}$ can be pruned to create a reduced set of *parcellations*. Because the hierarchical subdivision of the brain volume (by successive inclusions) is naturally identified as a tree $\mathcal{T}$, choosing a parcellation adapted to the prediction problem means optimizing a cut of the tree. Each sub-tree created by the cut represents a region whose average signal is used for prediction. As no optimal solution is currently available to solve this problem, we consider two approaches to perform such a cut (see Fig. 5.2). In order to have Δ parcels, these two methods start from the root of the tree $\mathcal{T}$ (one unique parcel for the whole brain), and iteratively refine the parcellation:

- The first solution consists in using the *inertia criterion* from Ward algorithm: the cut consists of a subdivision of the Ward's tree into its Δ main branches. As this does not take into account the target information $\mathbf{y}$, we call it *unsupervised cut (UC)*.

- The second solution consists in initializing the cut at the highest level of the hierarchy and then successively finding the new sub-tree cut that maximizes a prediction score ζ (*e.g. explained variance*), while using a prediction function $\mathcal{F}$ (*e.g. SVM*) instantiated with the parcels-based signal averages at the current step. As in a greedy approach, successive cuts iteratively create a finer parcellation of the search volume, yielding the set of parcellations $\mathcal{P}_1, \ldots, \mathcal{P}_\Delta$. More specifically, one parcel is split at each step, where the choice of the split is driven by the prediction problem. After δ such steps of exploration, the brain is divided into $\delta + 1$ parcels. This procedure, called *supervised cut (SC)*, is detailed in algorithm 3.

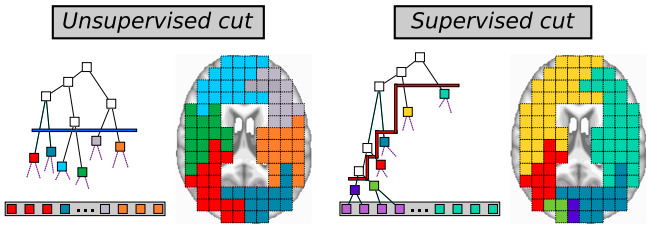


Figure 5.2: *Top-Down step (Pruning of the tree)*. In the *unsupervised cut* approach, (left) Ward's tree is divided into 6 parcels through a horizontal cut (blue). In the *supervised cut* approach (right), by choosing the best cut (red) of the tree given a score function ζ_e, we focus on some specific regions of the tree that are more informative.

Let us recall the following prediction scores. In the case of regression analysis, one can use the explained variance ζ :

$$\zeta(\mathbf{y^t}, \hat{\mathbf{y}}) = \frac{\text{var}(\mathbf{y^t}) - \text{var}\left(\mathbf{y^t} - \hat{\mathbf{y}}\right)}{\text{var}(\mathbf{y^t})} \tag{5.4}$$

For classification analysis, one can use the classification score κ:

$$\kappa(\mathbf{y^t}, \hat{\mathbf{y}}^\mathbf{t}) = \frac{\sum_{i=1}^{n^t} \delta(y_i^t, \hat{y}_i^t)}{n^t} \tag{5.5}$$

where n^t is the number of samples in the test set.

Model Selection step: optimal sub-tree $\widehat{\mathcal{T}}$

In both cases, a set of nested parcellations $\mathcal{P}_1, \ldots, \mathcal{P}_\Delta$ is produced, and the optimal model among the available cuts still has to be chosen. We select the sub-tree $\widehat{\mathcal{T}}$ that yields the optimal prediction score $\widehat{\zeta}$. The corresponding optimal parcellation is then used to create parcels on both training and test sets. Finally, a prediction accuracy is computed using these parcels.

5.1.3 Algorithmic considerations

Internal *cross-validations*

The *pruning of the tree* and the *model selection* step are included in an internal cross-validation procedure within the training set. However, this internal cross-validation scheme rises different issues. First, it is very time consuming to include the two steps within a complete internal cross-validation. A second, and more crucial issue, is that performing an internal cross-validation over the two steps yields many sub-trees (one by fold). However, it is not easy to combine these different sub-trees in order to obtain a average sub-tree that can be used for prediction on the test set [Oliver 95]. Moreover, the different optimal

sub-trees are not constructed using all the training set, and thus can be subject to specific choice of the internal cross-validation. Consequently, we choose an empirical, and potentially biased, heuristic that consists of using sequentially two separate cross-validation schemes C_e and C_s for the *pruning of the tree* and the *model selection* step.

Computational considerations

Our algorithm can be used to search informative regions in very high-dimensional data, where other algorithms do not scale well. Indeed, the highest number of features considered by our approach is Δ, and we can use any given prediction function $\mathcal{F}$, even if this function is not well-suited for high dimensional data. The computational complexity of the proposed *supervised clustering* algorithm depends thus on the complexity of the prediction function $\mathcal{F}$, and on the two cross-validation schemes C_e and C_s. At the current iteration $\delta \in [1, \Delta]$, $\delta + 1$ possible features are considered in the regression model, and the regression function is fit $n(\delta+1)$ times (in the case of a leave-one-out cross-validation with n samples). Assuming the cost of fitting the prediction function $\mathcal{F}$ is $\mathcal{O}(\delta^\alpha)$ at step δ, the overall cost complexity of the procedure is $\mathcal{O}(n\Delta^{(2+\alpha)})$. In general $\Delta \ll p$, and the cost remains affordable as long as $\Delta < 10^3$, which was the case in all our experiments. Higher values for Δ might also be used, but the complexity of $\mathcal{F}$ has to be lower.

The benefits of parcellation come at a cost regarding CPU time, the construction of the tree raising CPU time to 207 seconds and the parcels definition raising CPU time (*Intel(R) Xeon(R), 2.83GHz*) to 215 seconds on a subject of the dataset on the mental representation of size (with a non optimized Python implementation though) . Nevertheless, all this remains perfectly affordable for standard neuroimaging data analyzes.

5.2 Illustration of *Supervised clustering* on simulated data

In this section, we illustrate the *Supervised clustering* on simulated data. We compare the proposed approach to the *unsupervised cut*, and to the reference methods.

5.2.1 Illustration on simulated 1-dimensional data

We illustrate the *supervised clustering* on a simple simulated data set. Data are simulated as follows, where the informative features have a block structure:

$$\begin{aligned} \mathbf{X} &\sim \mathcal{N}(0,1) \text{ with } \epsilon \sim \mathcal{N}(0,1) \\ \mathbf{y} &= \sum_{i=0}^{p} \mathbf{w}\mathbf{X_i} + \epsilon \end{aligned}$$

Algorithm 3: Pseudo-code for *supervised cut*

Set a number of exploration steps Δ, a score function ζ, a prediction function $\mathcal{F}$, and two cross-validation schemes C_e and C_s.
Let $\mathcal{P}_\delta$ be the parcellation defined at the current iteration δ and $\mathbf{P}_\delta$ the corresponding *parcel-based* averages.
Construct $\mathcal{T}$ using Ward algorithm.
Start from the root of the tree $\mathcal{T}$, *i.e.* $\mathcal{P}_0 = \{P_0\}$ has only one parcel P_0 that contains all the voxels.
Pruning of the tree $\mathcal{T}$
for $\delta \leftarrow 1$ **to** Δ **do**
  foreach $P_i \in \mathcal{P}_{\delta-1}$ **do**
    - Split $P_i \rightarrow \{{P_i}^1, {P_i}^2\}$ according to $\mathcal{T}$.
    - Set $\mathcal{P}_{\delta,i} = \{\mathcal{P}_{\delta-1} \backslash P_i\} \cup \{{P_i}^1, {P_i}^2\}$.
    - Compute the corresponding *parcel-based* signal averages $\mathbf{P}_{\delta,i}$.
    - Compute the cross-validated score $\zeta_{e,i}(\mathcal{F})$ with the cross-validation scheme C_e.
  - Perform the split $i^\star$ that yields the highest score $\zeta_{e,i^\star}(\mathcal{F})$.
  - Keep the corresponding parcellation $\mathcal{P}_\delta$ and sub-tree $\mathcal{T}_\delta$.
Selection of the optimal sub-tree $\widehat{\mathcal{T}}$
for $\delta \leftarrow 1$ **to** Δ **do**
  - Compute the cross-validated score $\zeta_{s,\delta}(\mathcal{F})$ with the cross-validation scheme C_s, using the parcellation $\mathcal{P}_\delta$.
Return the sub-tree $\widehat{\mathcal{T}}_{\delta^\star}$ and corresponding parcellation $\widehat{\mathcal{P}}_{\delta^\star}$, that yields the highest score $\zeta_{s,\delta^\star}(\mathcal{F})$.

and $\mathbf{w}$ is defined as:

$$\begin{aligned} w_i &\sim \mathcal{U}_{0.75}^{1.25} \text{ for } 20 \leq i \leq 30 \\ w_i &\sim \mathcal{U}_{-1.25}^{-0.75} \text{ for } 50 \leq i \leq 60 \\ w_i &= 0 \text{ elsewhere} \end{aligned}$$

where $\mathcal{U}_a^b$ is the uniform distribution between a and b. We have $p = 200$ features and $n = 150$ images. The *supervised cut* is used with $\Delta = 50$, *Bayesian Ridge Regression (BRR)* as prediction function $\mathcal{F}$, and procedures C_e and C_s are set to 4-fold cross-validation.

5.2.2 Illustration on simulated neuroimaging data

We compare the *supervised clustering* approach with the *unsupervised clustering* and the two reference algorithms, *Elastic net* and *SVR*. The two reference methods are optimized by 4-fold cross-validation within the training set in the range described in Appendix B.1. We also compare the methods to a *searchlight* approach [Kriegeskorte 06] (radius of 2 and 3 voxels, combined with a *SVR* approach ($C = 1$)), which has emerged as a reference approach for decoding

local fine-grained information within the brain. Both *supervised cut* and *unsupervised cut* algorithms are used with $\Delta = 50$, *Bayesian Ridge Regression (BRR)* as prediction function $\mathcal{F}$, and optimized with an internal 4-fold cross-validation.

5.2.3 Results on 1-dimensional simulated data

The results of the *supervised clustering* algorithm are given Fig. 5.3. On the top, we give the tree $\mathcal{T}$, where the parcels found by the *supervised clustering* are represented by red squares, and the bottom row are the input features. The features of interest are represented by green dots. We note that the algorithm focuses the parcellation on two sub-regions, while leaving other parts of the tree unsegmented. The weights found by the prediction function based on the optimal parcellation (bottom) clearly outlines the two simulated informative regions. The predicted weights are normalized by the number of voxels in each parcel.

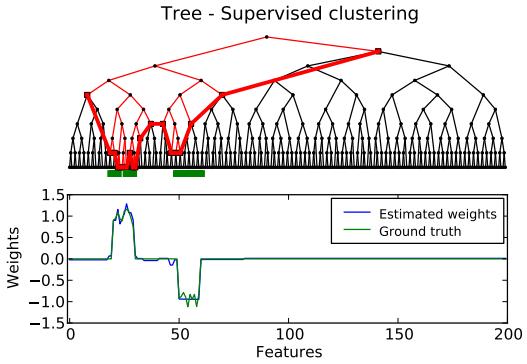
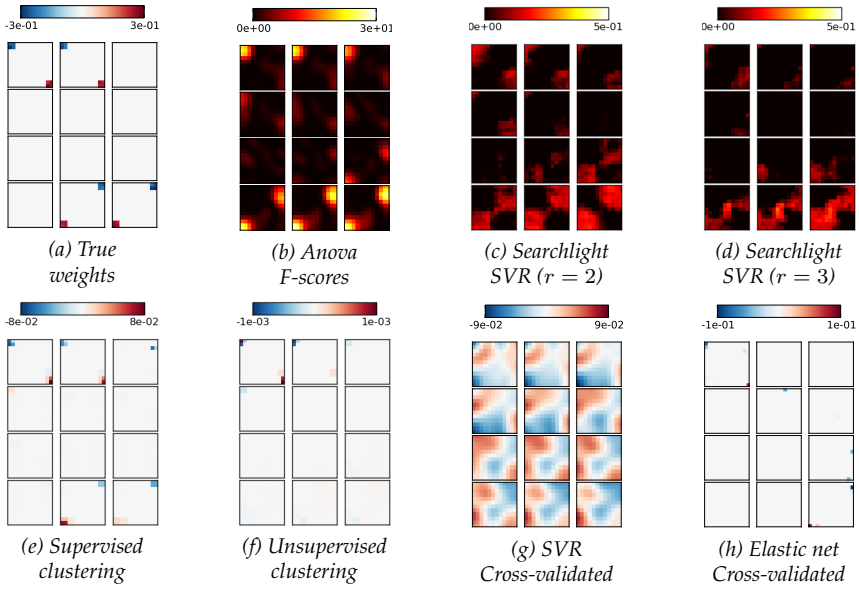


Figure 5.3: Illustration of the *supervised clustering* algorithm on a simple simulated data set. The cut of the tree (top, red line) focuses on the regions of interest (top, green dots), which allows the prediction function to correctly weight the informative features (bottom).

5.2.4 Results on simulated neuroimaging data

We compare different methods on the simulated data, see Fig.5.4. The predicted weights of the two parcel-based approaches are normalized by the number of voxels in each parcel. Only the *supervised clustering* (e) extracts the simulated discriminative regions. The *unsupervised clustering* (f) does not retrieve the whole support of the weights, as the created *parcels* are constructed based only on the signal and spatial information, and thus do not consider the target to be predicted. *Elastic net* (h) only retrieves part of the support of the weights, and yields an overly sparse solution which is not easy to interpret. *SVR* (g) approach yields weights in the primal space that are dependent on the smoothness of the images. The searchlight approach (c,d), which is a commonly used brain mapping techniques, shows here its limits: it does not cope with the long range multivariate structure of the weights, and yields very blurred informative maps, because this method naturally degrades data resolution.



(a) True weights *(b) Anova F-scores* *(c) Searchlight SVR ($r = 2$)* *(d) Searchlight SVR ($r = 3$)*

(e) Supervised clustering *(f) Unsupervised clustering* *(g) SVR Cross-validated* *(h) Elastic net Cross-validated*

Figure 5.4: Comparisons of the weights given by the different procedures (b-h) with the true weights (a). Only the *supervised cut* algorithm (e) retrieves the regions of interest. For the searchlight approach (c, f), the images show the explained variance obtained using the voxels within a sphere centered on each voxel.

5.3 *Supervised clustering* for *fMRI-based inverse inference*

We compare the *supervised clustering* to reference methods on different inter-subject data sets, as the method is well-suited for such analyzes. Indeed, the voxel-to-voxel correspondence is usually not problematic in intra-subject analysis, so that averaging neighboring voxels does not increase prediction accuracy compared to voxel-based analyzes. Similar results as voxel-based analyzes can be found using the *supervised clustering* combined with a sparsity promoting prediction function as *Lasso*. However, the extracted parcels are small and sparse, and there is no interest in such an approach compared to voxel-based analyzes. Thus, we do not further detail intra-subject analysis.

5.3.1 Details on real data

Details on real data - mental processing of size

In this experiment, we assess the performance of *supervised clustering* in the inter-subject regression analysis on the mental representation of size (see Ap-

pendix B.2)

The *supervised clustering* and *unsupervised clustering* are used with *Bayesian Ridge Regression (BRR)* (as described in section 3.3 in [Bishop 07]) as prediction function $\mathcal{F}$. Internally, a *leave-one-subject-out* cross-validation is used, and we set $\Delta = 75$. A major asset of *BRR* is that it adapts the regularization to the data at hand, and thus can cope with the different dimensions of the problem: in the first steps of the *supervised clustering* algorithm, we have more samples than features, and for the last steps, we have more features than samples.

Details on real data - mental processing of shape

In this experiment, we assess the performance of *supervised clustering* in the inter-subject classification analysis on the mental representation of shape (see Appendix B.2).

In this experiment, the *supervised clustering* and *unsupervised cut* are used with *Support Vector Classification (SVC)* ($C = 0.01$) as prediction function $\mathcal{F}$. Such value of C yields a good regularization of the weights in the proposed approach, and the results are not too sensitive to this parameter (68.3% for $C = 0.001$ and 67.5% for $C = 10$).

Details on real data - mental processing of dot sets cardinalities

We use a part of a real dataset on the mental processing of quantities (see [Eger 09]. During the experiment, ten healthy volunteers (6 males and 4 females, mean age 21.2 +/- 3.0 years) viewed dot patterns with different numbers of dots ($Y = 2$, 4, 6 and 8) with 4 repetitions of each stimulus in each one of 8 sessions : so that we have a total of $N_p = 32$ images per subject. We aimed at predicting the values of Y from the fMRI data through regression. Functional images were acquired on a 3 Tesla MR system with 12-channel head coil (Siemens Trio TIM) as T2* weighted echo-planar image (EPI) volumes using a high-resolution EPI-sequence. 26 oblique-transverse slices covering parietal and superior parts of frontal lobes were obtained in interleaved acquisition order with a TR of 2.5 s (FOV 192 mm, fat suppression, TE 30 ms, flip angle 78° , $1.5 \times 1.5 \times 1.5$ mm voxels). Standard pre-processings and the fit of the general linear model were performed with the SPM5 software. We used images of parameter estimates, one per condition and repetition.

We run the different methods in an inter-subjects analysis. For each subject, we first compute a fixed-effects activation image that represents the average effect of each stimulus, one for each condition (then, we have 4 images by subjects in 10 subjects). We evaluate the performance of the method by cross-validation (leave-one-subject-out), which yields an average rate of explained variance across subjects. This analysis in launched on the intersection of the masks of all the subjects, which roughly corresponds to the whole brain volume.

In this experiment, the *supervised clustering* and *unsupervised cut* are used with *Elastic net*, parametrized by $\lambda_1 = 0.05\tilde{\lambda}$ ($\tilde{\lambda} = \|\mathbf{P}_\delta^T \mathbf{y}\|_\infty$ is computed at each

step δ of the algorithm) and $\lambda_2 = 1$, as prediction function $\mathcal{F}$. We choose *Elastic net* as we are expecting a spatial layout more sparse than the one obtained in the study of the mental processing of size, as the processing of dot sets cardinalities is a more high-level cognitive process. The parameters of *Elastic net* are choose in the middle of the range used for voxel-based *Elastic net*.

We give in Fig.5.5 the results of a sensitivity studied performed on the first subject of the dataset. We can see that the prediction accuracy is sensitive to this choice of parameters. One should performed an internal cross-validation to optimize these parameters, but this can be computationally costly. Another solution should be to used an *oracle* to choose the parameters of *Elastic net*, *e.g.* by performing an internal cross-validation every 10 or 20 steps, in order to adapt the amount and relative sparsity of the regularization to the data.

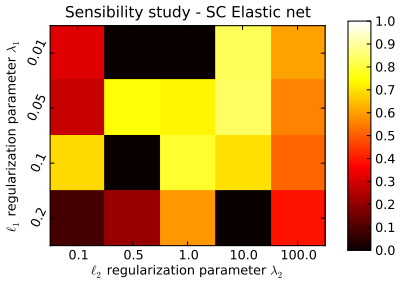


Figure 5.5: *Mental representation of dot sets cardinalities*. Sensitivity study performed on the first subject of the dataset. The prediction accuracy is sensitive to this choice of parameters for *Elastic net*.

5.3.2 Results on real data

Results for the mental representation of size

The results of the inter-subjects analysis are given in Tab.5.1. Both parcel-based methods perform better than voxel-based reference methods. Parcels can be seen as an accurate method for compressing information without loss of prediction performance. Fig. 5.6 gives the weights found for the *supervised cut*, the two reference methods and the searchlight (SVR with $C = 1$ and a radius of 2 voxels), using the whole data set. For the *supervised clustering* approach, the predicted weights are normalized by the number of voxels in each parcel. As one can see, the proposed algorithm yields clustered loadings map, compared to the maps yielded by the voxel-based methods, which are very sparse and difficult to represent. Compared to searchlight, the *supervised clustering* algorithm creates more clusters that are also easier to interpret as they are well separated. Moreover, the proposed approach also yields a prediction accuracy for the whole brain analysis, a contrario to the searchlight that only gives a local measure of information.

The parcels are found within the occipital cortex. The majority of informative parcel are located in the posterior part of the occipital cortex, most likely corresponding to primary visual cortex, with few additional slightly more anterior parcels in posterior lateral occipital cortex. This is consistent with the previous findings [Eger 08] where a gradient of sensitivity to size was observed across object selective lateral occipital *ROIs*, while the most accurate discrimination of sizes is obtained in primary visual cortex.

Methods	mean ζ	std ζ	max ζ	min ζ	p-value to UC
SVR	0.77	0.11	0.97	0.58	0.0817
Elastic net	0.78	0.1	0.97	0.65	0.0992
UC - BRR	0.83	0.08	0.97	0.73	-
SC - BRR	0.82	0.08	0.93	0.7	0.8184

Table 5.1: Explained variance ζ for the different methods in the *Size prediction experiment*. The p-values are computed using a paired t-test. The *unsupervised cut (UC)* algorithm yields the best prediction accuracy in leave-one-subject-out cross-validation. The *supervised cut (SC)* yields similar results as *UC* (the difference is not significant). The two voxel-based approaches yield lower prediction accuracy than the parcel-based approaches.

Results for the mental representation of shape

The results of the inter-subjects analysis are given in Tab.5.2. The *supervised cut* method outperforms the other approaches. In particular, the classification score is 21% higher than with voxel-based *SVC* and 27% higher than with voxel-based *SMLR*. Both parcel-based approaches are significantly more accurate and more stable than the voxel-based approaches.

Methods	mean κ	std κ	max κ	min κ	p-value to SC
SVC	48.33	15.72	75.0	25.0	0.0063 **
SMLR	42.5	9.46	58.33	33.33	0.0008 **
UC - SVC	65.0	8.98	75.0	50.0	0.1405
SC - SVC	70.0	10.67	83.33	50.0	-

Table 5.2: Classification performance κ for the different methods in the *Object prediction experiment*. The p-values are computed using a paired t-test. The *supervised cut (SC)* algorithm yields the best prediction accuracy in leave-one-subject-out cross-validation. Both parcels-based approaches are significantly more accurate and more stable than the voxel-based approaches.

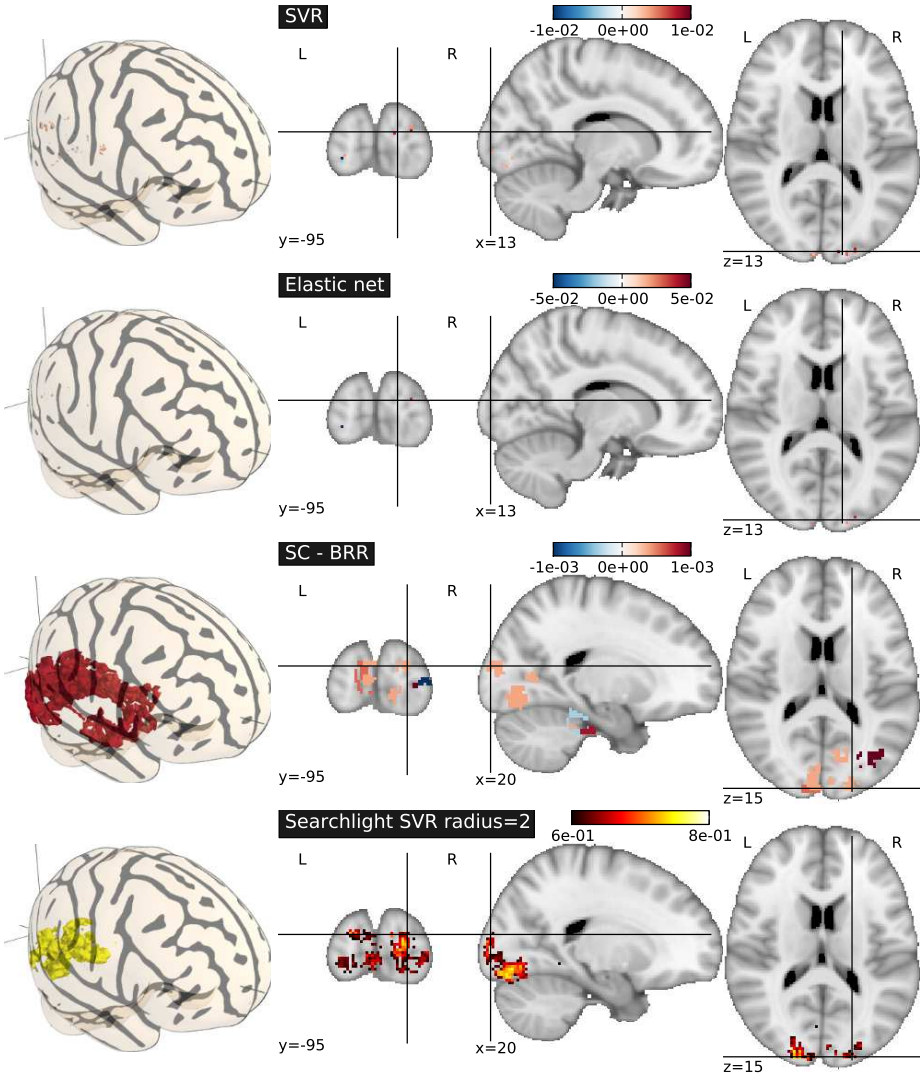


Figure 5.6: *Mental representation of size*. Maps of weights found by *supervised cut* and the two reference voxel-based methods and the searchlight. The proposed algorithm creates very interpretable clusters, compared to the reference methods, which is related to the fact that they do not consider the spatial structure of the image. Moreover, the *supervised clustering* yields similar maps as *searchlight*, but also retrieves some additional clusters.

Results on real data - mental representation of dot sets cardinalities

The results of the inter-subjects analysis are given in Tab. 5.3. The fact that a significant proportion of the stimulus variance can be fit using brain activation across subjects was not expected. However, this results is probably re-

lated to the fact that for small numbers of dots as used here (but not for larger cardinalities or symbolic numbers [Eger 09]) parametric activity increases can be observed in relatively extended and contiguous parietal regions. Whether these reflect special mechanisms for processing small numbers of objects, or secondary factors not related to numerical representation per se (e.g., increased effort when attempting to count), is currently not clear. However, even if these data have a particular confound that is not clearly representative of the mental representation of dot sets cardinalities, the *supervised clustering* approach still yields a good prediction and interpretable maps (Fig. 5.7) compared to reference methods.

Methods	mean ζ	std ζ	max ζ	min ζ	p-value to SC
SVR	0.55	0.2	0.87	0.09	0.3741
Elastic net	0.53	0.2	0.88	0.14	0.3208
UC - Elastic Net	0.56	0.26	0.91	0.08	0.3957
SC - Elastic Net	0.61	0.29	0.95	0.04	-

Table 5.3: Explained variance ζ for the different methods in the *dot sets cardinalities prediction experiment*. The p-values are computed using a paired t-test. The two parcels-based methods yield the best prediction accuracy in leave-one-subject-out cross-validation.

5.3.3 Discussion

In this chapter, we have presented a new method for enhancing brain activity prediction from *fMRI* brain images. The proposed approach constructs *parcels* (groups of connected voxels) within the whole brain, and allows to take into account both the spatial structure and the multivariate information within the whole brain.

Given that an fMRI brain image has typically 10^4 to 10^5 voxels, it is perfectly reasonable to use intermediate structures such as parcels (*i.e. feature agglomeration*) for reducing the dimensionality of the data. We also confirmed by different experiments that parcels are a good way to tackle the spatial variability problem in inter-subjects studies [Tahmasebi 10, Tucholka 10]. Thus *feature agglomeration* is an accurate approach for the challenging inter-subject generalization of *brain-reading* [Norman 06, Haynes 06]. This can be explained by the fact that considering *parcels* allows to localize functional activities across subjects and thus find a common support of neural codes of interest (see Fig. 5.8). On the contrary, voxel-based methods suffer from the inter-subject spatial variability and their performances are relatively lower.

The results for the dot sets cardinalities prediction experiment confirm the intuition that massive activity correlates with small dot sets cardinality in some parietal regions. It remains to be decided whether a population code can be defined, *i.e.* whether there is a spatial gradient between regions activating for

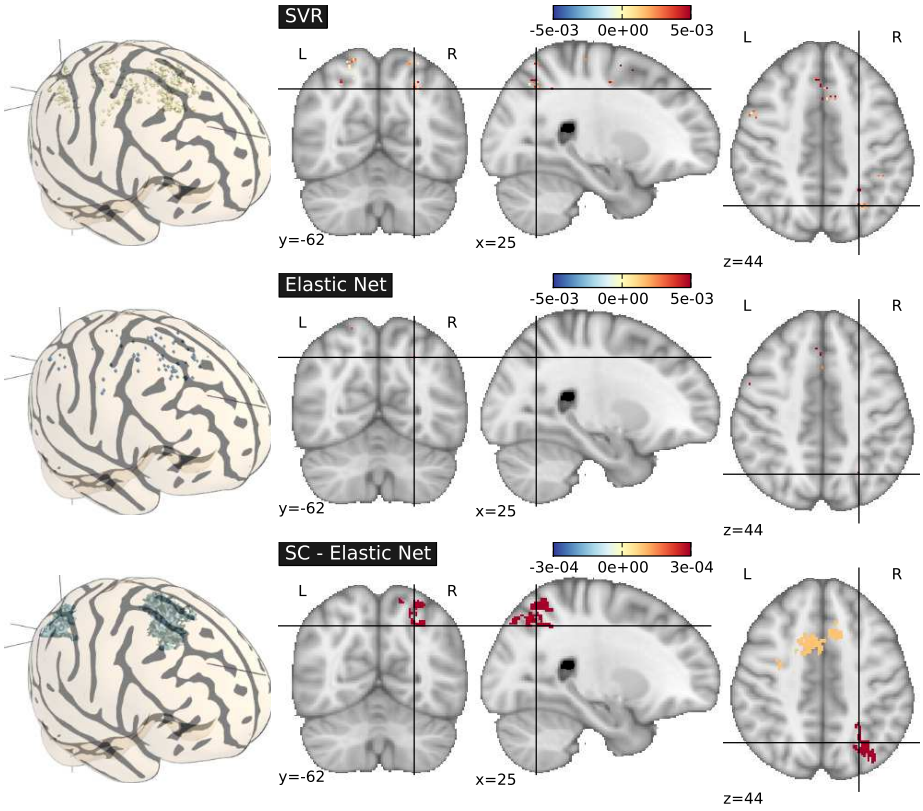


Figure 5.7: *dot sets cardinalities prediction experiment*. Maps of weights found by *supervised cut* and the two reference voxels-based methods. For *SVR* (top) and *Elastic net* (middle) (both used on 500 voxels selected by *Anova*), the voxels are spread all over the brain, without any emerging coherence. In comparison, the *supervised cut* approach extract very few parcels starting from a whole-brain analysis, and thus creates more interpretable map.

large versus low quantities. This is hard to conclude given the resolution and SNR limits of the data, but our parcellation scheme may help in that respect.

Our approach entails the technical difficulty of optimizing the parcellation with respect to the spatial organization of the information within the image. To break the combinatorial complexity of the problem, we have defined a recursive parcellation of the volume using Ward algorithm, which is furthermore constrained to yield spatially connected clusters. Note that it is important to define the parcellation on the training database to avoid data overfit. The sets of possible volume parcellations is then reduced to a tree, and the problem reduces to finding the optimal cut of the tree. We propose a *supervised cut* approach that attempts to optimize the cut with respect to the prediction task. Although finding an optimal solution is infeasible, we adopt a greedy strategy

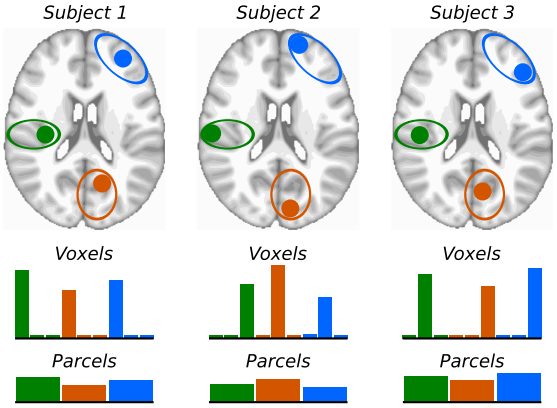


Figure 5.8: Illustration of *feature agglomeration* to cope with inter-subject variability. The regions implied in the cognitive task are represented by disks of different colors. The populations of active neurons are not exactly at the same position across subjects (top), and the across subjects mean signal in informative voxels (bottom) carries very weak information. Thus, it is clear that, in this case, *voxel-based* decoding approaches will perform poorly. However, the mean of informative voxels within each region across subjects (bottom) carries more information and should yield an accurate inter-subject prediction.

that recursively finds the splits that most improve the prediction score. However, there is still no guarantee that the optimal cut might be reached with this strategy. Model selection is then performed a posteriori by considering the best generalizing parcellation among the available models. Additionally, our method is tractable on real data and runs in a very reasonable of time (a few minutes).

In terms of *prediction accuracy*, the proposed methods yield better results for the inter-subjects study on the different experiments, compared to state of the art approaches (*SVR*, *Elastic net*, *SVC* and *SMLR*). The *supervised cut* yields similar or higher prediction accuracy than the *unsupervised cut*. On the experiment on mental representation of size, the information is not very fine-grained (a contrario to the experiment on mental representation of shape), and thus the simple heuristic of *unsupervised cut* yields a good prediction accuracy.

In terms of *interpretability*, we have shown on simulations and real data that this approach has the particular capability to highlight regions of interest, while leaving uninformative regions unsegmented, and it can be viewed as a multi-scale segmentation scheme [Michel 10]. The proposed scheme is further useful to accurately locate contiguous predictive regions and to create interpretable maps, and thus can be viewed as an intermediate approach between brain mapping and inverse inference. Moreover, compared to a state of the art approach for fine-grained decoding, namely the searchlight, the proposed method yields similar maps, but additionally, takes into account non-local information and yields only one prediction score corresponding to whole brain

analysis. From a neuroscientific point of view, a proposed approach retrieves well-known results, *i.e.* that processings of visual information about sizes are performed in early occipital cortex, with some extent in more parietal regions.

5.4 Conclusion - Supervised clustering

In conclusion, we propose a new feature building method for extracting information from brain images. Contrarily to classic methods, the supervised clustering we propose builds relevant features by agglomeration rather than simple selection. The method is validated in the context of inter-subject inference. A particularly important property of this approach is its ability to focus on relatively small but informative regions while leaving vast but uninformative areas unsegmented. Additional, this approach is not restricted to a given prediction function and can be used with many different classification/regression methods. Experimental results demonstrate that this algorithm performs well for inter-subjects analysis where the accuracy of the prediction is tested on new subjects. Indeed, the spatial averaging of the signal induced by the parcellation appears as a powerful way to deal with inter-subject variability.

Publications

The contributions developed in this chapter have been published in:

- V. Michel, E. Eger, C. Keribin, J.-B. Poline and B. Thirion. *A supervised clustering approach for extracting predictive information from brain activation images*. In IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA10) - IEEE Conference on Computer Vision and Pattern Recognition. 2010.
- V. Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin and B. Thirion. *A supervised clustering approach for fMRI-based inference of brain states*. Submitted to Pattern Recognition - Special Issue on 'Brain Decoding'. 2010

6

Total variation regularization

In the previous chapters, we have seen that regularization schemes (see chapter 4) and using spatial information (see chapter 5) can improve the prediction accuracy. In this chapter, we develop an approach combining these two results, in order to obtain more informative and interpretable results. We propose to use the ℓ_1 norm of the image gradient, *a.k.a.* its *Total Variation* (or *TV*), as regularization.
After introducing some notions on convex optimization, we give the mathematical and implementation details of *TV regression/classification*. As far as we know, the present contribution is the first one that uses *TV* in the context of image classification, which we find to be a novel and powerful tool for image-based machine learning. In a third part, we apply both *TV regression* and *TV classification* to the *fMRI* paradigm on a real data set.

Contents

6.1 Convex optimization for regularized regression

In this section we introduce some generic notions on convex optimization.

We recall the following predictive linear model, which has already been systematically used in the previous chapters:

$$\mathbf{y} = f(\mathbf{X}, \mathbf{w}, b) \ , \tag{6.1}$$

where $\mathbf{y}$ represents the behavioral variable and $(\mathbf{w}, b)$ are the parameters to be estimated on a learning set. The matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix. Each row is a p-dimensional sample, *i.e.*, an activation map related to the observation. A vector $\mathbf{w} \in \mathbb{R}^p$ can be seen as an image; p is the number of features (or voxels) and $b \in \mathbb{R}$ is called the *intercept*.

As described in previous chapters (see chapters 3 and 4), a standard approach to performing the estimation of $\mathbf{w}$ with regularization uses penalization of a maximum likelihood estimator. It leads to the following minimization problem:

$$\hat{\mathbf{w}} = \underset{\mathbf{w},b}{\operatorname{argmin}} \ \ell(\mathbf{y}, f(\mathbf{X}\mathbf{w} + b)) + \lambda J(\mathbf{w}) \ \ , \ \ \lambda \geq 0 \tag{6.2}$$

where $\lambda J(\mathbf{w})$ is the regularization term and $\ell(\mathbf{y}, f(\mathbf{X}\mathbf{w}+b))$ is the loss function. The parameter λ balances the loss function and the penalty $J(\mathbf{w})$. Note that the intercept b is not included in the regularization term.

Here we focus on convex functions for ℓ and J, in order to make problem 6.2 well posed, with unique solution. Next, we detail how large scale convex problems can be solved.

6.1.1 Convexity and duality

Definition 6.1 (Convex set)**.** *Let a set C be a collection of elements from a vector space. If, for each pair of points in a set C, every point of the line segment between the two points is also within C, the set C is said to be a* convex set*, i.e.:*

$$\forall \ (x_1, x_2) \in C, \ \ \forall \ \theta \in [0, 1] \ , \ \theta x_1 + (1 - \theta) x_2 \in C \tag{6.3}$$

The empty set and the singleton sets are convex sets, and the intersection of any two convex sets is a convex set.

Definition 6.2 (Convex function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex on a convex set C if:*

$$\forall \ (x, y) \in C, \ \forall \ \theta \in \ [0, 1] \ , \ f(\theta x + (1 - \theta) y) \leq \theta f(x) + (1 - \theta) f(y) \tag{6.4}$$

The function f is strictly convex if:

$$\forall \ (x, y) \in C, x \neq y, \ \forall \ \theta \in]0, 1[\ , \ f(\theta x + (1 - \theta) y) < \theta f(x) + (1 - \theta) f(y) \tag{6.5}$$

Additionally, f is concave if $-f$ is convex.

For example, power functions ($x^\alpha,\ \alpha \geq 1$), exponential functions ($\exp^{ax},\ a \in \mathbb{R}$), and norm functions are convex functions. The composition with an affine function preserves the convexity and $f(Ax+b)$ is convex if f is convex (in particular, for any norm $\|.\|$, $f(x) = \|Ax+b\|$ is a convex function). We will note **dom** f the domain of a function f.

Definition 6.3 (Conjugate function)**.** *The conjugate of a function f is:*

$$f^*(y) = \sup_{x \in \textbf{\textit{dom}}\ f} \left(y^T x - f(x)\right)\ ,\ y \in \mathbb{R}^n \tag{6.6}$$

f^* is convex (even if f is not). See illustration Fig.6.1.

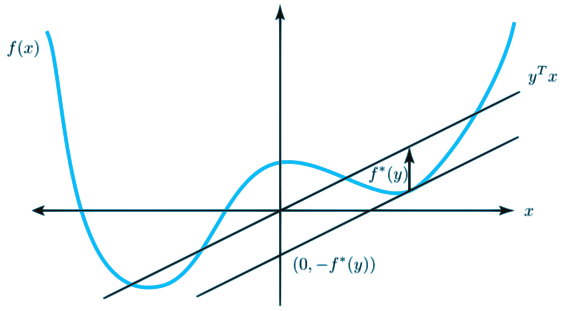


Figure 6.1: Example of function and its conjugate. Adapted from http://www.ece.ucsb.edu/∼roy/

Definition 6.4 (Dual norms)**.** *Let $\|\cdot\|$ be a norm on V. The associated dual norm $\|\|_*$ is defined as:*

$$\|y\|_* = \sup\{y^T x \ \mid\ \|x\| \leq 1\}\ ,\ y \in \mathbb{R}^n \tag{6.7}$$

which yields the inequality: $y^T x \leq \|x\|\|y\|_*$. The dual of the dual norm is the original norm.

Lets introduce the ℓ_p norms as:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \quad \text{for } p \geq 1 \tag{6.8}$$

with the particular case $\|x\|_\infty = \max_k |x_k|$. By using the Hölder's inequality, we can prove that the dual of ℓ_p is ℓ_q, with $1/p + 1/q = 1$, so that p and q are said to be *Hölder conjugates*.

Definition 6.5 (Conjugate function of a norm)**.** *Let $\|\cdot\|$ be a norm on V, and let f be the function $f(x) = \|x\|$, the conjugate function of f is defined as:*

$$f^*(y) = \chi_{\|\cdot\|_* \leq 1}(y) = \begin{cases} 0 & \text{if } \|y\|_* \leq 1 \\ \infty & \text{otherwise} \end{cases} \tag{6.9}$$

where χ is the indicator function, and $\|\cdot\|_*$ the dual norm of $\|\cdot\|$.
Thus, the conjugate of a norm is the indicator function of the unit ball for the dual norm.

Sketch of the proof. Let $\|\cdot\|$ be a norm on V, with the associated dual norm $\|\cdot\|_*$, and let f be the function $f(x) = \|x\|$. We search the associated conjugate function $f^*(y)$.

- If $\|y\|_* > 1$, using def. 6.4, there exists $z \in V$ such that $\|z\| \leq 1$ and $y^T z > 1$. Taking $x = tz$ and letting $t \to \infty$, we have:

$$y^T x - \|x\| = t(y^T z - \|z\|) \to \infty \tag{6.10}$$

  and thus, $f^*(y) = \infty$.

- If $\|y\|_* \leq 1$, we have $\forall x \in V\ ,\ y^T x \leq \|x\|\|y\|_*$, and thus, $y^T x - \|x\| \leq 0$. The maximum of $y^T x - \|x\|$ is 0, and is obtained for $x = 0$.

□

We have the particular case that the dual of ℓ_1 is ℓ_∞:

$$f(x) = \|x\|_1 \Rightarrow f^*(y) = \chi_{\|\cdot\|_\infty \leq 1}(y) \tag{6.11}$$

6.1.2 Proximity operator

Lets us recall the minimization problem that we want to solve:

$$\hat{\mathbf{w}} = \underset{\mathbf{w},b}{\operatorname{argmin}}\ \ell(\mathbf{y}, f(\mathbf{X}\mathbf{w} + b)) + \lambda J(\mathbf{w})\ \ ,\ \ \lambda \geq 0 \tag{6.12}$$

When $J(\mathbf{w})$ is non-smooth (*i.e.* not differentiable), an analytical solution does not exist and the optimization can not be performed with simple algorithms such as Gradient descent or Newton method. This is for example the case with $J(\mathbf{w}) = \|\mathbf{w}\|_1$ (ℓ_1 norm *a.k.a. Lasso* penalty) which requires advanced optimization strategies. A recently studied strategy [Daubechies 04, Combettes 05, Nesterov 07, Beck 09a] is based on iterative procedures involving the computation of *proximity operators* [Moreau 65]. Such approaches are adapted to composite problems with both a smooth term and a non-smooth term as it is the case here (see [Tseng 09] for a recent review). In the context of neuroimaging, such optimization schemes have been proposed recently in order to solve the inverse problem of magneto- and electro-encephalography (collectively M/EEG) when considering non ℓ_2 priors [Gramfort 09b, Gramfort 09a].

Definition 6.6 (Proximity operator)**.** *Let $J : \mathbb{R}^p \to \mathbb{R}$ be a convex function. The proximity operator associated with J and $\lambda \in \mathbb{R}_+$ denoted by $prox_{\lambda J} : \mathbb{R}^p \to \mathbb{R}^p$ is given by:*

$$prox_{\lambda J}(\mathbf{w}) = \underset{\mathbf{v} \in \mathbb{R}^p}{arg\min} \left(\frac{1}{2}\|\mathbf{v} - \mathbf{w}\|_2^2 + \lambda J(\mathbf{v}) \right) \tag{6.13}$$

For $J(\mathbf{w}) = \|\mathbf{w}\|_1$ (*Lasso* penalty), we have the proximal operator:

$$\text{prox}_{\lambda\|\cdot\|_1}(\mathbf{w}) = \arg\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{v}-\mathbf{w}\|_2^2 + \lambda\|\mathbf{v}\|_1 \tag{6.14}$$

which yields the solution v^* (also known as *soft thresholding*) for the i^{th} coordinate of $\mathbf{v}$:

$$\mathbf{v}_i^* = \mathbf{w}_i\left(1-\frac{\lambda}{|\mathbf{w}_i|}\right)^+ = \mathbf{w}_i \max\left(1-\frac{\lambda}{|\mathbf{w}_i|}\ ,\ 0\right) \tag{6.15}$$

6.1.3 Iterative procedures

We detail two iterative procedures for convex optimization.

***Iterative Shrinkage-Thresholding Algorithm* (*ISTA*)**

The iterative procedure known as *ISTA* (*Iterative Shrinkage-Thresholding Algorithm*, *a.k.a Forward-Backward iterations*) [Daubechies 04, Combettes 05], is based on the alternate minimization of the *loss* term $\ell(\mathbf{w})$, by gradient descent, and the penalty $J(\mathbf{w})$, by computing a *proximity operator*. One can show (a sketch of the proof is given below), that this can be done in one single step by iterating:

$$\mathbf{w}^{(k+1)} = \text{prox}_{\lambda J/L}\left(\mathbf{w}^{(k)} - \frac{1}{L}\nabla\ell(\mathbf{w}^{(k)})\right) \quad , \tag{6.16}$$

where $\frac{1}{L}\nabla\ell(\mathbf{w}^{(k)})$ is the gradient descent term with a stepsize $\frac{1}{L}$, $\text{prox}_{\lambda J/L}$ is the *proximity operator* of the penalty and the scalar L is an upper bound on the *Lipschitz constant* $\mathcal{L}$ of the gradient of the loss function. The *Lipschitz constant* $\mathcal{L}$ is a positive constant, such that , for a *Lipschitz continuous* function f:

$$\forall x_1, x_2 \ \ , \ \ |f(x_1) - f(x_2)| \leq \mathcal{L}|x_1 - x_2| \tag{6.17}$$

The pseudo code of the *ISTA* procedure is given in Algo. 4.

Algorithm 4: *ISTA* procedure

Compute the Lipschitz constant $\mathcal{L}$ of the operator $\nabla\ell$.
Initialize $\mathbf{w}^{(0)} \in \mathbb{R}^p$
repeat
  $\mathbf{w}^{(k+1)} = \text{prox}_{\lambda J/L}\left(\mathbf{w}^{(k)} - \frac{1}{L}\nabla\ell(\mathbf{w}^{(k)})\right)$ with $L > \mathcal{L}$
until *convergence* ;
return $\mathbf{w}$

Sketch of the proof. We give the sketch of proof of Eq. 6.16. The loss $\ell(\mathbf{w})$ being differentiable, the second-order linearization of $\ell(\mathbf{w})$ reads:

$$\ell(\mathbf{w}) \approx \ell(\mathbf{w}^{(k)}) + (\mathbf{w}-\mathbf{w}^{(k)})^T\nabla\ell(\mathbf{w}^{(k)}) + \frac{1}{2}(\mathbf{w}-\mathbf{w}^{(k)})^T\nabla^2\ell(\mathbf{w}^{(k)})(\mathbf{w}-\mathbf{w}^{(k)}) \tag{6.18}$$

Let $\mathcal{L}$ be the Lipschitz constant of $\nabla\ell$, *i.e.*:

$$\|\nabla\ell(\mathbf{w}) - \nabla\ell(\mathbf{w}^{(k)})\| \leq \mathcal{L}\|\mathbf{w} - \mathbf{w}^{(k)}\| \tag{6.19}$$

Let L be an upper bound on the *Lipschitz constant* $\mathcal{L}$. As in [Ortega 00], Eq. 6.16 yields:

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\operatorname{argmin}}\ \ell(\mathbf{w}^{(k)}) + \frac{L}{2}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2 + (\mathbf{w} - \mathbf{w}^{(k)})^T\nabla\ell(\mathbf{w}^{(k)}) + \lambda J(\mathbf{w}) \tag{6.20}$$

Ignoring constant terms, this can be rewritten as [Daubechies 04]:

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\operatorname{argmin}}\ \frac{1}{2}\|\mathbf{w} - (\mathbf{w}^{(k)} - \frac{1}{L}\nabla\ell(\mathbf{w}^{(k)}))\|^2 + \frac{1}{L}\lambda J(\mathbf{w}), \tag{6.21}$$

Finally, using definition Eq. 6.6 of the proximity operator for $J(\mathbf{w})$, we obtain the result of Eq. 6.16:

$$\mathbf{w}^{(k+1)} = \text{prox}_{\lambda J/L}\left(\mathbf{w}^{(k)} - \frac{1}{L}\nabla\ell(\mathbf{w}^{(k)})\right) \tag{6.22}$$

□

***Fast Iterative Shrinkage-Thresholding Algorithm* (*FISTA*)**

Inspired by previous findings [Nesterov 07], the *FISTA* (*Fast Iterative Shrinkage-Thresholding Algorithm*) procedure [Beck 09a, Beck 09b] has been developed to speed up the convergence of *ISTA*. While *ISTA* converges in $\mathcal{O}(1/K)$, *FISTA* is proved to converge in $\mathcal{O}(1/K^2)$, where K is the number of iterations. The pseudo code of the *FISTA* procedure is given in Algo. 5. The main improvement in *FISTA* is to compute the next descent direction using the previous one. Such an idea is also present in the well known conjugate gradient algorithm that uses all previous iterates to compute the next descent direction.

Algorithm 5: *FISTA* procedure

Compute the Lipschitz constant $\mathcal{L}$ of the operator $\nabla\ell$.
Initialize $\mathbf{w}^{(0)} \in \mathbb{R}^p$, $\mathbf{v}^{(1)} = \mathbf{w}^{(0)}$ and $t_1 = 1$.
repeat

$$\begin{aligned}
\mathbf{w}^{(k)} &= \text{prox}_{\lambda J/L}\left(\mathbf{v}^{(k)} - \frac{1}{L}\nabla\ell(\mathbf{v}^{(k)})\right) \\
t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2} \\
\mathbf{v}^{(k+1)} &= \mathbf{w}^{(k)} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)})
\end{aligned}$$

until *convergence* ;
return $\mathbf{w}$

6.2 *Total Variation* regularization

We first detail the notations of the problem. We then develop the *TV regularization* framework. Finally, we detail the algorithm used for regression and classification.

6.2.1 Spatial structure and *TV*

We have already described (see Chapter 3) some reference methods for regularized regression: *Elastic net, Lasso* and *Ridge*. However, the associated penalizations do not take into account the underlying structure of $\mathbf{w}$, *i.e.*, a 3-dimensional grid in the case of brain images. As previously stated (see Chapter 5), spatial information can be used to yield more accurate prediction.

TV regularization

In this section, we develop an approach for regularized prediction based on *Total Variation (TV)*, $J(\mathbf{w}) = TV(\mathbf{w})$. *TV*, mathematically defined as the ℓ_1 norm of the image gradient, has been primarily used for image denoising [Rudin 92, Chambolle 04] as it preserves edges. The motivation for using TV for brain imaging is that it promotes estimates $\hat{\mathbf{w}}$ of $\mathbf{w}$ with a block structure, therefore outlining the brain regions involved in the target behavioral variable.

Let us define $\Omega \subset \mathbb{R}^3$ the 3D image domain. In a continuous formulation, the coefficients $\mathbf{w}$ define a function from Ω to $\mathbb{R}$, *i.e.*, $\mathbf{w} : \Omega \to \mathbb{R}$. Its TV reads:

$$TV(\mathbf{w}) = \int_{\omega \in \Omega} \|\nabla \mathbf{w}\|(\omega) d\omega \tag{6.23}$$

$$= \int_{\omega \in \Omega} \sqrt{\nabla_x \mathbf{w}(\omega)^2 + \nabla_y \mathbf{w}(\omega)^2 + \nabla_z \mathbf{w}(\omega)^2} d\omega \tag{6.24}$$

Gradient and divergence

An issue specific to *fMRI* data is the computation of the gradient and divergence over a mask of the brain with correct border conditions. We denote M the mask of the brain, which is a $p_i \times p_j \times p_k$ three dimensional grid, with:

$$\begin{cases} M_{i,j,k} = 1 & \text{if the voxel is in the mask} \\ M_{i,j,k} = 0 & \text{if the voxel is not in the mask} \end{cases}$$

with $\sum_{i,j,k} M_{i,j,k} = p$. Additionally, we define grad : $\mathbb{R}(\Omega) \to \mathbb{R}^3(\Omega)$ a gradient operator and div : $\mathbb{R}^3(\Omega) \to \mathbb{R}(\Omega)$ the associated adjoint divergence operator. Let K the convex set defined by:

$$K = \{g : \Omega \to \mathbb{R}^3 \mid \forall \omega \in \Omega, \|g(\omega)\| \, \omega \leq 1\} \tag{6.25}$$

and Π_K the projection operator onto the set K:

$$\begin{cases} \Pi_K(g)(\omega) = g(\omega) & \text{if } \|g(\omega)\| \leq 1 \\ \Pi_K(g)(\omega) = g(\omega)/\|g(\omega)\| & \text{otherwise.} \end{cases}$$

With $I \in \mathbb{R}^{p_i \times p_j \times p_k}$ an image, the gradient operator is defined by:

$$(\text{grad } I)_x^{i,j,k} = \begin{cases} I_{i+1,j,k} - I_{i,j,k} & \text{if } M_{i,j,k} = M_{i+1,j,k} = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{grad } I)_y^{i,j,k} = \begin{cases} I_{i,j+1,k} - I_{i,j,k} & \text{if } M_{i,j,k} = M_{i,j+1,k} = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{grad } I)_z^{i,j,k} = \begin{cases} I_{i,j,k+1} - I_{i,j,k} & \text{if } M_{i,j,k} = M_{i,j,k+1} = 1 \\ 0 & \text{otherwise} \end{cases}$$

The divergence operator for a gradient p is defined by:

$$\begin{aligned} (\text{div } p)^{i,j,k} = & \begin{cases} p_{i,j,k}^x - p_{i-1,j,k}^x & \text{if } M_{i,j,k} = M_{i-1,j,k} = 1 \\ p_{i,j,k}^x & \text{if } M_{i,j,k} \neq M_{i-1,j,k} = 0 \\ -p_{i-1,j,k}^x & \text{if } M_{i,j,k} \neq M_{i-1,j,k} = 1 \end{cases} \\ & + \begin{cases} p_{i,j,k}^y - p_{i,j-1,k}^y & \text{if } M_{i,j,k} = M_{i,j-1,k} = 1 \\ p_{i,j,k}^y & \text{if } M_{i,j,k} \neq M_{i,j-1,k} = 0 \\ -p_{i,j-1,k}^y & \text{if } M_{i,j,k} \neq M_{i,j-1,k} = 1 \end{cases} \\ & + \begin{cases} p_{i,j,k}^z - p_{i,j,k-1}^z & \text{if } M_{i,j,k} = M_{i,j,k-1} = 1 \\ p_{i,j,k}^z & \text{if } M_{i,j,k} \neq M_{i,j,k-1} = 0 \\ -p_{i,j,k-1}^z & \text{if } M_{i,j,k} \neq M_{i,j,k-1} = 1 \end{cases} \end{aligned}$$

6.2.2 Convex optimization of *Total Variation*

Proximity operator of the *Total Variation*

We now give some details in the particular case of the proximity operator $\text{prox}_{\lambda TV}$ known as the *ROF (Rudin Osher Fatemi)* problem in the image processing literature [Rudin 92]. The computation of $\text{prox}_{\lambda TV}$ and the associated duality gap requires the derivation of a Lagrange dual problem [Boyd 04].

Proposition 6.7 ($\text{prox}_{\lambda TV}$ Dual problem)**.** *A dual problem associated with* $\text{prox}_{\lambda TV}$ *is given by*

$$\mathbf{z}^* = \underset{\mathbf{z} \in K}{arg\max} -\|div\, \mathbf{z} + \mathbf{w}/\lambda\|_2^2 \ , \tag{6.26}$$

where $\mathbf{z}$ *is the dual variable that satisfies* $\mathbf{v} = \mathbf{w} + \lambda\, div\, \mathbf{z}$*.*

This result is adapted from [Chambolle 04] (see a sketch of the proof below). The problem Eq. 6.26 consist in maximizing a smooth concave function over a

convex set. As shown in [Beck 09b], it can be solved with the *FISTA* iterative procedure. The resolution of the *ROF* problem is therefore achieved by solving the dual problem. Once $\mathbf{z}^*$ is obtained, $\mathbf{v}^* = \text{prox}_{\lambda TV}(\mathbf{w})$ is given by $\mathbf{v}^* = \mathbf{w} + \lambda \operatorname{div} \mathbf{z}^*$.

Duality gap of the *Total Variation*

Let us introduce now the notion of *duality gap*. The *duality gap* is a natural stopping condition for iterative convex optimization solvers, such as *ISTA* and *FISTA*. In practice, if the *duality gap* is below a value $\epsilon > 0$, it guarantees that the solution obtained is ϵ-optimal, *i.e.*, that the value of the cost-function reached by the algorithm is not greater than ϵ more the globally optimal value. A comprehensive presentation of this notion [Boyd 04] is beyond the scope of this chapter, and we give some details in the particular case of the *ROF* problem. The latter result also gives an estimates of the duality gap (see a sketch of the proof below).

Proposition 6.8 (Duality gap)**.** *The* duality gap δ_{gap} *associated with the* ROF *problem is given by:*

$$\delta_{gap}(\mathbf{v}) = \frac{1}{2}\|\mathbf{w} - \mathbf{v}\|_2^2 + \lambda TV(\mathbf{v}) - \frac{1}{2}(\|\mathbf{w}\|_2^2 - \|\mathbf{v}\|_2^2) \geq 0 \ , \tag{6.27}$$

where the primal variable $\mathbf{v}$ *is obtained during the iterative procedure from the current estimate of the dual variable* $\mathbf{z}$ *with* $\mathbf{v} = \mathbf{w} + \lambda \, div \, \mathbf{z}$*.*

This *duality gap* will be used as a stopping criterion for the *FISTA* procedure solving the *ROF* problem. At each iteration of the *FISTA* procedure, we stop the iterative loop if the *duality gap* is below a given threshold ϵ. In practice, ϵ is set to $10^{-4} \times \|\mathbf{w}\|_2^2$ to be invariant to the scaling of the data.

Sketch of the proof. We give the sketch of proofs of propositions 6.7 and 6.8. We recall [Boyd 04] that the duality between the ℓ_1 norm and the ℓ_∞ norm yields:

$$TV(\mathbf{v}) = \|\nabla \mathbf{v}\|_1 = \max_{\|\mathbf{z}\|_\infty \leq 1} \langle \nabla \mathbf{v}, \mathbf{z} \rangle \tag{6.28}$$

and that the adjoint relation between the gradient and the divergence operator reads:

$$\langle \nabla \mathbf{v}, \mathbf{z} \rangle = -\langle \mathbf{v}, \operatorname{div} \mathbf{z} \rangle \tag{6.29}$$

Using Eq. 6.28 and Eq. 6.29, we minimize:

$$\begin{aligned}
\min_{\mathbf{v}} \left(\frac{1}{2}\|\mathbf{w} - \mathbf{v}\|_2^2 + \lambda TV(\mathbf{v}) \right) &= \lambda \min_{\mathbf{v}} \left(\frac{1}{2\lambda}\|\mathbf{w} - \mathbf{v}\|_2^2 + \max_{\|\mathbf{z}\|_\infty \leq 1} \langle \nabla \mathbf{v}, \mathbf{z} \rangle \right) \\
&= \lambda \max_{\|\mathbf{z}\|_\infty \leq 1} \left(\min_{\mathbf{v}} \left(\frac{1}{2\lambda}\|\mathbf{w} - \mathbf{v}\|_2^2 + \langle \nabla \mathbf{v}, \mathbf{z} \rangle \right) \right) \\
&= \lambda \max_{\|\mathbf{z}\|_\infty \leq 1} \left(\min_{\mathbf{v}} \left(\frac{1}{2\lambda}\|\mathbf{w} - \mathbf{v}\|_2^2 - \langle \mathbf{v}, \operatorname{div} \mathbf{z} \rangle \right) \right)
\end{aligned}$$

The computation of the minimum and the maximum above can be exchanged because the optimization over $\mathbf{v}$ is convex and the optimization over $\mathbf{z}$ is concave [Boyd 04].
By setting the derivative with respect to $\mathbf{v}$ to 0 one gets the resulting solution of the minimization problem over $\mathbf{v}$:

$$\min_{\mathbf{v}} \left(\frac{1}{2\lambda} \|\mathbf{w} - \mathbf{v}\|_2^2 - \langle \mathbf{v}, \operatorname{div} \mathbf{z} \rangle \right) \Rightarrow \mathbf{v}^* = \mathbf{w} + \lambda \operatorname{div} \mathbf{z} \tag{6.30}$$

Replacing $\mathbf{v}$ by $\mathbf{v}^*$ in the previous expression leads to:

$$\begin{aligned}
\min_{\mathbf{v}} \left(\frac{1}{2} \|\mathbf{w} - \mathbf{v}\|_2^2 + \lambda TV(\mathbf{v}) \right) &= \lambda \max_{\|\mathbf{z}\|_\infty \leq 1} \left(\frac{\lambda}{2} \|\operatorname{div} \mathbf{z}\|_2^2 - \langle \mathbf{w}, \operatorname{div} \mathbf{z} \rangle - \lambda \|\operatorname{div} \mathbf{z}\|_2^2 \right) \\
&= \lambda \max_{\|\mathbf{z}\|_\infty \leq 1} \left(-\frac{\lambda}{2} \|\operatorname{div} \mathbf{z}\|_2^2 - \langle \mathbf{w}, \operatorname{div} \mathbf{z} \rangle \right) \\
&= \frac{1}{2} \max_{\|\mathbf{z}\|_\infty \leq 1} \left(-\lambda^2 \|\operatorname{div} \mathbf{z}\|_2^2 - 2\lambda \langle \mathbf{w}, \operatorname{div} \mathbf{z} \rangle \right) \\
&= \frac{1}{2} \max_{\|\mathbf{z}\|_\infty \leq 1} \left(\|\mathbf{w}\|_2^2 - \|\lambda \operatorname{div} \mathbf{z} + \mathbf{w}\|_2^2 \right)
\end{aligned}$$

This gives the proof of Prop. 6.7. Also, given a variable $\mathbf{z}$ satisfying $\|\mathbf{z}\|_\infty \leq 1$ and an associated $\mathbf{w}$ such that $\mathbf{v} = \mathbf{w} + \lambda \operatorname{div} \mathbf{z}$, one can guarantee that

$$\frac{1}{2} \|\mathbf{w} - \mathbf{v}\|_2^2 + \lambda TV(\mathbf{v}) \geq \frac{1}{2} (\|\mathbf{w}\|_2^2 - \|\mathbf{v}\|_2^2) \tag{6.31}$$

The strict convexity of the problem guarantees that, at the optimum, the equality holds. This last derivation proves the proposition 6.8. □

6.2.3 Prediction framework

For $J(\mathbf{w}) = TV(\mathbf{w})$, the global algorithm for solving the minimization problem defined in Eq. 6.12 consists in a *FISTA* procedure (resolution of the *ROF* problem) nested inside an *ISTA* procedure (resolution of the main minimization problem). The *FISTA* procedure is performed at each step of *ISTA* with a *warm restart* on the dual variable $\mathbf{z}$. We do not use *FISTA* for solving the main minimization problem, as this procedure requires an exact *proximity operator*. The resolution of the *ROF* problem only leads to an ϵ-optimal solution. The pseudo-code of the global algorithm for the *TV regularization* is provided in Table 6. Moreover, the upper bound $\tilde{L}$ for the Lipschitz constant of the *FISTA* procedure also needs to be estimated on each input data. To do this we use a power method that is classically used to estimate the spectral norm of a linear operator, here equal to the Laplacian $\Delta : \Omega \to \Omega$ defined by $\Delta(\omega) = \operatorname{div}(\operatorname{grad}(\omega))$.

Algorithm 6: *TV regularization* solver

Set maximum number of iterations K (*ISTA*).
Set the threshold ϵ on the *dual gap* (*FISTA*).
Set $L = 1.1\mathcal{L}$ where $\mathcal{L}$ is the Lipschitz constant of $\nabla\ell$.
Set $\tilde{L} = 1.1\tilde{\mathcal{L}}$ where $\tilde{\mathcal{L}}$ is the Lipschitz constant of the Laplacian operator $\Delta : w \in \mathbb{R}(\Omega) \rightarrow \text{div}(\text{grad}(w))$.
Initialize $\mathbf{z} \in \mathbb{R}(\Omega^3)$ with zeros.
ISTA loop
for $k = 1 \dots K$ **do**
  $\mathbf{v} = \mathbf{w} - \frac{1}{L}\nabla\ell(\mathbf{w})$
  ### FISTA loop ###
  Initialize $\mathbf{z}_{aux} = \mathbf{z}$, $t = 1$
  repeat
    $\mathbf{z}_{old} = \mathbf{z}$
    $\mathbf{z} = \Pi_K\left(\mathbf{z}_{aux} - \frac{1}{\lambda\tilde{L}}\text{grad}(L\mathbf{v} - \lambda\text{div}(\mathbf{z}_{aux}))\right)$
    $t_{old} = t$
    $t = (1 + \sqrt{1 + 4t^2})/2$
    $\mathbf{z}_{aux} = \mathbf{z} + \frac{t_{old}-1}{t}(\mathbf{z} - \mathbf{z}_{old})$
  until $\delta_{gap}(a) \leq \epsilon$;
  $\mathbf{w} = \mathbf{v} - \lambda\text{div}(\mathbf{z})$
return $\mathbf{w}$

TV regression

The regression version of the *TV* is called *TV regression*. In this case, we use the least-squares *loss*:

$$\begin{cases} \ell(\mathbf{w}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \\ \nabla\ell(\mathbf{w}) = -\frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) \end{cases} \tag{6.32}$$

The Lipschitz constant $\mathcal{L}$ of the operator $\nabla\ell$ is $\mathcal{L} = |||\mathbf{X}^T\mathbf{X}|||/n$, where $|||.|||$ stands for the spectral norm equal to largest singular value. The constant L is set in practice to $L = 1.1\mathcal{L}$.

TV classification

The classification version of the *TV* is called *TV classification*, and is based on a logistic *loss* (see Chapter 3). The corresponding *loss* and the *loss* gradient read:

$$\begin{cases} \ell(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + \exp^{-y_i(\mathbf{X_i}^T\mathbf{w})}\right) \\ \nabla\ell(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{y_i\mathbf{X_i}}{1+\exp^{y_i(\mathbf{X_i}^T\mathbf{w})}} \end{cases} \tag{6.33}$$

The Lipschitz constant $\mathcal{L}$ of the operator $\nabla\ell$ is $\mathcal{L} = 1/(4n)$. The classification framework developed in this chapter treats the binary case with a logistic

model, *a.k.a.*, binomial model. In our analysis, we expand this framework to multi-class classification using a one-versus-one voting heuristic. The number of classifiers used is $(k-1) \times (k-2)/2$, where k is the number of classes. The predicted class is then selected as the class which yields the highest probability across the predictions of all of the classifiers, as defined by the *logistic* regression.

6.3 *TV* for *fMRI-based inverse inference*

We first illustrate *Total Variation* on simulated data. Then, we give the results obtained on a real *fMRI* dataset, in both regression and classification settings.

6.3.1 Illustration on simulated neuroimaging data

We create a set of simulated neuroimaging data, as described in Appendix B.1. We compare *TV regression* with a value of λ cross-validated in the range $\{0.01, 0.05, 0.1, \}$, with the two reference algorithms, *Elastic net* and *SVR*. All three methods are optimized by 4-folds cross-validation within the training set as described in Appendix B.1.

We compare the different methods on the simulated data: see the results in Fig. 6.3. The true weights (a) and resulting *Anova* F-scores (b) are shown. Only *TV regression* (e) extracts the simulated discriminative regions. The reference methods also find the *ROIs*, but *Elastic net* (d) only retrieves part of the support of the weights, and yields an overly sparse solution. We note that the weights in the primal space estimated by *SVR* (c) are non-zero everywhere and do not retrieve the support of the ground truth.

6.3.2 Sensitivity study on real data

Before any further analysis on real data, we have performed a sensitivity analysis of our model, with regards to the parameter λ. In the inter-subject analysis on the *mental representation of size*, we compute the cross-validated prediction accuracy for twelve different values of λ between 10^{-4} and 0.95. The results are detailed in Fig. 6.2, and are extremely stable in a wide range of values $[10^{-4}, 10^{-1}]$. Based on these results, we set the parameter $\lambda = 0.05$ for the following analyzes.

6.3.3 Results for regression analysis

In a first set of analyzes, we assess the performance of *TV regression* in both intra-subject and inter-subject cases, where the aim is to predict the size of an object seen by the subject during the experiment (see Appendix B.2).

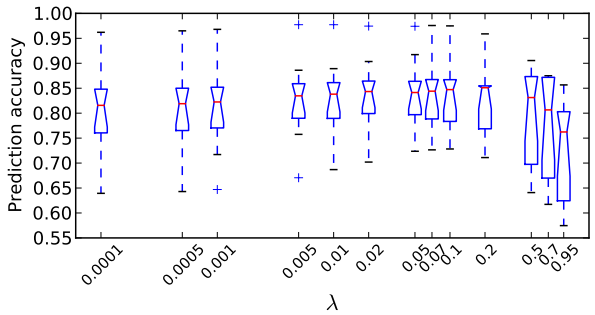


Figure 6.2: Explained variance ζ for different values of λ, in the inter-subjects regression analysis. The accuracy is very stable regarding to λ in the range $[10^{-4}, 10^{-1}]$.

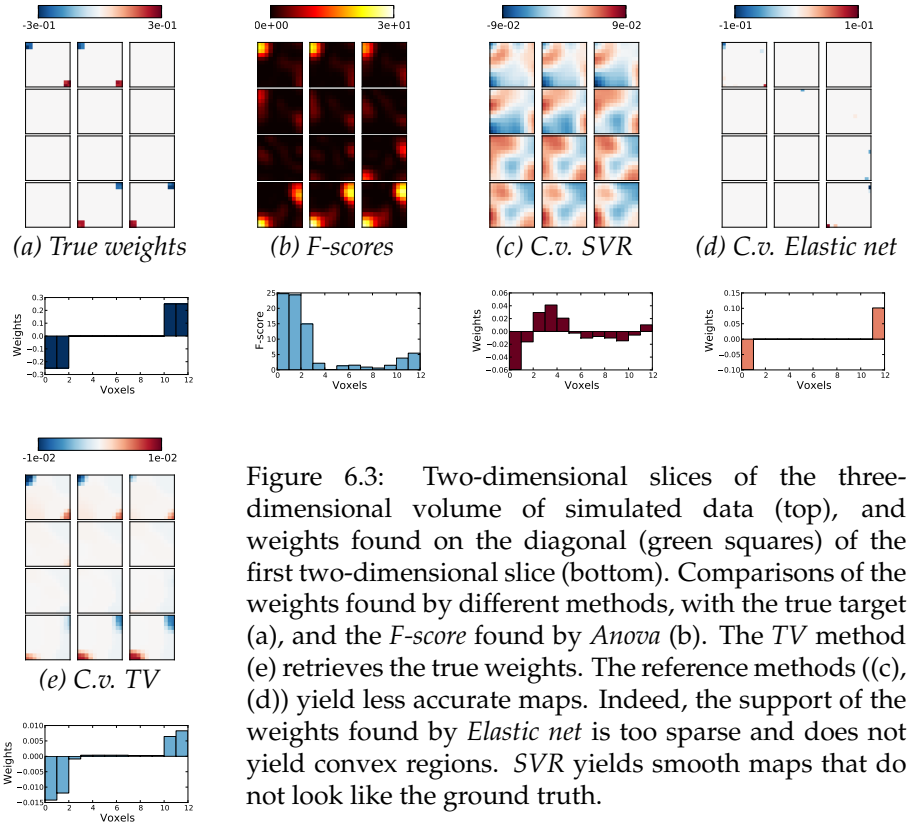


Figure 6.3: Two-dimensional slices of the three-dimensional volume of simulated data (top), and weights found on the diagonal (green squares) of the first two-dimensional slice (bottom). Comparisons of the weights found by different methods, with the true target (a), and the *F-score* found by *Anova* (b). The *TV* method (e) retrieves the true weights. The reference methods ((c), (d)) yield less accurate maps. Indeed, the support of the weights found by *Elastic net* is too sparse and does not yield convex regions. *SVR* yields smooth maps that do not look like the ground truth.

Intra-subject analysis

The results obtained by the three methods are given in Table. 6.1. *TV regression* outperforms the two alternative methods, yielding an average explained variance of 0.92 across the subjects. The difference with *SVR* is significant, but not with *Elastic net*. Moreover, the results of the regularized methods (*TV, Elastic net*) are more stable (standard deviation three times smaller) across subjects, than the results of the *SVR*.

Methods	mean ζ	std ζ	max ζ	min ζ	p-value to TV
SVR	0.82	0.07	0.9	0.67	0.0015 **
Elastic net	0.9	0.02	0.93	0.85	0.0672 *
TV $\alpha = 0.05$	0.92	0.02	0.95	0.88	-

Table 6.1: *Regression - Mental representation of size - Intra-subject analysis*. Explained variance ζ for the three different methods. The p-values are computed using a paired t-test. *TV regression* yields the best prediction accuracy, while being more stable than the two reference methods (standard deviation of ζ three times smaller than *SVR*).

Inter-subject analysis

The results obtained with the three methods are given in Table. 6.2. As in the intra-subject analysis, *TV regression* outperforms the two alternative methods, yielding an average explained variance of 84%, and also more stable predictions. Such stability can be illustrated on the subject 3, where both reference methods yield poor results, while *TV regression* yields an explained variance 0.2 higher.

Methods	mean ζ	std ζ	max ζ	min ζ	p-value to TV
SVR	0.77	0.11	0.97	0.58	0.0277 **
Elastic net	0.78	0.1	0.97	0.65	0.0405 **
TV $\lambda = 0.01$	0.83	0.07	0.98	0.69	0.236
TV $\lambda = 0.05$	0.84	0.07	0.97	0.72	-
TV $\lambda = 0.1$	0.84	0.07	0.97	0.73	0.5544

Table 6.2: *Regression - Mental representation of size - Inter-subject analysis*. Explained variance ζ for the three different methods. The p-values are computed using a paired t-test. *TV regression* still yields the best prediction accuracy, with an explained variance 0.06 higher than the best reference method (*elastic net*).

The average positions and the sizes of the three main clusters found by the *TV* algorithm, using all the subjects, are given Table. 6.3. *TV regression* adapts the regularization to tiny regions, yielding *ROIs* from 25 to 193 voxels. The clusters are found within the occipital cortex. The majority of informative voxels are located in the posterior part of the occipital cortex ($y \leq -90$

mm), most likely corresponding to primary visual cortex, with one additional slightly more anterior cluster in posterior lateral occipital cortex. This is consistent with the previous findings [Eger 08] where a gradient of sensitivity to size was observed across object selective lateral occipital *ROIs*, and the most accurate discrimination of sizes in primary visual cortex.

x (mm)	y (mm)	z (mm)	Sizes (voxels)
24	-92	-16	25
-26	-96	-10	103
16	-96	12	193

Table 6.3: *Mental representation of size - Inter-subject analysis*: positions and sizes of the three main clusters for the *TV regression* method.

The maps of weights found by different values of the regularization parameter are detailed in Fig. 6.5. One can notice that, as λ increases, the spatial support of these maps tends to be aggregated in a few clusters within the occipital cortex, and that the maps have a nearly constant value on these clusters. By contrast, both reference methods yield uninterpretable (*i.e.* more complex) maps (see Fig. 6.4), with a few informative voxels scattered in the whole occipital cortex.

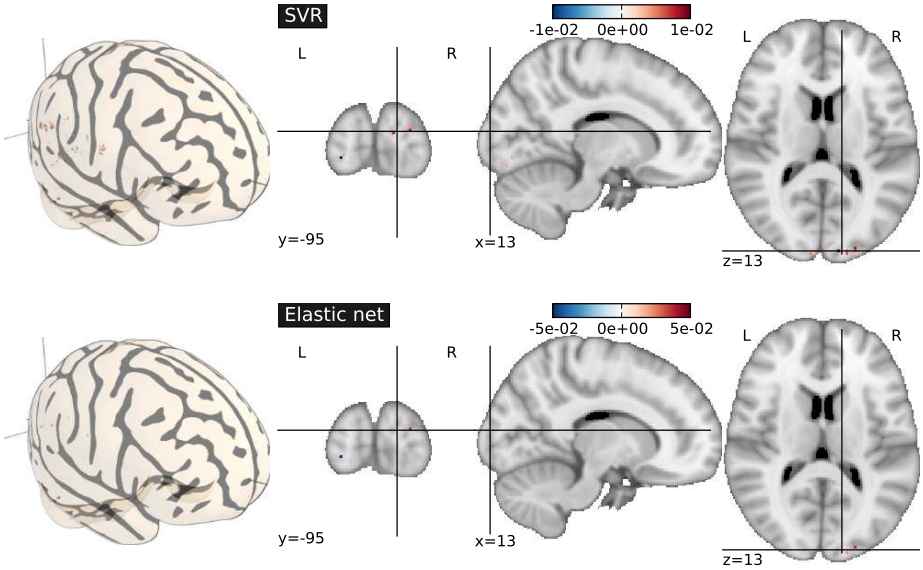


Figure 6.4: *Regression - Mental representation of size - Inter-subject analysis*. Maps of weights found by the *SVR* (up) and *elastic net* (bottom) methods. The optimal number of voxels selected by *Anova* is 500, but *Elastic net* further reduces this set to 21 voxels.

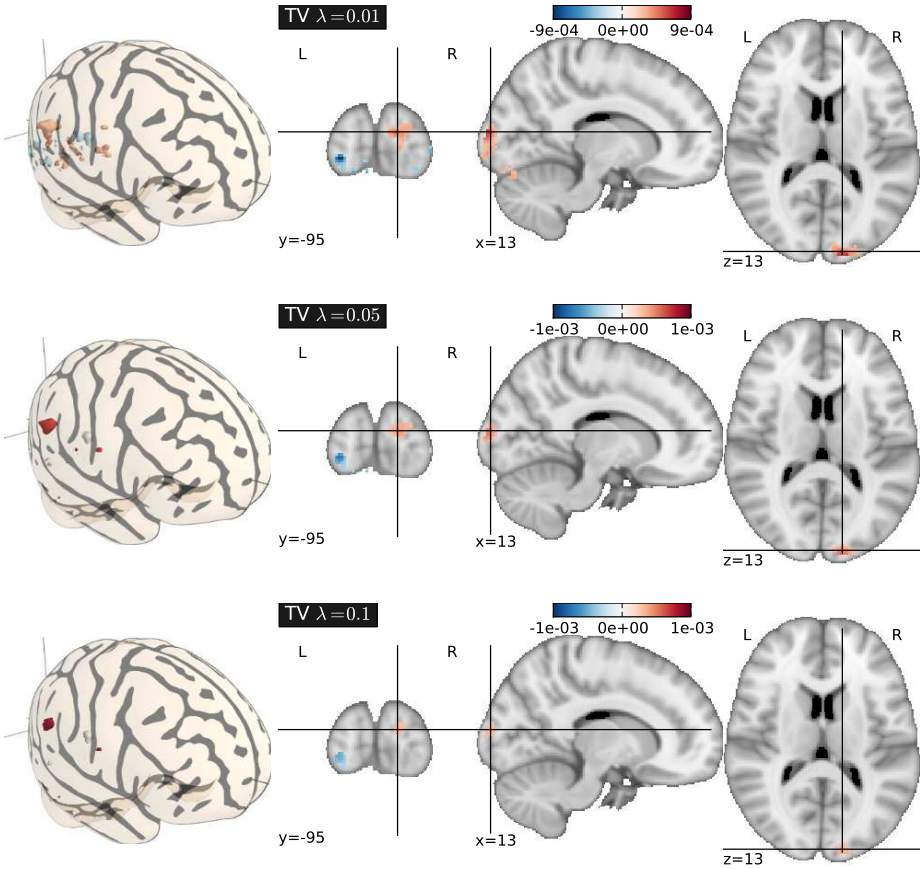


Figure 6.5: *Regression - Mental representation of size - Inter-subject analysis*. Maps of weights found by *TV regression* for various values of the regularization parameter λ. When λ decreases, the *TV regression* algorithm creates different clusters of weights with constant values. These clusters are easily interpretable, compared to voxel-based maps (see below). The *TV regression* algorithm is very stable for different values of λ, has shown by the explained variance ζ.

6.3.4 Results for classification analysis

In a second analysis, we study the *TV classification* method in an intra-subject and inter-subject classification analysis, in which the aim is to predict which object among 4 is seen by the subject (see Appendix B.2).

Intra-subject analysis

The results found by the three methods are given in Table 6.4. The highest prediction accuracy is obtained with the *SVC* approach. The proposed approach

yields prediction accuracy higher than chance, but is not as good as the reference methods. This can be explained by the fact that the information used for objects discrimination is a relatively high level cognitive information: the discriminative information might be encoded in a more finer grain activation pattern (see [Haxby 01]) than the information on the sizes (*sparse coding*). Thus, within a subject, the voxel-to-voxel correspondence between different volumes of a same acquisition is not problematic, and voxel-based approach such as *SVC* performs well. By contrast, our approach, that seeks clusters of activations, can loose a part of the fine grain information, and thus does not perform as well as *SVC* in this particular context.

Methods	mean κ	std κ	max κ	min κ	p-value to SVC
SVC	92.22	5.7	98.61	79.17	-
SMLR	86.81	7.64	94.44	70.83	0.0171 *
TV $\lambda = 0.05$	63.75	12.19	87.5	45.83	0.0001 **

Table 6.4: *Classification - Mental representation of shape - Intra-subject analysis*. Classification score κ for the three different methods. The p-values are computed using a paired t-test. *SVC* yields the best prediction accuracy.

Inter-subject analysis

The results (averaged across the two categories) found by the three methods are given in Table. 6.5. As in the inter-subject regression analysis, the *TV*-based method outperforms the *SMLR* method. Moreover, it yields an average classification score similar to the *SVC* while being more stable. Compared to the intra-subject analysis, the *TV*-based method performs better, due to the fact there is no longer a voxel-to-voxel correspondence. Seeking clusters of activation thus seems a reasonable way to cope with inter-subject variability.

Methods	mean κ	std κ	max κ	min κ	p-value to SVC
SVC	48.33	15.72	75.0	25.0	-
SMLR	42.5	9.46	58.33	33.33	0.2419
TV $\lambda = 0.05$	45.83	14.55	66.67	25.0	0.7128

Table 6.5: *Classification - Mental representation of shape - Inter-subject analysis*. Classification score κ for the three different methods. The p-values are computed using a paired t-test. *SVC* yields the best prediction accuracy.

The maps of weights found by *TV classification* for different binaries classifiers, averaged across the two categories of objects, are given Fig. 6.6. We can notice that *TV* extracts predictive patterns in the primary visual cortex, but also in a more lateral and anterior part of the cortex, as expected [Eger 08].

The average number of selections of each voxel within one of the three larger clusters for each one-versus-one map are given Fig. 6.7. The informa-

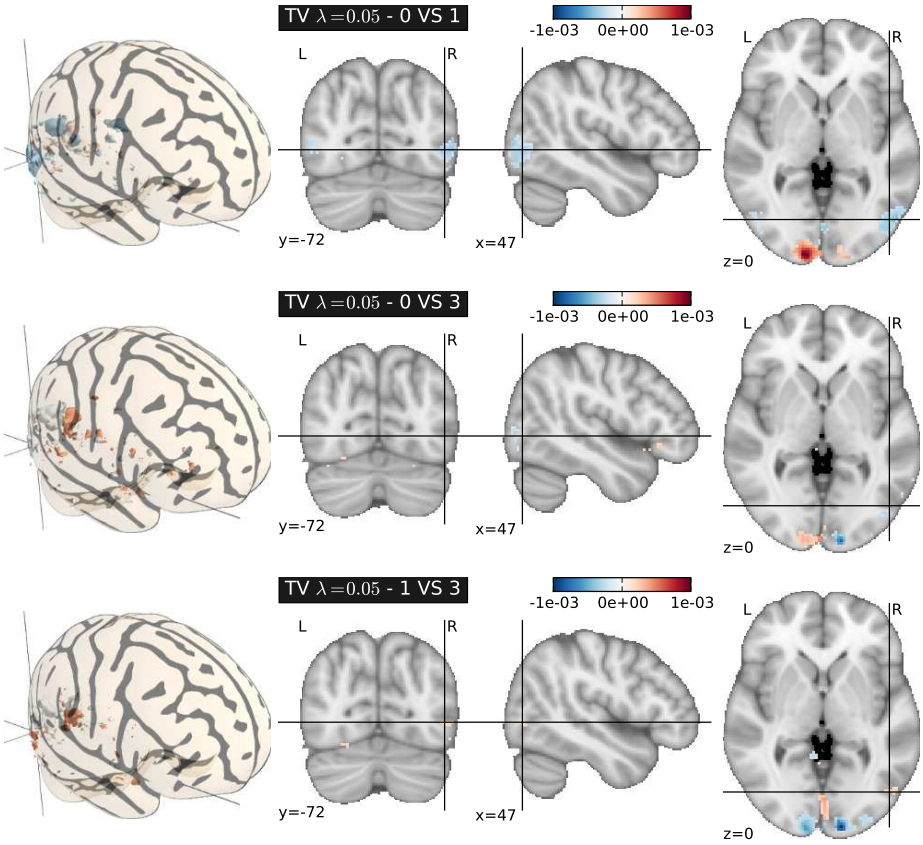


Figure 6.6: *Classification - Mental representation of shape - Inter-subject analysis*. Maps of weights found by *TV classification* for different binaries classifiers, averaged across the two categories of objects. Some predictive voxels are found in a more lateral and anterior part of the cortex, as expected [Eger 08].

tive clusters are more anterior and more ventral than the ones found within the sizes prediction paradigm. We thus confirm the results found by classical brain mapping approach, such as *Anova* (see results in Appendix B), while providing a classification score based on cross-validation on independent data which allows to check the actual implication of these regions in the cognitive process.

6.3.5 Discussion

In this chapter, we present the first use of *TV* regularization for brain decoding. This method outperforms the reference methods with regards to prediction accuracy, and yields sparse brain maps with clear informative foci. Moreover, in

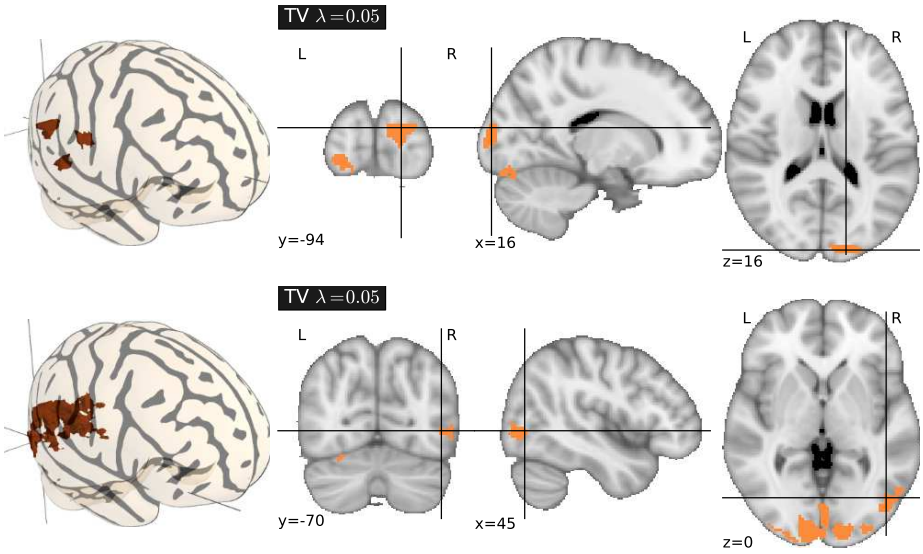


Figure 6.7: *Mental representations of shape and size* - Inter-subject analysis. Top - voxels selected within one of the three main clusters by *TV regression*, for the *Sizes prediction experiment*. Bottom - voxels selected at least one time within one of the three main clusters for each of the one-vs-one *TV classification*, for the *Objects prediction experiment*. Some clusters found in the *Object prediction experiment* are more anterior (their center of mass are at $[50, -72, -2]$ mm and $[46, -80, -2]$ mm) than the ones found for the *Size prediction experiment* (center of mass at $[16, -96, 10]$ mm and $[-26, -96, -10]$ mm). This is coherent with the hypothesis that the processing of shapes is done at a higher level in the processing of visual information, and thus the implied regions are found further in the ventral pathway [Eger 08].

classification settings, we integrate *TV* in a logistic regression framework. This approach which, to our knowledge, has not been used before, yields high prediction accuracy, is a promising method for machine learning problems beyond the scope of neuroimaging.

One major benefit of the proposed method is that, in the case of a multi-subject studies, the use of regions with a spatial extent compensates for spatial misalignment, hence it generalizes better than voxel-based methods. As shown on both inter-subject analyzes, the proposed *TV* approach yields significantly higher prediction accuracy than reference voxel-based methods. In addition, as the proposed approach takes into account the spatial neighborhood of the images, it yields weight maps very similar to the maps obtained by a classical brain mapping approach (such as *Anova*). We note that the solution found by our method is sparse but sufficient for good prediction accuracy, which explains the fact that the regions observed may be more localized than the ones with *Anova*. Thus, the *TV* approach benefits from the power of a predictive

framework, while leading to accurate brain maps similar to classical *SPMs*.

Moreover, *TV regression* allows to consider the whole brain in the analysis, without requiring any prior feature selection. As many accurate dimension reduction approaches such as *Recursive Feature Elimination* [Guyon 02] can be extremely costly in computational time, avoiding this step is a major asset. An important feature of our implementation is thus that it reduces computation time to a reasonable amount, so that it is not significantly more costly than SVR or elastic net in practical settings (*i.e.*, including the cross-validation loops). In the inter-subject regression analysis, the average computational time is 185 seconds for *TV regression*, 131 seconds for *Anova + SVR* and 121 seconds for *Anova + Elastic net*, on a *Intel(R) Xeon(R) CPU* at 2.83GHz.

Regularization of the voxel weights significantly increases the generalization ability in regression problems, because it performs feature selection and training of the prediction function jointly. However, to date, regularization has most often been performed without using the spatial structure of the images. By applying a penalization on the gradient of the weight and thus taking into account the spatial structure of the image-based information, our approach performs an adaptive and efficient regularization, while creating sparse weight maps with regions of quasi constant weights. *TV regularization* method fulfills thus the two requirements that make it suitable for neuroimaging brain mapping: a good prediction accuracy (better than the reference methods for regression experiments, and equal for classification, with the exception of intra-subject classification), and a set of interpretable features, made of clusters of similarly-tuned voxels. In that sense, it can be seen as the first method for performing *multivariate brain mapping*.

From a neuroscientific point of view, the regions extracted from the whole brain analysis in the size discrimination task are concentrated in the early visual cortex. This is consistent with the fact that early visual cortex yields highly reliable signals that are discriminative of feature/shape differences between object exemplars, which holds as long as no high-level generalization across images is required (see *e.g.* [Cox 03] and [Eger 08]). This is expected, given the small receptive fields of neurons in these regions that will reliably detect differences in the spatial envelop or other low-level structure of the images. Most importantly, the predictive spatial pattern is stable enough across individuals to make reliable predictions in new subjects. In fact our method compares best with regards to the state of the art in the inter-subjects setting, as it selects predictive regions that are not very sensitive to anatomo-functional variability. In the object discrimination task, the clusters found by our approach are also in the visual cortex, but including more anterior ones (probably corresponding to *lateral occipital* region) compared to size discrimination, which is consistent with the fact that shape discrimination requires intermediate/higher level visual areas. The finding that large parts of early visual cortex are also discriminative is explained by the fact that we do not perform generalization across viewing condition and classification can therefore be driven by lower-level features. Even if similar maps as the ones found by our method can be obtained using *Anova*, these do not provide a measure of the quantity of information

(*i.e.* prediction score) shared by these regions, and thus a measure of their implication in the *neural coding* of the cognitive process. A further advantage of our approach is that the regions obtained in this approach are more spatially coherent and therefore provide a simpler description of the data.

6.4 Conclusion - *Total variation* regularization

In this chapter, we have introduced some concepts for *convex optimization*. In particular, the notion of *proximity operator* allows to develop some iterative procedures such as *ISTA* and *FISTA*, in order to solve convex minimization problems. In the specific case of *Total Variation regularization*, this optimization is done within a double loop of *ISTA* and *FISTA* procedures. The algorithm detailed here can be used for both regression and classification.

TV regularization can be used for extracting information from brain images, both in regression or classification settings. Feature selection and model estimation are performed jointly and capture the predictive information present in the data better than alternative methods. A particularly important property of this approach is its ability to create spatially coherent regions with similar weights, yielding simplified and informative sets of features. Experimental results show that this algorithm performs well on real data, and is far more accurate than voxel-based reference methods for multi-subject analysis. In particular, the segmented regions are robust to inter-subject variability. These observations demonstrate that *TV regularization* is a powerful tool for understanding brain activity and spatial mapping of cognitive process, and is the first method that derives meaningful statistical weight maps, as in the standard *SPM* approach, within the inverse inference framework.

Publications

The contributions developed in this chapter have been published in:

- V. Michel, A. Gramfort, G. Varoquaux and B. Thirion. *Total Variation regularization enhances regression-based brain activity prediction*. In 1st ICPR Workshop on Brain Decoding - Pattern recognition challenges in neuroimaging - 20th International Conference on Pattern Recognition. 2010.
- V. Michel, A. Gramfort, G. Varoquaux, E. Eger and B. Thirion. *Total variation regularization for fMRI-based prediction of behavior*. Submitted to IEEE Transactions on Medical Imaging. 2010.

Conclusion

In this thesis, we have presented contributions to the *inverse inference* framework for *fMRI* analysis, which relies on a *pattern recognition* approach. It can be used for *decoding* brain activity, and more precisely, learning the *spatial layout* of the *neural coding*, from brain images.

Experimental contributions - Many *statistical learning* algorithms can be used for prediction and *dimension reduction* in the *pattern recognition* step. We have presented state of the art algorithms, and implemented and evaluated them on real data. We have systematically investigated their performance in the challenging context of *fMRI* data, and highlighted the required properties for *machine learning* algorithm to be well-suited to *fMRI*-based inverse inference. Studies on experimental data were performed in collaboration with neuroscientists and yielded significant prediction results in fields such as mental representations of quantities or preferences, as well as result on the more challenging aspect of cortical recycling in high-level cognitive functions.

Methodological contributions - Our research is focused on methods that can increase the *interpretability* of the resulting maps:

- A first contribution is a *Bayesian* framework for sparsity-inducing regularization, called *Multi-Class Sparse Bayesian Regression – MCBR*. This approach is a generalization of the two principal Bayesian regularizations, *Bayesian Ridge Regression* and *Automatic Relevance Determination*.
- A second axis of our research was motivated by the fact that *fMRI* data have a spatial structure that has rarely been taken into account within the different state of the art approaches. Thus, we proposed an approach, called *supervised clustering*, that includes spatial information in the prediction framework, and yields clustered weighted maps. It can be used with any prediction functions for highly dimensional data.
- Our last contribution aims at implementing both sparsity and spatially-informed regularization within the same framework. We proposed a generalization of the *Total Variation regularization* for prediction task, and we showed its good performance in the case of *fMRI* data analysis.

These different approaches have been tested on real data on the mental representations of size and shape of objects, and yield accurate maps for decoding specific parts of the visual system.

Software contributions - In addition to the methodological directions that we have described in this thesis, we have also focused on the implementation of the algorithms studied an detailed in this thesis. A high-quality implementation is critical as the high dimensionality of *fMRI* data can be challenging.

We have also contributed to *Scikit-learn*, an open-source library for statistical learning. In these developments, we were more specifically implied in *generative models* (*GNB*), *dimension reduction* methods (*univariate feature selection, RFE*), *model selection* schemes, and we were the referent developer for *Bayesian regularizations*.

Research Perspectives

Intra-subject and inter-subject analyzes: a statistical learning point of view

The prediction accuracy of the reference methods, as well as the prediction accuracy of the methods developed in this thesis, are given for intra-subject analysis in Tab. 7.6, and for inter-subject analysis in Tab. 7.7. We can see that the methods perform differently on intra-subject and inter-subject analyzes. This variability can be explained by the difference of spatial layout of neural coding in intra and inter-subject settings. Indeed, this spatial layout can be very sparse and fine-grained at the single subject level, but has a larger spatial extent in inter-subject analysis, due to the lack of voxel-to-voxel correspondence. Thus, sparsity-promoting approaches are likely to perform well in intra-subject analysis, where clusters-promoting approaches are likely to perform well in inter-subject analysis. There is thus a great interest in methods that can adapt their level of sparsity to the data.

Extensions A possible extension is the addition of a ℓ_1 norm regularization to *Total Variation* regularization, thus yielding the following minimization problem:

$$\hat{\mathbf{w}}^l = \underset{\mathbf{w},b}{\operatorname{argmin}} \ \ell(\mathbf{w}) + \lambda_1 \|\mathbf{w}\|^1 + \lambda_2 TV(\mathbf{w}) \ \ , \ \ \lambda_1 \geq 0 \ , \ \lambda_2 \geq 0 \tag{7.34}$$

By optimizing the two parameters λ_1 and λ_2 by *internal cross-validation*, we can adapt the model between sparsity promoting or clusters promoting regularizations. The problem defined in Eq. 7.34 is very similar to *smooth Lasso* [Hebiri 10] that is based on the ℓ_2 norm of the gradient. However, TV regularization is more adapted for extracting clusters, by penalizing instead the ℓ_1 norm of the gradient. Another promising prospect should be to consider within the same framework both intra and inter-subject information, using structured regularizations based on mixed norms (*e.g. group Lasso* [Yuan 06, Bach 08]).

Bayesian versus classical discriminative approaches

The methods presented in this thesis can be roughly classified in two groups: *Bayesian approaches* (*e.g. Bayesian Ridge Regression, Automatic Relevance Determination* or *Multi-Class Sparse Bayesian Regression*) and *Discriminative approaches* (*e.g. Lasso, Elastic net, SVC* or *Total Variation framework*).

In term of computation time, Bayesian approaches are not very efficient compared to discriminative approaches. Although the bayesian framework

Methods	mean ζ	std ζ	max ζ	min ζ	p-val to VB-MCBR
SVR	0.82	0.07	0.9	0.67	0.0003 **
Elastic net	0.9	0.02	0.93	0.85	0.0002 **
BRR	0.92	0.02	0.96	0.88	0.0011 ***
ARD	0.89	0.03	0.95	0.85	0.0003 **
Gibbs-MCBR	0.93	0.01	0.95	0.92	0.0099 **
VB-MCBR	0.94	0.01	0.96	0.92	-
TV $\lambda = 0.05$	0.92	0.02	0.95	0.88	0.0002 **

Table 7.6: *Mental representation of size - **Intra-subject** analysis*. Explained variance ζ for the different methods used in this thesis. The p-values are computed using a paired t-test.

Methods	mean ζ	std ζ	max ζ	min ζ	p-value to TV
SVR	0.77	0.11	0.97	0.58	0.0277 *
Elastic net	0.78	0.1	0.97	0.65	0.0405 *
BRR	0.72	0.1	0.94	0.6	0.0008 **
ARD	0.52	0.33	0.93	-0.28	0.0085 **
Gibbs-MCBR	0.79	0.1	0.97	0.62	0.0289 *
VB-MCBR	0.78	0.1	0.97	0.65	0.0151 *
SC - BRR	0.82	0.08	0.93	0.7	0.5816
TV $\lambda = 0.05$	0.84	0.07	0.97	0.72	-

Table 7.7: *Mental representation of size - **Inter-subject** analysis*. Explained variance ζ for the different methods used in this thesis. The p-values are computed using a paired t-test.

automatically adapts the parameters of the model to the data, it comes at a computational cost, that is often more expensive than an internal cross-validation over the parameters of a discriminative approaches. An exception is the *supervised clustering*, where *Bayesian Ridge Regression* is a well-suited approach for adapting the regularization to the variable complexity of the problem when pruning the tree. In this specific case, the degree of sparsity can vary between the high and low levels of the tree, and *Bayesian Ridge Regression* tunes precisely the regularization to the specific sparsity of each cut of the tree.

In term of prediction accuracy, the two types of approaches performed similarly, with a slight advantage of the bayesian methods in the intra-subject analysis. Indeed, such approaches can more finely tune the regularization to the fine-grained spatial layout of the neural coding specific to the subject. In the inter-subject analysis, discriminative approaches perform slightly better, because, by tuning their parameters by internal cross-validation, they are less prone to overfit a particular training set of subjects.

In conclusion, discriminative approaches only aim at performing an accu-

rate prediction, whereas bayesian methods can be used to construct a more interpretable models that consider different hypothesis on *fMRI* data. The bayesian models currently used are not dedicated to *fMRI* data analysis, and, in that sense, it is interesting to take into account some hypothesis about neural coding in the model, as in *MCBR* where we assume a population coding hypothesis (*i.e.* different groups of voxels are implied in the coding) (see also [Friston 08]). Thus, Bayesian approaches seem promising as they can more easily use priors on *fMRI* data than discriminative approaches.

Extensions *Gaussian processes (GPs)* [Rasmussen 05] models have been successfully used in *fMRI* [Marquand 10], and implement complex priors over the mean and the covariance of the weights. In particular, they can be used to introduce the spatial information in bayesian models, in a similar way as [Friston 08].

Another extension can be to consider the construction of activation maps and the predictive model within a same framework. The *joint detection-estimation* framework developed in [Vincent 07] or the hierarchical model proposed in [Lashkari 10] are both interesting alternative for combining the identification of functional pattern in the brain and using them for prediction. In order to take into account the spatial information in a whole Bayesian framework, *Markov Random Field (MRF)* is also as a promising approach (see [Ou 10] for an example of use of anatomical information).

Spatial information and *feature agglomeration*

In this thesis, we relied on the notion of *feature agglomeration*, and demonstrated that including spatial information, within voxel-based analysis with *Total Variation regularization*, or by creating intermediate structure as *parcels*, yields both accurate and interpretable results for *inverse inference*. Thus, there is a great interest in using *spatial information* in inverse inference, and we believe that it is a promising prospect for statistical learning frameworks in neuroimaging.

We have introduced the *supervised clustering* approach in chapter 5, that yields accurate prediction in inter-subject analysis, and outperforms state-of-the-art approaches. More generally, this method is not restricted to brain images, and might be used in any dataset where multi-scale structure is considered as important (*e.g.* medical or satellite images). Additionally, such clustering is well-suited for constructing an anatomo-functional atlas, as it jointly considers both spatial and functional information.

Extensions Among other research extension, one can develop a approach similar to *Random Forests* [Breiman 01], by aggregating different tree of parcellations created by *bootstrap* on the training set. The resulting weighted parcellations are thus combined by averaging the weights. Preliminary works show some increase in prediction accuracy, but additional work have to be done to preserve the spatial structure of the parcels.

Moreover, one major limitation of the proposed *supervised clustering* algorithm is that it relies on a greedy exploration of the tree, and optimality is not ensured. Thus, one can introduce the hierarchical structure of the tree within convex optimization problem, following for example the work detailed in [Jenatton 10].

Appendices

A

A short introduction to *Magnetic Resonance Imaging*

In this appendix, we briefly explain the physical basis of *Magnetic Resonance Imaging (MRI)*. To keep the presentation as simple and tight as possible, we do no detail the mathematical and physical expressions. We focus our explanations on the hydrogen atom H^1, that is the most studied element in *MRI*, due to its very high abundance in the human body (63% of the atoms [Foster 84]).

A.1 Notions of magnetism

Spin and energy

MRI is based on an intrinsic and fundamental magnetic property of the particles called the *spin* s (positive or negative multiple of $-\frac{1}{2}$, and thus possibly integer), a proton having a *spin* $s = \frac{1}{2}$. A particle with a spin has a *spin magnetic moment* $\vec{\mu}$. This moment has a random orientation in the absence of magnetic field. However, when a population of *spins* is placed in a magnetic field $\vec{B_0}$ (assumed to be aligned with the z axis), few *spins* have a moment $\vec{\mu}$ aligned in the field direction, and many of them precess (*i.e.* rotate) around $\vec{B_0}$. This precession is performed with a frequency ν_0, called *Larmor's frequency*, which is typically in the radio-frequency domain ($\nu_0 = \gamma B_0$, where $\gamma = 42.576$ MHz/T for the proton, is the gyro-magnetic ratio).

A proton in a magnetic field $\vec{B_0}$ has two possible states and associated energy levels:

- a **low-energy** state, in which $\vec{\mu}$ is parallel to $\vec{B_0}$, with: $E = -\vec{\mu}.\vec{B_0}$
- a **high-energy** state, in which $\vec{\mu}$ is anti-parallel to $\vec{B_0}$, with: $E = +\vec{\mu}.\vec{B_0}$

A transition, called **state transition**, is possible between the two states, by emitting or receiving an energy ΔE proportional to ν_0.

Packets of spins and *net magnetization*

In the context of *MRI*, it is more meaningful to represent the spins by packets, each packet being a group of neighboring spins experiencing the same magnetic field. We define the **net magnetization** $\vec{M}$ as the sum of all the individuals magnetic moments of spins in both states. Due to the fact that the protons in a given packet can be parallel or anti-parallel to the magnetic field, some contributions will vanish. Thus, $\vec{M}$ is proportional to the difference between the populations of those two states (given by the Boltzmann's statistics), and to the magnetic field. As for a single proton, $\vec{M}$ precesses around $\vec{B_0}$, and has two components. The first component is a longitudinal component $\vec{M_z}$ that is aligned with the magnetic field and increases with the concentration of protons in the volume. The second component is a transverse component $\vec{M_{xy}}$ that usually vanishes by averaging (see Fig. A.1).

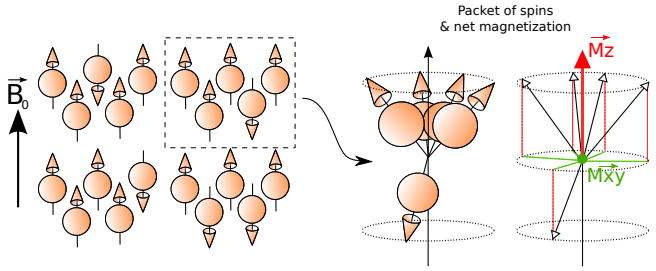


Figure A.1: Illustration of the effect of a magnetic field $\vec{B_0}$ on a population of spins. All the spins are aligned with $\vec{B_0}$, a majority of them having the same direction. One can represent the population of spins at the mesoscopic scale using packets of spins, each of them having a *net magnetization* $\vec{M}$ which can be decomposed in the two components $\vec{M_z}$ and $\vec{M_{xy}}$.

A.2 *Nuclear Magnetic Resonance - NMR*

Nuclear Magnetic Resonance allows to access some information about the composition of the sample by measuring M_z that is proportional to the concentration of protons in the sample. However, as $\vec{M}$ is parallel to the strong magnetic field $\vec{B_0}$, we have to excite the system (i.e., to perturb $\vec{M}$) to access M_z. The main idea is to use an oscillating magnetic field, called a *Radio Frequency – RF* – pulse (noted $\vec{B_1}$), with a frequency at the *Larmor's frequency* of the proton.

RF pulse and Relaxation

The *RF* pulse $\vec{B_1}$ is applied perpendicularly to $\vec{B_0}$ (say along the x axis), and rotates M around the x axis. The rotation angle depends on the duration τ of

the application of the *RF* pulse $\vec{B_1}$. A pulse of $90°$ will rotate $\vec{M}$ by $90°$ around x and a pulse of $180°$ will rotate $\vec{M}$ by $180°$ around x (this is called a *population inversion*: $M_z = -M_{z0}$, where M_{z0} is the initial value of M_z). Moreover, due to the resonance phenomenon at *Larmor's frequency*, the field $\vec{B_1}$ re-phases all the spins together, and thus, creates a non null transverse component M_{xy} (non null average of the magnetic moments μ). When the field $\vec{B_1}$ is stopped, $\vec{M}$ has a given angle (called *flip angle*) with $\vec{B_0}$, and still precesses around the z axis. The net magnetization returns progressively to the equilibrium, *i.e.* its initial state (see Fig.A.2):

- For the *longitudinal component*, this return to equilibrium is due to the readjustment of the spin population that has changed due to the $\vec{B_1}$ field. This realignment with $\vec{B_0}$ has a time constant T_1.
- For the *transverse component*, this return to equilibrium is due to the fact that each spin packet experiences a slightly different magnetic field $\vec{B_0}$ ($\vec{B_0}$ is not homogeneous in the sample), and thus has a different Larmor's frequency. The transverse component vanishes with a time constant T_2. We always have $T_1 \geq T_2$. Moreover, one can define a combined time constant $T_2^* \leq T_2$, that also takes into account the effect of the molecular interactions, and T_2^* is subject to additional losses above the normal T_2 decay.

These time constants are related to the nature of the tissue that has received the *RF* pulse, and thus can be used to characterize the sample.

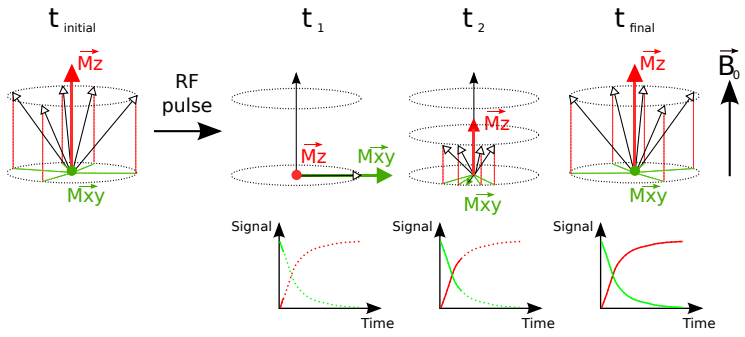


Figure A.2: Illustration of the effect of a *RF pulse* of $90°$. After the application of the pulse, the two components of the *net magnetization*, $\vec{M_{xy}}$ (green) and $\vec{M_z}$ (red), return to equilibrium, each of them with a time constant depending of the tissue surrounding the protons.

The T_1*-weighted* scans provide a good gray matter/white matter contrast. The T_2*-weighted* scans are sensitive to water content. The T_2^**-weighted* scans can increase contrast for certain types of tissue, such as venous blood, and thus, are well-suited for *functional Magnetic Resonance Imaging*.

Free Induction Decay – *FID*

During the relaxation, the transverse component M_{xy} creates an electromagnetic induction in a coil (*a.k.a receiver coil*) that is placed around the sample. This current is a signal that represents the fundamental magnetic resonance data: at a given time t, the voltage in the coil is proportional to the sum of all the transverse components M_{xy} that are precessing in the sample. This current is a sinusoid with an amplitude that decreases exponentially with time, and is called *FID* (*Free Induction Decay*) [Farrar 71]. The Fourier Transform (FT) of the *FID* will give different peaks corresponding to the different frequencies of resonance of the protons in the sample (multiple frequencies due to the inhomogeneities in $\vec{B_0}$). It is thus possible to obtain a signal carrying information about the tissues in the sample, using *NMR*. However, this signal is a unique response that contains the contributions of all the spins within the whole sample (e.g. for the whole brain). A spatial encoding is thus needed to produce localized signals, hence images of the sample, and is called *Magnetic Resonance Imaging – MRI*.

A.3 *Magnetic Resonance Imaging – MRI*

In order to create an image reflecting the spatially varying structure of the sample, we have to separate the different contributions of the different regions of the sample in the *FID*. The main idea is to use magnetic gradients to encode position in the *Larmor's frequency*.

Slice selection

First, we use a gradient of magnetic field $\vec{b_z}$ along the z axis. The magnetic field is not uniform, and each spin packet along this axis will be subject to a different static magnetic field, and thus will have different *Larmor's frequencies*. By using the relationship between frequency and position along the z axis, we can select a *slice* in the volume, as only the slice with a *Larmor's frequency* equal to the *RF* pulse's frequency will be excited by the pulse (see Fig.A.3). The imaging process of the whole sample is thus done slice by slice.

Phase and frequency codings

One the slice is selected during the excitation process, the previous gradient is turned off, and we use a frequency and phase combination to encode a specific region of the previously selected slice (see Fig.A.3).

A gradient of magnetic field is applied along another direction (e.g. x), that make the spins precess around this axis with different frequencies. When this gradient is stopped, all the spins in the given slice experience the same static magnetic field, and thus precess with the same frequency again. But, due to the difference of frequency during the application of the gradient, they have acquired a different phase: this is the *phase encoding*. During the read-out

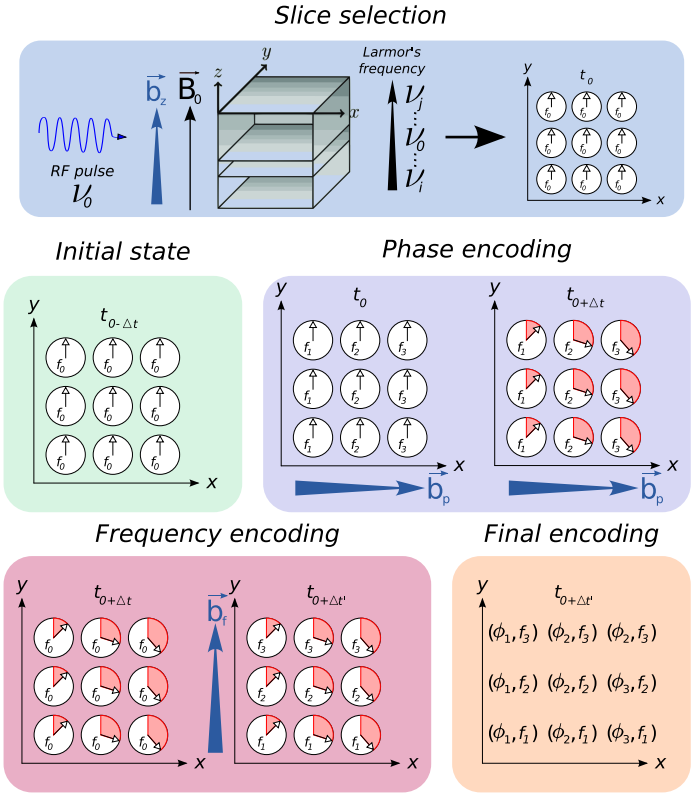


Figure A.3: Illustration of the different steps of *MRI*. The gradient $\vec{b_z}$ along the z axis selects a given slice with an *RF* pulse of a given frequency ν_0. By combining a phase-encoding and a frequency-encoding gradients, each position in the x, y space is encoded by an unique combination of phase and frequency.

(acquisition of the *MRI* signal), we apply a gradient to encode the last axis (e.g. y) in term of frequency. On a given slice, this gradient makes the spins precess with different frequencies along y: this is the *frequency encoding*.

The temporal scheme defining the starting times and durations of each gradient *RF pulse* is called a *sequence*, and the time between the repetitions of the sequence is called the repetition time (TR). Through the use of these gradients, *MRI* data is acquired directly in the *k-space* (*i.e.* the frequency-phase space). We need to have one phase-encoding gradient step for each location in the phase-encoding gradient direction. Thus, if we wish to resolve 256 locations in the phase encoding direction we need 256 different magnitudes of the phase-encoding gradient and record 256 different *FID* signals. One can use a *Fourier transform* on the data (*FID*) to retrieve the contribution of each location in the *FID*. By taking the magnitude and converting them to pixels intensities, an image can be constructed.

B

Description of the data sets

B.1 Details on simulated data sets

The simulated data set $\mathbf{X}$ consists of $n = 100$ images (size $12 \times 12 \times 12$ voxels) with a set of four square Regions of Interest (ROIs) (size $2 \times 2 \times 2$). We call $\mathcal{R}$ the support of the ROIs (*i.e.* the 32 resulting voxels of interest). Each of the four ROIs has a fixed weight in $\{-0.5, 0.5, -0.5, 0.5\}$. We call $w_{i,j,k}$ the weight of the (i, j, k) voxel. The resulting images are smoothed with a Gaussian kernel with a standard deviation of 2 voxels, to mimic the correlation structure observed in real fMRI data. To simulate the spatial variability between images (inter-subject variability, movement artifacts in intra-subject variability), we define a new support of the ROIs, called $\tilde{\mathcal{R}}$ such as, for each image l^{th}, 50% (randomly chosen) of the weights $\mathbf{w}$ are set to zero. Thus, we have $\tilde{\mathcal{R}} \subset \mathcal{R}$. We simulate the target $\mathbf{y}$ for the l^{th} image as:

$$y_l = \sum_{(i,j,k)\in\tilde{\mathcal{R}}} w_{i,j,k} X_{i,j,k,l} + \epsilon_l \tag{B.1}$$

with the signal in the (i, j, k) voxel of the l^{th} image simulated as:

$$X_{i,j,k,l} \sim \mathcal{N}(0, 1) \tag{B.2}$$

and $\epsilon_l \sim \mathcal{N}(0, \gamma)$ is a Gaussian noise with standard deviation $\gamma > 0$. We choose γ in order to have a signal-to-noise ratio of 5 dB.

Competing methods

In our experiments, different methods are compared to state of the art methods. For regression settings:

- *Elastic net* regression. A cross-validation procedure within the training set is used to optimize the parameters. We use $\lambda_1 \in \{0.2\tilde{\lambda}, 0.1\tilde{\lambda}, 0.05\tilde{\lambda}, 0.01\tilde{\lambda}\}$, with $\tilde{\lambda} = \|\mathbf{X}^T\mathbf{y}\|_\infty$, and $\lambda_2 \in \{0.1, 0.5, 1., 10., 100.\}$.
- *SVR* with a linear kernel.The C parameter is optimized by cross-validation in the range 10^{-3} to 10^1 in multiplicative steps of 10.

For classification settings:

- *SMLR* classification. A cross-validation procedure within the training set is used to optimize the parameters. We use $\lambda_1 \in \{0.2\tilde{\lambda}, 0.1\tilde{\lambda}, 0.05\tilde{\lambda}, 0.01\tilde{\lambda}\}$, where $\tilde{\lambda} = \|\mathbf{X}^T\mathbf{y}\|_\infty$, and $\lambda_2 \in \{0.1, 0.5, 1., 10., 100.\}$.
- *SVC* with a linear kernel. The C parameter is optimized by cross-validation in the range 10^{-3} to 10^1 in multiplicative steps of 10.

B.2 Data set on mental representations of size and shape of objects

This real fMRI dataset is related to an experiment studying the representation of objects. It has been acquired by Evelyn Eger (Neurospin/Unicog), and more details can be found in [Eger 08].

Data description

Participants and data acquisition

Ten healthy volunteers (1 left-handed) with normal or corrected vision (3 men and 9 women; mean age, 25.4 ± 3.2 yr) gave written informed consent. Functional images were acquired at the Brain Imaging Center of Frankfurt University, Frankfurt, Germany, on a 3-T MR system with eight-channel head coil (Siemens Trio, Erlangen, Germany) as T2*-weighted echo-planar image (EPI). Twenty transverse slices were obtained with a repetition time of 2 s (echo time, 30 ms; flip angle, $70°$; $2 \times 2 \times 2$-mm voxels; 0.5-mm gap).

Stimuli and design

Subjects were presented visual stimuli representing objects. Object stimuli (2 categories, 4 chairs and 4 teapots) were the same as the ones described in [Eger 08]. For each object, three sizes were created with size 2 corresponding along each axis to 150% of size 1 and size 3 to 150% of size 2 (or size 2 to 225% of the area of size 1 and size 3 to 225% of the area of size 2), yielding a total of 12 experimental conditions (4 exemplars $\times$ 3 sizes).

Experimental protocol and task

Stimuli were back-projected onto a translucent screen above the subjects' head and viewed via a mirror on the head coil. Pictures subtended ~ 3.3, ~ 5, and $\sim 7.5°$ of visual angle for the three sizes. Objects were presented in short blocks of four identical (in exemplar and size) pictures each (1-s stimulus, 0.5-s blank), followed by a fixation baseline of 4 s, with pseudo-randomized order of conditions. Each stimulus randomly appeared in a red or green hue, and participants reported the color of each stimulus via keypad. This task was

performed in six experimental sessions of 8.2-min length each, encompassing 24 blocks altogether per experimental condition.

An additional scanning session of ~ 5-min length mapped object responsive areas for each participant using a standard LOC localizer, comparing pictures of various common objects to mosaic-scrambled versions of the same images (20×20 fragments). Objects and scrambled images were alternated in blocks with 500 ms per picture every 1 s and block length of 12 s (6-s fixation baseline).

Image processing and data analysis

The initial analysis of the imaging data used SPM5 (http://www.fil.ion.ucl.ac.uk/spm/software/spm5). After motion correction and normalization to an EPI-template in MNI space, the unsmoothed EPI images were entered into a general linear model, modeling separately the effect of each of the 12 conditions convolved with a standard hemodynamic response function, while accounting for serial autocorrelation with an AR(1) model and removing low-frequency drift terms by a high-pass filter with a cut-off of 128 s. This analysis yielded six independent estimates of fMRI signal change (corresponding to the 6 sessions), which were subsequently used for pattern recognition analysis.

Experiments

We used the resulting session-wise parameter estimate images for different analysis, and all the analysis are performed on the whole volume.

Regression experiments

First, we perform an intra-subject regression analysis. The four different shapes of objects (for the two categories) were pooled across for each one of the three sizes, and we are interested in finding discriminative information between sizes. This reduces to a regression problem, in which our goal is to predict a simple scalar factor (size of an object) (see Fig. B.1). Each subject is evaluated independently, in a 12-fold cross-validation. The dimensions of the real data set for one subject are $p \sim 7 \times 10^4$ and $n = 72$ (divided in 3 different sizes, 24 images per size). We evaluate the performance of the method by cross-validation (leave-one-condition-out, *i.e.*, leave-6-images-out). The parameters of the reference methods are optimized with a nested leave-one-condition-out cross-validation within the training set, in the ranges given before.

Additionally, we perform an inter-subject regression analysis on the sizes. The inter-subject analysis relies on subject-specific fixed-effects activations, *i.e.* for each condition, the 6 activation maps corresponding to the 6 sessions are averaged together. This yields a total of 12 images per subject, one for each experimental condition. The dimensions of the real data set are $p \sim 7 \times 10^4$ and $n = 120$ (divided in 3 different sizes). We evaluate the performance of the method by cross-validation (leave-one-subject-out). The parameters of the

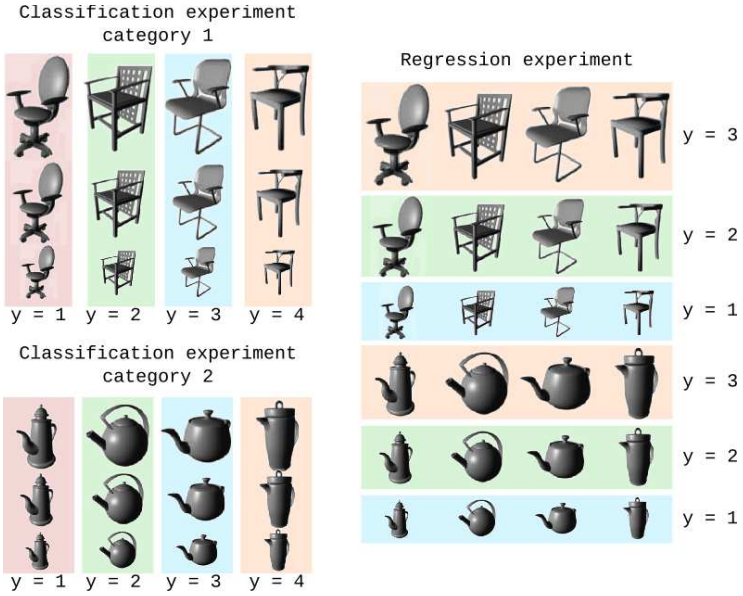


Figure B.1: Experiment paradigm for the classification of object in each of the category (left) and regression (right) experiments. Each color represents the stimuli which are pooled together in one of the three experiments (classification category 1, classification category 2 and regression).

reference methods are optimized with a nested leave-one-subject-out cross-validation within the training set, in the ranges given before.

Classification experiments

We evaluate the performance on a second type of discrimination task which is object classification (see Fig. B.1). In that case, we collapse the conditions across the three sizes and are interested in discriminating between individual object exemplars/shapes. For each of the two categories, this can be handled as a classification problem, where we aim at predicting the shape of an object corresponding to a new fMRI scan. The inter-subject analysis is performed in the same way as described for the regression study, except that now, we perform two analyzes corresponding to the two categories used, each one including 5 subjects.

There is also a difference in the validation procedure: in the intra-subject analysis, in order to have balanced classes, we evaluate the performance of the method using a leave-one-session-out cross-validation, which boils down to a 6-fold cross-validation. The parameters of the reference methods are optimized with a leave-one-session-out cross-validation within the training set, in the ranges given before.

Statistical Parametric Maps

For comparison purposes, the corresponding maps of *Anova* (*F-score*), or *SPMs*, for the inter-subject analysis are given Fig. B.2, for the discrimination of sizes (top) and discrimination of objects for the two categories (middle and bottom). As expected, the sizes are most significantly discriminated in primary visual cortex, while for objects, discrimination at lower significance levels is observed in additional lateral occipital regions (*LOC*) [Eger 08].

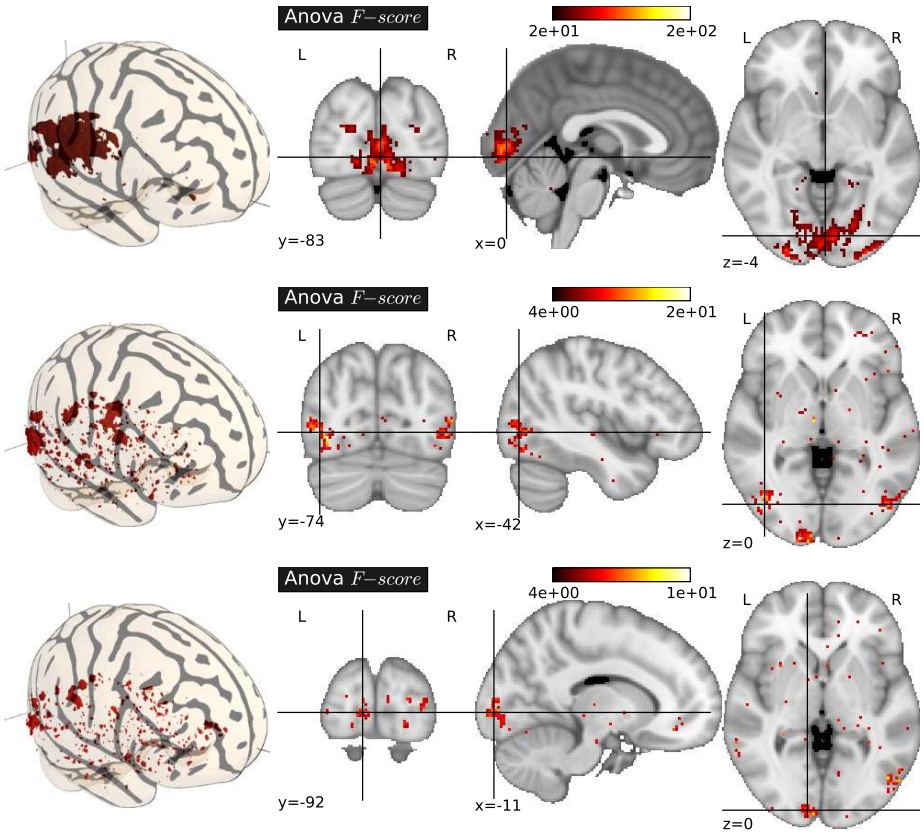


Figure B.2: *Mental representations of size and shape - Inter-subject analysis* - Maps of *Anova* (*F-score*) for the sizes prediction experiment (up) and the objects identifications for category 1 (middle) and category 2 (bottom).

Competing methods

In our experiments, different methods are compared to state of the art methods. For regression settings:

- *Elastic net* regression. A cross-validation procedure within the training

set is used to optimize the parameters. We use $\lambda_1 \in \{0.2\tilde{\lambda}, 0.1\tilde{\lambda}, 0.05\tilde{\lambda}, 0.01\tilde{\lambda}\}$, with $\tilde{\lambda} = \|\mathbf{X}^T\mathbf{y}\|_\infty$, and $\lambda_2 \in \{0.1, 0.5, 1., 10., 100.\}$.

- *SVR* with a linear kernel.The C parameter is optimized by cross-validation in the range 10^{-3} to 10^1 in multiplicative steps of 10.

For classification settings:

- *SMLR* classification. A cross-validation procedure within the training set is used to optimize the parameters. We use $\lambda_1 \in \{0.2\tilde{\lambda}, 0.1\tilde{\lambda}, 0.05\tilde{\lambda}, 0.01\tilde{\lambda}\}$, where $\tilde{\lambda} = \|\mathbf{X}^T\mathbf{y}\|_\infty$, and $\lambda_2 \in \{0.1, 0.5, 1., 10., 100.\}$.
- *SVC* with a linear kernel. The C parameter is optimized by cross-validation in the range 10^{-3} to 10^1 in multiplicative steps of 10.

All these methods are used after an *Anova*-based feature selection, as this maximizes their performance. This selection is performed on the training set of each fold in an internal cross-validation loop, and the optimal number of voxels is selected within the range $\{50, 100, 250, 500\}$.

The implementation of *Elastic net* is based on *coordinate descent* [Friedman 10], while *SVR* and *SVC* are based on LibSVM [Chang 01]. Methods are used from *Python* via the *Scikit-learn* open source package [sci 10].

C

Scikit-learn for *fMRI inverse inference*

In this appendix, we detail another aspect of our contribution, which are developments in the *Scikit-learn* . The *Scikit-learn* (http://scikit-learn.sourceforge.net/) is a Python module integrating classic machine learning algorithms in a well-designed API, as it yields algorithm-like formulations. It is open-source and aims at providing simple and efficient solutions to learning problems. It can be easily used for *fMRI inverse inference* and we give here the principle functions that can be used in such case. The description of the functions follows the organization of the chapter 3.

C.1 Global framework

Evaluation of the decoding

Explained variance can be computed with the *Scikit-learn* , using the Listing. C.1.

Listing C.1: Evaluation by explained variance

```
>>> from scikits.learn.metrics import explained_variance
>>> ### y_test is the true target, y_pred is the predicted target
>>> score = explained_variance(y_t, y_pred)
```

Classification score can be computed using the Listing. C.2.

Listing C.2: Evaluation by classification score

```
>>> from scikits.learn.metrics import zero_one
>>> ### y_test is the true target, y_pred is the predicted target
>>> classif_rate = 100 * zero_one(y_t, y_pred) / y_t.shape[0]
```

Model selection and validation

Leave-one-out cross-validation can be done with the *Scikit-learn* , using the Listing. C.3.

Listing C.3: Leave-one-out cross-validation

```
>>> from scikits.learn import cross_val
>>> loo = cross_val.LeaveOneOut(X.shape[0])
>>> for train_index, test_index in loo:
>>>     X_train, y_train = X[train], y[train]
>>>     X_test, y_test = X[test], y[test]
```

K-fold cross-validation can be done with the *Scikit-learn* , using the Listing. C.4.

Listing C.4: 4-fold cross-validation

```
>>> from scikits.learn import cross_val
>>> kfold = cross_val.KFold(X.shape[0], k=4)
>>> for train_index, test_index in kfold:
>>>     X_train, y_train = X[train], y[train]
>>>     X_test, y_test = X[test], y[test]
```

Leave-one-subject-out cross-validation can be done with the *Scikit-learn* , using the Listing. C.5.

Listing C.5: Leave-one-subject-out cross-validation

```
>>> from scikits.learn import cross_val
>>> ### All the subjects are concatenated in a single set (X,y)
>>> loso = cross_val.LeaveOneLabelOut(subjects)
>>> for train_index, test_index in loso:
>>>     X_train, y_train = X[train], y[train]
>>>     X_test, y_test = X[test], y[test]
```

C.2 The historical approaches

Support Vector Classification – SVC

Prediction using *SVC* can be done with the *Scikit-learn* , using the Listing. C.6.

Listing C.6: Classification using *Support Vector Machine*

```
from scikits.learn.svm import SVC
clf = SVC(kernel='linear',C=1)
clf.fit(Xl, yl)
ypred = clf.predict(Xt)
```

Support Vector Regression – SVR

Prediction using *SVC* can be done with the *Scikit-learn* , using the Listing. C.6.

Listing C.7: Regression using *Support Vector Machine*

```
from scikits.learn.svm import SVR
clf = SVR(kernel='linear',C=1)
clf.fit(Xl, yl)
ypred = clf.predict(Xt)
```

Generative models

Prediction using *Gaussian Naive Bayes* can be done with the *Scikit-learn* , using the Listing. C.8.

Listing C.8: Prediction using *Gaussian Naive Bayes*

```
>>> from scikits.learn.naive_bayes import GNB
>>> clf = GNB()
>>> clf.fit(X_train, y_train)
>>> y_pred = clf.predict(X_test)
```

Prediction using *LDA* can be done with the *Scikit-learn* , using the Listing. C.9.

Listing C.9: Prediction using *Linear Discriminant Analysis*

```
>>> from scikits.learn.lda import LDA
>>> clf = LDA(priors=None,use_svd=True)
>>> clf.fit(X_train, y_train)
>>> y_pred = clf.predict(X_test)
```

Prediction using *QDA* can be done with the *Scikit-learn* , using the Listing. C.10.

Listing C.10: Prediction using *Quadratic Discriminant Analysis*

```
>>> from scikits.learn.qda import QDA
>>> clf = QDA(priors=None)
>>> clf.fit(X_train, y_train)
>>> y_pred = clf.predict(X_test)
```

C.3 Regularization

Ridge Regression - ℓ_2 regularization

Prediction using *Ridge Regression* can be done with the *Scikit-learn* , using the Listing. C.11.

Listing C.11: Prediction using *Ridge Regression*

```
>>> from scikits.learn import glm
>>> ridge = glm.Ridge(alpha=1.0)
>>> ridge.fit(X_train, y_train)
>>> pred = ridge.predict(X_test)
```

Lasso - ℓ_1 Regularization

Prediction using *Lasso*, based on coordinate descent [Friedman 07], can be done with the *Scikit-learn* , using the Listing. C.12.

Listing C.12: Prediction using *Lasso*

```
>>> from scikits.learn import glm
>>> lasso = glm.Lasso(alpha=0.1)
>>> lasso.fit(X_train, y_train)
>>> pred = lasso.predict(X_test)
```

Elastic net - $\ell_1 + \ell_2$ Regularization

Prediction using *Elastic net*, based on coordinate descent [Friedman 07], can be done with the *Scikit-learn* , using the Listing. C.13.

Listing C.13: Prediction using *Elastic net*

```
>>> from scikits.learn import glm
>>> clf = glm.ElasticNet(alpha=0.1,rho=0.5)
>>> clf.fit(X_train, y_train)
>>> pred = clf.predict(X_test)
```

C.4 Bayesian regularization

Bayesian Ridge Regression – BRR

Prediction using *Bayesian Ridge Regression* can be done with *Scikit-learn* , using the Listing. C.14.

Listing C.14: Prediction using *Bayesian Ridge Regression*

```
>>> from scikits.learn.glm import BayesianRidge
>>> clf = BayesianRidge()
>>> clf.fit(X_train, y_train)
>>> y_pred = clf.predict(X_test)
```

Automatic Relevance Determination – ARD

Prediction using *Automatic Relevance Determination* can be done with the *Scikit-learn* , using the Listing. C.15.

Listing C.15: Prediction using *Automatic Relevance Determination*

```
>>> from scikits.learn.glm import ARDRegression
>>> clf = ARDRegression()
>>> clf.fit(X_train, y_train)
>>> y_pred = clf.predict(X_test)
```

C.5 Dimension reduction

Univariate feature selection

Based on *Scikit-learn* , we give in Listing. C.16 an example of *univariate feature selection* for classification settings.

Listing C.16: Univariate feature selection for classification - 500 features

```
>>> from scikits.learn.feature_selection import SelectKBest, f_classif
>>> univariate_filter = SelectKBest(f_classif, k=500)
>>> ###  X_r are the reduced data.
>>> X_r = univariate_filter.fit(X_train, y_train).transform(X_train)
>>> selected_voxels = univariate_filter.get_support()
```

We give in Listing. C.17 an example of *univariate feature selection* for regression settings.

Listing C.17: Univariate feature selection for regression - 500 features

```
>>> from scikits.learn.feature_selection import SelectKBest, f_regression
>>> univariate_filter = SelectKBest(f_regression, k=500)
>>> ###  X_r are the reduced data.
>>> X_r = univariate_filter.fit(X_train, y_train).transform(X_train)
>>> selected_voxels = univariate_filter.get_support()
```

Recursive feature elimination – RFE

We give an example of *SVM-RFE* with model selection by cross-validation in the Listing. C.18, based on *Scikit-learn* .

Listing C.18: SVM - RFE with model selection by 4-folds cross-validation

```
>>> from scikits.learn.svm import SVC
>>> from scikits.learn.feature_selection.rfe import RFECV
>>> from scikits.learn.cross_val import KFold
>>> from scikits.learn.metrics import zero_one
>>> svc = SVC(kernel='linear', C=1)
>>> rfecv = RFECV(estimator=svc, n_features=5,
                  percentage=0.05, loss_func=zero_one)
>>> rfecv.fit(X_train, y_train, cv=KFold(yl.shape[0], 4))
```

Bibliography

Publications of the author

Articles in peer-reviewed journal

• V. Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin and B. Thirion. *A supervised clustering approach for fMRI-based inference of brain states*. Submitted to Pattern Recognition - Special Issue on 'Brain Decoding'. 2010
• V. Michel, A. Gramfort, G. Varoquaux, E. Eger and B. Thirion. *Total variation regularization for fMRI-based prediction of behaviour*. Submitted to IEEE Transactions on Medical Imaging. 2010.
• M. Lebreton, S. Jorge, V. Michel, B. Thirion and M. Pessiglione. *An automatic valuation system in the human brain : evidence from functional neuroimaging*. Neuron 64, 3, 2009.
• E. Eger, V. Michel, B. Thirion, A. Amadon, S. Dehaene and A. Kleinschmidt. *Deciphering Cortical Number Coding from Human Brain Activity Pattern*. Current Biology. 2009, 19:1608.
• A. Knops, B. Thirion, E.M. Hubbard, V. Michel and S. Dehaene. *Recruitment of an area involved in eye movements during mental arithmetic*. Science. 2009 Jun 19;324(5934):1583-5.
• A. Bachrach, A. Gramfort, V. Michel, E. Cauvet, B. Thirion and C. Pallier. *Decoding of syntactic trees*. In prep.

Peer-reviewed conference with proceedings

• V. Michel, C. Damon, and B. Thirion. *Mutual information-based feature selection enhances fMRI brain activity classification*. In 5th Proc. IEEE ISBI, pages 592-595, 2008.
• V. Michel, E. Eger, C. Keribin and B. Thirion. *Adaptive multi-class bayesian sparse regression - an application to brain activity classification*. In MICCAI'09 Workshop on Analysis of Functional Medical Images, 2009.
• V. Michel, E. Eger, C. Keribin and B. Thirion. *Multi-Class Sparse Bayesian Regression for Neuroimaging data analysis*. In International Workshop on Machine Learning in Medical Imaging (MLMI) In conjunction with MICCAI 2010.
• V. Michel, E. Eger, C. Keribin, J.-B. Poline and B. Thirion. *A supervised clustering approach for extracting predictive information from brain activation images*. In IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA10) - IEEE Conference on Computer Vision and Pattern Recognition. 2010.
• V. Michel, A. Gramfort, G. Varoquaux and B. Thirion. *Total Variation regularization enhances regression-based brain activity prediction*. In 1st ICPR Workshop on Brain Decoding - Pattern recognition challenges in neuroimaging - 20th International Conference on Pattern Recognition. 2010.
• R. Genuer, V. Michel, E. Eger, and B. Thirion. *Random forests based feature selection for decoding* fMRI *data*. In COMPSTAT 19th International Conference on Computational Statistics, pages 372 , 2010.

Other conference

• V. Michel, C. Damon, and B. Thirion. *Mutual information-based feature selection enhances fMRI brain activity classification*. In Organization for Human Brain Mapping, 2008.
• V. Michel, E. Eger, C. Keribin, J.-B. Poline and B. Thirion. *A supervised clustering approach for extracting predictive information from brain activation images*. In Organization for Human Brain Mapping, 2010.
• E. Eger, V. Michel, B. Thirion, A. Amadon, S. Dehaene and A. Kleinschmidt. *Information on individual numerosity in fMRI patterns of human parietal cortex*. 38th annual meeting of the Society for Neuroscience, 2008.

Bibliography

[Aizerman 64] A. Aizerman, E. M. Braverman & L. I. Rozoner. *Theoretical foundations of the potential function method in pattern recognition learning*. Automation and Remote Control, vol. 25, pages 821–837, 1964.

[Ashburner 99] J. Ashburner & K.J. Friston. *Spatial Normalization*. In A.W. Toga, editeur, Brain Warping, pages 27–44. Academic Press, 1999.

[Bach 08] Francis R. Bach. *Consistency of the Group Lasso and Multiple Kernel Learning*. J. Mach. Learn. Res., vol. 9, pages 1179–1225, 2008.

[Bandettini 92] Peter A. Bandettini, Eric C. Wong, R. Scott Hinks, Ronald S. Tikofsky & James S. Hyde. *Time course EPI of human brain function during task activation*. Magnetic Resonance in Medicine, vol. 25, no. 2, pages 390–397, 1992.

[Barlow 72] H. B. Barlow. *Single units and sensation: a neuron doctrine for perceptual psychology?* Perception, vol. 1, no. 4, pages 371–394, 1972.

[Beck 09a] A. Beck & M. Teboulle. *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*. SIAM Journal on Imaging Sciences, vol. 2, no. 1, pages 183–202, 2009.

[Beck 09b] Amir Beck & Marc Teboulle. *Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems*. Trans. Img. Proc., vol. 18, no. 11, pages 2419–2434, 2009.

[Bishop 07] Christopher M. Bishop. Pattern recognition and machine learning (information science and statistics). Springer, 1st ed. 2006. corr. 2nd printing edition, 2007.

[Blokland 08] Gabriëlla A. M. Blokland, Katie L. McMahon, Jan Hoffman, Zhu Gu, Matthew Meredith, Nicholas G. Martin, Paul M. Thompson, Greig I. De Zubicaray & Margaret J. Wright. *Quantifying the heritability of task-related brain activation and performance during the N-back working memory task : A twin fMRI study*. Biological psychology, vol. 79, no. 1, pages 70–79, 2008.

[Bottou 94] L. Bottou, C. Cortes, J.S. Denker, H. Drucker, I. Guyon, L.D. Jackel, Y. LeCun, U.A. Muller, E. Sackinger,

P. Simard & V. Vapnik. *Comparison of classifier methods: a case study in handwritten digit recognition*. In Proc. of the International Conference on Pattern Recognition, volume II, pages 77–82, 1994.

[Boyd 04] Stephen Boyd & Lieven Vandenberghe. Convex optimization. Cambridge University Press, 2004.

[Brammer 04] Michael Brammer. *Brain scam ?* Nature Neuroscience, no. 7, page 1015, 2004.

[Breiman 95] Leo Breiman. *Better subset regression using the nonnegative garrote*. Technometrics, vol. 37, no. 4, pages 373–384, 1995.

[Breiman 01] Leo Breiman. *Random Forests*. Machine Learning, vol. 45, no. 1, pages 5–32, 2001.

[Buxton 98] R. B. Buxton, E. C. Wong & Frank. L. R. *Dynamics of blood flow and oxygenation changes during brain activation: the balloon model*. Magnetic Resonance in Medicine, vol. 39, pages 855–864, 1998.

[Carlson 03] Thomas A. Carlson, Paul Schrater & Sheng He. *Patterns of Activity in the Categorical Representations of Objects*. Journal of Cognitive Neuroscience, vol. 15, no. 5, pages 704–717, 2003.

[Carroll 09] Melissa K. Carroll, Guillermo A. Cecchi, Irina Rish, Rahul Garg & A. Ravishankar Rao. *Prediction and interpretation of distributed neural activity with sparse models*. NeuroImage, vol. 44, no. 1, pages 112 – 122, 2009.

[Chambolle 04] Antonin Chambolle. *An Algorithm for Total Variation Minimization and Applications*. J. Math. Imaging Vis., vol. 20, no. 1-2, pages 89–97, 2004.

[Chang 01] Chih-Chung Chang & Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[Chen 10] Y. Chen, A. Wiesel, Y. C. Eldar & A. O. Hero. *Shrinkage Algorithms for MMSE Covariance Estimation*. Signal Processing, IEEE Transactions on, vol. 58, no. 10, pages 5016 –5029, 2010.

[Chib 01] S. Chib & I. Jeliazkov. *Marginal Likelihood From the Metropolis-Hastings Output*. Journal of the American Statistical Association, vol. 96, pages 270–281, 2001.

[Childress 99] A.R. Childress, P.D. Mozley, W. McElgin, J. Fitzgerald, M. Reivich & C.P. O'Brien. *Limbic activation during cue-induced cocaine craving.* Am J Psychiatry, vol. 156, no. 1, pages 11–8, 1999.

[Chklovskii 04] Dmitri B. Chklovskii & Alexei A. Koulakov. *MAPS IN THE BRAIN: What Can We Learn from Them?* Annual Review of Neuroscience, vol. 27, no. 1, pages 369–392, 2004.

[Chu 10] Carlton Chu, Yizhao Ni, Geoffrey Tan, Craig J. Saunders & John Ashburner. *Kernel regression for fMRI pattern prediction*. NeuroImage, vol. In Press, Corrected Proof, pages –, 2010.

[Ciuciu 03] P. Ciuciu, J.-B. Poline, G. Marrelec, J. Idier, Ch. Pallier & H. Benali. *Unsupervised robust non-parametric estimation of the hemodynamic response function for any fMRI experiment*. IEEE Trans. Med. Imag., vol. 22, no. 10, pages 1235–1251, 2003.

[Cohen 00] Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff & François Michel. *The visual word form area*. Brain, vol. 123, no. 2, pages 291–307, 2000.

[Combettes 05] P. L. Combettes & V. R. Wajs. *Signal recovery by proximal forward-backward splitting*. Multiscale Modeling and Simulation, vol. 4, no. 4, pages 1168–1200, 2005.

[Cooper 75] R. Cooper, D. Papakostopoulos & HJ. Crow. Rapid changes of cortical oxygen associated with motor and cognitive function in man. Harper M, Bennett B, Miller D, Rowan J. Edinburgh, London, New York, 1975.

[Cordes 02] D. Cordes, V. M. Haughtou, J. D. Carew, K. Arfanakis & K. Maravilla. *Hierarchical Clustering to Measure Connectivity in fMRI Resting-State Data*. Magnetic resonance imaging, vol. 20, no. 4, pages 305–317, 2002.

[Cortes 95] Corinna Cortes & Vladimir Vapnik. *Support-Vector Networks*. Machine Learning, vol. 20, no. 3, pages 273–297, 1995.

[Cox 03] D. D. Cox & R. L. Savoy. *Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex.* Neuroimage, vol. 19, pages 261–270, 2003.

[Daubechies 04] I. Daubechies, M. Defrise & C. De Mol. *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*. Commun. Pure Appl. Math., vol. 57, no. 11, pages 1413 – 1457, 2004.

[Davatzikos 05] C. Davatzikos, K. Ruparel, Y. Fan, D.G. Shen, M. Acharyya, J.W. Loughead, R.C. Gur & D.D. Langleben. *Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection*. NeuroImage, vol. 28, no. 3, pages 663 – 668, 2005.

[Dayan 01] Peter Dayan & L. F. Abbott. Theoretical neuroscience: Computational and mathematical modeling of neural systems. The MIT Press, 2001.

[De Mol 09] Christine De Mol, Ernesto De Vito & Lorenzo Rosasco. *Elastic-net regularization in learning theory*. J. Complex., vol. 25, no. 2, pages 201–230, 2009.

[Dehaene 98] Stanislas Dehaene, Gurvan Le Clec'H, Laurent Cohen, Jean-Baptiste Poline, Pierre-Francois van de Moortele & Denis Le Bihan. *Inferring behavior from functional brain images*. Nature Neuroscience, vol. 1, page 549, 1998.

[Demirci 08] Oguz Demirci, Vincent P. Clark & Vince D. Calhoun. *A projection pursuit algorithm to classify individuals using fMRI data: Application to schizophrenia*. NeuroImage, vol. 39, no. 4, pages 1774 – 1782, 2008.

[Desimone 91] Robert Desimone. *Face-Selective Cells in the Temporal Cortex of Monkeys*. Journal of Cognitive Neuroscience, vol. 3, no. 1, pages 1–8, 1991.

[Downing 01] Paul E. Downing, Yuhong Jiang, Miles Shuman & Nancy Kanwisher. *A Cortical Area Selective for Visual Processing of the Human Body*. Science, vol. 293, no. 5539, pages 2470–2473, 2001.

[Downing 06] P. E. Downing, Chan, M. V. Peelen, C. M. Dodds & N. Kanwisher. *Domain Specificity in Visual Cortex*. Cereb. Cortex, vol. 16, no. 10, pages 1453–1461, October 2006.

[Eger 08] E. Eger, C. Kell & A. Kleinschmidt. *Graded size sensitivity of object exemplar evoked activity patterns in human LOC subregions*. J. Neurophysiol., vol. 100(4):2038-47, 2008.

[Eger 09] Evelyn Eger, Vincent Michel, Bertrand Thirion, Alexis Amadon, Stanislas Dehaene & Andreas Kleinschmidt. *Deciphering Cortical Number Coding from Human Brain*

Activity Patterns. Current biology, vol. 19, no. 19, pages 1608–1615, 2009.

[Ettlinger 68] G. Ettlinger, E. Iwai, M. Mishkin & H. E. Rosvold. *Visual discrimination in the monkey following serial ablation of inferotemporal and preoccipital cortex.* Journal of Comparative and Physiological Psychology, vol. 65, no. 1, pages 110–117, 1968.

[Evans 93] A.C. Evans, D.L. Collins, S.R. Mills, E.D. Brown, R.L. Kelly & T.M. Peters. *3D statistical neuroanatomical models from 305 MRI volumes*. In Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record., volume 3, pages 1813 –1817, 1993.

[Fan 06] Yong Fan, Dinggang Shen & Christos Davatzikos. *Detecting Cognitive States from fMRI Images by Machine Learning and Multivariate Classification*. In CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop, page 89, 2006.

[Farah 04] Martha J. Farah. *Neuroethics: the practical and the philosophical*. Trends in Cognitive Sciences, vol. 9, no. 1, pages 34 – 40, 2004.

[Farrar 71] Thomas C. Farrar & Edwin D. Becker. Pulse and fourier transform nmr; introduction to theory and methods. Academic Press, 1971.

[Farwell 01] LA Farwell & SS Smith. *Using brain MERMER testing to detect knowledge despite efforts to conceal*. J Forensic Sci., no. 46, pages 135–43., 2001.

[Filzmoser 99] Peter Filzmoser, Richard Baumgartner & Ewald Moser. *A hierarchical clustering method for analyzing functional MR images*. Magnetic Resonance Imaging, vol. 17, no. 6, pages 817 – 826, 1999.

[Flandin 02] G. Flandin, F. Kherif, X. Pennec, G. Malandain, N. Ayache & J.-B. Poline. *Improved Detection Sensitivity in Functional MRI Data using a Brain Parcelling Technique*. In Medical Image Computing and Computer-Assisted Intervention (MICCAI'02), volume 2488 of *LNCS*, pages 467–474, 2002.

[Flandin 04] Guillaume Flandin. *Utilisation d'informations géométriques pour l'analyse statistique des données d'IRM fonctionnelle*. PhD thesis, Université de Nice-Sophia Antipolis, 2004.

[Ford 03] James Ford, Hany Farid, Fillia Makedon, Laura A. Flashman, Thomas W. McAllister, Vasilis Megalooikonomou & Andrew J. Saykin. *Patient Classification of fMRI Activation Maps*. In in Proc. of the 6th Annual International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'03, pages 58–65, 2003.

[Foster 84] M.A. Foster. Magnetic resonance in medicine and biology. Pergamon Press, New York, 1984.

[Fox 86] P. T. Fox, M. A. Mintun, M. E. Raichle, F. M. Miezin, J. M. Allman & D. C. Van Essen. *Mapping human visual cortex with positron emission tomography.* Nature, vol. 323, pages 806 – 809, 1986.

[Fox 91] PT. Fox. *Physiological ROI definition by image subtraction.* J Cereb Blood Flow Metab, vol. 11, no. 2, pages 79–82, 1991.

[Frank 93] Ildiko E. Frank & Jerome H. Friedman. *A Statistical View of Some Chemometrics Regression Tools*. Technometrics, vol. 35, no. 2, pages 109–135, 1993.

[Friedman 96] Jerome H. Friedman. *Another approach to polychotomous classification*. Rapport technique, Department of Statistics, Stanford University, 1996.

[Friedman 07] Jerome Friedman, Trevor Hastie, Holger Höfling & Robert Tibshirani. *Pathwise coordinate optimization*. Annals of Applied Statistics, no. 2, 2007.

[Friedman 10] Jerome Friedman, Trevor Hastie & Rob Tibshirani. *Regularization paths for generalized linear models via coordinate descent*. Journal of Statistical Software, vol. 33, no. 1, 2010.

[Friston 95] K.J. Friston, A.P. Holmes, K.J. Worsley, J.B. Poline, C. Frith & R.S.J. Frackowiak. *Statistical Parametric Maps in Functional Imaging: A General Linear Approach*. Human Brain Mapping, vol. 2, pages 189–210, 1995.

[Friston 96] K.J. Friston, J.B. Poline, A.P. Holmes, C. Frith & R.S.J. Frackowiak. *A multivariate analysis of PET activation studies*. Human Brain Mapping, vol. 4, pages 140–151, 1996.

[Friston 08] Karl Friston, Carlton Chu, Janaina Mourão-Miranda, Oliver Hulme, Geraint Rees, Will Penny & John Ashburner. *Bayesian decoding of brain images*. NeuroImage, vol. 39, no. 1, pages 181 – 205, 2008.

[Frostig 90] R. D. Frostig, E. E. Lieke, D. Y. Ts'o & A. Grinvald. *Cortical functional architecture and local coupling between neuronal activity and the microcirculation revealed by in vivo high-resolution optical imaging of intrinsic signals.* Proceedings of the National Academy of Sciences of the United States of America, vol. 87, no. 16, pages 6082–6086, August 1990.

[Ganesh 08] G. Ganesh, E. Burdet, M. Haruno & M. Kawato. *Sparse linear regression for reconstructing muscle activity from human cortical fMRI*. NeuroImage, vol. 42, no. 4, pages 1463 – 1472, 2008.

[Geman 87] Stuart Geman & Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. Morgan Kaufmann Publishers Inc., 1987.

[Genuer 10] Robin Genuer, Vincent Michel, Evelyn Eger & Bertrand Thirion. *Random forests based feature selection for decoding fmri data.* In COMPSTAT 19th International Conference on Computational Statistics, page 372, 2010.

[George 93] Edward I. George & Robert E. McCulloch. *Variable Selection Via Gibbs Sampling*. Journal of the American Statistical Association, vol. 88, no. 423, pages 881–889, 1993.

[Georgopoulos 86] AP Georgopoulos, AB Schwartz & RE Kettner. *Neuronal population coding of movement direction*. Science, vol. 233, no. 4771, pages 1416–1419, 1986.

[Ghebreab 08] Sennay Ghebreab, Arnold Smeulders & Pieter Adriaans. *Predicting Brain States from fMRI Data: Incremental Functional Principal Component Regression*. In Advances in Neural Information Processing Systems, pages 537–544. MIT Press, 2008.

[Glover 99] Gary H. Glover. *Deconvolution of Impulse Response in Event-Related BOLD fMRI*. NeuroImage, vol. 9, no. 4, pages 416 – 429, 1999.

[Golland 03] Polina Golland & Bruce Fischl. *Permutation Tests for Classification: Towards Statistical Significance in Image-Based Studies*. In Information Processing in Medical Imaging, volume 2732 of *Lecture Notes in Computer Science*, pages 330–341. Springer Berlin / Heidelberg, 2003.

[Golland 07] P. Golland, Y. Golland & R. Malach. *Detection of Spatial Activation Patterns as Unsupervised Segmentation of fMRI Data*. In Med Image Comput Comput Assist Interv. MICCAI 2007, pages 110–118, 2007.

[Gramfort 09a] A Gramfort. *Mapping, timing and tracking cortical activations with MEG and EEG: Methods and application to human vision*. PhD thesis, Telecom ParisTech, 2009.

[Gramfort 09b] Alexandre Gramfort & Mathieu Kowalski. *Improving M/EEG source localization with an inter-condition sparse prior*. In IEEE International Symposium on Biomedical Imaging, 2009.

[Grazia 08] Maria Grazia, Di Bono & Marco Zorzi. *Decoding Cognitive States from fMRI Data Using Support Vector Regression*. PsychNology Journal, vol. 6, no. 2, pages 189–201, 2008.

[Gross 92] Charles G. Gross & S. De Schonen. *Representation of Visual Stimuli in Inferior Temporal Cortex*. Philosophical Transactions: Biological Sciences, vol. 335, no. 1273, pages 3–10, 1992.

[Gross 02] Charles G. Gross. *Genealogy of the Grandmother Cell*. Neuroscientist, vol. 8, no. 5, pages 512–518, 2002.

[Gunn 98] S.R. Gunn. *Support Vector Machines for Classification and Regression*. Rapport technique, Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, 1998.

[Guyon 02] Isabelle Guyon, Jason Weston, Stephen Barnhill & Vladimir Vapnik. *Gene Selection for Cancer Classification using Support Vector Machines*. Machine Learning, vol. 46, no. 1-3, pages 389–422, 2002.

[Hanson 04] Stephen J. Hanson, Toshihiko Matsuka & James V. Haxby. *Combinatorial codes in ventral tempoal lobe for object recognition: Haxby (2001) revisited: is there a face area?* NeuroImage, vol. 23, no. 1, pages 156–166, 2004.

[Hanson 08] Stephen José Hanson & Yaroslav O. Halchenko. *Brain Reading Using Full Brain Support Vector Machines for Object Recognition: There Is No Face Identification Area*. Neural Computation, vol. 20, no. 2, pages 486–503, 2008.

[Harris 80] Charles S. Harris. Visual coding and adaptability. L. Erlbaum Associates, Hillsdale, N.J., 1980.

[Hastie 03] T. Hastie, R. Tibshirani & J. H. Friedman. The elements of statistical learning. Springer, 2003.

[Haxby 01] James V. Haxby, Ida M. Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten & Pietro Pietrini. *Distributed and overlapping representations of faces and objects in ventral temporal cortex*. Science, vol. 293, no. 5539, pages 2425–2430, 2001.

[Hayasaka 04] S. Hayasaka, K.L. Phan, I. Liberzon, K.J. Worsley & T.E. Nichols. *Non-Stationary Cluster Size Inference with Random Field and Permutation Methods*. NeuroImage, vol. 22, pages 676–687, 2004.

[Haynes 05a] John-Dylan Haynes & Geraint Rees. *Predicting the orientation of invisible stimuli from activity in human primary visual cortex*. Nature Neuroscience, vol. 8, no. 5, pages 686–691, 2005.

[Haynes 05b] John-Dylan Haynes & Geraint Rees. *Predicting the Stream of Consciousness from Activity in Human Visual Cortex*. Current Biology, vol. 15, no. 14, pages 1301 – 1307, 2005.

[Haynes 06] John-Dylan Haynes & Geraint Rees. *Decoding mental states from brain activity in humans*. Nature Reviews Neuroscience, vol. 7, no. 7, pages 523–534, 2006.

[He 08] Lili He & Ian R. Greenshields. *An MRF spatial fuzzy clustering method for fMRI SPMs*. Biomedical Signal Processing and Control, vol. 3, no. 4, pages 327 – 333, 2008.

[Hebiri 10] Mohamed Hebiri & Sara A. Van De Geer. The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. http://hal.archives-ouvertes.fr/hal-00462882/PDF/SLasso.pdf, 2010.

[Hoerl 70] Arthur E. Hoerl & Robert W. Kennard. *Ridge Regression: Biased Estimation for Nonorthogonal Problems*. Technometrics, vol. 12, no. 1, pages 55–67, 1970.

[Horwitz 00] B. Horwitz, K. J. Friston & J. G. Taylor. *Neural modeling and functional brain imaging: an overview*. Neural Networks, vol. 13, no. 8-9, pages 829 – 846, 2000.

[Hsu 02] Chih-Wei Hsu & Chih-Jen Lin. *A comparison of methods for multiclass support vector machines*. IEEE Transactions on Neural Networks, vol. 13, no. 2, pages 415–425, 2002.

[Hubel 62] D. H. Hubel & T. N. Wiesel. *Receptive fields, binocular interaction and functional architecture in the cat's visual cortex*. J Physiol, vol. 160, pages 106–154, 1962.

[Hughes 68] G. Hughes. *On the mean accuracy of statistical pattern recognizers*. Information Theory, IEEE Transactions on, vol. 14, no. 1, pages 55–63, 1968.

[Hung 05] Chou P. Hung, Gabriel Kreiman, Tomaso Poggio & James J. DiCarlo. *Fast Readout of Object Identity from Macaque Inferior Temporal Cortex*. Science, vol. 310, no. 5749, pages 863–866, 2005.

[Hutchinson 09] Rebecca A. Hutchinson, Radu Stefan Niculescu, Timothy A. Keller, Indrayana Rustandi & Tom M. Mitchell. *Modeling fMRI data generated by overlapping cognitive processes with unknown onsets using Hidden Process Models*. NeuroImage, vol. 46, no. 1, pages 87 – 104, 2009.

[Iadecola 97] Costantino Iadecola, Guang Yang, Timothy J. Ebner & Gang Chen. *Local and Propagated Vascular Responses Evoked by Focal Synaptic Activity in Cerebellar Cortex*. J Neurophysiol, vol. 78, no. 2, pages 651–659, 1997.

[Jenatton 10] R. Jenatton, J. Mairal, G. Obozinski & F. Bach. *Proximal Methods for Sparse Hierarchical Dictionary Learning*. In International Conference on Machine Learning (ICML), 2010.

[Ji 04] Ye Ji, Hong-Bo Liu, Xiu-Kun Wang & Yi-Yuan Tang. *Cognitive states classification from fMRI data using support vector machines*. In Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on, volume 5, pages 2919 – 2923, 2004.

[Jiang 04] Fang Jiang, Alice J. O'Toole, Hervé Abdi & James V. Haxby. *Partially distributed representations of objects and faces in ventral temporal cortex: evidence from the structure of the object categories and neural response patterns*. Journal of Vision, vol. 4, no. 8, page 903, 2004.

[Johnson 67] S. C. Johnson. *Hierarchical Clustering Schemes*. Psychometrika, vol. 2, pages 241–254, 1967.

[Kamitani 05] Yukiyasu Kamitani & Frank Tong. *Decoding the visual and subjective contents of the human brain*. Nature Neuroscience, vol. 8, no. 5, pages 679–685, 2005.

[Kay 08] Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger & Jack L. Gallant. *Identifying natural images from human brain activity*. Nature, vol. 452, pages 352–355, 2008.

[Keller 09] M. Keller, M. Lavielle, M. Perrot & A. Roche. *Anatomically Informed Bayesian Model Selection for fMRI Group Data Analysis*. In 12th MICCAI, 2009.

[Kim 93] SG Kim, J Ashe, K Hendrich, JM Ellermann, H Merkle, K Ugurbil & AP Georgopoulos. *Functional magnetic resonance imaging of motor cortex: hemispheric asymmetry and handedness*. Science, vol. 261, no. 5121, pages 615–617, 1993.

[Kirchhoff 06] Brenda A. Kirchhoff & Randy L. Buckner. *Functional-Anatomic Correlates of Individual Differences in Memory*. Neuron, vol. 51, no. 2, pages 263 – 274, 2006.

[Kjems 02] U. Kjems, L. K. Hansen, J. Anderson, S. Frutiger, S. Muley, J. Sidtis, D. Rottenberg & S. C. Strother. *The Quantitative Evaluation of Functional Neuroimaging Experiments: Mutual Information Learning Curves*. NeuroImage, vol. 15, no. 4, pages 772 – 786, 2002.

[Klein 09] A. Klein, J. Andersson, B.A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G.E. Christensen, D.L. Collins, J. Gee, P. Hellier, J.H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thomson, T. Vercauteren, R.P. Woods, J.J. Mann & R. Parsey. *Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration*. NeuroImage, vol. 46, no. 3, pages 786–802, 2009.

[Knerr 90] Stefan Knerr, Léon Personnaz & Gérard Dreyfus. *Single-Layer Learning Revisited: A Stepwise Procedure for Building and Training a Neural Network*. In F. Fogelman Soulié & J. Hérault, editeurs, Neurocomputing: Algorithms, Architectures and Applications, volume F68 of *NATO ASI Series*, pages 41–50. Springer-Verlag, 1990.

[Knops 09] André Knops, Bertrand Thirion, Edward Hubbard, Vincent Michel & Stanislas Dehaene. *Mathematics as cortical recycling : Recruitment of an area involved in eye movements during mental arithmetic*. Science, vol. 324, pages 1583 – 1585, 2009.

[Kohavi 97] Ron Kohavi & George H. John. *Wrappers for feature subset selection*. Artif. Intell., vol. 97, no. 1-2, pages 273–324, 1997.

[Koltchinskii 04] Vladimir Koltchinskii, Manel Martinez-ramon & Stefan Posse. *Optimal Aggregation of Classifiers and Boosting Maps in Functional Magnetic Resonance Imaging*. NIPS, 2004.

[Konorski 67] J. Konorski. *Integrative activity of the brain : An interdisciplinary approach*. JAMA, vol. 203, no. 5, pages 371–, 1967.

[Kontos 04] D. Kontos, V. Megalooikonomou, D. Pokrajac, A. Lazarevic, Z. Obradovic, O. B. Boyko, J. Ford, F. Makedon & A. J. Saykin. *Extraction of Discriminative Functional MRI Activation Patterns and an Application to Alzheimer's Disease*. In Med Image Comput Comput Assist Interv. MICCAI 2004, pages 727–735, 2004.

[Koutsouleris 09] Nikolaos Koutsouleris, Eva M. Meisenzahl, Christos Davatzikos, Ronald Bottlender, Thomas Frodl, Johanna Scheuerecker, Gisela Schmitt, Thomas Zetzsche, Petra Decker, Maximilian Reiser, Hans-Jurgen Moller & Christian Gaser. *Use of Neuroanatomical Pattern Classification to Identify Subjects in At-Risk Mental States of Psychosis and Predict Disease Transition*. Arch Gen Psychiatry, vol. 66, no. 7, pages 700–712, 2009.

[Kreiman 00] G. Kreiman & C Koch. *Category-specific visual responses of single neurons in the human medial temporal lobe*. Nature Neuroscience, no. 3, pages 946–953, 2000.

[Kriegeskorte 06] Nikolaus Kriegeskorte, Rainer Goebel & Peter Bandettini. *Information-based functional brain mapping*. Proceedings of the National Academy of Sciences of the United States of America, vol. 103, no. 10, pages 3863–3868, March 2006.

[Kriegeskorte 09] Nikolaus Kriegeskorte, W. Kyle Simmons, Patrick S. F. Bellgowan & Chris I. Baker. *Circular analysis in systems neuroscience: the dangers of double dipping*. Nature Neuroscience, vol. 12, no. 5, pages 535–540, 2009.

[Krishnapuram 05] Balaji Krishnapuram, Lawrence Carin, Mario A.T. Figueiredo & Alexander J. Hartemink. *Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, pages 957–968, 2005.

[Kwong 92] K K Kwong, J W Belliveau, D A Chesler, I E Goldberg, R M Weisskoff, B P Poncelet, D N Kennedy, B E Hoppel, M S Cohen & R Turner. *Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation*. Proceedings of the National Academy of Sciences of the United States of America, vol. 89, no. 12, pages 5675–5679, 1992.

[LaConte 03] Stephen LaConte, Jon Anderson, Suraj Muley, James Ashe, Sally Frutiger, Kelly Rehm, Lars Kai Hansen, Essa Yacoub, Xiaoping Hu, David Rottenberg & Stephen Strother. *The Evaluation of Preprocessing Choices in Single-Subject BOLD fMRI Using NPAIRS Performance Metrics*. NeuroImage, vol. 18, no. 1, pages 10 – 27, 2003.

[LaConte 05] Stephen LaConte, Stephen Strother, Vladimir Cherkassky, Jon Anderson & Xiaoping Hu. *Support vector machines for temporal classification of block design fMRI data*. NeuroImage, vol. 26, no. 2, pages 317 – 329, 2005.

[Langlebe 05] DD Langlebe, JW Loughead, WB Bilker, K Ruparel, AR Childress, SI Busch & Gur RC. *Telling truth from lie in individual subjects with fast event-related fMRI*. Hum Brain Mapp., vol. 4, pages 262–72, 2005.

[Langs 10] Georg Langs, Bjoern H. Menze, Danial Lashkari & Polina Golland. *Detecting stable distributed patterns of brain activation using Gini contrast*. NeuroImage, vol. In Press, Corrected Proof, pages –, 2010.

[Larsen 99] J. Larsen & C. Goutte. *On optimal data split for generalization estimation and model selection*. In Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop, pages 225 – 234, 1999.

[Lashkari 10] D. Lashkari, R. Sridharan, E. Vul, Po-Jang Hsieh, N. Kanwisher & P. Golland. *Nonparametric hierarchical Bayesian model for functional brain parcellation*. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pages 15 –22, 2010.

[Ledoit 03] Olivier Ledoit & Michael N. Wolf. *Honey, I Shrunk the Sample Covariance Matrix*. In UPF Economics and Business Working Paper, 2003.

[Leng 06] Chenlei Leng, Yi Lin & Grace Wahba. *A note on the Lasso and related procedures in model selection*. Statistica Sinica, vol. 16, no. 4, pages 1273–1284, 2006.

[Logothetis 01] N.K. Logothetis, J. Pauls, M.A. Augath, T. Trinath & A. Oeltermann. *Neurophysiological Investigation of the Basis of the fMRI signal.* Nature, vol. 412, pages 150–157, 2001.

[Luo 09] June Luo. *The discovery of mean square error consistency of a ridge estimator*. Statistics & Probability Letters, 2009.

[MacKay 92] David J. C. MacKay. *Bayesian interpolation*. Neural Comput., vol. 4, no. 3, pages 415–447, 1992.

[Mandeville 98] J B Mandeville, J J A Marota, B E Kosofsky, J R Keltner, R Weissleder, B R Rosen & R M Weisskoff. *Dynamic functional imaging of relative cerebral blood volume during rat forepaw stimulation*. Magnetic Resonance in Medicine, vol. 39, no. 4, pages 615–624, 1998.

[Marquand 10] Andre Marquand, Matthew Howard, Michael Brammer, Carlton Chu, Steven Coen & Janaina Mourao-Miranda. *Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes*. NeuroImage, vol. 49, no. 3, pages 2178 – 2189, 2010.

[Martinez-Ramon 06] Manel Martinez-Ramon, Vladimir Koltchinskii, Gregory L. Heileman & Stefan Posse. *fMRI pattern classification using neuroanatomically constrained boosting*. NeuroImage, vol. 31, no. 3, pages 1129 – 1141, 2006.

[Martino 08] Federico De Martino, Giancarlo Valente, Noël Staeren, John Ashburner, Rainer Goebel & Elia Formisano. *Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns*. NeuroImage, vol. 43, no. 1, pages 44 – 58, 2008.

[McClure 04] Samuel M. McClure, Jian Li, Damon Tomlin, Kim S. Cypert, Latané M. Montague & P.Read Montague. *Neural Correlates of Behavioral Preference for Culturally Familiar Drinks*. Neuron, vol. 44, no. 2, pages 379 – 387, 2004.

[Michel 08] Vincent Michel, Cécilia Damon & Bertrand Thirion. *Mutual information-based feature selection enhances fMRI brain activity classification*. In 2008 5th IEEE international symposium on biomedical imaging: From nano to macro 2008 5th IEEE international symposium on biomedical imaging: From nano to macro, pages 592 – 595, 2008.

[Michel 10] Vincent Michel, Evelyn Eger, Christine Keribin, Jean-Baptiste Poline & Bertrand Thirion. *A supervised clustering approach for extracting predictive information from brain activation images.* MMBIA'10, 2010.

[Mitchell 03] Tom Mitchell, Rebecca Hutchinson, Marcel Adam Just, Radu S. Niculescu, Francisco Pereira & Xuerui Wang.

Classifying Instantaneous Cognitive States from fMRI Data. In In Proceedings of the 2003 Americal Medical Informatics Association Annual Symposium. Washington D.C, page 469, 2003.

[Mitchell 04] Tom M. Mitchell, Rebecca Hutchinson, Radu S. Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just & Sharlene Newman. *Learning to Decode Cognitive States from Brain Images*. Machine Learning, vol. V57, no. 1, pages 145–175, 2004.

[Mitchell 08] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason & Marcel Adam Just. *Predicting Human Brain Activity Associated with the Meanings of Nouns*. Science, vol. 320, no. 5880, pages 1191–1195, 2008.

[Miyawaki 08] Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C. Tanabe, Norihiro Sadato & Yukiyasu Kamitani. *Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders*. Neuron, vol. 60, pages 915–929, 2008.

[Mørch 97] Niels Mørch, Lars Kai Hansen, Stephen C. Strother, Claus Svarer, David A. Rottenberg, Benny Lautrup, Robert Savoy & Olaf B. Paulson. *Nonlinear versus Linear Models in Functional Neuroimaging: Learning Curves and Generalization Crossover*. In IPMI '97: Proceedings of the 15th International Conference on Information Processing in Medical Imaging, pages 259–270, 1997.

[Moreau 65] J.J. Moreau. *Proximité et dualité dans un espace hilbertien*. Bull. Soc. Math. France., vol. 93, pages 273–299, 1965.

[Mourao-Miranda 05] Janaina Mourao-Miranda, Arun L.W. Bokde, Christine Born, Harald Hampel & Martin Stetter. *Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data*. NeuroImage, vol. 28, no. 4, pages 980 – 995, 2005.

[Mourao-Miranda 07] Janaina Mourao-Miranda, Karl J. Friston & Michael Brammer. *Dynamic discrimination analysis: A spatial-temporal SVM*. NeuroImage, vol. 36, no. 1, pages 88 – 99, 2007.

[Neal 96] Radford M. Neal. Bayesian learning for neural networks (lecture notes in statistics). Springer, 1 edition, 1996.

[Nesterov 07] Y Nesterov. *Gradient methods for minimizing composite objective function*. Core discussion papers, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Sep 2007.

[Ng 02] Andrew Y. Ng & Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, volume 2. MIT, 2002.

[Ni 08] Yizhao Ni, C. Chu, C. J. Saunders & J. Ashburner. *Kernel methods for fMRI pattern prediction*. Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, pages 692–697, 2008.

[Nichols 03] Thomas Nichols & Satoru Hayasaka. *Controlling the familywise error rate in functional neuroimaging: a comparative review*. Statistical Methods in Medical Research, vol. 12, no. 5, pages 419–446, 2003.

[Nikolova 00] Mila Nikolova. *Local Strong Homogeneity of a Regularized Estimator*. SIAM Journal on Applied Mathematics, vol. 61, no. 2, pages 633–658, 2000.

[Norman 06] K. A. Norman, S. M. Polyn, G. J. Detre & J. V. Haxby. *Beyond mind-reading: multi-voxel pattern analysis of fMRI data.* Trends Cogn Sci, vol. 10, no. 9, pages 424–430, 2006.

[O'Craven 00] K. M. O'Craven & N. Kanwisher. *Mental Imagery of Faces and Places Activates Corresponding Stimulus-Specific Brain Regions*. J. Cogn. Neurosci., vol. 12, no. 6, pages 1013–1023, 2000.

[Ogawa 90a] S. Ogawa, T. M. Lee, A. R. Kay & 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, vol. 87, no. 24, pages 9868–9872, 1990.

[Ogawa 90b] Seiji Ogawa, Tso-Ming Lee, Asha S. Nayak & Paul Glynn. *Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields*. Magnetic Resonance in Medicine, vol. 14, no. 1, pages 68–78, 1990.

[Ogawa 92] S. Ogawa, D. W. Tank, R. Menon, J. M. Ellermann, S. G. Kim, H. Merkle & K. Ugurbil. *Intrinsic signal changes accompanying sensory stimulation: functional brain mapping*

with magnetic resonance imaging. Proc Natl Acad Sci U S A, vol. 89, no. 13, pages 5951–5955, July 1992.

[Oliver 95] Jonathan J. Oliver & David J. Hand. *On Pruning and Averaging Decision Trees*. In ICML, pages 430–437, 1995.

[Onut 04] Iosif-Viorel Onut & Ali A. Ghorbani. *Classifying cognitive states from fMRI data using neural networks*. In 2004 International Joint Conference on Neural Networks, page 3302, 2004.

[Op de Beeck 05] Hans P. Op de Beeck, Chris I. Baker, Sandra Rindler & Nancy Kanwiser. *An increased bold response for trained objects in object-selective regions of human visual cortex*. Journal of Vision, vol. 5, no. 8, 2005.

[Ortega 00] James M. Ortega & Werner C. Rheinboldt. Iterative solution of nonlinear equations in several variables. Society for Industrial and Applied Mathematics, 2000.

[Osborne 99] Michael R. Osborne, Brett Presnell & Berwin A. Turlach. *On the LASSO and Its Dual*. Journal of Computational and Graphical Statistics, vol. 9, pages 319–337, 1999.

[Osuna 97] Edgar Osuna, Robert Freund & Federico Girosi. *An Improved Training Algorithm for Support Vector Machines*. In Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop, pages 276–285, 1997.

[O'Toole 07] Alice J. O'Toole, Fang Jiang, Hervé Abdi, Nils Pénard, Joseph P. Dunlop & Marc A. Parent. *Theoretical, Statistical, and Practical Perspectives on Pattern-based Classification Approaches to the Analysis of Functional Neuroimaging Data*. Journal of Cognitive Neuroscience, vol. 19, no. 11, pages 1735–1752, 2007.

[Ou 10] Wanmei Ou, William M. Wells III & Polina Golland. *Combining spatial priors and anatomical information for fMRI detection*. Medical Image Analysis, vol. 14, no. 3, pages 318 – 331, 2010.

[Palatucci 07] Mark Palatucci & Tom Mitchell. *Classification in Very High Dimensional Problems with Handfuls of Examples*. In Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD). Springer-Verlag, September 2007.

[Pauling 36] Linus Pauling & Charles D. Coryell. *The Magnetic Properties and Structure of Hemoglobin, Oxyhemoglobin and Carbonmonoxyhemoglobin*. Proceedings of the National Academy of Sciences of the United States of America, vol. 22, no. 4, pages 210–216, 1936.

[Pereira 09] Francisco Pereira, Tom Mitchell & Matthew Botvinick. *Machine learning classifiers and fMRI: A tutorial overview*. NeuroImage, vol. 45, no. 1, Supplement 1, pages S199 – S209, 2009.

[Perez-Orive 02] Javier Perez-Orive, Ofer Mazor, Glenn C. Turner, Stijn Cassenaer, Rachel I. Wilson & Gilles Laurent. *Oscillations and Sparsening of Odor Representations in the Mushroom Body*. Science, vol. 297, no. 5580, pages 359–365, 2002.

[Perkel 68] DH. Perkel & TH. Bullock. *Neural coding*. Neurosciences Research Program Bulletin, vol. 221-348, 1968.

[Phothisonothai 08] Montri Phothisonothai & Masahiro Nakagawa. *EEG-Based Classification of Motor Imagery Tasks Using Fractal Dimension and Neural Network for Brain-Computer Interface*. IEICE - Trans. Inf. Syst., vol. E91-D, no. 1, pages 44–53, 2008.

[Platt 99] John C. Platt. Fast training of support vector machines using sequential minimal optimization. MIT Press, 1999.

[Poeppel 08] T.D. Poeppel & B.J. Krause. *Functional imaging of memory processes in humans: Positron emission tomography and functional magnetic resonance imaging*. Methods, vol. 44, no. 4, pages 315 – 328, 2008.

[Polyn 05] Sean M. Polyn, Vaidehi S. Natu, Jonathan D. Cohen & Kenneth A. Norman. *Category-Specific Cortical Activity Precedes Retrieval During Memory Search*. Science, vol. 310, no. 5756, pages 1963–1966, 2005.

[Pouget 00] Alexandre Pouget, Teter Dyan & Richard Zemel. *Information Processing with population codes*. Nature Reviews Neuroscience, vol. 1, pages 125–132, 2000.

[Qi 04] Yuan Qi, Thomas P. Minka, Rosalind W. Picard & Zoubin Ghahramani. *Predictive automatic relevance determination by expectation propagation*. In ICML '04: Proceedings of the twenty-first international conference on Machine learning. ACM Press, 2004.

[Quiroga 05] R. Quian Quiroga, L. Reddy, G Kreiman, C. Koch & I Fried. *Invariant visual representation by single neurons in the human brain*. Nature, vol. 435, pages 1102–1107, 2005.

[Raine 98] Adrian Raine, J. Reid Meloy, Susan Bihrle, Jackie Stoddard, Lori Lacasse & Monte S. Buchsbaum. *Reduced prefrontal and increased subcortical brain functioning assessed using positron emission tomography in predatory and affective murderers*. Behavioral Sciences & the Law, vol. 16, no. 3, pages 319–332, 1998.

[Rasmussen 05] Carl Edward Rasmussen & Christopher K. I. Williams. Gaussian processes for machine learning (adaptive computation and machine learning). The MIT Press, 2005.

[Reddy 06] Leila Reddy & Nancy Kanwisher. *Coding of visual objects in the ventral stream*. Current Opinion in Neurobiology, vol. 16, no. 4, pages 408–414, August 2006.

[Rissman 10] Jesse Rissman, Henry T. Greely & Anthony D. Wagner. *Detecting individual memories through the neural decoding of memory states and past experience*. Proceedings of the National Academy of Sciences, vol. 107, no. 21, pages 9849–9854, 2010.

[Rudin 92] L Rudin, S Osher & E Fatemi. *Nonlinear total variation based noise removal algorithms*. Physica D, Jan 1992.

[Rustandi 06] Indrayana Rustandi. *Hierarchical gaussian naive bayes classifier for multiple-subject fmri data*. In In NIPS Workshop: New Directions on Decoding Mental States from fMRI Data, 2006.

[Ryali 10] Srikanth Ryali, Kaustubh Supekar, Daniel A. Abrams & Vinod Menon. *Sparse logistic regression for whole-brain classification of fMRI data*. NeuroImage, vol. 51, no. 2, pages 752 – 764, 2010.

[Sato 09] Joao Ricardo Sato, André Fujita, Carlos Eduardo Thomaz, Maria da Graca Morais Martin, Janaina Mourao-Miranda, Michael John Brammer & Edson Amaro Junior. *Evaluating SVM and MLDA in the extraction of discriminant regions for mental state prediction*. NeuroImage, vol. 46, no. 1, pages 105 – 114, 2009.

[Sayres 06] Rory Sayres, David Ress & Kalanit Grill-Spector. *Identifying Distributed Object Representations in Human Extrastriate Visual Cortex*. Advances in Neural Information Processing Systems 18, pages 1169–1176, 2006.

[Schmah 08] Tanya Schmah, Geoffrey E. Hinton, Richard S. Zemel, Steven L. Small & Stephen C. Strother. *Generative versus discriminative training of RBMs for classification of fMRI images*. In NIPS, pages 1409–1416, 2008.

[Schölkopf 00] Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson & Peter L. Bartlett. *New Support Vector Algorithms*. Neural Comput., vol. 12, no. 5, pages 1207–1245, 2000.

[sci 10] *scikit-learn*. http://scikit-learn.sourceforge.net/, downloaded in Apr. 2010. version 0.2.

[Shawe-Taylor 04] John Shawe-Taylor & Nello Cristianini. Kernel methods for pattern analysis. Cambridge University Press, 2004.

[Shinkareva 06] Svetlana V. Shinkareva, Hernando C. Ombao, Bradley P. Sutton, Aprajita Mohanty & Gregory A. Miller. *Classification of functional brain images with a spatio-temporal dissimilarity map*. NeuroImage, vol. 33, no. 1, pages 63 – 71, 2006.

[Shinkareva 08] S. V. Shinkareva, R. A. Mason, V. L. Malave, W. Wang, T. M. Mitchell & M. A. Just. *Using FMRI brain activation to identify cognitive States associated with perception of tools and dwellings.* PLoS ONE, vol. 3, no. 1, 2008.

[Sidtis 03] John J Sidtis, Stephen C Strother & David A Rottenberg. *Predicting performance from functional imaging data: methods matter.* Neuroimage, vol. 20, no. 2, pages 615–24, 2003.

[Signe 09] Bray Signe, Chang Catie & Hoeft Fumiko. *Applications of multivariate pattern classification analyses in developmental neuroimaging of healthy and clinical populations*. Frontiers in Human Neuroscience, vol. 3, 2009.

[Smola 98] A. J Smola & B Schölkopf. *On a Kernel–based Method for Pattern Recognition, Regression, Approximation and Operator Inversion*. Algorithmica, vol. 22, pages 211–231, 1998. Technical Report 1064, GMD FIRST, April 1997.

[Smola 04] Alex J. Smola & Bernhard Schölkopf. *A Tutorial on Support Vector Regression*. Statistics and Computing, vol. 14, no. 3, pages 199–222, 2004.

[Spiers 07] Hugo J. Spiers & Eleanor A. Maguire. *Decoding human brain activity during real-world experiences*. Trends in Cognitive Sciences, vol. 11, no. 8, pages 356 – 365, 2007.

[Strother 02] Stephen C. Strother, Jon Anderson, Lars Kai Hansen, Ulrik Kjems, Rafal Kustra, John Sidtis, Sally Frutiger, Suraj Muley, Stephen LaConte & David Rottenberg. *The Quantitative Evaluation of Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework*. NeuroImage, vol. 15, no. 4, pages 747 – 771, 2002.

[Strother 04] Stephen Strother, Stephen La Conte, Lars Kai Hansen, Jon Anderson, Jin Zhang, Sujit Pulapura & David Rottenberg. *Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis*. NeuroImage, vol. 23, no. Supplement 1, pages S196 – S207, 2004.

[Tahmasebi 10] Amir M. Tahmasebi. *Quantification of Inter-subject Variability in Human Brain and Its Impact on Analysis of fMRI Data*. PhD thesis, Queen's University, 2010.

[Tesla 33] Nikola Tesla. *Tremendous New Power Soon to be Unleashed*. Kansas City Journal-Post, 1933.

[Thirion 06a] B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J. B. Poline, D. Lebihan & S. Dehaene. *Inverse retinotopy: inferring the visual content of images from brain activation patterns*. Neuroimage, vol. 33, no. 4, pages 1104–16, 2006.

[Thirion 06b] B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu & J.-B. Poline. *Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets*. Hum. Brain Mapp., vol. 27, no. 8, pages 678–693, 2006.

[Thulborn 82] Keith R. Thulborn, John C. Waterton, Paul M. Matthews & 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, vol. 714, no. 2, pages 265–270, February 1982.

[Thyreau 06] B. Thyreau, B. Thirion, G. Flandin & J.-B. Poline. *Anatomo-functional description of the brain: a probabilistic approach*. In Proc. 31th Proc. IEEE ICASSP, volume V, pages 1109–1112, 2006.

[Tibshirani 96] Robert Tibshirani. *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, pages 267–288, 1996.

[Tipping 00] M. Tipping. The relevance vector machine. Morgan Kaufmann, 2000.

[Tipping 01] Michael E. Tipping. *Sparse Bayesian Learning and the Relevance Vector Machine*. Journal of Machine Learning Research, vol. 1, pages 211–244, 2001.

[Treisman 96] Anne Treisman. *The binding problem*. Current Opinion in Neurobiology, vol. 6, no. 2, pages 171 – 178, 1996.

[Trevor 02] Bradley Efron Trevor, Trevor Hastie, Lain Johnstone & Robert Tibshirani. *Least Angle Regression*. Annals of Statistics, vol. 32, pages 407–499, 2002.

[Tsao 06] Doris Y. Tsao, Winrich A. Freiwald, Roger B. H. Tootell & Margaret S. Livingstone. *A Cortical Region Consisting Entirely of Face-Selective Cells*. Science, vol. 311, no. 5761, pages 670–674, 2006.

[Tseng 09] Paul Tseng. *Approximation Accuracy, Gradient Methods, and Error Bound for Structured Convex Optimization*. submitted to Math Program. B, 2009.

[Tucholka 08] A. Tucholka, B. Thirion, M. Perrot, P. Pinel, J.-F. Mangin & J.-B. Poline. *Probabilistic anatomo-functional parcellation of the cortex: how many regions?* In 11thProc. MICCAI, LNCS Springer Verlag, 2008.

[Tucholka 10] Alan Tucholka. *Prise en compte de l'anatomie cérébrale individuelle dans les études d'IRM fonctionnelle*. PhD thesis, Université Paris-Sud, 2010.

[Turner 91] Robert Turner, Denis Le Bihan, Chrit T. W. Moonen, Daryl Despres & Joseph Frank. *Echo-planar time course MRI of cat brain oxygenation changes*. J Neurophysiol, vol. 22, no. 1, pages 159–166, 1991.

[Tzourio-Mazoyer 02] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer & M. Joliot. *Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain*. NeuroImage, vol. 15, no. 1, pages 273–289, 2002.

[Ugurbil 03] Kâmil Ugurbil, Louis Toth & Dae-Shik Kim. *How accurate is magnetic resonance imaging of brain function?* Trends in Neurosciences, vol. 26, no. 2, pages 108 – 114, 2003.

[Vallabhaneni 04] A. Vallabhaneni & B. He. *Motor imagery task classification for brain computer interface applications using spatiotemporal principle component analysis*. Neurological Research, vol. 26, 2004.

[van Gerven 10] Marcel A. J. van Gerven, Floris P. de Lange & Tom Heskes. *Neural Decoding with Hierarchical Generative Models*. Neural Computation, vol. 0, no. 0, pages 1–16, 2010.

[Vapnik 82] Vladimir Vapnik. Estimation of dependences based on empirical data: Springer series in statistics (springer series in statistics). Springer-Verlag New York, Inc., 1982.

[Varoquaux 10] Gaël Varoquaux, Alexandre Gramfort, Jean Baptiste Poline & Bertrand Thirion. *Brain covariance selection: better individual functional connectivity models using population prior*. In Advances in Neural Information Processing Systems Advances in Neural Information Processing Systems, 2010.

[Vincent 07] T. Vincent, P. Ciuciu & J. Idier. *Spatial Mixture Modelling for the Joint Detection-Estimation of Brain Activity in fMRI*. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, volume 1, pages I–325 –I–328, 2007.

[Wada 01] Y Wada & T Yamamoto. *Selective impairment of facial recognition due to a haematoma restricted to the right fusiform and lateral occipital region*. Journal of Neurology, Neurosurgery & Psychiatry, vol. 71, no. 2, pages 254–257, 2001.

[Wang 04] Tao Wang, Jie Deng & Bin He. *Classifying EEG-based motor imagery tasks by means of time-frequency synthesized spatial patterns*. Clinical Neurophysiology, vol. 115, no. 12, pages 2744 – 2753, 2004.

[Wang 07] Ze Wang, Anna R. Childress, Jiongjiong Wang & John A. Detre. *Support vector machine learning-based fMRI data group analysis*. NeuroImage, vol. 36, no. 4, pages 1139 – 1151, 2007.

[Wang 09] Ze Wang. *A hybrid SVM-GLM approach for fMRI data analysis*. NeuroImage, vol. 46, no. 3, pages 608 – 615, 2009.

[Ward 63] Joe H. Ward. *Hierarchical Grouping to Optimize an Objective Function*. Journal of the American Statistical Association, vol. 58, no. 301, pages 236–244, 1963.

[Williams 07] Mark A A. Williams, Sabin Dang & Nancy G G. Kanwisher. *Only some spatial patterns of fMRI response are read out in task performance.* Nat Neurosci, 2007.

[Wipf 08] David Wipf & Srikantan Nagarajan. *A New View of Automatic Relevance Determination*. In Advances in Neural Information Processing Systems 20, pages 1625–1632. MIT Press, 2008.

[Yamashita 08] Okito Yamashita, Masa aki Sato, Taku Yoshioka, Frank Tong & Yukiyasu Kamitani. *Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns*. NeuroImage, vol. 42, no. 4, pages 1414 – 1429, 2008.

[Young 02] Thomas Young. On the theory of light and colors. The Society, London, 1802.

[Yuan 06] Ming Yuan & Yi Lin. *Model selection and estimation in regression with grouped variables*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 68, no. 1, pages 49–67, 2006.

[Zou 05] Hui Zou & Trevor Hastie. *Regularization and variable selection via the Elastic Net*. J. Roy. Stat. Soc. B, vol. 67, page 301, 2005.