Μπεϋζιανές Μέθοδοι Ανάλυσης και Επεξεργασίας Βιοϊατρικού Σήματος και Εικόνας

Η ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης

του Τμήματος Πληροφορικής Εξεταστική Επιτροπή

από τον

Ευάγγελο Οικονόμου

ως μέρος των Υποχρεώσεων για τη λήψη του

ΔΙΔΑΚΤΟΡΙΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

Ιούλιος 2010

Τριμελής Συμβουλευτική Επιτροπή

- Κωνσταντίνος Μπλέκας, Επίκουρος Καθηγητής του Τμήματος Πληροφορικής του Πανεπιστημιού Ιωαννίνων (επιβλέπων)
- Δημήτριος Ι. Φωτιάδης, Καθηγητής του Τμήματος Μηχανικών Επιστήμης Υλικών του Πανεπιστημίου Ιωαννίνων
- Ισαάκ Λαγαρής, Καθηγητής του Τμήματος Πληροφορικής του Πανεπιστημιού Ιωαννίνων

Επταμελής Εξεταστική Επιτροπή

- Κωνσταντίνος Μπλέκας, Επίκουρος Καθηγητής του Τμήματος Πληροφορικής του Πανεπιστημιού Ιωαννίνων (επιβλέπων)
- Δημήτριος Ι. Φωτιάδης, Καθηγητής του Τμήματος Μηχανικών Επιστήμης Υλικών του Πανεπιστημίου Ιωαννίνων
- Ισαάκ Λαγαρής, Καθηγητής του Τμήματος Πληροφορικής του Πανεπιστημιού Ιωαννίνων
- Αριστείδης Λύκας, Αναπληρωτής Καθηγητής του Τμήματος Πληροφορικής του Πανεπιστημιού Ιωαννίνων
- Χριστόφορος Νίκου, Επίκουρος Καθηγητής του Τμήματος Πληροφορικής του Πανεπιστημιού Ιωαννίνων
- Αναστάσιος Τέφας, Λέκτορας του Τμήματος Πληροφορικής του Αριστοτέλειου Πανεπιστημιού Θεσσαλονίκης
- Μιχάλης Ζερβάκης, Καθηγητής του Τμήματος Ηλεκτρονικών Μηχανικών & Μηχανικών Υπολογιστών, Πολυτεχνείο Κρήτης

Dedication

Στην μητέρα μου και τα αδέλφια μου.

Acknowledgements

Θα ήθελα να ευχαριστήσω τους επιβλέποντες της παρούσης διατριβής, κ. Δημήτριο Ι. Φωτιάδη, καθηγητή του τμήματος Μηχανικών Επιστήμης Υλικών του Πανεπιστημίου Ιωαννίνων, και κ. Κωνσταντίνο Μπλέκα, επίκουρο καθηγητή του τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων, για την αμέριστη βοήθεια τους και την επιστημονική τους καθοδήγηση καθ' όλη τη διάρκεια της εκπόνησης της. Επίσης, θα ήθελα να ευχαριστήσω τον κ. Ισαάκ Λαγαρή, καθηγητή του τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων, για την επιστημονική του καθοδήγηση. Ιδιαιτέρως, θα ήθελα να ευχαριστήσω τον κ. Νικόλαο Γαλατσάνο, καθηγητή του τμήματος Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογι- στών του Πανεπιστημίου Πατρών, που με ενεύπνεσαι μέσα από το μάθημα του και την επιστημονική του ταυτότητα, αλλά και την επιστημονική του καθοδήγηση στις αρχές της διατριβής. Επίσης, θα ήθελα να ευχαριστήσω τα μέλη της επταμελούς επιτροπής, κκ. Αριστείδη Λύκα, Χριστόφορο Νίκου, Αναστάσιο Τέφα και Μιχάλη Ζερβάκη, για την επιστημονική συμβολή τους στην εκπόνηση της παρούσης διατριβής.
Επιπλέον, θα ήθελα να ευχαριστήσω όλους τους φίλους και συναδέλφους για την υποστήριξη τους σε θέματα επιστημονικής φύσεως ή μη.
Τέλος, θα ήθελα να ευχαριστήσω την μητέρα μου και τα αδέλφια μου που τόσο πολύ βοηθήσαν να ολοκληρωθεί η παρούσα διατριβή.

Table of Contents

List of Figures

List of Tables

Abstract

Oikonomou, Vangelis, P. V..
PhD, Computer Science Department, University of Ioannina, Greece.
July, 2010.
Title: Bayesian Methods for Biomedical Signal and Image Processing.
Thesis Supervisor: Konstantinos Blekas.

This thesis is focused on the study and the development of intelligent methods for the processing of biomedical signal and image. Biomedical signals belong to the class of sequential data, i.e. data that are evolved in time or space. Their structure is complex and can be obtained in serial or batch mode. Finally, biomedical signals contain hidden characteristics and their detection constitutes a difficult task. All the above properties although bring some serious obstacles during the study of these signals, are issues of great research interest and challenges, in the sense of becoming the seeds of building effective and innovative mechanisms and tools for biomedical data analysis. Moreover, the necessity of these methods is further amplified with the fact that biomedical signals belongs to the kind of data from the real world applications. Under these prism, analyzing these pieces of information may significantly affect the human life, improve the understanding of the human body, as well as may become a light to the discovery of new perceptions and achievements within the medical world.

The scope of this thesis is to study and present powerful statistical models that incorporate various interesting properties of biomedical signals, such as spatial and temporal correlations between, their time-varying nature, and their environment, in order to achieve the improvement of the fidelity of analysis and the decision making procedure. A desired feature of the models that are presented thought this thesis is to describe the signal with a single and less complex, but powerful and efficient, formulation in a way of increasing their generalization capabilities. One such representation is the sparse representation, which constitutes a modern tendency to the statistical data analysis community with many applications to several others fields, such the Biomedical Engineering, Biology, Machine Learning etc. Variations of the generalized linear regression model and the state - space models, such as the Kalman Filter, are the main stochastic models that are presented for analyzing electroencephalograms, and time series from the heart and from function Magnetic Resonance Imaging.

In chapter 2 and 3, basic notions about the nature of data and problems that results from this are given. Biomedical signals that are studied in this thesis are derived from

the brain and the heart. In chapter 2 basic information about these two organs is given as well as information about the mechanism that generated the corresponding biomedical signals. For the study of these signals probabilistic models are used in conjunction with the Bayesian framework. Thus, basic tools from statistics and machine learning are presented, since these tools will be used to learn model parameters. Also, a review of various general approaches, used in biomedical signal community, is performed. In chapter 3, a description of various probabilistic models is given. More specifically, the linear regression model, the state-space model and the autoregressive model are presented. These models will be used latter in this thesis. Furthermore, a description of various preprocessing steps in the analysis of fMRI data is given.

In chapter 4, a method for the enhancement of epileptic spikes is proposed. Epileptic spikes are observed in the electroencephalogram. To deal with the non stationarity of EEG signal, a time - varying autoregressive model (TVAR) is used. The TVAR model parameters are estimated with the help of the Kalman Filter. The experimental results have shown that the proposed method is able to reduce the false alarms while at the same time keep at acceptable level the loose of true spikes.

One important aspect that must be taken into account is that the biomedical signal is observed with noise. The origin of noise can be some malfunction of hardware or other physiological process of the human body. In chapter 5, a method is proposed to remove the noise for the observations. This is achieved by using a useful prior over the signal of interest. The prior is characterized for its smooth nature and is based on the laplacian operator. Then, adopting the Bayesian framework the model parameters are estimated. The proposed method is used for the estimation of Event Related Potentials (observed in EEG) and the removal of drift from time series that described the heart rhythm. The results have shown accurate estimation of these two signals.

In chapter 6, a method is proposed for the analysis of fMRI time series when the noise is non - stationary. The basic building block of this method is a probabilistic approach of the linear regression model. A sparse representation is used for the weights of the linear regression model through a sparse prior. Sparsity can improve pattern recognition, compression, and noise reduction among others. The noise term of the linear regression model is non stationary and consists from two components, one component originates from the time series while the other from the images. Two versions of this model are used to describe the time series. The first is based on a voxel-by-voxel analysis of fMRI data and the second is based on a simultaneous use of all data. Both approaches are used an extended design matrix to model the drift component, while for the estimation procedure the Variational Bayesian Methodology is adopted resulting in two iterative algorithms. The results, based on real and simulated data, have shown the usefulness of the proposed methods to find the activated brain areas.

The time series arising from fMRI experiments contains correlation between them when come out from adjacent brain areas. In chapter 7, we proposed a method that take into account this information. More specifically, a probabilistic linear regression model with sparse and spatial properties is proposed. This is achieved by proposing an enhanced version of Gibbs distribution for the prior distribution of weights. The potential function of the Gibbs distribution is of specific purpose and has two components, one to model the sparsity between the weights of one time series and the other to model the spatial correlation between weights that belongs to adjacent time series. To perform inference over model parameters the Maximum A Posteriori (MAP) estimation framework is used. Also, an alternative view of the proposed model, using the Expectation - Maximization (EM) algorithm, is presented. The results, based on real and simulated data, have shown the ability of the proposed method to detect accurately the activated brain areas.

In chapter 8, we proposed a new probabilistic mixture modeling approach for clustering fMRI time series based on linear regression models where each cluster is described as a linear regression model. A sparse representation of every cluster regression model is used through the use of an appropriate sparse prior over the regression coefficients. Enforcing sparsity is a fundamental regularization principle and has been used to tackle several problems, such as model order selection. Also, spatial properties of data have been incorporated to the mixture model through the notion of Markov Random Field (MRF). Furthermore, to avoid sensitivity of the design matrix to the choice of kernel matrix, we have used a kernel composite design matrix constructed as linear combination of Gaussian kernel matrices with different scaling parameter. The clustering procedure is formulated as a Maximum A Posteriori (MAP) estimation problem where the Expectation - Maximization (EM) algorithm constitutes a powerful framework for solving it. To avoid problems with the initialization of the algorithm, an incremental strategy for building the mixture model is presented. Experiments using artificial and real fMRI dataset have shown that the proposed method offers very promising results with an excellent behavior in difficult and noisy environments.

ΕΚΤΕΤΑΜΕΝΗ ΠΕΡΙΛΗΨΗ ΣΤΑ ΕΛΛΗΝΙΚΑ

Ευάγγελος Οικονόμου του Παύλου και της Βαΐας.
PhD, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Ιούλιος, 2010.
Τίτλος: Μπεϋζιανές Μέθοδοι Ανάλυσης και Επεξεργασίας Βιοϊατρικού Σήματος και Εικόνας.
Επιβλέποντας: Κωνσταντίνος Μπλέκας.

Η παρούσα διατριβή εστιάζεται στην μελέτη και ανάπτυξη ευφυών μεθόδων επεξεργασίας βιοϊατρικού σημάτος και εικόνας. Τα βιοϊατρικά σήματα που αναλύονται ανήκουν στην περιοχή των χρονικά μεταβαλλόμενα δεδομένων καθώς αποτελούν ακολουθίες τιμων στο χρόνο. Η δομή τους είναι πολύπλοκη και μη ευδιάκριτη, ενώ μπορεί να εμφανίζονται σειριακά ή μαζικά. Τέλος, τα βιοϊατρικά σήματα περιέχουν κρυμμένα χαρακτηριστικά και η ανίχνευσή τους αποτελεί μια επίπονη διαδικασία. Οι παραπάνω ιδιαιτερότητες, παρ' όλη την δυσκολία που επιφέρουν, χρήζουν ιδιαίτερου ερευνητικού ενδιαφέροντος στην επιστημονική κοινότητα και τροφοδοτούν αποτελεσματικά την μελέτη και ανάπτυξη καινοτόμων εργαλείων και μηχανισμών επεξεργασίας βιοϊατρικών σημάτων. Μάλιστα η αναγκαιότητα των μεθόδων αυτών ενισχύεται ακόμα περισσότερο, καθώς τα βιοϊατρικά σήματα αποτελούν δεδομένα του πραγματικού κόσμου. Η εξαγωγή γνώσης που προκύπτει απο την ανάλυση τους μπορεί να προσφέρει σημαντικά στην βελτίωση της ανθρώπινης ζωής, στην καλύτερη κατανόηση της λειτουργίας του ανθρώπινου σώματος αλλά και στην ανακάλυψη νέων αντιλήψεων και επιτευγμάτων του ιατρικού κόσμου.

Σκοπός της διατριβής είναι η μελέτη και κατασκευή μοντέλων που θα λαμβάνουν υπόψη ιδιότητες των βιοϊατρικών σημάτων, όπως οι χωρικές και χρονικές εξαρτήσεις που υπάρχουν μεταξύ αυτών, η χρονική μεταβλητότητά τους, και το περιβάλλον που λαμβάνονται, με στόχο την βελτίωση της πιστότητας της ανάλυσης τους και την αποτελεσματική εξαγωγή συμπερασμάτων με το μικρότερο δυνατό σφάλμα. Ένας επιπλέον στόχος των προτεινόμενων μοντέλων είναι να παραχθεί μια αποτελεσματική και εύχρηστη αναπαράσταση του σήματος για διευκόλυνση της περαιτέρω ανάλυσης αυξάνοντας ταυτόχρονα την γενικευτική ικανότητα. Μια τέτοια αναπαράσταση είναι η αραιή αναπαράσταση η οποία αποτελεί μια σύγχρονη τάση της στατιστικής ανάλυσης δεδομένων με πλούσιες και ενδιαφέρουσες εφαρμογές στην Ιατρική Τεχνολογία, την Βιολογία, καθώς και σε άλλους τομείς της επιστήμης. Πιο συγκεκριμένα, προτείνονται παραλλαγές του γραμμικού μοντέλου παλινδρόμησης και των φίλτρων Kalman, για την επεξεργασία ηλεκτροεγκεφαλογραφήματος, χρονοσειρών που περιγράφουν τον ρυθμό της καρδιακής λειτουργίας, και χρονοσειρών που προέρχονται από εικόνες λειτουργικού μαγνητικού συντονισμού (fMRI). Για την μελέτη των μοντέλων και την ανάπτυξη των αλγορίθμων επιλέχθηκε το Μπευζιανό πλαίσιο εργασίας.

Αρχικά, στα κεφάλαια 2 και 3, παρουσιάζονται βασικές έννοιες της φύσης των δεδομένων και των προβλημάτων που απορρέουν από αυτά. Τα ιατρικά σήματα που μελετούνται προέρχονται από τον εγκέφαλο και την καρδιά. Στο κεφάλαιο 2 παρέχονται βασικές έννοιες και χαρακτηριστικά πληροφορίες που σχετίζονται με τα δυο αυτά ανθρώπινα όργανα καθώς επίσης και με τον τρόπο παραγωγής των αντίστοιχων σημάτων τους. Για την αναπαράσταση και επεξεργασία αυτών των σημάτων χρησιμοποιούνται κυρίως πιθανοτικά παραμετρικά μοντέλα σε ένα Μπεϋζιανό πλαίσιο ανάπτυξης. Έτσι, παρουσιάζονται βασικά θεωρητικά εργαλεία της επιστημονικής περιοχής της στατιστικής και της μηχανικής μάθησης (machine learning), τα οποία θα χρησιμοποιηθούν στη διαδικασία εκπαίδευσης των παραμέτρων του μοντέλου. Επίσης, επιτελείται μια ευρύτερη ανασκόπηση βασικών αρχών και μεθοδολογιών που χρησιμοποιούνται στην επεξεργασία βιοϊατρικού σήματος και εικόνας. Στο κεφάλαιο 3 περιγράφονται εισαγωγικές έννοιες που αφορούν τα στοχαστικά μοντέλα που θα χρησιμοποιηθούν στην παρούσα διατριβή. Πιο συγκεκριμένα, παρουσιάζονται το γραμμικό μοντέλο παλινδρόμηση, το state - space μοντέλο, και το μοντέλο αυτοσυσχέτισης καθώς και μια πιθανοτική αναπαράσταση αυτού. Τέλος, δίνεται μια γενική περιγραφή των βημάτων που αφορούν την στατιστική ανάλυση χρονοσειρών από εικόνες fMRI.

Στο κεφάλαιο 4, προτείνεται μια μέθοδος που αφορά τήν ενίσχυση των επιληπτικών δυναμικών (epileptic EEG spike) που παραρατηρούνται στο ηλεκτροεγκεφαλογράφημα, με στόχο την διεκόλυνση της διαδικασίας ανιχνευσή τους. Για να αντιμετωπιστεί η μη στασιμότητα (nonstationarity) του ηλεκτροεγκεφαλογραφήματος χρησιμοποιείται το χρονικά μεταβαλλόμενο μοντέλο αυτοσυσχέτισης (time varying autoregressive model), οι παράμετροι του οποίου εκτιμούνται με την χρήση των φίλτρων Kalman. Τα πειραματκά αποτελέσματα ήταν πολύ ενθαρρυντικά, καθώς προσφέρουν σημαντική μείωση του πλήθους των λανθασμένων προειδοποιήσεων (false alarms) που αφορούν την εμφάνιση ενός επιληπτικού δυναμικού.

Ένα σημαντικό στοιχείο κατά την λήψη των βιοιατρικών σημάτων ειναι το θορυβώδες περιβάλλον. Ο θόρυβος μπορεί να προέρχεται απο δυσλειτουργίες του υλικού ή από άλλες λειτουργίες του ανθρώπινου σώματος που συμβαίνουν ταυτόχρονα με τη λήψη του σήματος ενδιαφέροντος. Στο κεφάλαιο 5 προτείνεται μια μέθοδος που αφορά την ανάκτηση (ή την εκτίμηση) του σήματος ενδιαφέροντος όταν αυτό παρατηρείται μέσα σε ένα θορυβώδες περιβάλλον. Για να την επίτευξη αυτού του στόχου χρησιμοποείται μια κατάλληλη εκ των προτέρων κατανομή (prior distribution) πάνω στο σήμα ενδιαφέροντος. Καθώς η επιθυμητή ιδιότητα του σήματος είναι αυτή της ομαλότητας (smoothness) εφαρμόζεται μια κατανομή, που χαρακτηρίζεται για την ομαλότητα της (smoothness prior). Για την εκτίμηση των παραμέτρων του μοντέλου χρησιμοποιείται μια προσεγγιστική Μπεϋζιανή μεθοδολογία, που καλείται Variational Bayesian. Το προτεινόμενο μοντέλο χρησιμοποιείται για την εκτίμηση προκλητών δυναμικών σχετιζόμενα με ένα γεγονός (Event Related Potentials) και για την αφαίρεση του drift (συστατικό του σήματος που εμφανίζεται στις χαμηλές συχνότητες) από χρονοσειρές που εκφράζουν την μεταβλητότητα του καρδιακού ρυθμού (Heart Rate Variability).

Στο κεφάλαιο 6 προτείνεται μια μέθοδος που σχετίζεται με την ανάλυση χρονοσειρών που προέρχονται από εικόνες λειτουργικού μαγνητικού συντονισμού (fMRI). Το βασικό μοντέλο στην ανάλυση αυτού του είδους χρονοσειρών είναι το γραμμικό μοντέλο παλινδρόμησης και ο σκοπός της ανάλυσης είναι ο καθορισμός των περιοχών ενεργοποίησης του εγκεφάλου κατά την δάρκεια ενός ερεθίσματος. Η μέθοδος που προτείνεται λαμβάνει υπόψη χαρακτηριστικά των χρονοσειρων, όπως η μη στατιμότητα του θορύβου καθώς και η παρουσία του drift. Μελετάται η αποτελεσματικότητα δυο γραμμικών μοντέλων παλινδρόμησης για την ανάλυση των χρονοσειρών. Το πρώτο μοντέλο χρησιμοποιείται για ανάλυση μιας χρονοσειράς κάθε φορά. Το δεύτερο μοντέλο λαμβάνει υπόψη όλες τις χρονοσειρές με αποτέλεσμα να έχουμε μια χωροχρονική ανάλυση των χρονοσειρών. Και τα δυο μοντέλα παρέχουν αραιή αναπαράσταση των χρονοσειρών μέσω μιας εκ των προτέρων κατανομής αραιού τύπου (sparse prior) στα βάρη (ή συντελεστές παλινδρόμησης) του γραμμικού μοντέλου. Για την εκτίμηση των παραμέτρων του μοντέλου χρησιμοποιείται η Variational Bayesian μεθοδολογία.

Οι χρονοσειρές που προέρχονται από εικόνες λειτουργικού μαγνητικού συντονισμού παρουσιάζουν χωρικές εξαρτήσεις λόγω της φυσιολογίας του εγκεφάλου. Στο κεφάλαιο 7 προτείνεται ένα σύνθετο γραμμικό μοντέλο παλινδρόμησης εμπλουτίζοντας το με σημαντικές ιδιότητες που προέρχονται από τις χωρικές εξαρτήσεις ανάμεσα στις χρονοσειρές καθώς και αραιή αναπαράσταση του συναρτησιακού μοντέλου περιγραφής. Αυτό επιτυγχάνεται με τη χρήση κατάλληλης εκ των προτέρων κατανομής στα βάρη του γραμμικού μοντέλου που βασίζεται στο μοντέλο MRF (Markov Random Field). Πιο συγκεκριμένα προτείνεται μια σύνθετη Gibbs κατανομή η οποία ενσωματώνει τις δυο παραπάνω ιδιότητες στο μοντέλο. κατάλληλη για το πρόβλημα μας. Ακολουθείται το Μπεϋζιανό πλαίσιο δράσης. Τα βάρη του μοντέλου εκτιμούνται μέσω της μεγιστοποίησης της εκ των υστέρων πιθανοφάνειας (Maximum A Posteriori) παράγοντας επαναληπτικούς τύπους. Ενναλακτικά, η προτεινόμενη μεθοδολογία μπορεί να προσεγγιστεί ακολουθώντας τον αλγόριθμο Expectation - Maximization (EM) αν θεωρησουμε τα βάρη ως κρυμμένες μεταβλητές. Τα πειράματα που διεξήχθησαν τόσο σε τεχνητά όσο και σε πραγματικά δεδομένα ήταν πολύ σημαντικά και ανέδειξαν τη χρησιμότητα της μεθόδου, σε σύγκριση με άλλες μεθόδους της βιβλιογραφίας.

Στο κεφάλαιο 8 παρουσιάζεται το πρόβλημα της ανάλυσης fMRI δεδομένων ως ένα πρόβλημα ομαδοποίησης (clustering). Για τον σκοπό αυτό χρησιμοποιείται ένα μικτό μοντέλο γραμμικών παλινδρομητών, όπου κάθε ομάδα (cluster) περιγράφεται με ένα μοντέλο γραμμικής παλινδρόμησης (linear regression model). Οι καινοτομίες της προτεινόμενης μεθοδολογίας εντοπίζονται στα εξής σημεία: Αρχικά γίνεται η υπόθεση ότι η πληροφορία της ετικέτας της ομάδας που ανήκει κάθε χρονοσειρά είναι μια τυχαία μεταβλητή η τιμή της οποίας εξαρτάται από τη ευρύτερη γειτονιά στην οποία βρίσκεται πάνω στο χάρτη ενεργοποίησης. Έτσι, εφαρμόζοντας Μαρκοβιανά Τυχαία Πεδία (Markov Random Fields) και την κατανομή Gibbs πάνω στις ετικέτες, επιτυγχάνεται ο εμπλουτισμός του μικτού μοντέλου με τις χωρικές ιδιότητες που υπάρχουν και στη φύση των δεδομένων του προβλήματος. Για να εξασφαλιστεί μια περισσότερο γενικευμένη ικανότητα σε κάθε ομάδα περιοχή που κατασκευάζεται χρησιμοποιείται μια αραιού τύπου αναπαράσταση του γραμμικού μοντέλου παλινδρόμησης με κατάλληλη αραιή κατανομή στους συντελεστές κάθε ομάδας. Τέλος, για τον πίνακα σχεδίασης προτείνεται ένας γραμμικός συνδυασμός από πίνακες σχεδίασης με

Γκαουσιανές συναρτήσεις πυρήνα που έχουν διαφορετική παράμετρο διασποράς. Με τον τρόπο αυτό επιτυγχάνεται η εξάλειψη τους προβλήματος της εξάρτησης από την παράμετρο διασποράς, η οποίαεπηρεάζει σε σημαντικό βαθμό το κατάλληλο ταίριασμα των δεδομένων και κατ' επέκταση το αποτέλεσμα της ομαδοποίησης. Επίσης, προτείνεται μια αυξητική κατασκευή του μικτού μοντέλου εφαρμόζοντας μια συνεχή διαδικασία διάσπασης (splitting), καθώς και ένα κριτήριο τερματισμού με βάση το βαθμό συσχέτισης. Έτσι, παράλληλα προτείνεται και η εύρεση του κατάλληλου αριθμού των ομάδων, το οποίο αποτελεί και ένα πολύ σημαντικό πρόβλημα στην ομαδοποίση.

Chapter 1

Introduction

Many functions of the human body are associated with signals of electrical, chemical or acoustic origin. Such signals carry information which may not be obvious but it is hidden in the structure of the signal. This information must be decoded before the signals provide us with some meaningful interpretation. The signals reflect properties of associated biological systems and their analysis have been found to be helpful in explaining and identifying various pathological conditions. The decoding process is sometimes straightforward and may involve very limited manual effort such as visual inspection. However, the complexity of the signal is quite often considerable and therefore Biomedical Signal Processing becomes an important tool for extracting clinically significant information hidden in the signal. Biomedical Signal Processing is an interdisciplinary field since knowledge from various scientific topics is required.

Biomedical Signal Processing plays a crucial role in many aspects of human life. The analysis of biomedical signals is the central part of automated medical systems, aiming at finding disorders of human body. Also, biomedical signals play significant role to the design of Human - Computer Interfaces (HCI) and Brain - Computer Interfaces (BCI). Recent development in technology allow monitoring physiological processes inside our body, for which no natural interfaces exist. In particular, we can measure blood pressure, heart rate variability, muscular activity, and brain activity in efficient and noninvasive ways. It is natural to assume that such information can be used in a useful way for the human. Nowadays, this information is used to treat various pathophysiological disorders of human body, to understand the underlying mechanism of the human body, to design machines which communicate with humans. Machines, based mostly on brain signals, have developed for a variety of applications ranging from assistive technologies for patients with motor disabilities, to entertainment devices.

Biomedical signals are observations of physiological activities of organisms, ranging from gene and protein sequences, to neural and cardiac rhythms, to tissue and organ images. The processing of biomedical signal aims at extracting useful information from it. Biomedical signals carry information that is useful for the understanding of mechanisms underlying the behavior of living systems. However, such information is difficult to be obtained directly from the raw recorded signals. In most of the cases, it is masked by other biomedical signals which occur at the same time or buried in some additive noise. For such reasons, processing is usually required to enhance the relevant information and to extract from it parameters which quantify the behavior of the biological system under study, mainly for physiologic studies, or to define the degree of pathology for routine clinical procedures (diagnosis, therapy, rehabilitation or monitoring).

In the beginning, biomedical signals have been assessed manually leading to unreliable diagnostic conclusions. A fundamental goal of biomedical signal processing is to reduce the subjectivity of the manual measurements. The introduction of computer-based methods helps to objectively quantify the various characteristics of signals. Those improve accuracy of measurements and their reproducibility.

In addition, biomedical signal processing can be used to develop methods for feature extraction to help characterize and understand the information obtained from a signal. Such feature extraction methods can be designed to mimic the manual measurements, but can also designed to extract information which can not be extracted by visual examination. For example, small variations in the heart rate that cannot be perceived by the human eye have been found to contain valuable clinical information when quantified using a signal processing method.

In many cases, the recorded signal is corrupted by different types of noise and interference, sometimes originating from another physiological process of the body. For example, such situations may arise when the ocular activity interferes with the desired brain activity, when the electrodes are poorly attached to the body, or when external sources degrade the signal such as the 50/60 Hz powerline interference. Hence, signal denoising represents a crucial objective of biomedical signal processing.

Certain diagnostic procedures required the recording of signals for large time. Such situations may arise, for example when we record brain signals to study the brain function during sleep or when we study disturbances of the heart rhythm. Also, in many cases this procedure involves many channels. All these result to huge data size fill up the hard disk. Transmission of biomedical signals across public networks is another application which involve the size of biomedical data. For all these situations, data compression of digital biomedical signals is essential. General purpose methods of data compression do not perform particularly well since the characteristics of biomedical signals are not exploited.

Finally, signal modeling and simulation is another important field of research in biomedical signal processing. This helps us to better understand physiological processes. With suitable defined model it is possible to create signals which resemble the true signals. For example, models have been created for the head and the brain to localize sources of the neural activity. Signal modeling is also part of the branch of signal processing called ”model - based signal processing”, where algorithm development is based on the optimization of an appropriately selected performance function. Algorithms for processing biomedical signals constitute the central core of any medical system responsible for therapy, monitoring and diagnosis.

Several signal processing techniques can be used to analyze biomedical signals. These techniques can be performed either on time- or frequency-domain of the signal. Even if it is possible to deal with continuous time waveforms, it is usually convenient to convert them into a digital form before processing. The general framework for biomedical signal processing is presented in Fig. 1.1. First, the setup of the experiment must be carry out. Then, the acquisition of the signal is performed. After that, some preprocessing steps, such as filtering, are performed. Then, the signal is analyzed to obtain useful information and perform the physiological interpretation of the signal i.e. pathological or normal condition of the subject. This thesis deals with the last three stages of the framework. The preprocessing stage aims at making the signal of interest suitable for the subsequent analysis. At the end of the preprocessing stage we obtain a signal which contains the desired information of our experiment. The statistical analysis stage includes the analysis of the signal to obtain useful information related to the experiment. This stage includes the use of a model to explain the signal. Finally, the interpretation of the results is performed with the help of a medical expert.

A useful class of methods to process signals is the model - based approach, which is adopted in this thesis. A model is a simplified mathematical representation of a signal. Also, it depends on some parameters which are unknown and usually are estimated using the observations of the experiment. Learning the model parameters can be done by minimizing (or maximizing) an objective function, which in most cases, is a function of the unknown parameters. The model-based approach to analyze a signal can be thought as a compact scheme consisting of three main parts:

- The model,
- The objective function,
- The learning process.

The model formalizes the prior knowledge about the process that generates the observations. The objective function is related to a function with respect to model parameters that takes into account some natural constraints of the problem and the parameters. The learning process offers an optimization framework for the objective function where the estimation of the model parameters is performed. For example, a regression problem can be described with the linear regression model. Least Squares play the role of the objective function that can minimized with a local optimization algorithm such as Newton, Gradient Descent, etc. In this way, the model parameters will be estimated.

The mathematical treatment of the models and algorithms in this thesis is based on the Bayesian Framework. This means that all the results are treated with probability distributions, which helps in modeling the uncertainties in the model and the physical randomness. Also, in the Bayesian Framework we are able to introduce constraints on our model or prior knowledge about it with an elegant and natural way using appropriate prior distributions. Bayesian analysis of data has been greatly facilitated in the last decade

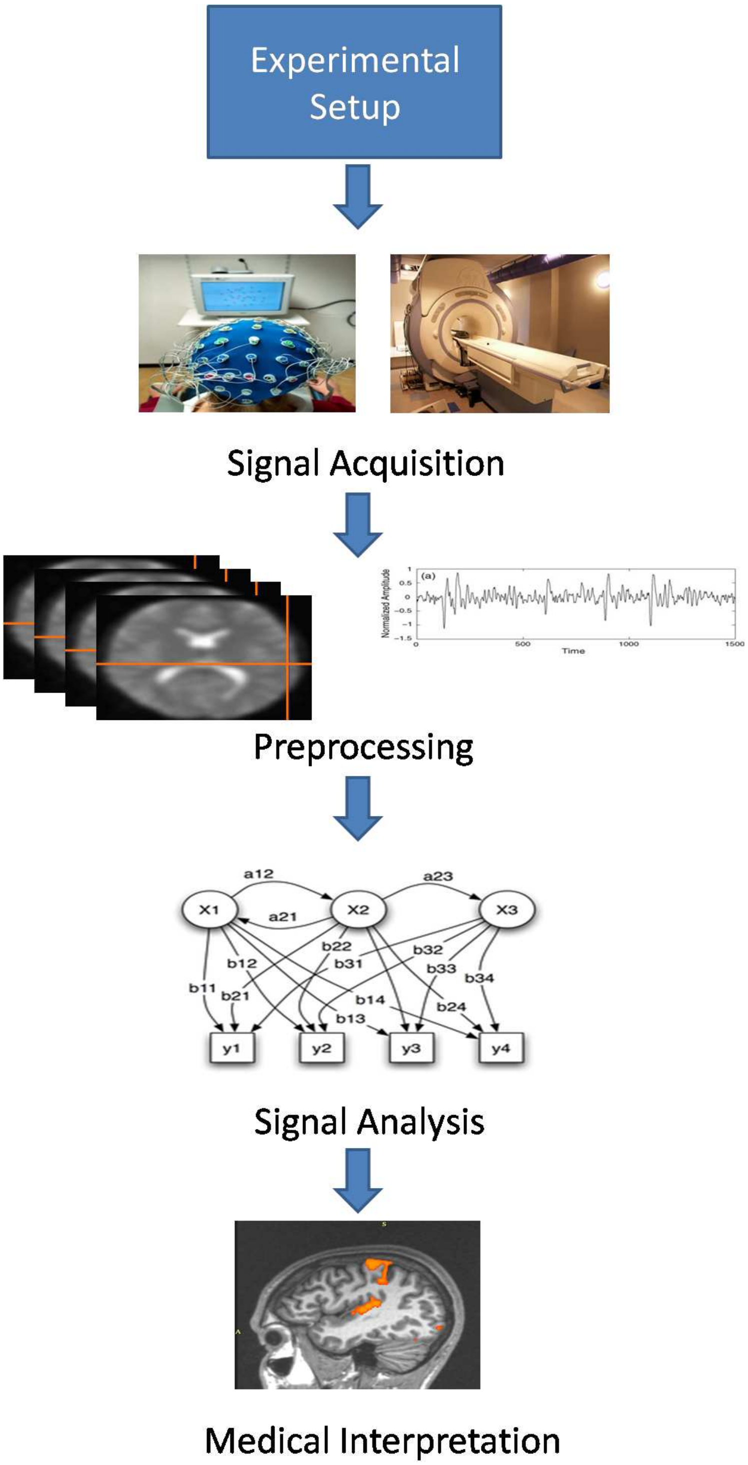


Figure 1.1: Biomedical Signal Processing framework.

by advances in computing power and improved scope for estimation via iterative sampling methods. The Bayesian approach allows to make probability statements about the model parameters and has a single tool, Bayes' theorem, which is used in all situations. Also, has a straightforward way of dealing with nuisance parameters, while the Bayes' theorem gives the way to find the predictive distribution of future observations. But, while it is easy to write the formula for the posterior distribution, a closed form exists only for simple cases, such as for a normal sample with a normal prior. In that case approximation techniques are applied.

1.1 Thesis Contribution

This thesis aims at providing innovative and efficient probabilistic models for biomedical signal processing. Throughout this thesis the target of methods which are proposed, is to incorporate appropriate medical knowledge and natural constraints of the problem to their body, so to become more effective and with more accurate results. This is achieved through the Bayesian Framework that supply a rich platform to naturally treat the physical properties of biomedical signals. This resulting probabilistic environment provides us with a way to introduce constraints into our problem through the use of prior distributions, to obtain an estimate of model parameters through their posterior distribution, and to predict future behavior through the predictive distribution. The use of constraints inside a stochastic model expands the capabilities of the model leading to the design of more complex and flexible models for the description of the signal. Since our goal is to create general - purpose methodologies, the proposed models are not restricted only to biomedical signals, but they can be applied to other application areas with sequential data such as image processing, computer vision, video analysis, bioinformatics, etc.

Chapters 2 and 3 provide introductory material for the rest of this thesis. More specifically, in chapter 2 we present the physiology and properties of biomedical signals which are used in this thesis. Also, a description of the basic tool used in this thesis, the Bayesian Framework, is provided. In addition, methods, that help us to make inference in a model, are described. In chapter 3, the basic model of this thesis, the linear regression model, is explained, as well as extensions of it. Furthermore, a description of the autoregressive model is provided since this model will be used in conjunction with the linear model.

In chapter 4, we present a methodology for epileptic spike enhancement in electroencephalographic (EEG) recordings. The goal of this method is to enhance the epileptic spikes so their detection be more easily performed. To achieve this the time varying autoregressive model (TVAR) is used. Using the Kalman Filter we can obtain estimates of the time varying AR coefficients and an enhanced version, with respect to the epileptic spikes, of the EEG signal. The results indicate that the proposed methodology reduce significantly the number of false alarms. Also, the proposed model can be used for time varying spectrum estimation.

In chapter 5, a method for the recovery of a biomedical signal from a noisy environment is proposed. The method is based on the model - based approach, where the signal is modeled through the use of a smoothness prior while the statistics of the noise are unknown. To make inference about the unknown quantities of the model, the Variational Bayesian Framework is used. The proposed method was applied for the estimation of Event Related Potentials and for the removal of the drift from Heart Rate Variability time series.

In chapter 6, two algorithms are proposed to deal with the non stationarity of the noise in the fMRI data. The first algorithm is based on the temporal analysis of the data and it is is based on the linear regression model, while the second algorithm is based on the spatio - temporal analysis where a spatio - temporal version of the linear model is used. Both algorithms estimate the variance of the noise across the images and the voxels. In the linear model, an extended design matrix is used to deal with the presence of the drift in the fMRI time series. To estimate the regression parameters of the GLM as well as the variance components of the noise, the Variational Bayesian (VB) Methodology is employed.

In chapter 7, an advanced Bayesian framework is presented for the analysis of functional Magnetic Resonance Imaging (fMRI) data that simultaneously employs both spatial and sparse properties. The basic building block of our method is the general linear regression model (GML) that constitutes a well-known probabilistic approach. By treating regression coefficients as random variables, we can apply an enhanced Gibbs distribution function that captures spatial constrains and at the same time allows sparse representation of fMRI time series. The proposed scheme is described as a maximum a posteriori (MAP) approach, where the known Expectation Maximization (EM) algorithm is applied offering closed form update equations for the model parameters.

In chapter 8, a new probabilistic mixture modeling approach is proposed for clustering fMRI time series based on linear regression models where each cluster is described as a linear regression model. A sparse representation of every cluster regression model is used through the use of an appropriate sparse prior over the regression coefficients. Enforcing sparsity is a fundamental regularization principle and has been used to tackle several problems, such as model order selection. Also, spatial properties of data have been incorporated to the mixture model through the notion of Markov Random Field (MRF). Furthermore, to avoid sensitivity of the design matrix to the choice of kernel matrix, we have used a kernel composite design matrix constructed as linear combination of Gaussian kernel matrices with different scaling parameter. The clustering procedure is formulated as a Maximum A Posteriori (MAP) estimation problem where the Expectation - Maximization (EM) algorithm constitutes a powerful framework for solving it. To avoid problems with the initialization of the algorithm, an incremental strategy for building the mixture model is presented. Experiments using artificial and real fMRI dataset have shown that the proposed method offers very promising results with an excellent behavior in difficult and noisy environments.

Finally, at chapter 9, concluding remarks and future direction of the proposed methods are provided.

Chapter 2

Biomedical Signal Processing

In this chapter we provide information about the physiology and the properties of biomedical signals which are used in this thesis. Also, the problems that we challenge, when these signals are analyzed, are described. Furthermore, details about the Bayesian Framework is provided, since it is the basic tool for the analysis of the models that are described in latter chapters. In addition, information about two methods, the Expectation - Maximization algorithm and the Variational Bayesian Methodology, are provided. These methods help us to make inference in a Bayesian approach of a problem.

2.1 Physiology of the brain

The human brain is the center of the human nervous system and a very complex organ. Enclosed in the cranium, it has the same general structure as the brains of the others mammals, but it larger from the brain of mammals with equivalent body size. Most of the expansion comes from the cerebral cortex, a convoluted layer of neural tissue that covers the surface of the forebrain. The cerebral cortex is symmetric, with left and right hemispheres and each hemisphere is divided into four parts, the frontal lobe, parietal lobe, temporal lobe and occipital lobe (see Fig. 2.1). This categorization does not actually arise from the structure of the brain itself, the lobes are named after the bones of the skull that overlie them.

The function of the cortex can be divided in three functional categories of areas. One consists of the primary sensory area, which receive information from the sensory nerves. Primary sensory area include the visual area of the occipital lobe, the auditory area in the temporal lobe and the somatosensory area of the parietal lobe. The second category is the primary motor area, which occupies the rear portion of the frontal lobe, directly in front of the somatosensory area. The primary motor area is responsible for the planning and execution of movements. Finally, the third category consists of the remaining parts of the cortex, which are called the association areas. These areas receive information from the sensory areas and are involved in the complex process that we call perception, thought

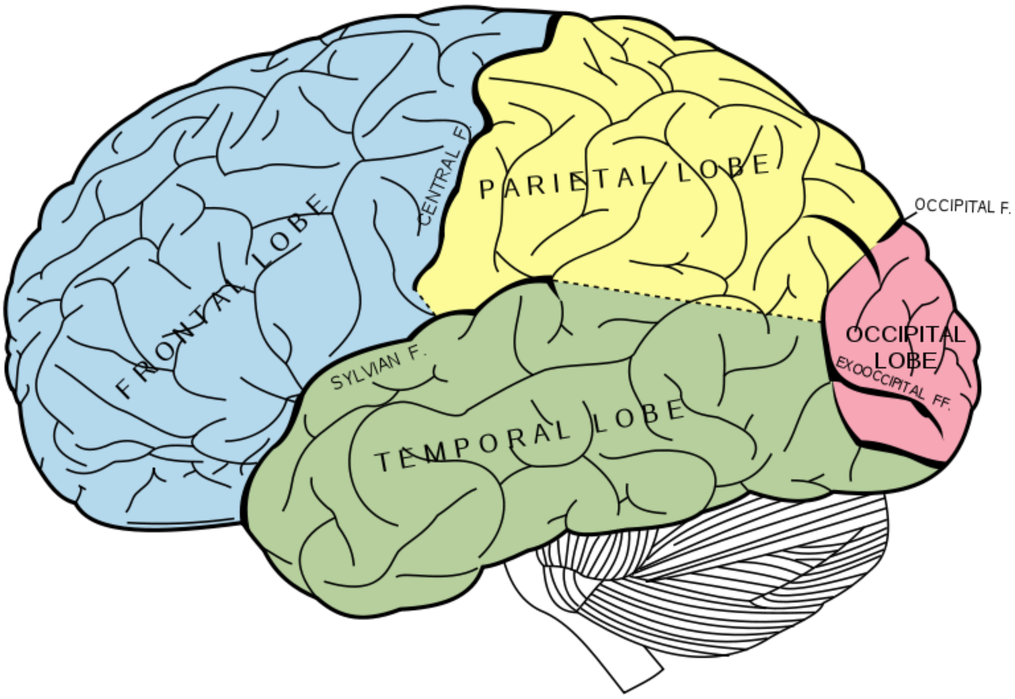


Figure 2.1: The four brain lobes (reprinted from wikipedia)

and decision making. Information about the structure and function of the human brain comes from a variety of methods known as functional neuroimaging. Functional neuroimaging is a general term for several brain imaging methods such as positron emission tomography (PET), single photon emission tomography (SPET), electroencephalography (EEG), magnetoencephalography (MEG) and functional Magnetic Resonance Imaging (fMRI) (for overview see [140, 142, 143, 9, 24, 25, 54]). All these methods, although are based in different principles, aim to reveal the function of the brain.

The nervous system gathers, communicate and processes information from various part of the body and assures that the responses are handled rapidly and accurately. The nervous system is divided into the central nervous system (CNS), consisting from the brain and the spinal cord, and the peripheral nervous system (PNS), connecting the parts of CNS to the body organs and sensory systems. The two systems are integrated because information from the PNS is sent for processing to the CNS, and responses are sent by the PNS to the organs of the body. The nerves transmitting information from the body to the CNS are called sensory nerves, while the nerves transmitting information from the CNS are called motor nerves.

The basic functional unit of the nervous system is the neuron, which transmit information to and from the brain. Neurons can be classified into three categories according to their functionality: sensory neurons, connected to sensor organs, motor neurons, connected to muscles, and interneurons, connected to other neurons. The neuron consists of

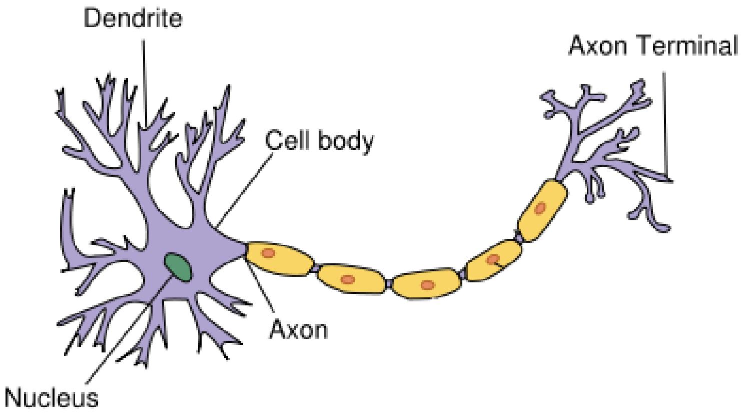


Figure 2.2: Main parts of a neuron.

the cell body, the dendrites and the axon (Fig. 2.2). Dendrites can consist of thousands of branches, where each branch receive information from another signal. The axon is usually a single branch responsible to transmit the information of the neuron to other parts of the nervous system. The transmission of information between the neurons take place at the synapse. The synapse is the part where one neuron contact to the other. The information is transmitted between the various part of the nervous system as an electrical or chemical signal. The currents generated by a single neuron are too weak to be detected noninvasively. However, the currents of individual neurons add up and the simultaneous activation of a population of neurons can result in a current that is large enough to be detectable on the surface of the brain. The recording of this electrical activity of the brain produces the electroencephalogram.

2.2 functional Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a medical imaging technique used to visualize the internal structure of the body. MRI provides much greater contrast between the different soft tissues of the body than computed tomography (CT) does. This fact makes MRI useful in neurological (brain), musculoskeletal, cardiovascular, and oncological (cancer) imaging. MRI uses a powerful magnetic field to align the nuclear magnetization of (usually) hydrogen atoms in water in the body. Radio frequency (RF) fields are used to systematically alter the alignment of this magnetization. This causes the hydrogen nuclei to produce a magnetic field detectable by the scanner. This signal can be manipulated by additional magnetic fields to build up enough information to construct an image of the body. An image can be constructed because the protons in different tissues return to their equilibrium state at different rates, which is a difference that can be detected. By changing the parameters on the MRI scanner, this effect is used to create contrast between different types of body tissue or between other properties, as in fMRI and diffusion MRI.

functional Magnetic Resonance Imaging (fMRI) is a type of specialized MRI scan. It measures the hemodynamic response (change in blood flow) related to neural activity in the brain. Since the 1990s, fMRI has become the dominated imaging technique in the brain mapping area due to its relatively low invasiveness, absence of radiation exposure and wide availability. The physical basis, which make the fMRI possible, is the Nuclear Magnetic Resonance (NMR) phenomenon. This phenomenon was discovered around 1920 and 1930. The magnetic field inside the scanner affects the magnetic nuclei of atoms. Normally, atomic nuclei are randomly oriented, but under the magnetic field the nuclei become aligned with the direction of the field. When the magnetic field is large enough, the tiny magnetic signals from the nuclei add up resulting in a signal that is large enough to measure. In fMRI it is the magnetic signal from hydrogen nuclei in water that is detected. The key to MRI is that the signal from hydrogen nuclei varies in strength depending on the surrounding area. This provides a means of discriminating between grey matter, white matter and cerebral spinal fluid in structural images of the brain. The fMRI is based on the observation that when neural activity increases there is an increased demand for oxygen, which leads in an increase in blood flow in regions of increased neural activity.

It is known that changes in blood flow and blood oxygenation in the brain (collectively known as hemodynamics) are closely linked to neural activity. When the nerve cells are active their consumption of oxygen is increased. The local response to the oxygen consumption is to increase blood flow to regions of increased neural activity, which occurs after a delay of approximately 15 seconds. This hemodynamic response rises to a peak over 45 seconds, before falling back to baseline. This leads to local changes in the relative concentration of oxyhemoglobin and deoxyhemoglobin and changes in local cerebral blood volume (CBV) and cerebral blood flow (CBF).

Blood Oxygen Level Dependent (BOLD) is the MRI contrast of blood deoxyhemoglobin. Through the hemodynamic response, blood releases oxygen to active neurons at a greater rate than to inactive neurons. Hemoglobin is diamagnetic when oxygenated but paramagnetic when deoxygenated. The magnetic resonance (MR) signal of blood is therefore slightly different depending on the level of oxygenation. Higher BOLD signal intensities arise from increases in the concentration of oxygenated hemoglobin since the blood magnetic susceptibility now more closely matches the tissue magnetic susceptibility. By collecting data in an MRI scanner with sequence parameters sensitive to changes in magnetic susceptibility one can assess changes in BOLD contrast. These changes can be either positive or negative depending upon the relative changes in both CBF and oxygen consumption. Increases in CBF that outstrip changes in oxygen consumption will lead to increased BOLD signal, conversely decreases in CBF that outstrip changes in oxygen consumption will cause decreased BOLD signal intensity. The signal difference is very small, but given many repetitions of a thought, action or experience, statistical methods can be used to determine the areas of the brain which reliably show more of this difference as a result, and therefore which areas of the brain are active during that thought, action or experience. The BOLD signal is an indirect indicator of the brain activity and an important question is how well it corresponds to the neural activity, which in general is taken as the definition of brain activity. In [109] show that the neural activity of the brain is well correlated to the blood oxygenation.

The ultimate goal of fMRI data analysis is to detect correlations between brain activation and the task the subject performs during the scan. The BOLD signature of activation is relatively weak, so other sources of noise in the acquired data must be carefully controlled. This means that a series of preprocessing steps must be performed on the acquired images before the actual statistical search for task-related activation can begin.

For a typical fMRI scan, the 3D volume of the subject's head is imaged every one or two seconds, producing a few hundred to a few thousand complete images per scanning session. The nature of MRI is such that these images are acquired in Fourier transform space, so they must be transformed back to image space to be useful. Because of practical limitations of the scanner the Fourier samples are not acquired on a grid, and scanner imperfections like thermal drift and spike noise introduce additional distortions. Small motions on the part of the subject and the subject's pulse and respiration will also affect the images.

The most common situation is that the researcher uses a pulse sequence supplied by the scanner vendor, such as an echo-planar imaging (EPI) sequence that allows for relatively rapid acquisition of many images [54, 24, 25]. Software in the scanner platform itself then performs the reconstruction of images from Fourier transform space. During this stage some information is lost (specifically the complex phase of the reconstructed signal). Some types of artifacts, for example spike noise, become more difficult to remove after reconstruction, but if the scanner is working well these artifacts are thought to be relatively unimportant. After reconstruction the fMRI data consists a series of 3D images of the brain. The most common corrections performed on these images are motion correction and correction for physiological effects. Outlier correction and spatial and/or temporal filtering may also be performed. A variety of methods are used to correlate these voxel time series with the task in order to produce maps of task-dependent activation. In Fig. 2.3 we see a diagram describing the overall procedure in the analysis of fMRI data, from the design of the experiment until the physiological interpretation of the data.

2.3 fMRI data analysis

In this section we will provide the general scheme for the analysis of fMRI data since this task covers a large part of this thesis.

2.3.1 Experimental design

The choice of the experimental design when setting up a fMRI study depends on the expected results and the target of the research. Two basic types of setups are used when concerning the appropriate and suitable designs: block design, event-related design [24, 25]

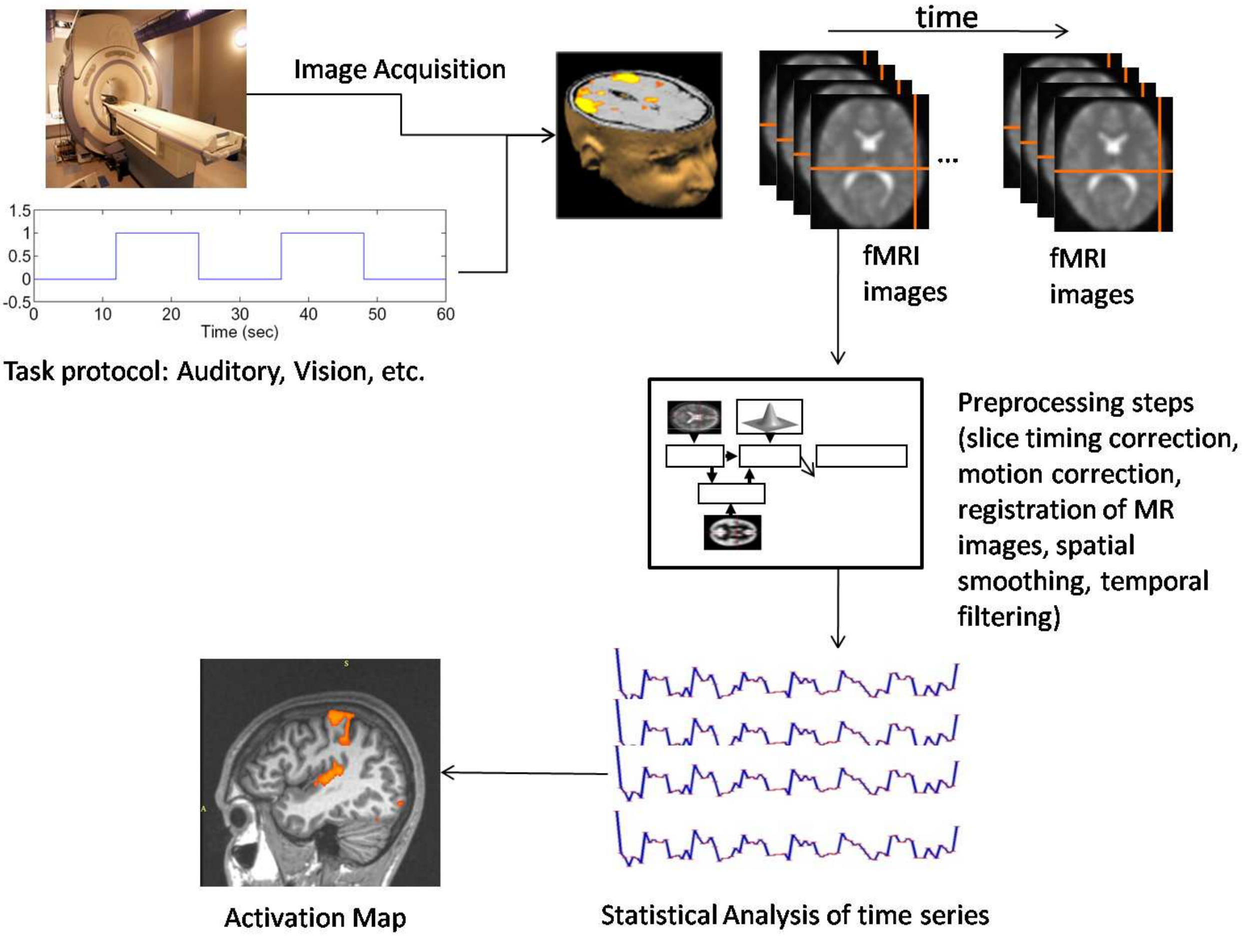

time
Image Acquisition
...
fMRI images
fMRI images
1.5
1
0.5
0
-0.5
0
10
20
30
40
50
60
Time (sec)
Task protocol: Auditory, Vision, etc.
Preprocessing steps (slice timing correction, motion correction, registration of MR images, spatial smoothing, temporal filtering)
Activation Map
Statistical Analysis of time series

Figure 2.3: Overall scheme in fMRI analysis.

or a combination of these two. All these designs have their weaknesses and strengths which should be considered when choosing the design.

Block design

In block design studies the activating stimuli are presented continuously during some time interval that is called a block. The blocks of activating stimuli are usually alternating with the so called baseline or resting blocks. During the baseline block no stimuli are presented. One active block may be consisted of only one long stimulus or several similar stimuli presented rapidly. It is also possible to study different stimuli by presenting each stimuli type in its own block. Therefore, several type of blocks might belong in one study and the order of the blocks may alternate randomly. The duration of the blocks may also vary.

Block designs are still in use nowadays. One reason for this is probably better signal-to-noise ratio (SNR) due to bigger amount of data to be averaged. This also ensures better detection power, to locate active cortical regions. The weakness of block designs is the poor estimation efficiency to estimate the hemodynamic response for a single stimulus. This is basically due to fast presenting rate of the stimuli so the responses overlap with each other. This overlapping is proved to be nonlinear, which complicates the estimation of the shape of the hemodynamic response. Block designs are also experimentally less demanding than more flexible designs. The possible inaccuracies in the experiment design are less serious in block designs than in event-related designs because the responses in one block are averaged.

Event-related design

In PET studies only block designs can be assessed because of the relatively long half-life of the used radioactive tracers. In fMRI, however, the origin of the response to a stimulus can be related to the cerebral hemodynamic changes, and these changes are detectable within seconds of the stimulus onset. The relatively fast response to stimuli enables the use of brief stimuli in studies of brain function.

In event-related design a brief stimuli are presented randomly. The term event related derives from electrophysiology and measuring the event-related potentials (ERPs). The design and the presentation of stimuli in fMRI is quite similar to technique used in measuring the ERPs. The stimuli are no longer presented in blocks of similar stimuli but one type of stimuli can be randomized so that different types of stimuli alter with each other and with baseline. The presentation rate may also vary i.e. a stimulus may occur twice a second or twice a minute. Event-related design has many virtues compared to block design. When the stimuli are presented in blocks, the subject's cognitive behavior may disrupt the response because the subject can guess when the next stimulus is presented and what kind of stimulus it is. The randomization of the stimuli prevents this kind of problems and also habituation. The responses can be post hoc categorized according to

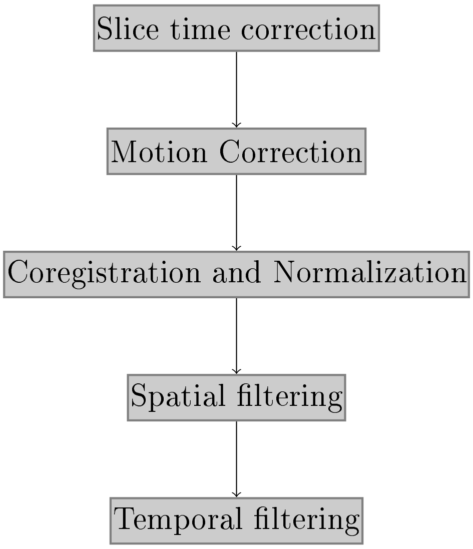


Figure 2.4: Preprocessing steps.

subjects performance and hence it is possible to study the difference between different responses caused by similar stimuli. Another advance of the event-related designs compared to block design is the ability to use the so called oddball-paradigm and study unpredicted stimuli. The advantages of the event-related design over block design encouraged research groups to study and compare the results obtained with both these design types [24, 25].

2.3.2 Preprocessing of fMRI data

Before analyzing the fMRI data several preprocessing steps can be applied in order steps to remove artifacts and validate the assumptions of the model [54, 24, 25, 108]. The main goals of data preprocessing are: a) to minimize the influence of the data acquisition and physiological artifacts, b) to validate the statistical assumptions and c) to standardize the brain regions across subjects. During the analysis of fMRI data is is assumed that all the voxels of the brain are acquired simultaneously and that each data point in a specific voxel's time series consists of a signal from that voxel (i.e. the participant does not move across measurements). Finally, all brains are assumed to be registered, so that each voxel is located in the same anatomical region for all subjects. However, these assumptions don't hold in reality and there is a need to make them more suitable for the statistical model. The major steps of preprocessing are: slice timing correction, realignment, coregistration of images, normalization, spatial smoothing and temporal filtering.

Slice timing correction

When analyzing 3D fMRI data it is typically assumed that the whole brain is measured simultaneously. However, this is not the case because the brain volume consists from multiple slices that are acquired sequentially, and therefore at different time points. Similar time points from different slices are shifted relative to one another. Slice time correction involves the correction of shift so that one can assume they are measured simultaneously. This is achieved by Fourier transforming each voxel's time series into the frequency domain, applying a phase shift to the data, and then applying the inverse Fourier transform to recover the corrected data. However, in the above solution there is a problem due to head motion. In SPM package [105] there is a note that this step will be remove in future.

Motion correction

An important issue involved in any fMRI study is proper handling of any subject movement that may have taken place during data acquisition. When movement occurs, the signal from a specific voxel is contaminated by the signal from neighboring voxels. The first step for motion correction is to find the best possible alignment between the input image and some target image. Usually, motion correction methods assume that the shape of the head does not change shape. This means that the correction involves only translations and rotations (rigid - body transformation). However, non - rigid shape changes can be occur in the brain tissue, for example due to the pulsation of the blood stream.

Coregistration and Normalization

fMRI data provides little anatomical detail. This is problem in the case we want to interpret the analysis results. To overcome this problem we need to map the results from the obtained fMRI data onto high resolution structural MR images. The process of aligning structural and functional images is called coregistration [54, 24, 25] and is performed using a rigid - body or an affine transformation. Also, individual brains have different shapes and features but there exists similarities between the brains. Normalization [54, 24, 25] attempts to register each brain anatomy to a common space defined by a template brain (e.g. the Talairach or Montreal Neurological Institute (MNI) brain). During the normalization non linear transformations are used to match the local features.

Spatial smoothing

It is useful to spatially smoothed the fMRI data prior to statistical analysis. There are several reasons why there is need to smooth the data. First, small amounts of smoothing improves the signal to noise ratio. The second reason is that the smoothing improve the quality of the data for statistical analysis by making them more appropriate for the model. A common approach to smooth the fMRI data is to blur them with a Gaussian filter [54, 24, 25]. The disadvantage of spatial smoothing is that we don't know if the size of the filter is the appropriate. Also, smoothing can also cause the merging of brain regions that are functionally different. These problems guide many researchers to investigate ways of combining spatial information in more sophisticated ways than simple blurring.

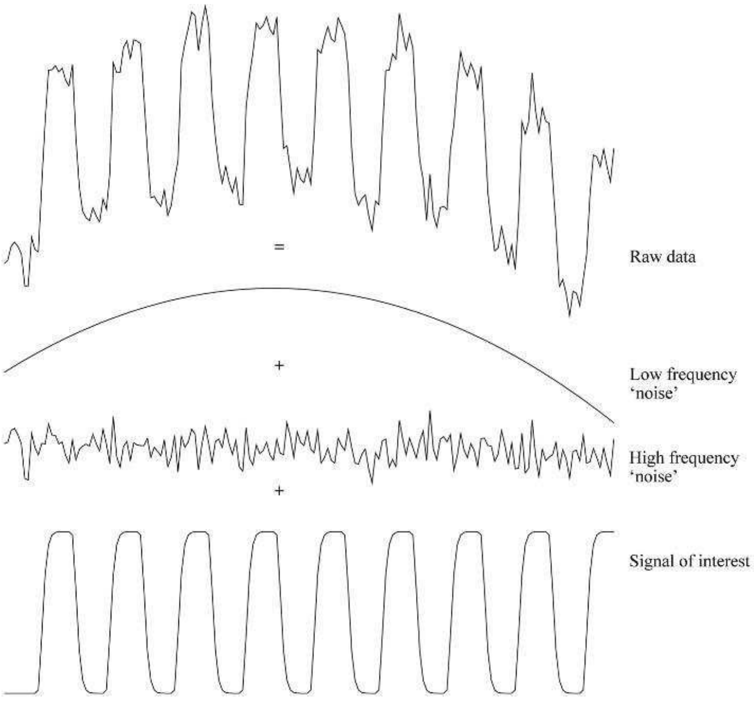


Figure 2.5: An example unfiltered time series from an activated voxel (reprinted from [24]).

Temporal filtering

Temporal filtering, instead of working in each image, such as the spatial smoothing, works in each voxel's time series. The main point of temporal filtering is to remove the unwanted components of a time series, without damaging the signal of interest. For example, if we applied a stimulation for 30 sec, followed by 30 sec rest, and this pattern is applied many times then the signal of interest is close to a square waveform of 60 sec period. Temporal filtering will attempt to remove components of the time series that are more slowly (high pass filtering) or more quickly varying (low pass filtering) than this 60 sec periodic signal. In Fig. 2.5, we show an example time series, decomposed into the various signal components. Temporal filtering is carried out using linear filters, such as FIR filter for high pass filtering and a Gaussian filter for low pass filtering [54, 24, 25]. We must have in mind that most statistical models are applied directly on voxel time series, so many aspects of temporal filtering can be incorporated into the statistical model [32].

While all the preprocessing steps outlined above are essential to the analysis of fMRI data, there is need to be a clear understanding of the effects they have on both the spatial and temporal correlation structure. More generally, it is necessary to study the interactions among the individual preprocessing steps. For example, is it better to perform slice timing correction first or realignment, and how this will impact the resulting data Ideally we want a model for both [108]. In last years there is a growing interest for generative models that incorporate many multiple steps at once.

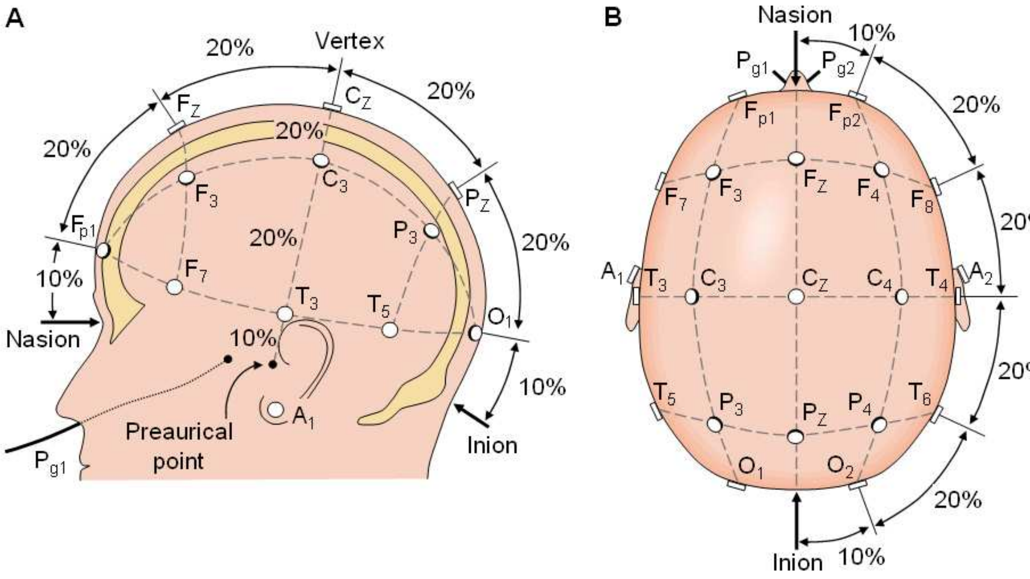


Figure 2.6: The international 10-20 system seen from (A) left and (B) above the head. (reprinted from [8])

2.4 Electroencephalogram

The electroencephalography concerns the recording and the interpretation of the electroencephalogram. Electroencephalogram (EEG) is a record of the electric signal generated by the cooperative action of brain cells, or more precisely, the time course of extracellular field potentials generated by their synchronous action. EEG can be measured by means of electrodes placed on the scalp or directly on the cortex. EEG recorded in the absence of an external stimulus is called spontaneous EEG, while if it is generated as a response to external or internal stimulus will be called event-related potential (ERP).

The EEG recording is obtained through electrodes located on the scalp, where some of them are used as references. Reference electrodes are either located on the scalp or on other parts of the body, e.g., the ear lobes. To ensure reproducibility among studies an international system for electrode placement, the 10-20 international system [9], has been defined (Fig. 2.6). It is based on anatomical location and on percentage of distance among these points giving the 10 or 20% in the system name. The original 10-20 system has only nineteen electrodes but has been extended to accommodate more than 200 electrodes. In this system the electrodes' locations are related to specific brain areas. For example, electrodes C3 and C4 are above the motor cortex. Each EEG signal can therefore be correlated to an underlying brain area. Of course this is only a broad approximation that depends on the accuracy of the electrode's placement.

The electroencephalogram can be roughly defined as the signal which corresponds to the mean electrical activity of the brain in different locations of the head. More specifically, it is the sum of the extracellular current flows in a large group of neurons. It can be acquired using either intracranial electrodes inside the brain or scalp electrodes on the surface of the head [9]. The EEG has been found to be a valuable tool in the diagnosis of numerous brain disorders. Nowadays, the EEG recording is a routine clinical procedure and is widely regarded as the physiological "gold standard" to monitor and quantify levels of drowsiness and wakefulness but also for detection of epileptic spikes and seizures and generally for the diagnosis of epilepsy [20]. The electric activity of the brain is usually divided into three categories: 1) bioelectric events produced by single neurons, 2) spontaneous activity, and 3) evoked potentials. EEG spontaneous activity is measured on the scalp or on the brain. Clinically meaningfull frequencies lie between 0.1Hz and 100Hz. In more restricted sense, the frequency range is classified into several frequency components, or delta rhythm (δ: 0.5-4Hz), theta rhythm (θ: 4-8Hz), alpha rhythm (α: 8-13Hz), beta rhythm (β: 13-30Hz), and gamma rhythm (γ: 30-60Hz) [9].

The properties of the EEG signal are complex [9, 19], due to the intricate neural system. Traditionally, the spontaneous EEG is characterized as a linear stochastic process with similarities to noise. From the signal processing view, EEG has the following properties [19]: (a) Noisy and pseudo-stochastic: The EEG is often between 10-300µV, which is easily affected by various physiological and electrical noises. Meanwhile, artefacts from electrocardiogram (ECG), electrooculogram (EOG), electromyogram (EMG), and recording systems can also contaminate the signals. Even the EEG shows a high degree of randomness and nonstationarity. (b) Time-varying and nonstationary: EEG is not a stationary process; it varies with the physiological states. The waveforms may include a complex of regular sinusoidal waves, irregular spikes/polyspikes, or spindles/polyspindles. In most pathological conditions, such as epileptic seizures, the EEG may show evident singularity or nonstationarity. In practice, EEG is considered as a stationary process over a relatively short period (approximately 3.5sec for routine spontaneous EEG) [18]. (c) High nonlinearity: Although the traditional linear models of EEG still play significant roles in EEG analysis and diagnosis, EEG is a nonlinear process [10]. This kind of nonlinearity is also time-, state-, and site-dependent [15].

One of the most important challenges of EEG analysis is the quantification of the manifestations of epilepsy [9, 19]. The main goal is to establish a correlation between the EEG and clinical or pharmacological conditions. One of the possible approaches is based on the properties of the inter-ictal EEG (electrical activity measured between seizures), which typically consists of linear stochastic background fluctuations interspersed with transient nonlinear spikes, sharp waves or spikes-and-wave complexes [20]. These transient potentials originate as a result of a simultaneous pathological discharge of neurons within a volume of at least several mm3. The traditional definition of a spike is based on its amplitude, duration, sharpness, and emergence from its background [21]. However, automatic epileptic spike detection systems based on this direct approach suffer from false detections in the presence of numerous types of artefacts and non-epileptic transients [20, 21]. This shortcoming is particularly acute for long-term EEG monitoring of epileptic patients, which became common in 1980s [22, 23]. In Fig. (2.7) we see an EEG segment contains four epileptic spikes.

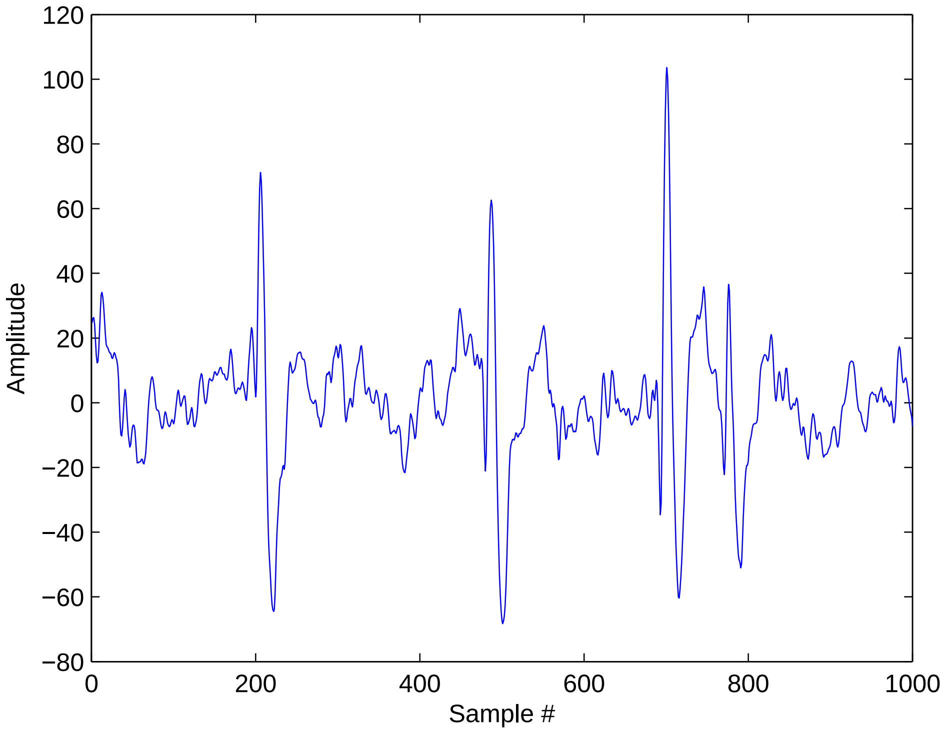


Figure 2.7: EEG signal contains four spikes.

There has also been a challenge to find functional cerebral activation indices for cognitive processes involved in a given task. The EEG is a continuous measure over time and can be used to study ongoing activity in the brain while subjects perform long-lasting and/or variable tasks. The alpha rhythm of the EEG is predominantly observed over the posterior cortex [17]. This rhythm correlates with relaxation, and for this reason it has been interpreted as a sign of inhibition of activity in the areas over which it has been recorded. Activation of the cortex causes a desynchronization of the alpha band, i.e. its amplitude decreases, while alpha synchronization denotes the increase of alpha activity ([13, 12]. When alpha desynchronization or synchronization is related to an internally or externally paced event, it is called as event-related desynchronization (ERD) [11] or event-related synchronization (ERS), respectively. The quantification of ERD/ERS requires the comparison of two different experimental conditions. ERD and ERS are defined as the relative difference in the EEG alpha power between the reference recorded before each event and the actual event. ERD/ERS is, thus, a 'within-subject' measure of cortical activation and is expressed as a percentage. ERD and ERS can be either externally (by stimuli) or internally (by voluntary behavior) paced and they have a specific topographical distribution depending upon the state of the brain, stimulus paradigm and modality [12]. ERD has been observed e.g. during complex auditory stimulation [16], during cognitive and attentional tasks, and during voluntary movement tasks [11]. The ERD/ERS of the lower alpha frequencies (8-10 Hz) has been claimed to reflect non-specific cognitive

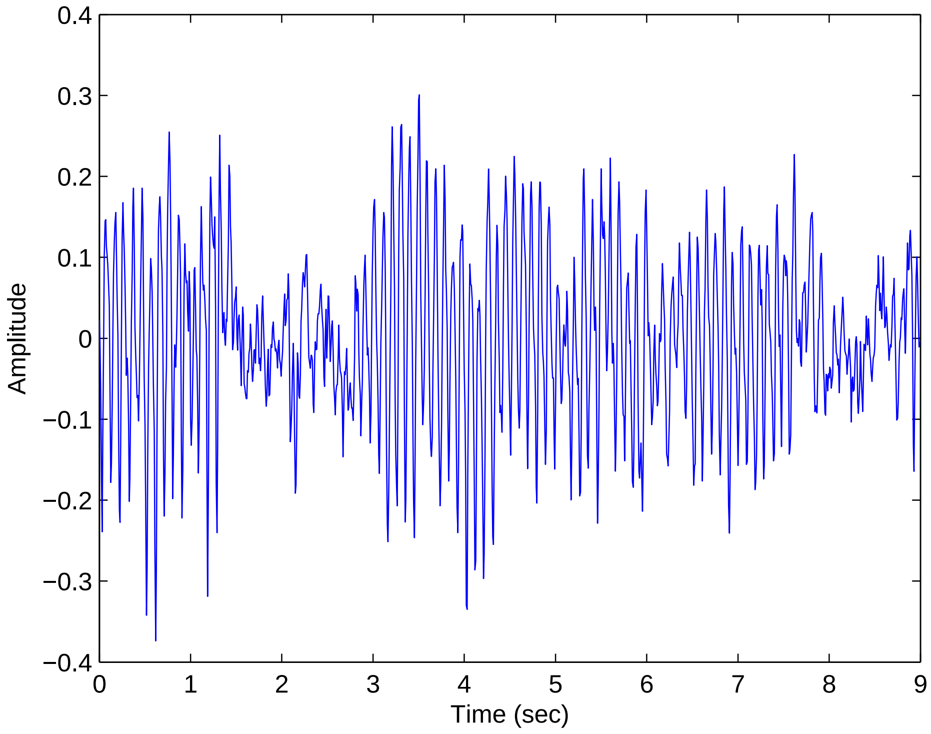


Figure 2.8: ERD/ERS signal.

functions, such as sustained attention, while that of the upper alpha frequencies (10-12 Hz) appears to reflect stimulus-related, i.e. task-specific cognitive processes. An example of EEG segment where we wish to find the ERD/ERS phenomenon is shown in Fig. (2.8).

Event-related potentials

Event-related potentials (ERPs) are the changes of spontaneous EEG activity related to a specific event [145]. ERPs triggered by particular stimuli, visual (VEP), auditory (AEP), or somatosensory (SEP), are called evoked potentials (EP). It is assumed that ERPs are generated by activation of specific neural populations, time-locked to the stimulus, or that they occur as the result of reorganization of ongoing EEG activity. The basic problem in analysis of ERPs is their detection within the larger EEG activity. ERP amplitudes are an order of magnitude smaller than that of the ongoing EEG. Averaging is a common technique in ERP analysis; it makes possible the reduction of background EEG noise. However, assumptions underlying the averaging procedure, namely (1) the background noise is a random process, (2) the ERP is deterministic and repeatable, and (3) EEG and ERP are independent, are not well justified. The ERP pattern depends on the nature of the stimulation, placement of the recording electrode, and the actual state of the brain. ERPs are usually described in terms of the amplitudes and latencies of their characteristic waves [176]. In Fig. (2.9) we see two trials of ERPs signal from channel Pz. The stimuli is presented at the time instant equal to 1 sec. At this time we also observe

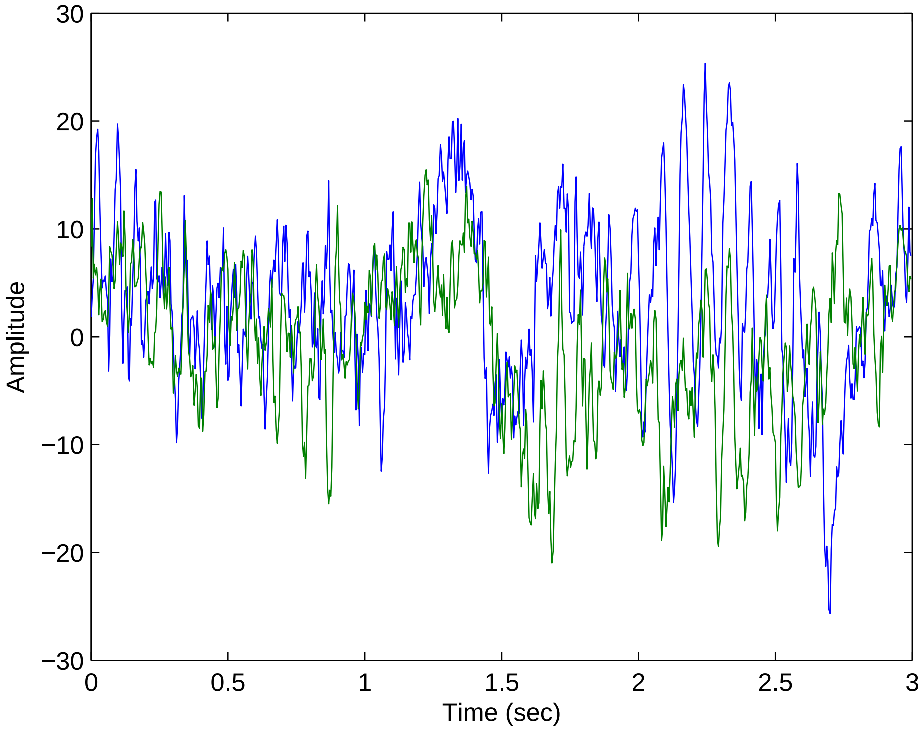


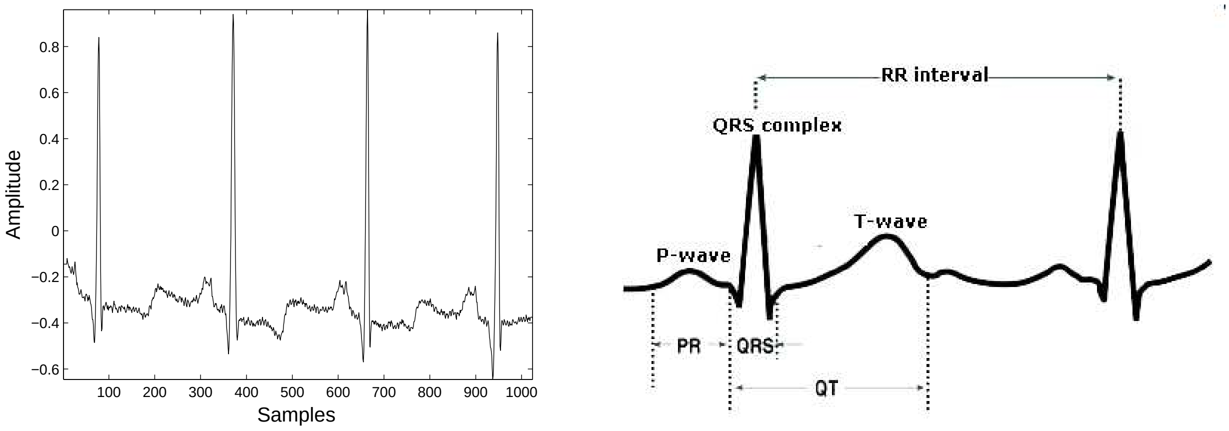
Figure 2.9: Two ERPs signals.

the distortion at the baseline.

2.5 Electrocardiogram and Heart Rate Variability

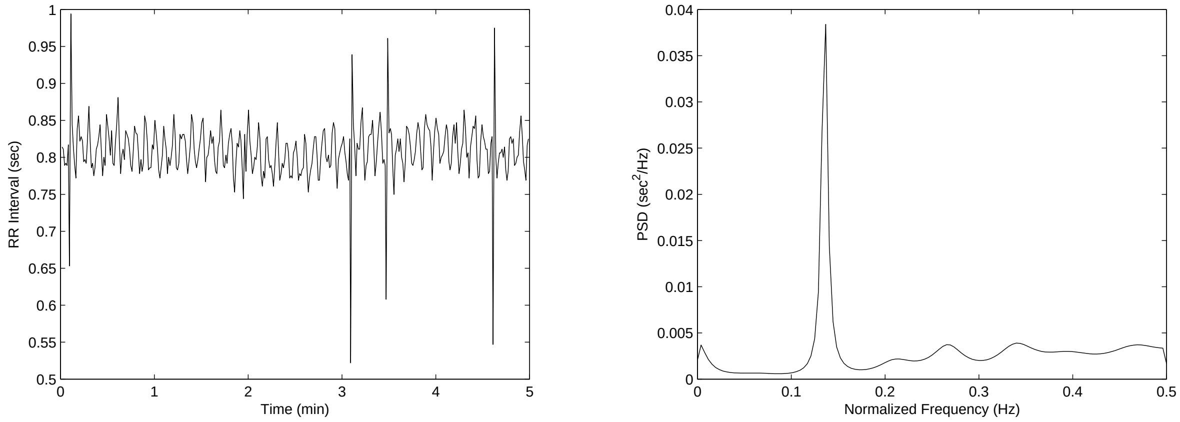
The electrical activity of the heart can be characterized by measurements acquired at the cellular level or from the body surface. The electrocardiogram (ECG) describes the electrical activity of the heart recorded by electrodes placed on the body surface. The voltage variations measured by the electrodes are caused by the action potentials of the cardiac cells. The resulting heartbeat is recorded to the ECG and consist of a series of waveforms whose morphology and timing convey information which are used for diagnosing diseases that are reflected by changes of the heart's electrical activity. In Fig. 2.10 a segment of ECG signal and the characteristics of a heart beat are illustrated.

The Heart Rate Variability (HRV) signal is obtained from the electrocardiogram (ECG) and describes the variations between consecutive cardiac beats. Studies have shown that this signal originates from the Autonomous Nervous System (ANS) [1]. Also, the HRV signal is strongly connected to respiration and blood pressure [2]. Thus the HRV signal can be used as a quantitative marker of the ANS and HRV parameters are used to evaluate the clinical condition of subjects in normal or pathological conditions. The HRV analysis methods can be divided into time-domain, frequency-domain, and nonlin-



(a) (b)

Figure 2.10: (a) ECG segment contains four beats (b) Characteristics of a heart beat.



(a) (b)

Figure 2.11: (a) HRV time series (b) Spectrum of HRV time series

ear methods [3]. The analysis of HRV is performed by studying various measures of the signal such as the standard deviation between two normal beats (SDNN) and the power spectral density (PSD) [4, 5]. However, the HRV signal contains artefacts which can be originate from other physiological processes, such as the breathing pattern of a human, or technical dysfunctions, such as QRS misdetection [6, 7]. These artefacts distort the HRV signal leading in erroneous calculations of various statistical measures. Usually, the HRV signal contains two oscillating components, the Low Frequency (LF) components and the High Frequency (HF) component. The LF component appears in the frequency range 0.05 - 0.2 Hz and the HF component appears in the frequency range 0.2-0.4 Hz. In Fig. (2.11) we see an example of HRV time series and the PSD of it.

2.6 Machine Learning approaches for signal processing

2.6.1 Bayesian inference

Bayesian inference provides a mathematical framework that can be used for modeling, where the uncertainties of the system are taken into account and the decisions are made according to logical principles. These main tools are random variables, the probability distributions and the rules of probability calculus.

Consider a dataset contains N samples, $Y = \{\mathbf{y}_n\}_{n=1}^N$, where we assume a distribution over them $p(\mathbf{y}_n|\theta)$, where θ is the set of parameters which are unknown and must be estimated. The choice of this distribution is very important since it must suite with the nature and particular characteristics of the observations. By assuming independent and identically distributed (i.i.d.) samples, a classical prcedure for estimating the parameter is through the Maximum Likelihood (ML) framework, where we maximize the joint probability of measurements, also called likelihood function:

$$L(\theta) = p(Y|\theta) = \prod_{n=1}^{N} p(\mathbf{y}_n|\theta)$$

For analytical purposes, it is easier to work with the logarithm of the likelihood function, because the logarithm is monotonically increasing, and thus maximizing the log-likelihood is equivalent to maximizing the likelihood. We can write more formal the ML estimation procedure as:

$$\theta_{ML} = \arg\max_{\theta}\{\log p(Y|\theta)\} = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\mathbf{y}_n|\theta).$$

The difference between the Bayesian inference and the ML method is that the former consider the parameters θ as a random variable. Then, the posterior distribution of parameters is computed by using the Bayes' rule:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} \tag{2.1}$$

where $p(\theta)$ is the prior distribution, which models the prior beliefs of parameters before we observe any measurement and $p(Y)$ is a normalization constant, independent of the parameter. In most situations the normalization constant is left out and since the measurements are conditionally independent given the parameters, the posterior distribution for parameters is written as:

$$p(\theta|Y) \propto p(Y|\theta)p(\theta) = \Big[\prod_{n=1}^{N} p(\mathbf{y}_n|\theta)\Big]p(\theta) \tag{2.2}$$

Now, that we have obtained the posterior distribution, we can use the most probable value as an estimate for parameters (Maximum A Posteriori estimate), which is given by the maximum of the posterior. Also, a candidate estimate is the posterior mean of parameters (MMSE estimate). There are many ways of choosing the point estimate from the distribution and the best way depends on the assumed loss function [40]. It is easy to see that ML estimate is equivalent to a MAP estimate when it is assume a uniform prior distribution over the parameter θ.

The basic components of a Bayesian model is the prior model encapsulating a preliminary knowledge of the shape and the range values of the parameters and the likelihood as a function.

Prior distribution The prior information consists of beliefs about the possible and impossible parameters values and their relative likelihoods before anything has been seen. The prior distribution is a mathematical representation of this information:

$$p(\theta) = \text{Information on parameter } \theta \text{ before arises any observations.}$$

The lack of prior information can be expressed by using a non-informative prior [125, 121].

Likelihood function Between the measurements and the parameters there is a noisy or inaccurate relationship. This relationship is modeled using the likelihood distribution:

$$p(\mathbf{y}|\theta) = \text{Distribution of observation } \mathbf{y} \text{ given the parameter } \theta.$$

Posterior Posterior distribution is the conditional distribution of parameters given the observation $\mathbf{y}$ and represents the information that we have after the observation $\mathbf{y}$ has been obtained. It can be computed by using the Bayes' rule:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} \tag{2.3}$$

where the normalization constant is given by:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta. \tag{2.4}$$

In the case we have multiple observations $Y = \{\mathbf{y}_n\}_{n=1}^N$ which are conditionally independent, the posterior distribution becomes:

$$p(\theta|Y) \propto \Big[\prod_{n=1}^{N} p(\mathbf{y}_n|\theta)\Big]p(\theta) \tag{2.5}$$

where the normalization term can be computed by integrating the right hand side over θ. If parameters are discrete variables then the integration is replaced by summation.

Predictive posterior distribution The predictive distribution is the distribution of the new observation $\mathbf{y}_{N+1}$:

$$p(\mathbf{y}_{N+1}|\mathbf{y}_1, \cdots, \mathbf{y}_N) = \int p(\mathbf{y}_{N+1}|\theta)p(\theta|\mathbf{y}_1, \cdots, \mathbf{y}_N)d\theta. \tag{2.6}$$

The predictive distribution can be used for computing the probability distribution of the $(N+1)^{th}$ observation, which has not been observed yet.

Maximum A Posteriori estimation

In the case we have a prior distribution over the parameters a simple approach is to use the Maximum A Posteriori (MAP) estimator. The MAP estimator is obtained by performing the following maximization:

$$\theta_{MAP} = \arg\max_{\theta}\{\log p(Y|\theta) + \log p(\theta)\}\ . \tag{2.7}$$

This estimator chooses the model with highest posterior probability density (the posterior mode). This approach provides us with point estimates, which contain the prior information, and can be seen as a penalized maximum likelihood estimator in the classical sense [121]. We can observe that as the sample size goes to infinity, $N \to \infty$, the likelihood function dominates over the prior distribution $p(\theta)$. Therefore, the MAP estimator is asymptotically equivalent to the ML estimator [121].

2.6.2 Expectation Maximization (EM) algorithm

The Expectation-Maximization (EM) algorithm introduced by Dempster et al [205] is a general method to solve ML estimation problems. The EM algorithm is the basis of many learning algorithms [45]. The objective of the algorithm is to maximize the likelihood of the observed data in the presence of hidden variables. Let us denote the observed data by $\mathbf{y}$, the hidden variables by $\mathbf{x}$ and the parameters of the model by θ. Maximizing the likelihood as a function of θ is equivalent to maximizing the log-likelihood:

$$L(\theta) = \log p(\mathbf{y}|\theta) = \log \int p(\mathbf{y},\mathbf{x}|\theta)d\mathbf{x} \tag{2.8}$$

Using any distribution q over the hidden variables, we can obtain a lower bound on L:

$$\begin{aligned}
\log \int p(\mathbf{y},\mathbf{x}|\theta) &= \log \int q(\mathbf{x}) \frac{p(\mathbf{y},\mathbf{x}|\theta)}{q(\mathbf{x})} d\mathbf{x} \\
&\geq \int q(\mathbf{x}) \log \frac{p(\mathbf{y},\mathbf{x}|\theta)}{q(\mathbf{x})} d\mathbf{x} \\
&= \int q(\mathbf{x}) \log p(\mathbf{y},\mathbf{x}|\theta) d\mathbf{x} - \int q(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x} \qquad (2.9) \\
&= F(q,\theta). \qquad (2.10)
\end{aligned}$$

The EM algorithm alternates between maximizing F with respect to the distribution q and the parameters θ, respectively, holding the other fixed.

$$\textbf{E-step:} \qquad q_{k+1} \leftarrow \arg\max_q F(q,\theta^k) \tag{2.11}$$

$$\textbf{M-step:} \qquad \theta^{k+1} \leftarrow \arg\max_\theta F(q_{k+1},\theta) \tag{2.12}$$

It is easy to show that the maximum in the E-step results when q is exactly the posterior distribution of the hidden variables $\mathbf{x}$, $q(\mathbf{x}) = p(\mathbf{x}|\mathbf{y},\theta^k)$, at which point the bound becomes an equality $F(q_{k+1}, \theta^k) = L(\theta^k)$. The maximum in the M-step is obtained by maximizing the first term in Eq. (2.9), since the entropy of q does not depend on θ:

$$\textbf{M-step: } \theta^{k+1} \leftarrow \arg\max_{\theta} \int p(\mathbf{x}|\mathbf{y}, \theta^k) \log p(\mathbf{y}, \mathbf{x}|\theta) d\mathbf{x} \tag{2.13}$$

At each iteration the EM algorithm guarantees that the log-likelihood does not decreased, $L(\theta^{t+1}) - L(\theta^t) \geq 0$. From Bayes' rule we have that:

$$\begin{aligned} \log p(\mathbf{x}|\mathbf{y}, \theta^{t+1}) &= \log p(\mathbf{y}, \mathbf{x}|\theta^{t+1}) - \log p(\mathbf{y}|\theta^{t+1}) \\ &= \log p(\mathbf{y}, \mathbf{x}|\theta^{t+1}) - L(\theta^{t+1}) \end{aligned} \tag{2.14}$$

Taking the expectation with respect to $q(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}, \theta^k)$ we obtain:

$$< \log p(\mathbf{x}|\mathbf{y}, \theta^{t+1}) >_q = < \log p(\mathbf{y}, \mathbf{x}|\theta^{t+1}) >_q - L(\theta^{t+1}) \tag{2.15}$$

since $< L(\theta^{t+1}) >_q = L(\theta^{t+1})$ because $L(\theta^{t+1})$ does not depends from hidden variables. The same holds for the term $L(\theta^t)$, i.e.

$$< \log p(\mathbf{x}|\mathbf{y}, \theta^{t}) >_q = < \log p(\mathbf{y}, \mathbf{x}|\theta^{t}) >_q - L(\theta^{t}) \tag{2.16}$$

From the above two equation we can obtain:

$$\begin{aligned} L(\theta^{t+1}) - L(\theta^t) &= - < \log p(\mathbf{x}|\mathbf{y}, \theta^{t+1}) >_q + < \log p(\mathbf{x}|\mathbf{y}, \theta^{t}) >_q \\ &\quad + < \log p(\mathbf{y}, \mathbf{x}|\theta^{t+1}) >_q - < \log p(\mathbf{y}, \mathbf{x}|\theta^{t}) >_q \end{aligned} \tag{2.17}$$

From the M - step, we have $< \log p(\mathbf{y}, \mathbf{x}|\theta^{t+1}) >_q - < \log p(\mathbf{y}, \mathbf{x}|\theta^{t}) >_q \geq 0$. Also, the difference $- < \log p(\mathbf{x}|\mathbf{y}, \theta^{t+1}) >_q + < \log p(\mathbf{x}|\mathbf{y}, \theta^{t}) >_q$ is the KL divergence between the distribution $p(\mathbf{x}|\mathbf{y}, \theta^{t+1})$ and $p(\mathbf{x}|\mathbf{y}, \theta^{t})$ and is greater or equal to zero. So, in each iteration of the EM we have $L(\theta^{t+1}) - L(\theta^t) \geq 0$.

The EM algorithm performs the M - step based on the ML estimator. However, we can change slightly this step to include the prior distribution of the parameters. Based on the MAP learning approach and the EM algorithm we are able to derive an EM-MAP algorithm [42] where the M-step of the classical EM is replaced by:

$$\textbf{M-step: } \theta^{k+1} \leftarrow \arg\max_{\theta} \Big\{ \log p(\theta) + \int p(\mathbf{x}|\mathbf{y}, \theta^k) \log p(\mathbf{y}, \mathbf{x}|\theta) d\mathbf{x} \Big\} \tag{2.18}$$

2.6.3 Variational Bayesian methodology

Variational Bayesian (VB) methodology is an approximate inference technique that proceeds by assuming an arbitrary approximation for the posterior distribution and inference is made using a EM-like algorithm. A brief introduction of the VB methodology follows. For more information on this subject one can see at [42, 45].

Assuming an arbitrary distribution for the hidden variables $\mathbf{x}$ and the model parameters θ $q(\mathbf{x}, \theta)$ the log of the evidence or the marginal likelihood can be written as:

$$\begin{aligned}
\log p(\mathbf{y}) &= \int q(\mathbf{x}, \theta) \log p(\mathbf{y}) d\theta d\mathbf{x} \\
&= \int q(\mathbf{x}, \theta) \log \big(p(\mathbf{y}) \frac{p(\mathbf{y}, \mathbf{x}, \theta)}{p(\mathbf{y}, \mathbf{x}, \theta)}\big) d\theta d\mathbf{x} \\
&= \int q(\mathbf{x}, \theta) \log \frac{p(\mathbf{y}, \mathbf{x}, \theta)}{p(\mathbf{x}, \theta|\mathbf{y})} d\theta d\mathbf{x} \\
&= \int q(\mathbf{x}, \theta) \log \frac{p(\mathbf{y}, \mathbf{x}, \theta)}{q(\mathbf{x}, \theta)} d\theta d\mathbf{x} \\
&\quad + \int q(\mathbf{x}, \theta) \log \frac{q(\mathbf{x}, \theta)}{p(\mathbf{x}, \theta|\mathbf{y})} d\theta d\mathbf{x} \\
&= F(q, \mathbf{x}, \theta) + KL(q(\mathbf{x}, \theta)||p(\mathbf{x}, \theta|\mathbf{y})). \qquad (2.19)
\end{aligned}$$

Maximizing $F(q, \mathbf{x}, \theta)$ is equivalent to minimizing the KL divergence between the true posterior and the arbitrary distribution $q(\cdot)$, which can be used as an approximation to the true posterior. The variational free energy $F(q, \mathbf{x}, \theta)$ is evaluated as:

$$\begin{aligned}
F(q, \mathbf{x}, \theta) &= \int q(\mathbf{x}, \theta) \log \frac{p(\mathbf{y}, \mathbf{x}, \theta)}{q(\mathbf{x}, \theta)} d\theta d\mathbf{x} \\
&= \int q(\mathbf{x}, \theta) \log \frac{p(\mathbf{y}|\mathbf{x}, \theta) p(\mathbf{x}, \theta)}{q(\mathbf{x}, \theta)} d\theta d\mathbf{x} \\
&= \int q(\mathbf{x}, \theta) \log p(\mathbf{y}|\mathbf{x}, \theta) d\theta d\mathbf{x} \\
&\quad - \int q(\mathbf{x}, \theta) \log \frac{q(\mathbf{x}, \theta)}{p(\mathbf{x}, \theta)} d\theta d\mathbf{x} \\
&= < \log p(\mathbf{y}|\mathbf{x}, \theta) >_{q(\mathbf{x}, \theta)} \\
&\quad - KL(q(\mathbf{x}, \theta)||p(\mathbf{x}, \theta)), \qquad (2.20)
\end{aligned}$$

where $< \cdot >_{q(\mathbf{x}, \theta)}$ is the expectation with respect to the approximate posterior of the parameters $\mathbf{x}$ and θ. We mention here that the KL divergence in Eq. (2.19) is between the approximate posterior of parameters and the true posterior, while in Eq. (2.20) is between the approximate posterior of the parameters and the prior of the parameters.

The goal in a variational approach is to choose a suitable form of $q(\mathbf{x}, \theta)$ so that the lower bound can be evaluated. In general, we choose a family of q-distributions and we seek the best approximation within this family by maximizing the lower bound. Since the true log-likelihood is independent of q this is equivalent to the minimization of the KL divergence. The KL divergence between the two distributions $q(\mathbf{x}, \theta)$ and $p(\mathbf{x}, \theta|\mathbf{y})$ is minimized when $q(\mathbf{x}, \theta) = p(\mathbf{x}, \theta|\mathbf{y})$ and, thus, the optimal solution for $q(\mathbf{x}, \theta)$ is the true posterior. This solution does not simplify the problem, so to make progress we consider a more restricted range of q-distribution. One approach is to consider a parametric form for $q(\mathbf{x}, \theta)$ such that $q(\mathbf{x}, \theta, \phi)$ is governed by a set of parameters ϕ [41]. We then minimize the KL divergence with respect to ϕ, finding the best approximation within this family.

An alternative approach is to restrict the functional form of $q(\mathbf{x},\theta)$ by assuming that it factorizes over the component variables in $\mathbf{x},\theta$ [42]:

$$q(\mathbf{x},\theta) = q(\mathbf{x})\prod_i q_i(\theta_i). \tag{2.21}$$

Minimizing the KL divergence over all the factorial distributions $q(\mathbf{x})$ and $q_i(\theta_i)$, we obtain:

$$\begin{aligned} q(\mathbf{x}) &\propto \exp < \ln p(\mathbf{y},\mathbf{x},\theta) >_{q(\theta)}, & (2.22)\\ q_i(\theta_i) &\propto \exp < \ln p(\mathbf{y},\mathbf{x},\theta) >_{q(\mathbf{x})q(\theta_{k\neq i})}, & (2.23)\end{aligned}$$

where $< \cdot >_{q(\cdot)}$ denotes expectation with respect to the distribution $q(\cdot)$. The above two equations consist the VB - E step and VB - M step respectively.

2.6.4 Sampling techniques

As we have seen the Bayesian inference includes calculations of very complicated integrals. A class of methods, that is used to calculate such integrals, is based on sampling techniques. These methods are applied to the computation of the evidence, the marginal density and moments and expectations. One such approach is the Monte Carlo integration method [177, 178].

The Monte Carlo integration method estimates the expectation of a function $\phi(\mathbf{y})$ under a probability distribution $p(\mathbf{y})$, by taking samples $\{\mathbf{y}^{(n)}\}_{n=1}^{N}$: $\mathbf{y}^{(n)} \sim p(\mathbf{y})$. An unbiased estimate, $\hat{\phi}$, of the expectation of $\phi(\mathbf{y})$ under $p(\mathbf{y})$, using N samples is given by:

$$\begin{aligned} \hat{\phi} &= \int \phi(\mathbf{y})p(\mathbf{y})d\mathbf{y}\\ &\approx \frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{y}^{(n)}) \end{aligned} \tag{2.24}$$

The Monte Carlo method returns more accurate and reliable estimates the more samples are taken. In cases where we can not produces samples from $p(\mathbf{y})$, we can use another probability distribution $q(\mathbf{y})$, where we can perform sampling, and correct for this by weighting the samples accordingly. This method is called importance sampling [177, 178]. The estimator is given by:

$$\begin{aligned} \hat{\phi} &= \int \phi(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y}\\ &\approx \frac{1}{N}\sum_{n=1}^{N}w^{(n)}\phi(\mathbf{y}^{(n)}) \end{aligned} \tag{2.25}$$

where $w^{(n)}$ are called the importance weights and are given by:

$$w^{(n)} = \frac{p(\mathbf{y}^{(n)})}{q(\mathbf{y}^{(n)})}. \tag{2.26}$$

Extension of the above approaches is the Markov Chain Monte Carlo techniques [177, 178]. Sampling methods based on the Markov Chains were first developed for applications in statistical physics. The classic paper of Metropolis [212] introduced what is now known as Metropolis algorithm. This method was popularized for Bayesian applications in the influential paper of Geman and Geman [98], who applied in image processing problems.

MCMC is Monte Carlo integration using Markov Chains [177]. As described above, Monte Carlo integration draw samples from the required distribution, and then forms sample averages to approximate expectations. Markov Chain Monte Carlo draws these samples by a more clever way based on Markov Chain. Suppose we generate a sequence of random variables $\{x_0, x_1, \cdots, x_N\}$ such that at each time $t \geq 0$ the next sample x_{t+1} is sampled from a distribution $k(x_{t+1}|x_t)$ which depends only on the current state. We see that the next state does not depends further form the history of sequence given the current state. This sequence is called Markov Chain and $k(\cdot|\cdot)$ is the transition kernel of the chain. It can be shown that after the passing of time the chain will forget the initial state and the transition kernel will converge to a unique stationary distribution $f(\cdot)$, which does not depends on time or the initial state [177]. Thus as time increases the samples $\{x_t\}$ will look like samples from $f(\cdot)$. Assuming that the converge to the stationary distribution is achieved after m iterations then we can obtain the samples $\{x_t, t = m, \cdots, N\}$ giving the estimator:

$$\hat{\phi} = \frac{1}{N-m} \sum_{t=m+1}^{N} \phi(x_t). \tag{2.27}$$

Eq. (2.27) show how a Markov Chain is used to compute the expectation when we have obtain the stationary distribution $f(\cdot)$. Now, the interesting part is how to construct a Markov Chain where its stationary distribution is precisely our distribution of interest $p(\cdot)$.

A useful method for this purpose is the Metropolis - Hastings (MH) algorithm [177]. For the MH algorithm at each time t, the next state x_{t+1} is chosen by first sampling a candidate point y from a proposal distribution $q(\cdot|x_t)$. The candidate point is then accepted with probability

$$a(x_t, y) = \min\Big(1, \frac{p(y)q(x_t|y)}{p(x_t)q(y|x_t)}\Big). \tag{2.28}$$

If the candidate is accepted then the next state becomes $x_{t+1} = y$ else the chain remains at the current state, $x_{t+1} = x_t$.

2.6.5 Mixture models

A mixture model is a linear combination of probability density functions of different sources and it is formulated as:

$$p(\mathbf{y}|\Theta) = \sum_{k=1}^{K} \pi_k p(\mathbf{y}|\theta_k) \tag{2.29}$$

where K is the number of mixture components, π_k are the mixing weights, $p(\mathbf{y}|\theta_k)$ are the component density functions with parameters θ_k and $\Theta = \{\pi_k, \theta_k\}_{k=1}^K$ is the set of parameters. Obviously, the component densities may be of different parametric form. The mixing coefficients π_k must satisfy the constraints:

$$0 \leq \pi_k \leq 1 \tag{2.30}$$

and

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2.31}$$

Suppose we have a set of observations $Y = \{\mathbf{y}_n, n = 1 \cdots, N\}$ and we want to model it with a mixture model. Assuming that the samples are drawn independently the likelihood of the data is given by:

$$p(Y|\Theta) = \prod_{n=1}^{N} \Big(\sum_{k=1}^{K} \pi_k p(\mathbf{y}|\theta_k) \Big) \tag{2.32}$$

By taking the logarithm we obtain:

$$\log p(Y|\Theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(\mathbf{y}|\theta_k) \tag{2.33}$$

Maximizing the above log-likelihood function is a difficult problem because the logarithm acts to summation and not directly over the Gaussian density. To overcome this problem the EM algorithm provides a useful framework for solving the optimization problem.

Let us introduce a binary vector $\mathbf{z}$ of dimension $K \times 1$ where a particular element of this z_k is equal to 1 and all others elements are zero. The values of z_k satisfy two conditions: $z_k \in \{0, 1\}$ and $\sum_{k=1}^{K} z_k = 1$. It is easy to see that there is K different conditions for the vector $\mathbf{z}$ depending to which element is nonzero. The marginal distribution of $\mathbf{z}$ is specified in term of mixing coefficients:

$$p(z_k = 1) = \pi_k$$

Because the $\mathbf{z}$ is a binary variable we have

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

Now, the conditional distribution of $\mathbf{y}$ given a particular element of $\mathbf{z}$ is:

$$p(\mathbf{y}|z_k = 1) = p(\mathbf{y}|\theta_k)$$

which can also be written as:

$$p(\mathbf{y}|\mathbf{z}) = \prod_{k=1}^{K} [p(\mathbf{y}|\theta_k)]^{z_k}$$

The joint distribution of $\mathbf{y}$ and $\mathbf{z}$ is given by $p(\mathbf{y}|\mathbf{z})p(\mathbf{z})$. Also, we can find the marginal distribution of $\mathbf{y}$ by summing the joint distribution over the $\mathbf{z}$:

$$p(\mathbf{y}) = \sum_{\mathbf{z}} p(\mathbf{y}|\mathbf{z})p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k p(\mathbf{y}|\theta_k)$$

In the case we have many observations $\mathbf{y}_n, n = 1, \cdots, N$, then for every observation $\mathbf{y}_n$ there is a corresponding variable $\mathbf{z}_n$. Now, to apply the EM algorithm we must define the complete data which in our study is the observed data $\mathbf{y}_n, n = 1, \cdots, N$ and the indicator variables $\mathbf{z}_n, n = 1 \cdots, N$, which plays the role of hidden variables. The likelihood of the complete data is:

$$p(\{\mathbf{y}_n, \mathbf{z}_n\}_{n=1}^{N}|\{\pi_k, \theta_k\}_{k=1}^{K}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}}[p(\mathbf{y}|\theta_k)]^{z_{nk}}$$

Taking the logarithm we obtain

$$\log p(\{\mathbf{y}_n, \mathbf{z}_n\}_{n=1}^{N}|\{\pi_k, \theta_k\}_{k=1}^{K}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\{\log \pi_k + \log p(\mathbf{y}|\theta_k)\}$$

We observe now, that the logarithm acts directly to the distribution, which leads to a simpler solution. In the E-step we need to find the posterior of the hidden variables, $p(\{\mathbf{z}_n\}_{n=1}^{N}|\{\mathbf{y}_n\}_{n=1}^{N}, \theta^{(t)})$, using the current model parameter values $\Theta^{(t)} = \{\pi_k^{(t)}, \theta_k^{(t)}\}_{k=1}^{K}$. Using the Bayes theorem we have:

$$\begin{aligned} p(\{\mathbf{z}_n\}_{n=1}^{N}|\{\mathbf{y}_n\}_{n=1}^{N}, \theta^{(t)}) &= \frac{p(\{\mathbf{y}_n\}_{n=1}^{N}|\{\mathbf{z}_n\}_{n=1}^{N}, \Theta^{(t)})p(\{\mathbf{z}_n\}_{n=1}^{N}|\theta^{(t)})}{p(\{\mathbf{y}_n\}_{n=1}^{N}|\theta^{(t)})} \\ &= \prod_{n=1}^{N} \frac{\prod_{k=1}^{K}[\pi_k^{(t)} p(\mathbf{y}_n|\theta_k^{(t)})]^{z_{nk}}}{\sum_{k=1}^{K} \pi_k^{(t)} p(\mathbf{x}_n|\theta_k^{(t)})} \\ &= \prod_{n=1}^{N}\prod_{k=1}^{K} \left[\frac{\pi_k^{(t)} p(\mathbf{y}_n|\theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} p(\mathbf{y}_n|\theta_k^{(t)})}\right]^{z_{nk}} \end{aligned} \tag{2.34}$$

We can see that the posterior of the hidden variables is a product of N multinomial distributions and the expectation of hidden variables in that case is given by:

$$E\{(z_{nk})\} = \frac{\pi_k^{(t)} p(\mathbf{y}_n|\theta_k^{(t)}))}{\sum_{k=1}^{K} \pi_k^{(t)} p(\mathbf{y}_n|\theta_k^{(t)})} . \tag{2.35}$$

Calculating the expected log-likelihood of the complete data we have:

$$E_{\{\mathbf{z}_n\}_{n=1}^{N}}\{\log p(\{\mathbf{y}_n, \mathbf{z}_n\}_{n=1}^{N}|\theta^{(t)})\} = \sum_{n=1}^{N}\sum_{k=1}^{K} E\{z_{nk}\}\{\log \pi_k + \log p(\mathbf{y}_n|\theta_k^{(t)})\} \tag{2.36}$$

The M - step of the algorithm consist from the maximization of the above expected log-likelihood with respect to the model parameters. Assuming that the components follow the Gaussian distribution, $\theta_k^{(t+1)} = \{\mathbf{m}_k^{(t+1)}, \mathbf{\Sigma}_k^{(t+1)}\}$, and maximizing the expected log-likelihood of the complete data with respect to model parameters we obtain:

$$\mathbf{m}_k^{(t+1)} = \frac{1}{N_k}\sum_{n=1}^{N} E\{z_{nk}\}\mathbf{y}_n \tag{2.37}$$

$$\mathbf{\Sigma}_k^{(t+1)} = \frac{1}{N_k}\sum_{n=1}^{N} E\{z_{nk}\}(\mathbf{y}_n - \mathbf{m}_k)(\mathbf{y}_n - \mathbf{m}_k)^T \tag{2.38}$$

$$\pi_k^{(t+1)} = \frac{N_k}{N} \tag{2.39}$$

where $N_k = \sum_{n=1}^{N} E\{z_{nk}\}$. When we perform the optimization over mixing coefficients, we must take into account the constraints by using Lagrange multipliers.

2.7 Useful priors distributions for modeling specific properties

2.7.1 Sparse priors

Let assume that we have a vector of weights $\mathbf{w} = \{w_1, \cdots, w_p\}$ which plays the role of model parameters. In many situations we aim at obtaining a sparse configuration of this vector, i.e most of the weights to be set to zero. The sparsity is a very helpful property, since the processing is faster and simpler in a sparse representation where few coefficients reveal the information we are looking for. From a signal processing perspective, the sparsity has found many applications, for example in signal compression and signal denoising [214]. In Bayesian inference, the sparsity is achieved through sparse priors which help us:

- to automatically adjust the order of the model,
- to reduce the complexity of the model and its decision part,
- to compute more easily the output of the model since few weights are non zero,
- and to determine which components of the model are relevant with the data, which may be very useful in many applications.

A natural choice of a sparse prior distribution for the weights $\mathbf{w}$ is a hierarchical prior described below. More specifically, the weights $\mathbf{w}$ are treated as a random variables that follow a Gaussian distribution with zero mean and variance $a_i^{-1}, i = 1, \cdots, p$:

$$p(\mathbf{w}|\mathbf{a}) = \prod_{i=1}^{p} N(0, a_i^{-1}). \tag{2.40}$$

The parameters a_i are called hyperparameters and control the prior distribution of the parameter vector $\mathbf{w}$. Hierarchical priors are often designed using conjugate distributions

[44]. This happens for analytical eases and because the previous knowledge can be readily expressed. The empirical Bayes refers to the practice of optimizing the hyperparameters of the priors, so as to maximize the marginal distribution of the dataset. This practice is suboptimal since it ignores the uncertainty of the hyperparameters. Alternatively, a more robust approach is to define priors over the hyperparameters. This leads us to a full Bayesian model. The prior distribution over each hyperparameter a_i is a gamma distribution:

$$p(a_i) = \Gamma(a_i; b_{a_i}, c_{a_i}). \tag{2.41}$$

The prior over one weight w_i depends on the hyperparameter a_i. The "true" prior is given by integrating over the hyperparameter:

$$p(w_i) = \int p(w_i \mid a_i) p(a_i) da_i. \tag{2.42}$$

Making the above integration we obtain for the parameter prior:

$$p(w_i) \propto \left(\frac{1}{b_{a_i}} + \frac{w_i^2}{2}\right)^{-(c_{a_i}+\frac{1}{2})} \tag{2.43}$$

which is the kernel of a Student-t density. If we allow $c_{a_i} \to 0$ and $b_{a_i} \to \infty$ then we obtain the hyperprior:

$$p(a_i) \propto \frac{1}{a_i}, \tag{2.44}$$

which is an noninformative prior[125]. Now, the true prior for one weight, w_i, is

$$p(w_i) \propto \frac{1}{|w_i|}, \tag{2.45}$$

and for all parameters:

$$p(\mathbf{w}) \propto \prod_{i=1}^{p} \frac{1}{|w_i|}. \tag{2.46}$$

This prior is recognized as sparse due to heavy tail and the sharp peak at zero [34, 157]. In Fig. (2.12) plots of the Student's t pdf using ν degrees of freedom are shown together with a plot of the Gaussian distribution. It is easy to observe that most of mass is concentrated around the point $x = 0$. Also, as the degrees of freedom are increased the distribution resembles the Gaussian distribution.

Another way to declare the sparsity over the weights $\mathbf{w}$ is to use a Laplacian distribution over them:

$$p(\mathbf{w}|\alpha) = \frac{1}{Z_w} \exp\{-\alpha \sum_{i=1}^{p} |w_i|\} \tag{2.47}$$

where α is the regularization parameter. However, this prior leads to a nonlinear optimization problem. To overcome this problem we use the Fan's approximation [158]:

$$|w_i| \approx \frac{1}{2}|\tilde{w}_i| + \frac{1}{2}\frac{w_i^2}{|\tilde{w}_i|} \tag{2.48}$$

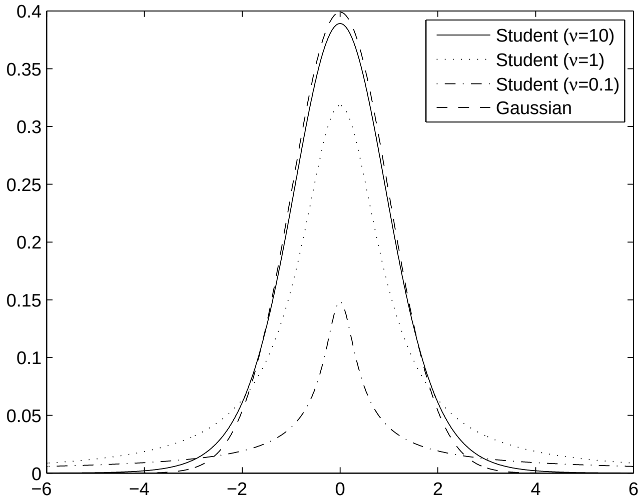


Figure 2.12: The Student's t pdf plots using 0.1, 1 and 10 degrees of freedom.

where $|\tilde{w}_i| \neq 0$ is a local approximation of $|w_i|$, such that $|w_i - \tilde{w}_i| < \epsilon$. The quantity ϵ is a small positive constant. Using the above approximation the prior distribution over the weights takes the following form:

$$p(\mathbf{w}|\alpha) = \frac{1}{Z_w} \exp\{-\alpha \mathbf{w}^T \mathbf{L} \mathbf{w}\}, \tag{2.49}$$

where $\mathbf{L} = [\frac{1}{|\tilde{w}_1|}, \frac{1}{|\tilde{w}_2|}, \cdots, \frac{1}{|\tilde{w}_p|}]$. This prior has been used in [158] for image denoising and wavelet coefficients thresholding. Also, in the same work an extension, based on the use of multiple precision components, is presented. Because in most of the cases, priors of this kind are used in an iterative fashion, we can use as local approximation of $|w_i|$, the estimated weight $|\hat{w}_i|$ of the previous iteration.

2.7.2 Spatial priors

There are problems where the data are spatially related to each other. A characteristic example is in the task of image analysis, where, apart from the intensity values, pixels positions constitute a significant piece of information that must be taken into account. The Markov Random Field (MRF) is a valuable tool to exploit the spatial characteristics of an image or the correlation between features in a classification problem. MRFs have found many application in image analysis, i.e. image denoising,image segmentation, and machine learning, i.e. classification and clustering problems.

In an MRF, the sites in $\mathcal{S}$, where $\mathcal{S}$ is the set of sites, are related to each other via a neighborhood system, which is defined as $\mathcal{N} = \{\mathcal{N}_i, i = 1, \cdots, N\}$, where $\mathcal{N}_i$ is the set of sites neighboring i, $i \notin \mathcal{N}_i$ and $i \in \mathcal{N}_j \Leftrightarrow j \in \mathcal{N}_i$. A random field $\mathcal{X}$ said to be an MRF on $\mathcal{S}$ with respect to a neighborhood system $\mathcal{N}$ if and only if

$$P(\mathbf{x}) > 0, \mathbf{x} \in \mathcal{X} \tag{2.50}$$

$$P(x_i | x_{\mathcal{S}-\{i\}}) = P(x_i | x_{\mathcal{N}_i}) \tag{2.51}$$

Note, the neighbourhood system can be multi-dimensional. The above property means that the probability in the site i depends only from the neighborhood (local characteristics of the field). It is easy to observe that the MRF is a generalization of the Markov process in which the time index is replaced by the space. According to the Hammersley-Clifford theorem [100], an MRF can equivalently be characterized by a Gibbs distribution. Thus,

$$P(\mathbf{x}) = \frac{1}{Z} \exp\{-U(\mathbf{x})\} \tag{2.52}$$

where

$$Z = \sum_{\mathbf{x} \in \mathcal{X}} \exp\{-U(\mathbf{x})\} \tag{2.53}$$

is a normalizing constant called the partition function, and $U(\mathbf{x})$ is an energy function of the form

$$U(\mathbf{x}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{x}) \tag{2.54}$$

which is a sum of clique potentials $V_c(\mathbf{x})$ over all possible cliques. A clique c is defined as a subset of sites in $\mathcal{S}$ in which every pair of distinct sites are neighbours, except for single-site cliques. The value of $V_c(\mathbf{x})$ depends on the local configuration on clique c. For more detail on MRF and Gibbs distribution see [100].

The properties of the distribution with respect to the neighborhood depends from the functional form of the potential function $V_c(\mathbf{x})$ [104]. An important special case arises when we consider cliques up to size two. Then the energy function takes the form:

$$U(\mathbf{x}) = \sum_{i \in \mathcal{S}} V_1(x_i) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(x_i, x_{i'}). \tag{2.55}$$

The first summation $\sum_{i \in \mathcal{S}} V_1(x_i)$ does not include any spatial information and for the moment is excluded for the subsequent analysis. The interesting part is the second summation $\sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(x_i, x_{i'})$ which includes spatial information through the inner summation. The potential function $V_2(x_i, x_{i'})$ specifies the relation between x_i and $x_{i'}$. In the literature many functional forms of this potential have been proposed, see for example [104]. A widely used function for this potential is:

$$V_2(x_i, x_{i'}) = (x_i - x_{i'})^2. \tag{2.56}$$

Another function, which is robust to outliers, is:

$$V_2(x_i, x_{i'}) = \frac{1}{1 + \frac{1}{(x_i - x_{i'})^2}}. \tag{2.57}$$

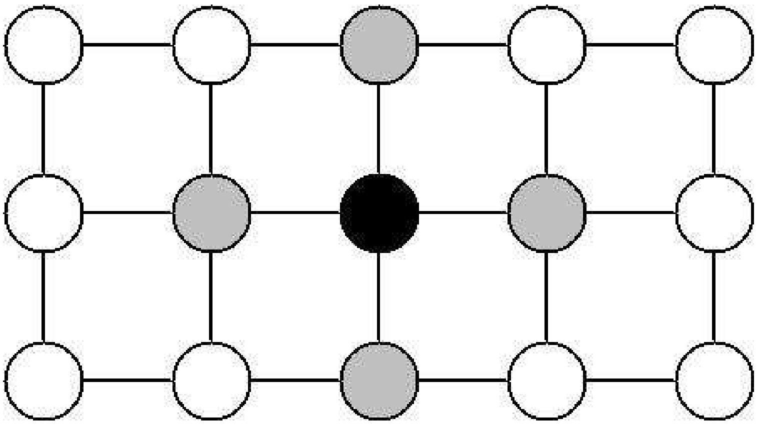


Figure 2.13: MRF example: Given the grey nodes, the black node is conditionally independent of all other nodes.

2.8 Transform - based signal processing

The purpose of a transform is to describe a signal or a system in terms of a combination of a set of elementary simple signals (such as sinusoidal signals) that lend themselves to relatively easy analysis, interpretation and manipulation. Transform-based signal processing methods include Fourier transform, Laplace transform, z-transform and wavelet transforms. The most widely applied signal transform is the Fourier transform, which is effectively a form of vibration analysis, in that a signal is expressed in terms of a combination of the sinusoidal vibrations that make up the signal. Fourier transform is employed in a wide range of applications, including popular music coders, noise reduction and feature extraction for pattern recognition. The Laplace transform, and its discrete-time version the z-transform, are generalizations of the Fourier transform. The wavelets are multi-resolution transforms in which a signal is described in terms of a combination of elementary waves of different durations. The set of basis functions in a wavelet is composed of contractions and dilations of a single elementary wave. This allows non-stationary events of various durations in a signal to be identified and analyzed.

A transform is an operation that performs on a signal. Also, a transform can have an inverse, which restores the original values and it can be thought of a different way of representing the same information. A natural question is, Why would we do this? One answer is so that we can analyze the transformed signal, for example to compress it. The discrete cosine transform have been used effectively to alter a signal for storing it in a compact form. Sometimes transforms are performed because things are easier to do in the transformed domain.

2.8.1 Fourier transform

Under mild conditions, the Fourier Transform describes a signal $x(t)$ as a linear superposition of sines and cosines characterized by their frequency f:

$$x(t) = \int X(f)e^{i2\pi ft}df \tag{2.58}$$

where

$$X(f) = \int x(t)e^{-i2\pi ft}dt \tag{2.59}$$

are complex valued coefficients that give the relative contribution of each frequency f. Equation (2.59) is the continuous Fourier Transform of the signal $x(t)$. It can be seen as an inner product of the signal $x(t)$ with the complex sinusoidal functions $e^{-i2\pi ft}$, i.e.

$$X(f) =< x(t), e^{-i2\pi ft} > \tag{2.60}$$

Its inverse transform is given by Eq. (2.58) and since the mother functions $e^{-i2\pi ft}$ are orthogonal, the Fourier Transform is nonredundant and unique.

2.8.2 Wavelet transform

A wavelet family $\psi_{a,b}$ is a set of elemental functions generated by dilations and translations of a unique admissible mother wavelet $\psi(t)$:

$$\psi_{a,b}(t) = |a|^{-\frac{1}{2}}\psi(\frac{t-a}{b}) \tag{2.61}$$

where $a, b \in \mathcal{R}$, $a \neq 0$, are the scale and translation parameters respectively. The mother wavelet is limited in time domain, has zero mean and is normalized. As a increases the wavelet becomes more narrow and by varying b, the mother wavelet is displaced in time. Thus, the wavelet family gives a unique pattern and its replicas at different scales and with variable localization in time. The continuous wavelet transform of a signal $x(t) \in L^2(\mathcal{R})$ (finite energy signals) is defined as the correlation between the signal and the wavelet functions $\psi_{a,b}$ i.e.

$$W_\psi x(a,b) = |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} x(t)\psi^*(\frac{t-a}{b})dt =< x(t), \psi_{a,b} > \tag{2.62}$$

where * denotes complex conjugation. Then, the different correlations $< x(t), \psi_{a,b} >$ indicates how precisely the wavelet function locally fits the signal at every scale a. Since the correlation is made with different scales of a single function, instead of with complex sinusoids characterized by their frequencies, wavelets give a time-scale representation. The inverse wavelet transform is:

$$x(t) = \frac{1}{C_\psi} \int_0^\infty \int_{-\infty}^{\infty} W_\psi x(a,b)|a|^{-\frac{1}{2}}\psi(\frac{t-a}{b})db\frac{da}{a^2} \tag{2.63}$$

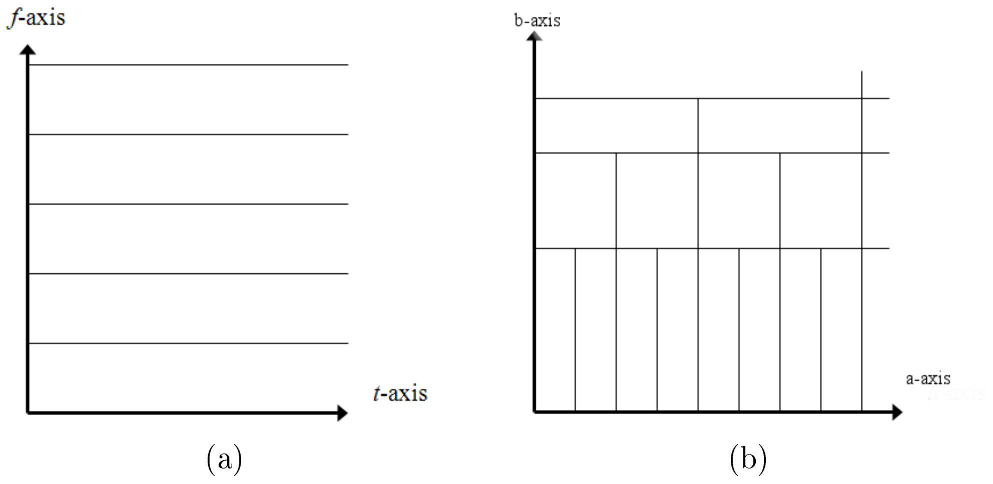


Figure 2.14: Time-frequency tile allocation of the two transforms: (a) Fourier transform and (b) wavelet transform.

where

$$C_\psi = \int_0^\infty \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty \tag{2.64}$$

$\Psi(\omega)$ is the Fourier Transform of the mother function $\psi(t)$.

The above wavelet transform is called Continuous Wavelet Transform because can operate at every scale, from that of the original signal up to some maximum scale that is determined by trading off our need for detailed analysis with available computational power. Calculating wavelet coefficients at every possible scale is time consuming. However, if we choose scales and positions based on powers of two then our analysis will be much more efficient and just as accurate. We obtain such an analysis from the discrete wavelet transform (DWT) [151].

The general scheme of Transform - based signal processing methods is illustrated in Fig. (2.15). First the signal is transformed into the new domain through the transform matrix **A**. Then, in the new domain coefficients an operation, linear or non-linear $f(\cdot)$, is performed. After that, the modified coefficients is transformed back into the original domain. For example, in wavelet denoising approach the signal is transformed into the wavelet domain and a thresholding operation is performed over wavelet coefficients which are transformed back into the original domain producing the new signal with the desired properties. Transform-based methods have been found many applications in biomedical signal processing [144].

2.9 Principal Component Analysis

Principal Component Analysis (PCA) is a tool in modern data analysis, where its application range from neuroscience to computer graphic. It is a simple, non-parametric

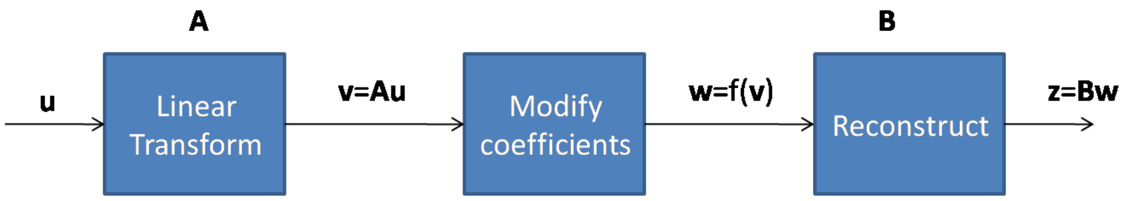


Figure 2.15: General scheme of Transform-based methods

method for extracting useful information from complex dataset. PCA help us to reduce the complexity of the original dataset and to reveal the structures that underlie it. It is also known as the Karhunen-Loève transform [45, 129]. In PCA we seek a linear transformation of the original dataset into a new dataset, where principal components with larger associated variance represent important structure of the dataset.

Consider the dataset of observations $\mathbf{y}_n, n = 1, \cdots, N$, where $\mathbf{y}_n$ is a vector of dimension $D \times 1$. The goal in the PCA is to find a projection of the data onto a space with smaller dimensionality than the original, $M < D$, while at the same time the variance of the projected data is maximized. First, we consider the projection onto a one dimensional space. We can define the direction of this space using a vector $\mathbf{u}_1$ of dimension $D \times 1$ with the constraint $\mathbf{u}_1^T\mathbf{u}_1 = 1$ (i.e. is the unit vector). Each data point is then projected onto the new space given the value $\mathbf{u}_1^T\mathbf{y}_n$. Now the variance of the projected data is given by

$$\mathrm{var}(\mathbf{u}_1^T\mathbf{y}_n) = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{u}_1^T\mathbf{y}_n - \mathbf{u}_1^T\bar{\mathbf{x}})^T(\mathbf{u}_1^T\mathbf{y}_n - \mathbf{u}_1^T\bar{\mathbf{y}}) = \mathbf{u}_1^T\mathbf{\Sigma}\mathbf{u}_1$$

where $\bar{\mathbf{y}} = \sum_{n=1}^{N}\mathbf{y}_n$ is the mean of the data and $\mathbf{\Sigma} = \frac{1}{N}(\mathbf{y}_n - \bar{\mathbf{y}})^T(\mathbf{y}_n - \bar{\mathbf{y}})$ is the data covariance.

Now, we want to maximize $\mathrm{var}(\mathbf{u}_1^T\mathbf{y}_n)$ with respect to $\mathbf{u}_1$ subject to the constraint $\mathbf{u}_1^T\mathbf{u}_1 = 1$, i.e.

$$\max_{\mathbf{u}_1} \mathbf{u}_1^T\mathbf{\Sigma}\mathbf{u}_1 \text{ s.t. } \mathbf{u}_1^T\mathbf{u}_1 = 1.$$

Introducing the Lagrange multiplier λ_1 and performing the resulting unconstrained maximization we obtain the solution:

$$\mathbf{\Sigma}\mathbf{u}_1 = \lambda_1\mathbf{u}_1.$$

It is easy to see from the above equation that the vector $\mathbf{u}_1$ is an eigenvector of the covariance matrix $\mathbf{\Sigma}$ and λ_1 the corresponding eigenvalue, which is also the variance of $\mathbf{u}_1^T\mathbf{y}_n$. So, to obtain the maximun variance the vector $\mathbf{u}_1$ must be the eigenvector of matrix $\mathbf{\Sigma}$ with the largest eigenvalue. We continue in the same way to introduce additional components until to use M eigenvectors, having in mind that each new direction maximize the projected variance, as before, while is orthogonal to the directions that have been already added. The appropriate choice of M is a difficult problem, however there are two simple approaches to choose M. The first approach is to choose M such that a large fraction d of the total variance is taken into account. Usually, the d is between 70% and

90%. The second approach is to examine the eigenvalue spectrum and see if there is a point where the values fall sharply before stay at small values [175]. We see that PCA is related to eigendecomposition of a matrix, which is symmetric and positive definite. When these properties are violated then the Singular Value Decomposition could be used to obtain a similar procedure. Finally, there is an extension of PCA based in probabilistic formulation of the problem [45]. This extension gives us the ability to determine the dimension, M, of the new space in a more flexible way by adopting the Bayesian Framework. Also, the use of EM algorithm provides us with an efficient approach in term of computational complexity, especially in high dimensional spaces [45]. As expected PCA has long history to the analysis of Biomedical Signals. It has been used for EEG monitoring [174], ERP analysis [120] and fMRI data analysis [47].

2.10 Independent Component Analysis

Consider the dataset of observations $\mathbf{y}_n, n = 1, \cdots, N$, where $\mathbf{y}_n$ is a vector of dimension $D \times 1$, we have D signals observed in time points n. In ICA we assume that each vector $\mathbf{y}_n$ is a linear mixture of K unknown sources:

$$\mathbf{x}_n = \mathbf{A}\mathbf{s}_n$$

where the matrix of mixing coefficients $\mathbf{A}$ is unknown. The goal in ICA is to find the sources $\mathbf{s}_n$ or to find the inverse of matrix $\mathbf{A}$. The sources are independently distributed with marginal distributions $p(\mathbf{s}_n) = p_i(s_n^{(i)})$. Following [152], we derive the ICA under the ML principle, where we assume that the number of observed signals is equal to the number of sources, $K = D$. The probability of the observations and sources given the matrix $\mathbf{A}$ is:

$$\begin{aligned} p(\{\mathbf{y}_n, \mathbf{s}_n\}_{n=1}^N|\mathbf{A}) &= \prod_{n=1}^N p(\mathbf{y}_n|\mathbf{s}_n, \mathbf{A})p(\mathbf{s}_n) \\ &= \prod_{n=1}^N \delta(\mathbf{y}_n - \mathbf{A}\mathbf{s}_n)p(\mathbf{s}_n) \end{aligned}$$

Performing the marginalization with respect to the sources we obtain the likelihood function for a single data point $\mathbf{x}_n$:

$$\begin{aligned} p(\mathbf{y}_n|\mathbf{A}) &= \frac{1}{|\mathbf{A}|}p(\mathbf{A}^{-1}\mathbf{y}_n) \\ &= \frac{1}{|\mathbf{A}|}\prod_{i=1}^K p(\sum_{j=1}^D A_{ij}^{-1}y_n^{(j)}) \end{aligned}$$

The log-likelihood of the mixing coefficients is:

$$\mathcal{L} = \log|\mathbf{W}| + \sum_{i=1}^K \log p(\sum_{j=1}^D W_{ij}y_n^{(j)}) \tag{2.65}$$

where we have made the convention $\mathbf{W} = \mathbf{A}^{-1}$. To find the optimum matrix $\mathbf{W}$ we maximize the log-likelihood with respect to it. The gradient is given by:

$$\frac{d\mathcal{L}}{d\mathbf{W}} = [\mathbf{W}^{-1}]^T + \mathbf{z}\mathbf{y}_n^T$$

where we have define $a_i = \sum_{j=1}^{D} W_{ij} y_n^{(j)}$, $\phi(a_i) = \frac{d \log p_i(z_i))}{dz_i}$ and $z_i = \phi(a_i)$. We can see that parameters a_i are the reconstructed sources. Since, we want to maximize the likelihood we adapt the matrix $\mathbf{W}$ by making small steps of the form:

$$\Delta\mathbf{W} \propto [\mathbf{W}^{-1}]^T + \mathbf{z}\mathbf{y}_n^T \ .$$

Until now, we have not discuss the function ϕ which defines the assumed prior distribution of sources. A popular choice is to use the *tanh* function. To conclude, the algorithm to find the independent components has three steps:

- Calculate an estimation sources through the mapping: $\mathbf{a} = \mathbf{W}\mathbf{x}$.
- Calculate a nonlinear mappping of the estimated sources $z_i = \phi(a_i)$.
- adjust the matrix $\mathbf{W}$ through $\Delta\mathbf{W} \propto [\mathbf{W}^{-1}]^T + \mathbf{z}\mathbf{y}^T$.

The above exposition of the ICA is based on the ML principle, however similar algorithms for ICA can be obtained by adopting other criteria for the independence. A useful introduction in ICA is presented [147], where a fast algorithm to perform ICA is also given. As expected Bayesian formulations of ICA - like model are presented to the literature [150, 148, 149]. From the perspective of biomedical signal processing the ICA has found many application among them to study the brain dynamics through EEG signals [118, 119] to identify the activated brain's areas in fMRI analysis [48] and to estimate the ERP signal from the EEG measurements [146]. A overview of ICA applied to EEG data is shown in Fig. (2.16). First the EEG data are decomposed in independent components, then by visual inspection some of these components are removed since contain artefacts (for example eyes blink), and finally the artifact-free EEG signals, is obtained by mixing and projecting back onto the scalp channels selected non-artifactual ICA components.

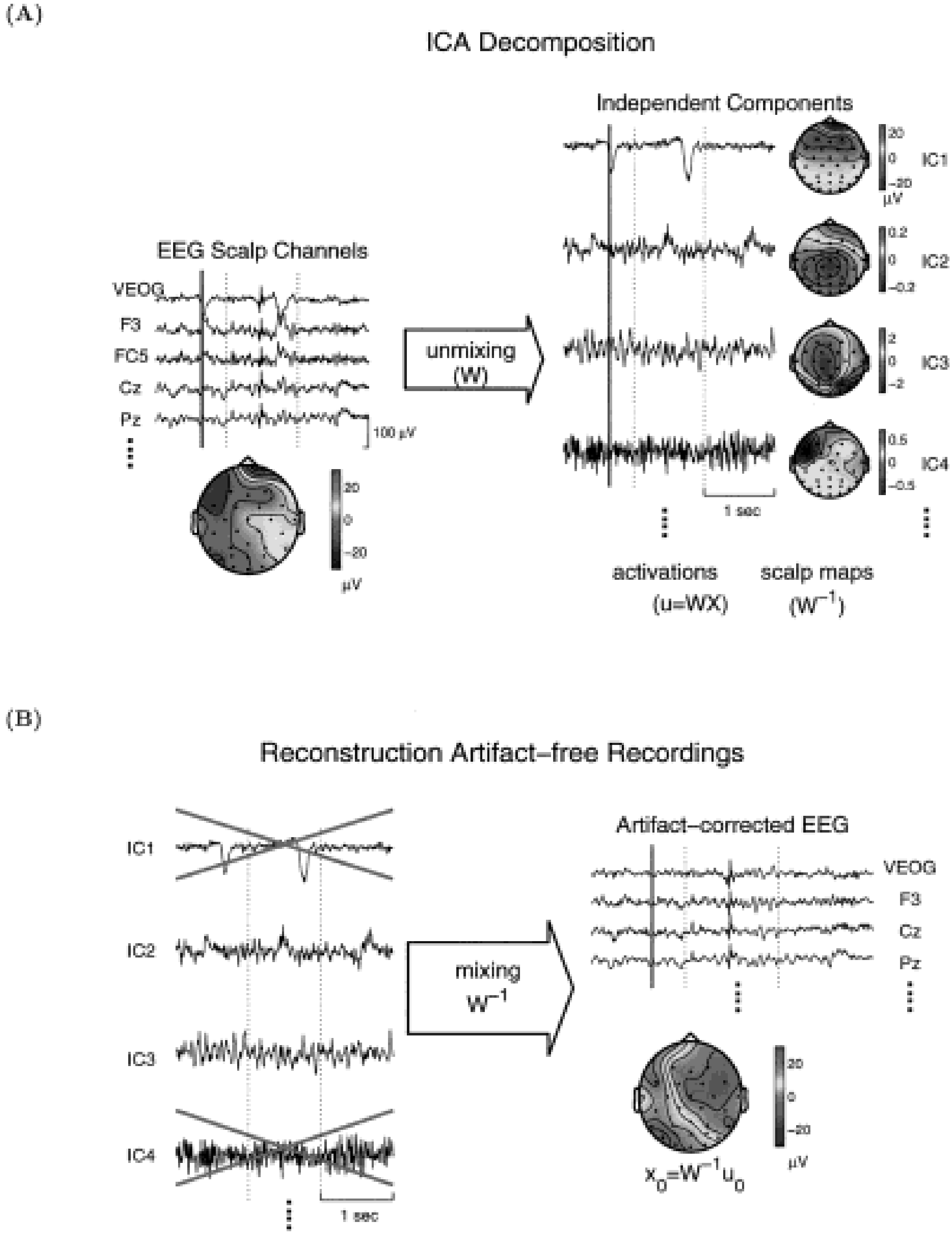


Figure 2.16: Schematic overview of ICA applied to EEG data. (Figure reprinted from [119])

Chapter 3

Statistical models for sequential data

Sequential data arise in many fields of engineering, physics and statistics. The data may either be a time series, or a sequence generated by a 1-dimensional spatial process, e.g., biosequences. One may be interested either in online analysis, where the data arrives in real-time, or in offline analysis, where all the data has already been collected. Also, the analysis of dynamic phenomena is a common problem related to sequential data. A time varying system can be represented through a dynamic model, defined by an observable component and unobservable state. The hidden state represents the desired information that we want to extrapolate.

In this chapter, we provide material related to the linear regression model. More specifically, the various components of the linear model, such as the design matrix and the weights (or regression coefficients), are described. In addition, information about the state-space model is provided. The state-space model is an extension of the linear model, which help us to include into our analysis data that have dynamic nature or arise sequentially. Finally, the autoregressive model is described. This model help us to analyze the correlation structure of a time series.

3.1 General Linear Model (GLM)

In the General Linear Model (GLM), the observations $\mathbf{y} = \{y_1, \cdots, y_N\}$ of an experiment are described as a linear combination of some predictors given by the equation:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \mathbf{e} \ . \tag{3.1}$$

where $\mathbf{\Phi}$ is the design matrix of size $N \times p$ and it is assumed to be known for the problem under study, $\mathbf{w}$ is the vector of weights of the linear combination and has size $p \times 1$ and $\mathbf{e}$ is the additive noise assumed to be zero mean and Gaussian distributed, $p(\mathbf{e}) = \mathcal{N}(0, \mathbf{C}_{\mathbf{e}}^{-1})$, where $\mathbf{C}_{\mathbf{e}}^{-1}$ is the inverse precision (covariance) matrix. The form of this matrix defined the properties of the additive noise. Usually, we assume that the error samples are independent and identically distributed (i.i.d), in that case a simple approach is to assumed $\mathbf{C}_{\mathbf{e}}^{-1} = \lambda\mathbf{I}$. Also, more general forms can be used such as a diagonal precision matrix, where we use for each observations y_n a separate precision λ_n. This form of the precision matrix help to use in indirect way more useful distributions such as the Student - t distribution [138]. Finally, the autoregressive (AR) model can be alternatively used to describe the autocorrelation between the error samples, where it can be written in the general form of the additive noise.

3.1.1 Design matrix

In this section the role of the design matrix and the various form of it will be described. The design matrix has the following general form:

$$\mathbf{\Phi} = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_p(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_p(\mathbf{x}_2) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \cdots & \phi_p(\mathbf{x}_N) \end{bmatrix}, \tag{3.2}$$

where $\{\mathbf{x}_n\}_{n=1}^N$ are the input variables and $\phi_j, j = 1, \cdots, p$ are the basis functions, both, the input variables and the basis functions, are assumed to be known. According to the linear model described previously each observation y_n is described as a linear combination of p basis functions:

$$y_n = \sum_{j=1}^{p} w_j \phi_j(\mathbf{x}_n) + e_n = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n) + e_n. \tag{3.3}$$

We see that the basis functions describe the relationship between the observations and the input variables. In the literature many forms for the basis functions have been proposed. In the case where a linear relationship between the observations and the input variables is assumed then the basis functions take the form $\boldsymbol{\phi}(\mathbf{x}_n) = \mathbf{x}_n$. It is important to observe here that, by using non linear functions we allow the model to be also non linear to the input variables while we keep the linearity with respect to the weights.

One possible choice is to use polynomial basis functions where the basis function has the form of powers of the input variables, i.e. $\phi_j(x) = x^j$. One other choice of basis functions is the Gaussian basis functions

$$\phi_j(x) = \exp\Big\{ - \frac{(x-\mu_j)^2}{2s^2} \Big\}, \tag{3.4}$$

where μ_j are the locations of the basis functions in the input space and s^2 their scale. Another possibility is the sigmoidal basis function of the form:

$$\phi_j(x) = \sigma\Big(\frac{x-\mu_j}{s}\Big), \tag{3.5}$$

where $\sigma(\alpha)$ is the logistic sigmoid function given by:

$$\sigma(\alpha) = \frac{1}{1+\exp(-\alpha)}. \tag{3.6}$$

Finally, another possible set of basis functions is the Fourier basis functions and the wavelet basis functions. The construction of the design matrix is not restricted only to the approach that we have described previously. The design matrix can have many other regressors (columns of the design matrix) related to the problem under study as we will see in latter chapter of this thesis.

3.1.2 Maximum Likelihood (ML) parameter estimation of the GLM

Assuming that the noise follows white Gaussian distribution, i.e. $\mathbf{e} \sim N(0, \lambda\mathbf{I})$, then the likelihood of the observations $\mathbf{y}$ is given by:

$$p(\mathbf{y}; \mathbf{w}, \lambda) = \Big(\frac{\lambda}{2\pi}\Big)^{N/2} \exp\Big\{-\frac{\lambda}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|^2\Big\} \tag{3.7}$$

Based on the above formulation, the training of the GLM becomes a maximum likelihood (ML) estimation problem for the regression model parameters $\Theta = \{\mathbf{w}, \lambda\}$, in the sense of maximizing the log-likelihood function given by

$$L_{ML}(\Theta) = \log p(\mathbf{y}; \mathbf{w}, \lambda) = \Big\{\frac{N}{2}\log\lambda - \frac{\lambda}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|^2\Big\} \ . \tag{3.8}$$

Setting the partial derivatives of the above function. with respect to the parameters, equal to zero, the following update rules for the model parameters are obtained

$$\hat{\mathbf{w}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y} \ , \tag{3.9}$$

$$\hat{\lambda} = \frac{N}{\|\mathbf{y} - \mathbf{\Phi}\hat{\mathbf{w}}\|^2} \ . \tag{3.10}$$

We want to mention here that we have make the assumption that the matrix $\mathbf{\Phi}^T\mathbf{\Phi}$ is invertible, which this is happened only when the design matrix $\mathbf{\Phi}$ is of full rank. A practical solution to this problem is to use a few columns on the design matrix which means that only few weights will be estimated.

3.2 State-Space Models

A useful family of models to study sequential data is the state-space model. In any state-space model three components play the most important role: the initial density $p(\mathbf{w}_1)$, the transition density of the states $p(\mathbf{x}_t|\mathbf{w}_{t-1})$ and the observation density $p(\mathbf{y}_t|\mathbf{w}_t)$. Suppose that the densities are the same for all time. There are many state-space models, the most known are the Hidden Markov Models (HMMs) and the Kalman Filters (KF). A state-space model is a model of how $\mathbf{w}_t$ generates $\mathbf{y}_t$ and $\mathbf{w}_{t+1}$ and our goal is to infer $\mathbf{w}_{1:t} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_t\}$ given $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_t\}$.

A graphical representation of a state-space model is illustrated in Fig. (3.1). In next two sections we will present the HMM and the KF. Also, in the KF model an inference procedure is provided for the states estimation.

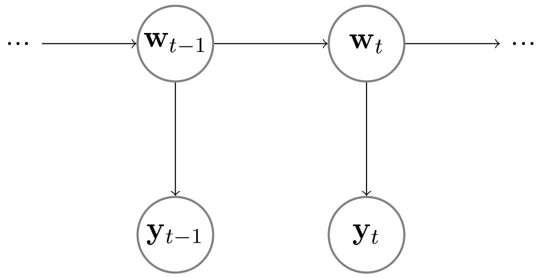


Figure 3.1: Graphical representation of a state-space model.

3.3 Hidden Markov Models

Consider a system which may be described as being in one of a set of N different states, $s_1, s_2, \cdots, s_N$ as depicted in Fig. (3.2). At discrete times, the system changes state (it can also remain in the same state) according to a set of probabilities associated with the state. We denote the time points, where the state changes, as $t = 1, 2, \cdots$, also we denote the state at time t as q_t. A full description of the system requires the current state as well as all previous states. However, in our exposition we will use only the first order Markov chain, which means that the current state and the previous state are used to describe the system:

$$P(q_t = s_j | q_t = s_i, q_{t-2} = s_k, \cdots) = P(q_t = s_j | q_t = s_i) \tag{3.11}$$

Furthermore we only consider those systems that the right hand side of (3.11) is independent of the time, i.e. the state transition probabilities does not change with time. The state transition probabilities a_{ji} has the form

$$a_{ij} = P(q_t = s_j | q_{t-1} = s_i), 1 \leq i, j \leq N. \tag{3.12}$$

Since the a_{ij} are probabilities they subject to the following constraints:

$$\begin{aligned} a_{ij} &\geq 0 \\ \sum_{j=1}^{N} a_{ij} &= 1 \end{aligned}$$

The above stochastic model is an observable Markov model since the output of the process is the set of states at each time instant and each state corresponds to an observable event. We will try to explain the above statistical model through an example. Probably the reader is familiar with foootball games. When a team plays a game the outcome (or observations) is being one of the following:

- State 1: Draw (D)
- State 2: Loose (L)
- State 3: Win (W)

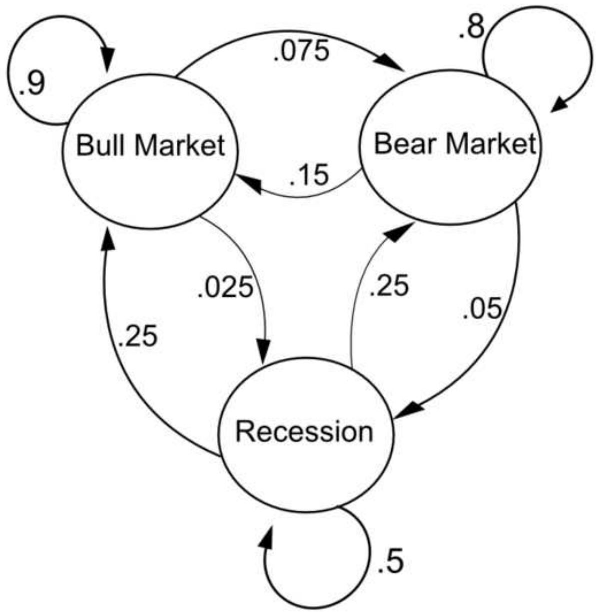


Figure 3.2: A simple Markov Model with 3 states.

In a year a team plays a number of games. Let us assume that we are at game number t and that the matrix of state transition probabilities is

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \tag{3.13}$$

Given that the outcome of the game 1 ($t = 1$) is Win we can ask the question: What is the probability that the team's results for the next 7 games will be "W-W-D-D-W-L-W"? This question can be stated more formally as: What is the probability to observe the sequence $o = \{s_3, s_3, s_3, s_1, s_1, s_3, s_2, s_3\}$ given the model $\mathcal{M}$. This probability can be evaluated as:

$$\begin{aligned} P(o|\mathcal{M}) &= P(s_3, s_3, s_3, s_1, s_1, s_3, s_2, s_3|\mathcal{M}) \\ &= P(s_3|s_2)P(s_2|s_3P(s_3|s_1)P(s_1|s_1)P(s_1|s_3)P(s_3|s_3)P(s_3|s_3)P(s_3) \\ &= a_{23}a_{32}a_{13}a_{11}a_{31}a_{33}a_{33}\pi_3 \\ &= (0.2)(0.1)(0.3)(0.4)(0.1)(0.8)(0.8)1 \\ &= 1.536 \times 10^{-4} \end{aligned}$$

where π_i are the initial state probabilities:

$$\pi_i = P(q_1 = s_i), 1 \leq i \leq N \ .$$

In the above Markov model, the state corresponded to an observable event. This model is too restrictive to be applicable in many problems. Hidden Markov Model (HMM) is an extension of classical MC, where the states are not deterministic but are stochastic

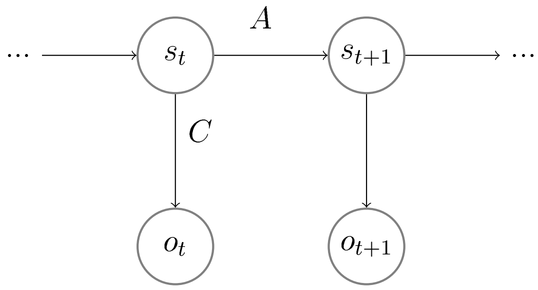


Figure 3.3: Graphical representation of a Hidden Markov Model.

A hidden Markov model is a bivariate discrete time process $\{o_t, s_t\}, t \geq 0$, where s_t is a Markov chain (sequence of states) and, conditional on this Markov chain, o_t is a sequence of independent random variables such that the conditional distribution of o_t only depends on s_t. The dependence structure of an HMM can be represented by a graphical model as in Fig. (3.3). According to the above, the complete likelihood of a sequence of length T is given by:

$$p(o_1, o_2, \cdots, o_T, s_1, s_2, \cdots, s_T) = p(s_1)p(o_1|s_1)\prod_{t=2}^{T} p(s_t|s_{t-1})p(o_t|s_t) \tag{3.14}$$

where $p(s_1)$ is the prior probability of the first state, $p(s_t|s_{t-1})$ denotes the transition probabilities from state s_{t-1} to s_t, and $p(o_t|s_t)$ are the emission probabilities for each symbol at each state. We can find the probability of observing the sequence $o_1, o_2, \cdots, o_T$ by summing over all possible hidden state, $p(O = \{o_1, o_2, \cdots, o_T\}) = \sum_{Q=\{s_1,s_2,\cdots,s_T\}} p(O, Q)$.

A HMM is described by the following features:

- the number of states in the model,$S = \{S_1, S_2, \cdots, S_N\}$
- the number of different observations symbols (alphabet), $V = \{v_1, v_2, \cdots, v_N$
- the state transition probabilities, $A = \{a_{ij}\}$,$a_{ij} = P(q_{t+1} = s_j|q_t = s_i)$ and $\sum_j a_{ij} = 1$
- the emission probabilities in state j, $C = \{c_{jk}\}$,$c_{jk} = p(v_k|s_j)$
- the initial state probabilities π_i, $pi_i = P(q_1 = s_i), 1 \leq i \leq N$ and $\sum_i pi_i = 1$

A complete specification of a HMM requires specification of the number of hidden states and observation symbols, and the specification of the three probability measures, $\lambda = (A, C, \pi)$. In the HMM literature there are three basic issues:

Problem 1 Given the observation sequence o and a model λ, how we efficiently compute the likelihood $P(o|\lambda)$ (forward backward algorithm).

Problem 2 Given the observation sequence o and the model λ, how do we define the most probable path of states (Viterbi algorithm).

Problem 3 How do we estimate the model parameters λ. Under the ML estimation framework the Baum - Welch algorithm can be applied. Also, an alternative framework is the EM algorithm which reach the same update rules for model parameters.

Extensions of HMMs are presented in [156]. The HMMs have found many interesting applications in biomedicine, especially in bioinformatics [153, 154]. An extension of HMMs, when we have continuous evolution, is the Kalman Filter, which is presented at the next section.

3.4 Kalman Filters

The Kalman Filter (KF) is a powerful tool in the analysis of the evolution of a dynamical model in time. The filter provides with a flexible manner to obtain recursive estimation of the parameters, which are optimal in the mean square error sense. The properties of KF along with the simplicity of the derived equations make it valuable in the analysis of signals. In this section an overview of the Kalman Filter, its properties and its applications are presented.

The Kalman Filter is an estimator with interesting properties like optimality in the Minimum Mean Square Error (MMSE). After its discovery in 1960 [160], this estimator has been used in many fields of engineering such as control theory, communication systems, speech processing, biomedical signal processing, etc. An analogous estimator has been proposed for the smoothing problem [161], which includes three different types of smoothers, namely fixed-lag, fixed-point and fixed interval [162, 163]. The difference between the two estimators, the Kalman Filter and the Kalman Smoother, it is related on how they use the observations to perform estimation. The Kalman Filter uses only the past and the present observations to perform estimation, while the Kalman Smoother uses also the future observations for the estimation. This means that the Kalman Filter is used for on - line processing while the Kalman Smoother for batch processing. The derivations of these two estimators is presented in [40, 164, 165]. Both estimators are recursive in nature. This means that the estimate of the present state is updated using the previous state only and not the entire past states. The Kalman Filter is not only an estimator but also a learning method [45, 164]. The observations are used to learn the states of the model. The Kalman Filter is also a computational tool and some problems may exist due to the finite precision arithmetic of the computers.

The Kalman Filter and the Kalman Smoother have been extensively used in biomedical signal processing. The general idea is to propose a model for the observations, in most cases linear, where some parameters must be estimated. To be able to apply the Kalman Filter or the Kalman Smoother the model for the observations must be written in a state-space form. A state-space model is represented by two equations: One equation, which describes the evolution of the parameters, and another equation, which describes the relation of the parameters with the observations:

$$\mathbf{w}_t = A\mathbf{w}_{t-1} + \mathbf{v}_t \tag{3.15}$$

$$\mathbf{y}_t = C\mathbf{w}_t + \mathbf{e}_t \tag{3.16}$$

These two equations represent a state-space model. In the above model $\mathbf{w}_t$ is the states vector in time t of dimension $p \times 1$, $\mathbf{y}_t$ is the vector of observations of dimension $M \times 1$, $\mathbf{v}_t$ is the state noise with zero mean and covariance matrix $\mathbf{C}_v$, $\mathbf{e}_t$ is the observation noise with zero mean and covariance matrix $\mathbf{C}_e$, A is the state transition matrix of dimension $p \times p$ and C is the observation matrix of dimension $M \times p$. All the noise processes are assumed to be independent between the time instants. In the above model the matrices A and C are assumed to be known, as well as the covariance matrices $\mathbf{C}_v$ and $\mathbf{C}_e$. However, in reality we are not able to know exactly the above matrices. In that case some assumptions are considered for the model. For example we can assume that the evolution of the parameters is a random walk procedure [166], i.e. $A = I$, where I is the identity matrix, or we restrict the matrix A to be a diagonal one [170]. Also, these matrices can be estimated through an estimation procedure like the EM algorithm [168, 169].

In [167] the authors proposed a non linear model for the electrocardiogram (ECG) signal. They use the model for ECG denoising and compression. To estimate the model parameters they use a modified version of the Kalman Filter, the Extended Kalman Filter (EKF) [165]. In [171] the authors use the Kalman Filter to detect and extract periodic noise from the ECG. In [172] they assumed that the Evoked Potentials in the Electroencephalogram can be represented as a linear combination of basis functions. The coefficients of the basis functions are assumed to change with time. This assumption lead to the use of the Kalman Filter to estimate the coefficients of the basis functions.

Besides these applications of the Kalman Filter and the Kalman Smoother for Biomedical Signal Processing, there is a particular application which has been attracted special interest, especially because at the end a time varying spectrum is obtained. This application concerns the use of parametric models such as the AR and ARMA models. The time varying autoregressive (TVAR) model is an AR model where the AR coefficients evolve in time. The parametric spectrum analysis is used to overcome the limited frequency resolution of FFT based methods. The spectral density can be calculated at each frequency point using the model parameters. The TVAR model has been used for EEG spike detection [170], for time varying - spectrum estimation of Event Related Synchronization (ERS) and Desynchronization (ERD) [168], for the calculation of coherence in the analysis of biomedical signals such EEG and ECG [84] and for time varying spectrum estimation of intracranial pressure signals from patients with traumatic brain injury [173]. In the above studies the TVAR coefficients have been estimated using the Kalman Filter or the Kalman Smoother, while in [168] the EM algorithm is used to estimate the parameters of the model.

The Kalman filtering problem is stated as follows:

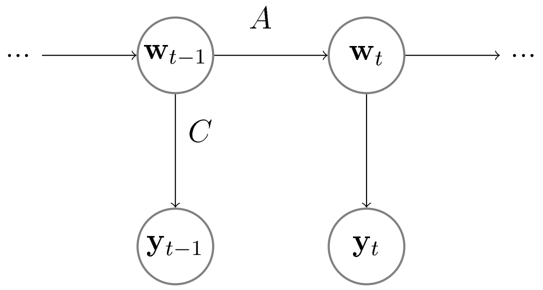


Figure 3.4: Graphical representation of the Kalman Filter.

- Use the entire observed data, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_k\}$, find for each $k \geq 1$ the minimum mean-square error estimate of the state $\mathbf{w}_t$.

The problem is called filtering if $t = k$, prediction if $t > k$ and smoothing if $1 \leq t < k$. The joint probability distribution of states and observations is given by:

$$p(\mathbf{y}_{t=1}^{N}, \mathbf{w}_{t=1}^{N}) = p(\mathbf{w}_1)p(\mathbf{y}_1|\mathbf{w}_1)\prod_{t=2}^{T} p(\mathbf{y}_t|\mathbf{w}_t)p(\mathbf{w}_t|\mathbf{w}_{t-1}) \tag{3.17}$$

and a graphical model for the above factorization is given in Fig. 3.4. We can see that is the same model as in the case of HMM.

The MMSE estimator of $\mathbf{w}_t$ based on observed data $\mathbf{Y}$ can be calculated sequentially using the following set of equations:

$$\begin{aligned}
\mathbf{w}_{t|t-1} &= A\mathbf{w}_{t|t-1}, & (3.18)\\
P_{t|t-1} &= A^T P_{t|t-1} A + \mathbf{C}_v, & (3.19)\\
K_t &= P_{t|t-1} C^T (\mathbf{C}_e + C^T P_{t|t-1} C)^{-1}, & (3.20)\\
\mathbf{w}_{t|t} &= \mathbf{w}_{t|t-1} + K_t(\mathbf{y}_t - C\mathbf{w}_{t|t-1}), & (3.21)\\
P_{t|t} &= (I - K_t C) P_{t|t-1}. & (3.22)
\end{aligned}$$

with initial condition $\mathbf{w}_{1|0} = \mu$ and $P_{1|0} = \Sigma$, where μ and Σ are the initial conditions for the states. For more information on how these equations have been derived the interested reader can look in [40, 165]. Of course the above set of equations is not in the most general form. Extensions can be made by letting the state and transition matrices to be time varying, as well as the covariance matrix of the noise processes.

From these equations we can observe how the Kalman Filter is working. To estimate the current state $\mathbf{w}_{t|t}$ a prediction step to obtain the predictive state $\mathbf{w}_{t|t-1}$ based only on the previous state $\mathbf{w}_{t-1|t-1}$ is performed. After that a correction step takes place using the present observation $\mathbf{y}_t$ and the predictive state. Also, we can observe that the update equation for the covariance matrix $P_{t|t}$ is calculated as the difference of two matrices. This can lead to numerical problems and destroy the symmetry of the matrix. To avoid these problems the update equation of covariance $P_{t|t}$ can be replaced with the so called Joseph form [163]:

$$P_{t|t} = (I - K_t C) P_{t|t-1} (I - K_t C)^T + K_t \mathbf{C}_e K_t^T. \tag{3.23}$$

3.4.1 Kalman Smoother and EM

Until now we have present the solution to the filtering problem. However, in some cases we have all the available data, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_K\}$, before the estimation of states. In that case we deal with the smoothing problem.

$$
\begin{aligned}
J_{t-1} &= P_{t-1|t-1} A^T P_{t-1|t-1}^{-1}, &(3.24)\\
\mathbf{w}_{t-1|K} &= \mathbf{w}_{t-1|t-1} + J_{t-1}(\mathbf{w}_{t-1|K} - A\mathbf{w}_{t-1|t-1}), &(3.25)\\
P_{t-1|K} &= P_{t-1|t-1} + J_{t-1}(P_{t|K} - P_{t|t-1})J_{t-1}^T. &(3.26)
\end{aligned}
$$

The derivation of those equations is explained in [165]. The equations of Kalman Filter, together with the above smoothing equations, consist the Kalman Smoother. In general to apply the Kalman Filter or the Kalman Smoother to a model, we must write the model in a state ὐ space form. After that the above equations can be applied easily. However, there are several parameters which are assumed known before the application of the update equations. These parameters are the covariance matrix of noise processes, $\mathbf{C}_e$ and $\mathbf{C}_v$, the state transiotion matrix A, the observation matrix C and the initial conditions, μ and Σ, i.e. $\theta = \{\mathbf{C}_e, \mathbf{C}_v, A, C, \mu, \Sigma\}$. To find the model parameters θ the EM algorithm can be used, where the states consist the hidden variables. The EM algorithm is an iterative scheme consisting of two steps, the E-step and the M-step. In the E-step the expected values of the hidden variables are evaluated and in the M-step the maximization is performed with respect to the model parameters. To perform the E-step the expected complete log-likelihood, $\mathcal{L} = \mathcal{E}\Big\{\log p(\mathbf{Y}, \mathbf{w}_{1:K|K}); \theta|\mathbf{Y}\Big\}$, must be calculated. The expected likelihood depends on three quantities:

$$
\begin{aligned}
\mathbf{w}_{t|K} &= \mathcal{E}\Big\{\mathbf{w}_t|\mathbf{Y}\Big\}, &(3.27)\\
S_{t|K} &= \mathcal{E}\Big\{\mathbf{w}_t\mathbf{w}_t^T|\mathbf{Y}\Big\} = P_{t|K} + \mathbf{w}_{t|K}\mathbf{w}_{t|K}^T, &(3.28)\\
S_{t,t-1|K} &= \mathcal{E}\Big\{\mathbf{w}_t\mathbf{w}_{t-1}^T|\mathbf{Y}\Big\} = P_{t,t-1|K} + \mathbf{w}_{t|K}\mathbf{w}_{t-1|K}^T. &(3.29)
\end{aligned}
$$

The first two quantities can be calculated using the Kalman Smoother equations, while for the calculation of the last quantity we can use the following equation:

$$
P_{t,t-1|K} = J_{t-1}P_{t|K}. \tag{3.30}
$$

The joint log - likelihood of the complete data $\{\mathbf{w}_0, \{\mathbf{w}_t, \mathbf{y}_t\}_{t=1}^K\}$ can be written as:

$$
\begin{aligned}
\mathcal{L} = & -\frac{1}{2}\mathcal{E}\Big\{(\mathbf{w}_0 - \mu)^T\Sigma^{-1}(\mathbf{w}_0 - \mu)^T\Big\} - \frac{1}{2}\log|\Sigma| \\
& -\sum_{t=1}^{K}\mathcal{E}\Big\{\frac{1}{2}(\mathbf{y}_t - C\mathbf{w}_t)^T\mathbf{C}_e^{-1}(\mathbf{y}_t - C\mathbf{w}_t)^T\Big\} - \frac{K}{2}\log|\mathbf{C}_e| \\
& -\sum_{t=1}^{K}\mathcal{E}\Big\{\frac{1}{2}(\mathbf{w}_t - A\mathbf{w}_{t-1})^T\mathbf{C}_v^{-1}(\mathbf{w}_t - A\mathbf{w}_{t-1})^T\Big\} - \frac{K}{2}\log|\mathbf{C}_V| \qquad (3.31)
\end{aligned}
$$

The M - step involves direct differentiation of $\mathcal{L}$ with respect to the parameters θ. The estimates for model parameters θ are given by:

$$\begin{aligned}
A_{new} &= \Big[\sum_{t=2}^{N} S_{t,t-1|N}\Big]\Big[\sum_{t=2}^{N} S_{t-1|N}\Big]^{-1} && (3.32)\\
\mathbf{C}_v &= \frac{1}{N-1}\sum_{t=2}^{T}\Big(S_{t|N} - A_{new}S_{t,t-1|N} - S_{t,t-1|N}A_{new}^T + A_{new}S_{t|N}A_{new}^T\Big) && (3.33)\\
C_{new} &= \Big[\sum_{t=1}^{N} \mathbf{y}_t\mathbf{w}_{t|N}^T\Big]\Big[\sum_{t=1}^{N} S_{t|N}\Big]^{-1} && (3.34)\\
\mathbf{C}_e &= \frac{1}{N}\sum_{t=1}^{N}(\mathbf{y}_t\mathbf{y}_t^T - C_{new}\mathbf{w}_{t|N}\mathbf{y}_t^T) && (3.35)\\
\mathbf{w}_0 &= \mathbf{w}_{1|N} && (3.36)\\
\Sigma &= P_{1|N} && (3.37)
\end{aligned}$$

The EM algorithm is consisted of two iterative steps. First applied the Kalman Smoother, using the parameters from previous step, to obtained the expected statistics, and then maximize the expected log - likelihood with respect to the parameters. These two step applied iteratively until the convergence of the likelihood.

3.5 AR and ARMA models

3.5.1 AR model

The autoregressive (AR) model is used in a diverse area of applications such as data forecasting, speech coding and recognition, model - based spectral analysis, signal restoration and biomedical signal processing and analysis. The AR model is also known as linear prediction model [113]. With the AR model we assume that the observed data have been generated by the difference equation:

$$y[n] = \sum_{k=1}^{p} a[k]y[n-k] + e[n], \qquad (3.38)$$

where $y[n]$ is the observed data, $e[n]$ is the driving noise, and the $a[k]$ are the AR coefficients. The driving noise $e[n]$ is a zero mean white noise process with variance σ_e^2, and p is the order of the model. This model is usually abbreviated as AR(p). When the model of $y[n]$ is an AR(p) model then the Power Spectrum Density (PSD) is given by [159]:

$$P_{AR}(f) = \frac{\sigma_e^2}{|1 - \sum_{k=1}^{p} a[k]e^{-j2\pi fk}|^2}. \qquad (3.39)$$

Thus, to find the PSD of an AR model we need to know the AR coefficients as well as the variance of the driving noise.

It is useful to derive the probability distribution of the AR model in more compact form since this will help us to include the AR model in probabilistic models more easily. For a signal block of N samples $[x[0], x[1], \cdots, x[N-1]]$ the N error equations can be written as:

$$\begin{bmatrix} e[0] \\ e[1] \\ e[3] \\ \cdots \\ e[N-1] \end{bmatrix} = \begin{bmatrix} y[0] \\ y[1] \\ y[3] \\ \cdots \\ y[N-1] \end{bmatrix} - \begin{bmatrix} y[-1] & y[-2] & y[-3] & \cdots & y[-p] \\ y[0] & y[-1] & y[-2] & \cdots & y[1-p] \\ y[1] & y[0] & y[-1] & \cdots & y[2-p] \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y[N-1] & y[N-2] & y[N-3] & \cdots & y[N-p-1] \end{bmatrix} \begin{bmatrix} a[0] \\ a[1] \\ a[3] \\ \cdots \\ a[p] \end{bmatrix} \tag{3.40}$$

where $[y[-1], y[-2], y[-3], \cdots, y[-p]]$ are the initial conditions. The above set of equations can be written in vector/matrix form:

$$\mathbf{e} = \mathbf{y} - \mathbf{Y}\mathbf{a}. \tag{3.41}$$

The pdf of the signal $\mathbf{y}$ given the AR coefficients and the initial conditions is equal to the pdf of the driving noise $\mathbf{e}$. Assuming that the driving noise follows a white Gaussian distribution with zero mean and variance σ_e^2 the pdf of the signal $\mathbf{y}$ is given:

$$p(\mathbf{y}|\mathbf{a}) = \Big(\frac{1}{2\pi\sigma_e^2}\Big)^{N/2} \exp\Big\{-\frac{1}{2\sigma_e^2}(\mathbf{y} - \mathbf{Y}\mathbf{a})^T(\mathbf{y} - \mathbf{Y}\mathbf{a})\Big\}. \tag{3.42}$$

The Eq. (3.40) can be written in an alternative form as

$$\begin{bmatrix} e[0] \\ e[1] \\ e[3] \\ \cdots \\ e[N-1] \end{bmatrix} = \begin{bmatrix} -a[p] & -a[p-1] & \cdots & -a[1] & 1 & \cdots & 0 & 0 & 0 \\ 0 & -a[p] & \cdots & -a[2] & -a[1] & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \cdots & -a[2] & -a[1] & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -a[p] & -a[p-1] & \cdots & -a[1] & 1 \end{bmatrix} \begin{bmatrix} y[-p] \\ y[-p+1] \\ y[-p+2] \\ \cdots \\ y[N-1] \end{bmatrix} \tag{3.43}$$

In vector/matrix notation we have:

$$\mathbf{e} = \mathbf{A}\mathbf{y}. \tag{3.44}$$

Using the above equation we can write the pdf of the signal $\mathbf{x}$ in an alternative form:

$$p(\mathbf{y}|\mathbf{a}) = \Big(\frac{1}{2\pi\sigma_e^2}\Big)^{N/2} \exp\Big\{-\frac{1}{2\sigma_e^2}\mathbf{y}^T\mathbf{A}^T\mathbf{A}\mathbf{y}\Big\}. \tag{3.45}$$

The above two versions of the AR process pdf will be used latter in this thesis.

3.5.2 ARMA model

An extension of the AR model is the Autoregressive Moving Average (ARMA) model. In this model the time series is described:

$$y[n] = \sum_{k=1}^{p} a[k]y[n-k] + \sum_{l=1}^{q} b[l]e[n-l] + e[n]. \tag{3.46}$$

It is easy to see that this model has two parts, the AR part and MA part, hence the name ARMA. In the above equation p is the order of the AR part and q is the order of the MA part. This model is usually abbreviated as ARMA(p, q). When the model of $y[n]$ is an ARMA(p, q) model then the PSD is given by [159]:

$$P_{ARMA}(f) = \sigma_e^2 \frac{|1 + \sum_{l=1}^{q} b[l] e^{-j2\pi fl}|^2}{|1 - \sum_{k=1}^{p} a[k] e^{-j2\pi fk}|^2}. \tag{3.47}$$

3.6 Gaussian Processes

A Gaussian process is a generalization of the Gaussian probability distribution. Whereas a probability distribution describes random variables which are scalars or vectors (for multivariate distributions), a stochastic process governs the properties of functions. A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution [179].

A Gaussian process is completely specified by its mean and covariance functions. We define mean function $m(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \mathbf{x}^{'})$ of a real process $f(\mathbf{x})$ as:

$$\begin{aligned} m(\mathbf{x}) &= \mathcal{E}\{f(\mathbf{x})\}, & (3.48) \\ k(\mathbf{x}, \mathbf{x}^{'}) &= \mathcal{E}\{(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}^{'}) - m(\mathbf{x}^{'}))\}. & (3.49) \end{aligned}$$

where $\mathbf{x}$ is the input vector and $\mathcal{E}\{\cdot\}$ denotes the expectation. A Gaussian process can be writen as:

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}^{'})) \tag{3.50}$$

The random variables represent the value of the function $f(\mathbf{x})$ at location $\mathbf{x}$. Often, Gaussian processes are defined over time, i.e. where the index set of the random variables is time.

The linear regression model can be seen as a Gaussian process. Assume that we have $f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w}$ where over the weights we have the prior $\mathbf{w} \sim \mathcal{N}(0, \Sigma_w)$. Then for the mean and the covariance we have:

$$\begin{aligned} m(\mathbf{x}) &= \mathcal{E}\{f(\mathbf{x})\} = \phi(\mathbf{x})\mathcal{E}\{\mathbf{w}\} = 0, & (3.51) \\ k(\mathbf{x}, \mathbf{x}^{'}) &= \mathcal{E}\{(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}^{'}) - m(\mathbf{x}^{'}))\} = \mathcal{E}\{f(\mathbf{x})f(\mathbf{x}^{'})\} \\ &= \phi(\mathbf{x})\mathcal{E}\{\mathbf{w}\mathbf{w}^T\}\phi(\mathbf{x}^{'}) = \phi(\mathbf{x})\Sigma_w\phi(\mathbf{x}^{'}). & (3.52) \end{aligned}$$

3.7 Bayesian networks

Bayesian inference is fairly simple when it involves small number of variables. However, it becomes much more complex when we want to do inference with many variables. In such problems the Bayesian networks provide a solution by adopting the Markov condition in order to represent the problem in a more efficient way. Bayesian networks are a combination of two areas: graph theory and probability theory.

A Bayesian network is a specific type of probabilistic graphical model called direct acyclic graph (DAG), where all the edges of the graph are directed and there are no

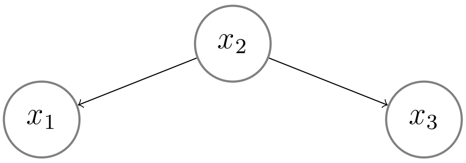


Figure 3.5: A Bayesian network.

cycles. It is a graphical model that efficiently encodes the joint probability distribution for a large set of variables. More formally, a Bayesian network for a set of variables $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ consists of a network structure S that encodes a set of conditional independence assertions about variables in $\mathbf{x}$, and a set P of local probability distributions associated with each variables. Together, these components define the joint distribution for $\mathbf{x}$. The nodes in S are in one - to - one correspondence with the variables in $\mathbf{x}$. Given the structure S, the joint distribution for $\mathbf{x}$ is given by:

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|pa(x_i))$$

where $pa(x_i)$ denotes the parents of variable x_i. The joint probability of all variables is the product of the probabilities of each variable given its parents.

In Fig. 3.5 a Bayesian network is depicted. The set of edges is $E = \{(x_2, x_1), (x_2, x_3)\}$. This is a DAG since there are no undirected edges and cycles. Further, since x_1 and x_2 are conditionally independent of each other we have:$p(x_1|x_2, x_3) = p(x_1|x_2)$. Similar conclusions can be drawn about the variable x_2. Finally, the joint distribution, as factorized by this Bayesian network, is given by: $p(x_1, x_2, x_3) = p(x_1|x_2)p(x_2)p(x_3|x_2)$.

There are three main tasks concern a Bayesian network: 1) inferring unobserved variables, 2) learning the model parameters and 3) learning the structure of the network. Efficient algorithms exist that perform inference and learning in Bayesian networks [45]. Also, Bayesian networks are used to model sequences of variables (e.g. speech signals or protein sequences), in that case are called dynamic Bayesian networks. Bayesian networks have recently been introduced as a tool for determining the dependencies between brain regions from fMRI data [215, 216].

3.8 Statistical analysis of fMRI time series

After the preprocessing of the fMRI data to meet the requirements of model assumptions, the statistical analysis is performed. In the statistical analysis, there is need to describe the data and based on this description to make a decision about the state of brain regions (activated or not). In the literature two approaches are use to for the description of the data. The first is the model -based approach, where a generative model is used to describe the data. The learning task is to estimate the model that optimally fit the data. These approaches use mainly the Generalized Linear Model (GLM) [26]. The second approach is data driven and to this approach the Independent Component Analysis (ICA) and the Principal Component Analysis (PCA) [47, 48, 49] belong. The data driven approaches do not assume a particular model. The general idea of PCA and ICA approaches is to decompose the dataset in principal or independent components and then to find a empirical relation of components with the activated area. However, the need to explore whole datasets leads to high computational costs. On the other hand, the model based approaches make an assumption for the generative model. Usually, they require less computational effort. In this thesis the generative model approach is adopted.

3.8.1 Modeling the response to the stimulus

In this section, we will describe the model for the brain response in the presence of a stimuli. In most cases the relationship between stimuli and BOLD response, $x(t)$, is modeled using a linear time invariant (LTI) system, where the stimulus, $s(t)$, acts as the input to the system and the HRF, $h(t)$, as the impulse response function. So, the BOLD response can be written as:

$$x(t) = \int_0^{\infty} h(u)s(t-u)du. \tag{3.53}$$

In the literature there exist many works which are concerned with the finding of the HRF. These works includes convolutive models, temporal basis functions, FIR models and non linear models [24, 25]. However, a simple and elegant approach, which is justifies by many studies [26, 24, 28], is to model the HRF with the difference of two gamma functions. This formulation of HRF captures the small dip after the HRF return to zero.

3.8.2 Data analysis

After the determination of the HRF and hence the BOLD signal, the time series of a voxel can be described by:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \mathbf{e}\ . \tag{3.54}$$

where $\mathbf{y}$ is the $N \times 1$ vector of voxel's time series, $\mathbf{\Phi}$ is a known design matrix of size $N \times p$ depending from the experiment, $\mathbf{w}$ is the $p \times 1$ vector of magnitudes response (or regression weights since we have a linear regression model) and $\mathbf{e}$ is the $N \times 1$ vector of noise.

After the determination of the model, we need a method to obtain estimates of weights $\mathbf{w}$. A useful, simple and widely accepted approach is the method of Ordinary Least Squares (OLS) [40]. In this method the weights $\mathbf{w}$ are found by minimizing the residuals sum-of-squares. In that case the estimates $\hat{\mathbf{w}}$ are obtained by:

$$\hat{\mathbf{w}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y}\ . \tag{3.55}$$

We want to mention here that we have make the assumption that the matrix $\mathbf{\Phi}^T\mathbf{\Phi}$ is invertible, which it is if, and only if, the design matrix $\mathbf{\Phi}$ is of full rank. Also, for the

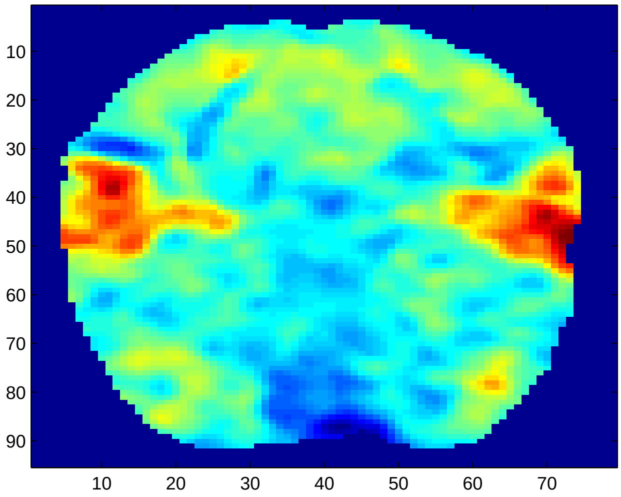


Figure 3.6: An example of statistical map.

GLM, the LS estimates are the ML estimates, and are the Best Linear Unbiased Estimates (BLUE) [40]. It can be shown that the parameters estimates are normally distributed: if the design matrix is of full rank then $\hat{\mathbf{w}} \sim \mathcal{N}(\mathbf{w}, \sigma^2(\mathbf{\Phi}^T\mathbf{\Phi})^{-1})$. This result will be used latter in the detection analysis.

After the estimation procedure, we need to use a statistic to explore the existence of an effect or not (activation or not). In neuroimaging studies two statistics have been used extensively: the t-statistics and the posterior distribution [53], which are based on the classical and the Bayesian approach, respectively.

In the classical approach the estimate of the parameters $\mathbf{w}$ is used to calculate a t-statistic for each voxel. The t-statistic is defined as:

$$t = \frac{\mathbf{c}^T\hat{\mathbf{w}}}{\sqrt{\mathbf{c}^T\mathbf{C}_{\hat{\mathbf{w}}}\mathbf{c}}}, \tag{3.56}$$

where $\hat{\mathbf{w}}$ is the estimate of the parameters $\mathbf{w}$, $\mathbf{C}_{\hat{\mathbf{w}}}$ is the covariance of the estimate $\hat{\mathbf{w}}$ and $\mathbf{c}$ is the contrast vector which specifies particular differences between the parameters $\mathbf{w}$. Then, these values of t-statistic are mapped on one brain image to produce the statistical parametric map (SPM) [54]. An image of t-values from an acoustic experiment is shown in Fig. 3.6. We can observe in this example that large values of t-statistic are concentrated at the auditory cortex, something that we expect due to the acoustic stimulus.

Under the Bayesian perspective, we can create maps of the brain based on the posterior distribution. A map of the activation regions on the brain can be obtained by computing the posterior probability that a voxel is activated or the probability that an effect is greater than some threshold value. Once, we obtain the mean and the covariance of the posterior distribution of the parameters $\mathbf{w}$, we can calculate the posterior probability, given the effect size γ, using the following equation:

$$pp = 1 - \Psi\Big(\frac{\gamma - \mathbf{c}^T\hat{\mathbf{w}}}{\sqrt{\mathbf{c}^T\mathbf{C}_{\hat{\mathbf{w}}}\mathbf{c}}}\Big), \tag{3.57}$$

where $\Psi(\cdot)$ is the normal cumulative distribution function (CDF), while $\hat{\mathbf{w}}$ and $\mathbf{C}_{\hat{\mathbf{w}}}$ are the mean and the covariance of the posterior distribution of the parameters $\mathbf{w}$. Then, these values of posterior probabilities are mapped on one brain image to produce the posterior probabilities map (PPM) [54]. The major difference between the two statistics is that the t - statistic has uniform specificity over all voxels [28]. For the OLS estimates of weights, we have $\mathbf{C}_{\hat{\mathbf{w}}} = \sigma^2(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}$ for both procedures, SPM and PPM.

Chapter 4

EEG spike detection using Kalman Filtering Techniques

4.1 Introduction

In this chapter we present a methodology for epileptic spike enhancement in electroencephalographic (EEG) recordings. The proposed approach takes advantage of the non stationarity nature of the EEG signal using a time varying autoregressive (TVAR) model. The time varying coefficients of AR model are estimated using the Kalman Filter. The results show considerably improvement in signal - to - noise ratio and significant reduction of the number of false positives. The general procedure of our approach is shown in Fig. (4.1). The EEG signal is fed to a KF, then on the output of KF a detection procedure is performed to provide us with a desicion. An EEG signal which contains four spikes is shown in Fig. (4.2).

Electroencephalography (EEG) is one of the clinical tools used in diagnosis, monitoring and management of neurophysiological disorders related to epilepsy. Epilepsy is characterized by sudden recurrent and transient disturbances of mental function and/or movements of body due to excessive discharge of brain. The presence of epileptiform activity in the EEG confirms the diagnosis of epilepsy which sometimes can be confused with other disorders producing similar seizure - like activity [60].

During the seizures (ictal activity) the scalp EEG of patients who suffer from epilepsy is usually characterized by high amplitude synchronized periodic waveforms reflecting abnormal discharge of a large groups of neurons. Between, before or after seizures (interictal activity), the EEG might contain occasional epileptiform transient waveforms. As a result relatively short recordings can still be useful in the diagnosis of epilepsy [61]. These transient waveforms, isolated spikes, sharp waves and spike wave complexes are clearly distinguished from background activity. More specifically, spikes are defined as having duration from 20 - 70 ms, while sharp waves have duration from 70 - 200 ms. On the other hand, spike and wave complexes are defined as spikes followed by slow waves and have duration from 150 - 350 ms [62, 63]. Throughout this paper, no distinction is made among

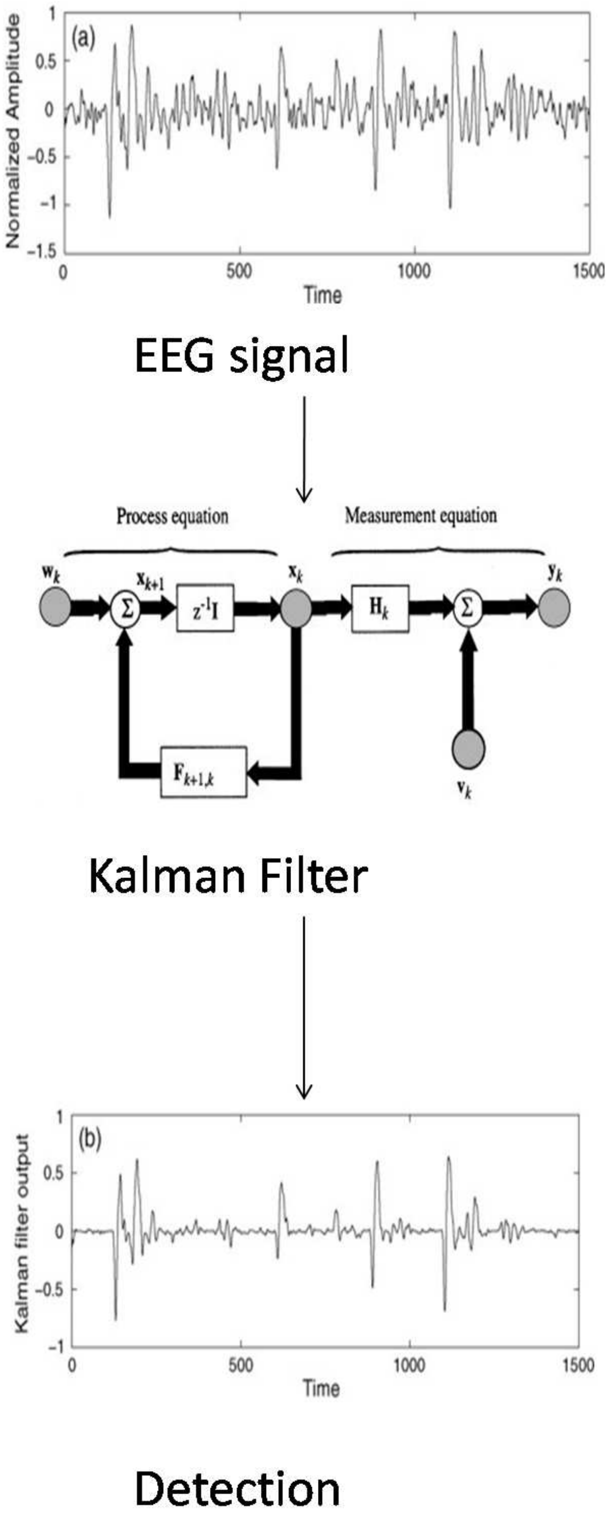


Figure 4.1: General Procedure for epileptic spike detection.

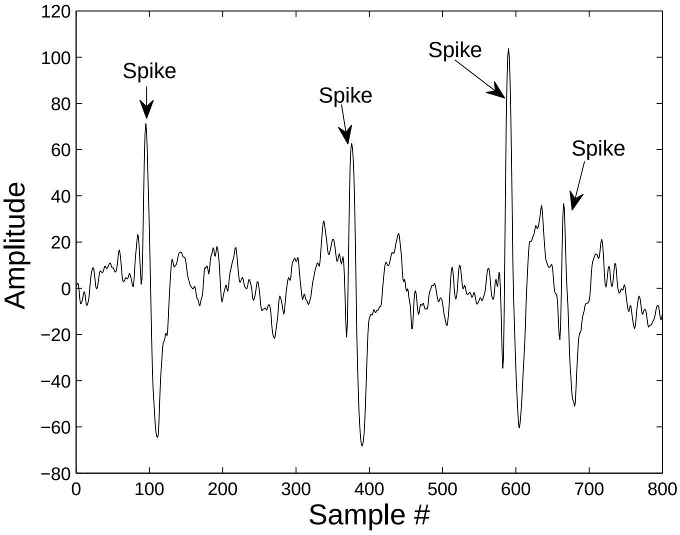


Figure 4.2: EEG signal which contains four spikes.

spike, sharp waves and spike - wave complexes and therefore they are collectively termed spikes. In general, the detection of epilepsy includes visual scanning of EEG recordings for spike by an experienced EEGer. This process, however, is time consuming, especially in the case of long recordings [62, 64]. In addition, the detection of epileptiform activity in the EEG is far from straightforward due to the variety of morphology of spikes and their similarities to waves which are part of the background activity and to artefacts (i.e. muscle activity, eye blinking activity, etc.) [21].

Several methods for spike detection have been proposed based on single and multichannel approaches. Those methods can be classified into five categories: (a) methods based on traditional recognition techniques, known as mimetic techniques [65, 66, 67], (b) methods using template matching algorithms [68], (c) methods based on parametric approaches [69], (d) methods based on artificial neural networks (ANNs) [61, 62, 63, 64, 70, 71, 72, 73, 74, 75, 76] and (e) methods utilizing knowledge-based rules [64, 77, 78, 79]. The methods belonging to the first category imitate the visual analysis followed by an expert. In particular, the features of EEG waveforms, such as duration, slope, sharpness, and amplitude, are compared with values which are provided by the experts. In the second category template matching is used for a priori known spike waveforms. The user selects manually spikes from a set of test data, which are averaged to create a template. Recent approaches use wavelets. The EEG signal is filtered using wavelets to obtain features of the signal energy which are used in the detection of spikes. The methods belonging to the third category assume local stationarity of the background activity and use single-channel or multichannel predictive filtering. Spikes are detected as deviation from stationarity. Implicit in these approaches is that non-stationarity behaviour comes only from Spikes. In the fourth category ANNs are used to recognize patterns, which are learnt by the network during the training phase. Supervised and unsupervised ANNs have been used in the diagnosis of epilepsy, either to study sleep behaviour, to detect seizures, to predict seizures or to classify and analyze waveforms in the EEG recordings. The majority of the methods, mainly those belonging to the first two categories treat single channel data only. In the fifth category, knowledge-based reasoning in addition to the above mentioned methods is widely used. This arises from the need to incorporate knowledge of the experts which takes the form of rules including temporal rules. Essentially, the spike detection problem can be simply transfered to the detection of the presence of spikes in the multichannel EEG recording with high sensitivity and selectivity. That is, a high proportion of true events must be detected with a minimum number of false detections.

Thus, a balance must be obtained between having high sensitivity and high selectivity. It is relatively easy to adjust system parameters to obtain performance where all spikes are found in a given patient but this would usually be accompanied by an unacceptably large number of false detections. On the other hand, it is also relatively easy to have a system with very low false detection rate but then this would usually be accompanied by an unacceptably large number of missed events. Many researchers argue that it is better to have a high sensitivity, minimize missed events and suffer more false detections which can be checked by the EEGer rather than missing events altogether. If we look at the system from the point of view of minimizing the number of false detections then the number of missed events will increase. However, if possible spikes can be enhanced prior to the use of a spike detector it should be possible to increase the sensitivity minimizing missed events, while maintaining the selectivity at a satisfactory level.

Thereby, a spike enhancer would not be a detector but would simply aim to enhance anything vaguely spike like. This means that real spikes, as well as spike like artefacts and background will be enhanced, i.e. a large number of unwanted waveforms will be enhanced along with real spikes. This is quite acceptable as long as the spike detection system has high selectivity. To our knowledge, there exist only a few methods that perform spike enhancement. James et al. [80] make use of multireference adaptive noise cancelling (MRANC) in which the background EEG on adjacent channels in the multichannel EEG recording is used to adaptively cancel the background EEG on the channel under investigation. In addition, adaptive noise cancelling has been applied to enhance somatosensory evoked potentials [81] and in cancelling the presence of EOG in the EEG [82]. The above methods assumed that EEG signal is a stationary one. However, it is well known that EEG contains non - stationarities. In chapter we propose a novel method for EEG spike enhancement, which combines the AR model with the KF.

4.2 Methodology

4.2.1 Time - Varying Autoregressive Model

Let the vector $\mathbf{y}$ be the one channel EEG signal. We assume that the EEG can be modeled by an autoregressive model (AR). In general, AR model found many applications in EEG analysis [83, 84], although EEG is a nonstationary signal. It can be described with the following equation:

$$y(t) = \sum_{i=1}^{p} s(i)y(t-i) + v(t), \tag{4.1}$$

where p is the order of the model, $s(i)$ the AR parameters, $y(t)$ the observations and $v(t)$ the Gaussian noise with zero mean and variance σ^2 ,i.e. $v(t) \sim N(0, \sigma^2)$. Since the EEG is non - stationary signal we let the AR parameters to vary in time:

$$y(t) = \sum_{i=1}^{p} s_t(i)y(t-i) + v(t), \tag{4.2}$$

or in vector notation:

$$y(t) = C(t)^T \mathbf{s}(t) + v(t), \tag{4.3}$$

where $C(t) = [y(t-1), y(t-2), \cdots, y(t-p)]^T$ is a px1 vector containing the p past observations. The vector $\mathbf{s}(t) = [s_t(1), \cdots, s_t(p)]^T$ contains the AR parameters and varies in time according to:

$$\mathbf{s}(t) = A\mathbf{s}(t-1) + \mathbf{w}(t), \tag{4.4}$$

where $\mathbf{w}(t)$ is Gaussian noise with zero mean and covariance Q. This describes an autoregressive model for the EEG signal with time varying coefficient in a state - space form. To estimate those coefficients we use the Kalman Filter approach (as described in chapter 3), which provides us with the set of equations:

$$\begin{aligned}
\hat{\mathbf{s}}^{t-1}(t) &= A\hat{\mathbf{s}}^{t-1}(t-1), & (4.5)\\
P_t^{t-1} &= AP_{t-1}^{t-1}A^T + Q, & (4.6)\\
\hat{\mathbf{s}}^{t}(t) &= \hat{\mathbf{s}}^{t-1}(t) + K(t)(y(t) - C(t)^T\hat{\mathbf{s}}^{t-1}(t-1)), & (4.7)\\
P_t^{t} &= (I - K(t)C(t)^T)P_t^{t-1}, & (4.8)\\
K(t) &= P_t^{t-1}C(t)(R + C(t)^T P_t^{t-1}C(t))^{-1}, & (4.9)
\end{aligned}$$

where $R = \sigma^2$. The signal that is used to the detection procedure is $z(t) = C\hat{\mathbf{s}}^t(t), t = 1, \cdots, T$.

4.2.2 Detection step

After the KF step, peaks from the output of the filter which are higher than a predefined threshold are considered as an indication of the existence of an epileptic spike at that location in the time series. In any spike detection algorithm the threshold is optimized to minimize missing of true peaks, while keeping the number of falsely detected peaks within a reasonable limit. For the proposed method the threshold value is chosen as:

$$Th = \lambda \frac{1}{N} \sum_{t=1}^{N} b(y_t) \tag{4.10}$$

where $b(y_t)$ is a segment of background EEG activity, N is the length of the segment and λ is a scaling factor.

Table 4.1: The characteristics of the EEG segments used in the evaluation of our methodology

Patient	Duration (sec)	# Epileptic Spikes
patient 1	60	44
patient 2	20	31
patient 3	20	35
patient 4	20	18
patient 5	30	25
patient 6	20	19
patient 7	20	9
patient 8	30	17
patient 9	30	47
patient 10	20	16
patient 11	40	16
patient 12	40	24
patient 13	40	32

4.3 Experimental results

4.3.1 Dataset description

All EEGs were recorded by placing electrodes on the scalp according to the International 10-20 system [85]. Sixteen channels were recorded from five bipolar montages where each electrode is referenced to an adjacent electrode. The EEGs are acquired while the patient is awake but resting and include periods of eyes open, eyes closed, hyperventilation and photic stimulation. Amplification was provided by Medelec Profile EEG machine. In order to reduce undesired noise, the recordings were sampled at 256 Hz and bandpass filtered from 1.6 - 70Hz with 12 bit resolution. Our methodology was tested on the EEG's of 13 patients who were diagnosed with epilepsy or were under evaluation at the University Hospital of Ioannina, Greece. Segments of EEG were chosen from each patient, containing spikes identified by an expert neurologist who had access to the full multichannel EEG i.e. could rate spikes based on spatial and temporal contextual information. Table 4.1 summarises the EEG characteristics of each patient.

4.3.2 Choice of the parameters

Special attention must be paid in the choice of parameters entering our methodology. Those parameters are: the variance of observation noise, the variance of the state noise, the order, p, of the time varying AR model and the matrix A. The form of matrix A reflects the correlation of coefficients between them and between different time instants. We assume that there is no correlation between the coefficients in time instant n, with those in time instant $n-1$. In addition, we assume a low degree of correlation between AR coefficients in adjusted time instants. Thus, the diagonal elements of A must have values << 1. We choose matrix A as:

$$A = \begin{pmatrix} 0.1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0.1 \end{pmatrix}.$$

The order of the AR model is $p = 15$ [69]. The variance of the observation noise, R, is $R = 1.5$ x (mean absolute value of the EEG signal). The covariance matrix, Q, of the the state noise is chosen as:

$$Q = \begin{pmatrix} 0.1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0.1 \end{pmatrix}.$$

Those values appeared to give the best results and have been chosen after long experimentation. In all experiments the raw signal has been normalized in the range $[-1, 1]$.

4.3.3 Evaluation

Theoretically SNR at time t is defined as the ratio between the amplitude of the signal at time t and the standard deviation of the noise. Such a time dependent definition is not particularly useful in neurophysiology, where SNR can be viewed as a single number which characterizes the noisiness of a spike train. In our problem the signal and noise are represented by the spikes and the background EEG, respectively. So SNR can be defined as the ratio of the peak-to-peak value to the root mean square (RMS) value of the background EEG for a number of samples on either side of the spike, excluding the spike itself (Fig. 4.3).

The spike is initially identified by the location of its maximum peak. A typical duration of 135 ms is assumed for a spike, which corresponds to 35 samples at a sampling rate 256 samples per second. The minimum sample within the range ±17 samples from the maximum peak is chosen to be the minimum peak of the spike and the peak - to - peak value S_{pp} is calculated accordingly. Finally, 35 samples (135 ms) on either side side of a 32 sample spike are chosen to describe the background EEG and its RMS value, B_{RMS}, is calculated. Thus, SNR is defined as:

$$SNR = \frac{S_{pp}}{B_{RMS}}. \tag{4.11}$$

Using this SNR definition the primary performance index used is the percentage increase in SNR defined as:

$$\Delta SNR = \frac{SNR_{new} - SNR_{old}}{SNR_{old}} \cdot 100\%, \tag{4.12}$$

where subscripts "old" and "new" refer to before and after filtering respectively.

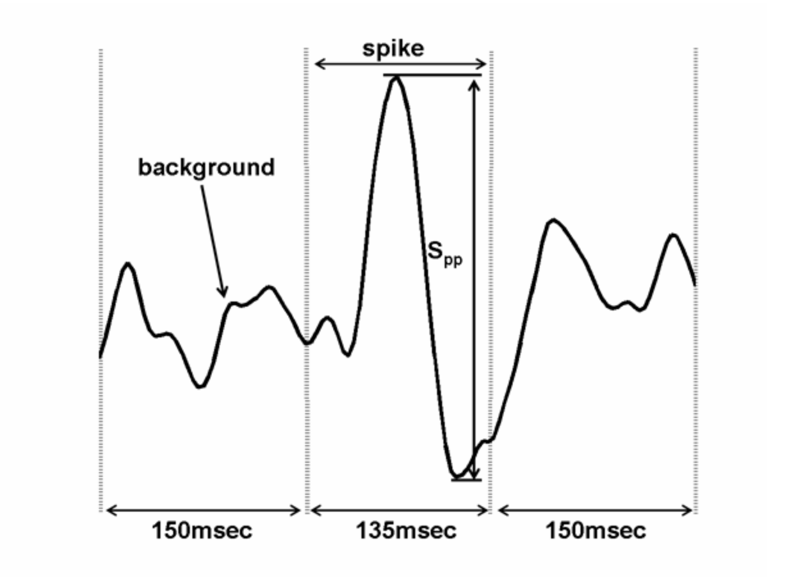


Figure 4.3: The SNR is defined as the ratio of peak-to-peak amplitude of the spike to the RMS of 35 samples on either side of the spike

Table 4.2 shows the average achieved increase of SNR for each patient. We can see that the proposed approach enhances considerably the epileptic spike with respect to background EEG activity. Fig. 4.4 depicts a segment of raw EEG signal, which contains a noisy spike, the signal after KF processing and the signal after AR processing. It is clear that in this case the AR processing produces a noisy signal, which make hard the detection of the spike. In contrast, the KF efficiently cancels the background activity and noise to produce a clear spike. Fig. 4.5 depicts spikes with low amplitude, compared to background activity. As we observe the KF was able to clearly distinguish the low amplitude spike (third and fifth spike) from background activity in contrast to AR. A different situation can be seen in Fig. 4.6. In this case we observe that the raw EEG signal contains spikes that are close to each other. The output of AR processing is noisy, especially in the time points from $t = 800 - 1000$.

A spike enhancer would not be a detector but would simply aim to enhance anything vaguely spike like. The aim of a spike enhancer is to maximize the selectivity (i.e. to decrease false detections). The lack of a proper definition of a spike other than "transients" clearly distinguished from background activity means that what constitutes the ideal spike varies. Using the above definition and making use of expert's knowledge we select as scaling factor $\lambda = 1.5$.

In Fig. 4.7 the signal before and after preprocessing is shown, as well as the attenuation of the background process after the application of KF. By a more carefull investigation of the signal after processing we can observe that the false detections have been considerably reduced.

The performance of our methodology is evaluated in terms of specificity and sensitivity. Table 4.3 shows the four possibilities which exist for each decision made by the system. In the case of a true positive the system identifies an EEG segment as spike which was

Table 4.2: Comparison of average (%) increase in SNR between Kalman Filter and inverse AR filtering

Patient	Kalman Filter	Inverse AR Filter
patient 1	111.89	28.33
patient 2	58.14	24.71
patient 3	53.05	16.27
patient 4	63.53	-12.56
patient 5	139.06	30.12
patient 6	163	32.52
patient 7	145.45	-7.07
patient 8	44.97	21.38
patient 9	373.54	-11.58
patient 10	58.48	28.27
patient 11	167.23	-11.2
patient 12	221.13	-12.62
patient 13	137.85	26.96
Average	133.64	11.81

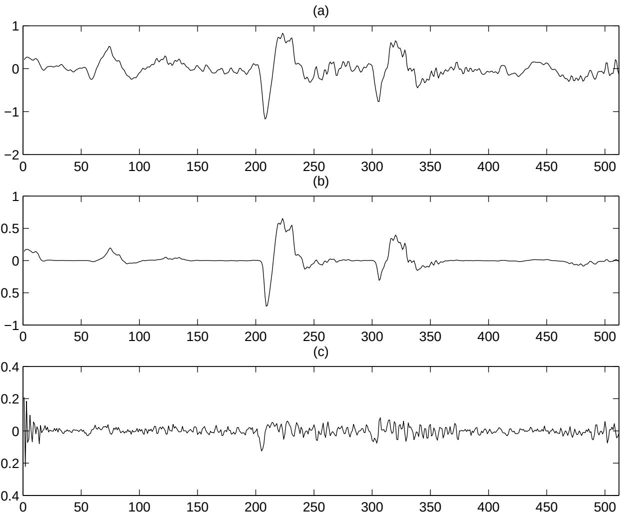


Figure 4.4: (a) raw EEG signal (b) signal after KF processing (c) signal after AR processing

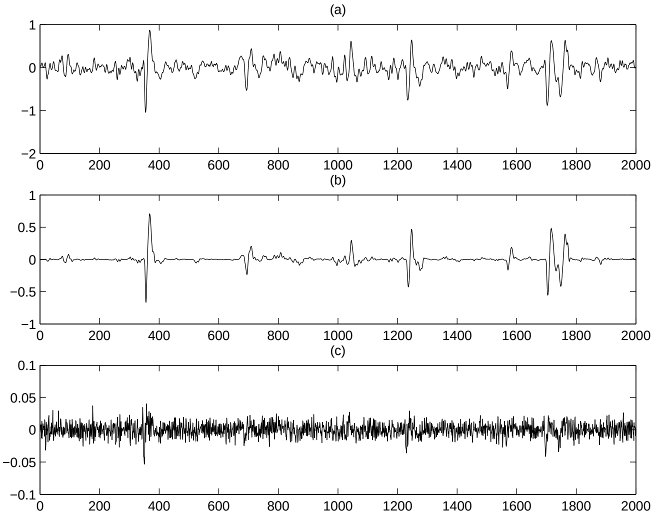


Figure 4.5: (a) raw EEG signal (b) signal after KF processing (c) signal after AR processing

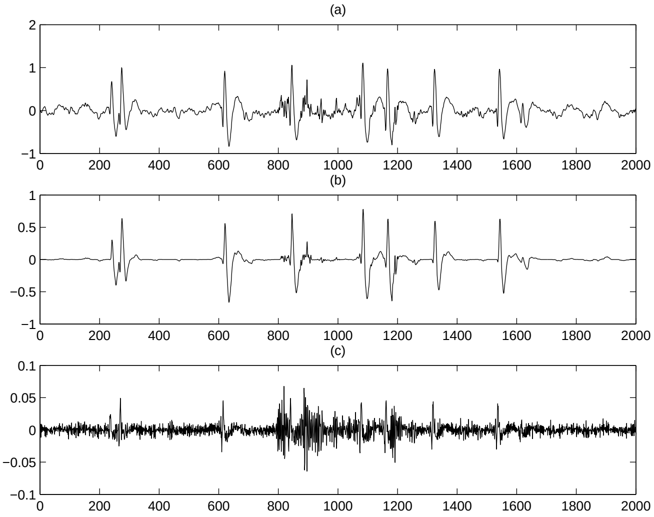


Figure 4.6: (a) raw EEG signal (b) signal after KF processing (c) signal after AR processing

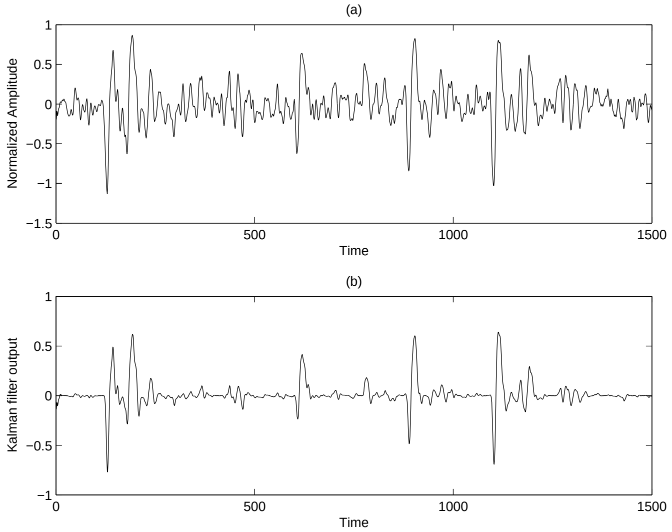


Figure 4.7: (a) raw EEG signal and (b) signal after KF processing

Table 4.3: Confusion Matrix

	System = spike	System = no - spike
Label = spike	True Positive (TP)	False Negative (FN)
Label = no-spike	Fasle Positive (FP)	True Negative (TN)

annotated such as by the expert. A false positive is the detection of a spike which is annotated as normal by the expert. A false negative indicates that the system has missed a spike. Finally, in the case of a true negative the system and the expert both agree that the EEG segment is normal. In table 4.4 the results from the detection procedure are shown. As we can see the use of a spike enhancer decreases the false detections.

In Fig. 4.8 the Power Spectral Density of an EEG segment using Welch's averaged, modified periodogram method with window length equal to 512 is shown. As we can see in Fig. (4.8a) the signal before KF enhancement exists a lobe in the frequency range from 50 - 80 Hz. This frequency range corresponds to EEG components that are irrelevant to the spike components. The application of KF in the EEG signal attenuates these components as we observe in Fig. (4.8b). Based in this observation we can conclude that the application of KF corresponds to a low pass filter.

Table 4.4: Detection Perfomance

	FN	FP	TP
without enhancer	38	1425	295
with enhancer	49	680	284

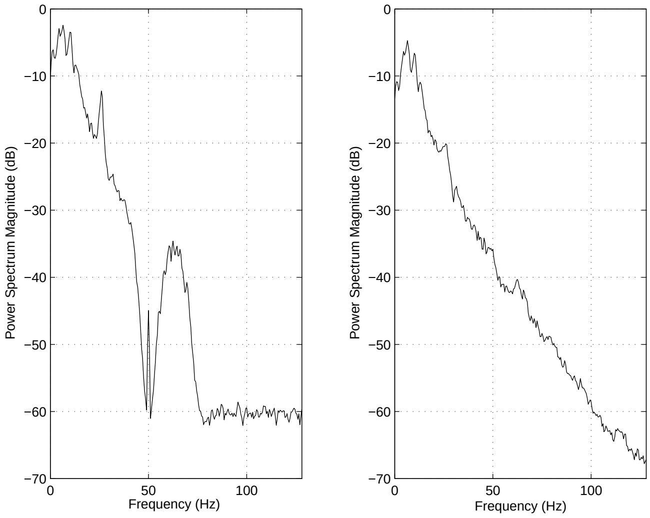


Figure 4.8: (a) PSD before KF preprocessing and (b) PSD after KF preprocessing

4.4 Discussion and Conclusion

In this chapter we have presented a methodology for EEG recordings spike enhancement. It is based on the assumption that EEG consists of an underlying background activity, which is assumed to be stationary, and superimposed transient non - stationarities. The method uses a time varying AR model for the enhancement of spikes. The parameters of the model are estimated by Kalman filter. The use of time vaying AR model enhance spikes. The sensitivity of the detection process was increased compared to the case without any preprocessing stage, i.e. when the raw EEG is used as input in detection stage. However, further analysis is required for the classification of the enhanced transients into epileptic spikes or other events.

Using the time varying AR model allows the EEG to be modeled as a time - varying process. Using this formulation we are able to enhance existing spikes and other events which are similar to spikes. Usually the published works on spike detection use a preprocessing stage to enhance spikes in EEG recordings [62]. However, only in [80] spike enhancement is explicitly addressed. They use Multireference Adaptive Noise Cancelling (MRANC). The EEG on nearby channels in the multichannel EEG recordings is used adaptively to cancel the background activity. The MRANC uses spatial and temporal information to enhance the spikes but as reported in [80] the presence of signal crosstalk between the primary and reference channel affects its performance. Another factor affecting MRANC is the correlation between the noise source in different channels. In contrast, our method uses the temporal information and the time varying nature of EEG components to enhance the spikes. With the use of the Kalman Filter we are able to suppress the background activity. Issues related to the correlation between noise or signal crosstalk do not enter the proposed approach.

One factor affecting the performance of our method is the variance of background activity compared to the amplitude of the spike. Spikes having similar amplitude with

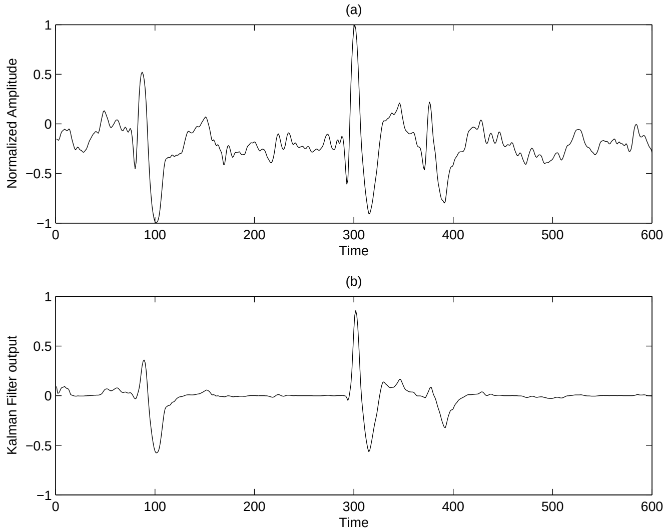


Figure 4.9: (a) Raw EEG segment (b) EEG segment after processing

the background EEG are supressed. This is shown in Figure 4.9, where a spike exists in $t = 400$ and its amplitude is less than the background activity. This happens because apart from the spike detection on a single channel itself, other contextual information is also used by the expert when he classifies events as epileptic or non epileptic. This information is related to other channel activity which takes places at the same time. The proposed method doesn't take advantage of the spatial information but ”inspects” each recording channel individually.

Our future work will focus on the use of such information in making the final diagnosis. Specifically, the use of multichannel information guides us to extend the Kalman Filter to the multichannel case. Another approach is the use of spatial combiner which utilizes the epileptic spikes across channels to detect the presence of epileptic events. However, such information must be included in an automated diagnosis system. More specifically, Kalman filtering must be applied in multichannel recordings. Alternatively, diagnosis must be assisted by a module which combines information about epileptic spikes from different channels.

Chapter 5

Biomedical Signal Denoising using the Variational Bayesian Approach with Applications to ERP Estimation and HRV Analysis

5.1 Introduction

In this chapter a Bayesian approach is proposed for the removal of the noise in biomedical signals. The biomedical signal is assumed to be smooth and it is observed with additive noise. The smoothness over the signal is achieved through a capable smoothness prior, while the statistics of the noise are unknown and must be estimated. The estimation is based on an hierarchical approach. The hyperparameters, which contain the degree of smoothness of the signal and the noise statistics, and the signal, are estimated using the Variational Bayesian (VB) Methodology. Results for single trial Event Related Potential (ERP) estimation are presented. The performance of the proposed method is evaluated in simulated and real ERP data and compared to the well known wavelet denoising approach and the Generalized Cross - Validation (GCV) criterion. The use of the proposed method results in a 4% increase in the classification rate. Also, the proposed method is used to estimate and remove the trend from Heart Rate Variability (HRV) signals.

Biomedical Signal Denoising attempts to improve one or more perceptual aspects of the signals corrupted by noise [110, 111]. Denoising is refereed to the process of recovering the clean signal from noisy observations. The removal of noise is a crucial step for any system which processes biomedical signals. The accuracy of all subsequent steps, e.g. detection, classification, etc., strongly depends on the quality of the noise reduction process. For example in [170] the authors use signal enhancement techniques for the detection of spike waves in the Electroencephalogram (EEG). However, signal denoising is not only met in biomedical signals. Denoising of a signal in a noisy environment can be employed for other types of signals such as speech signals [113]. In general, Signal denoising is related to the

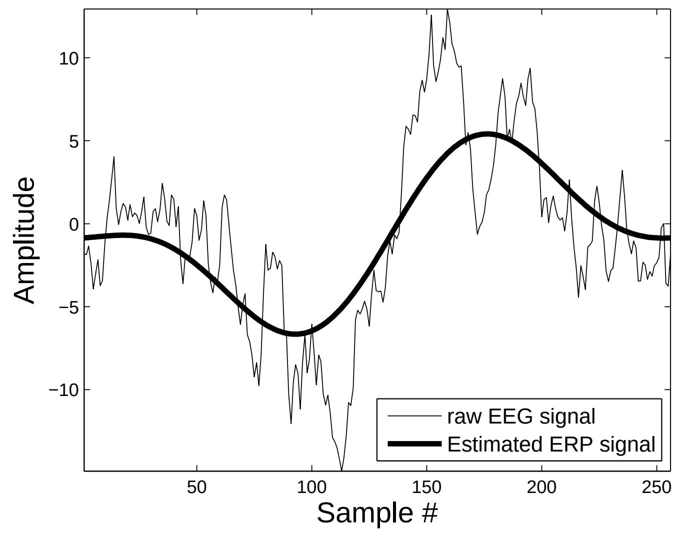


Figure 5.1: Raw EEG signal (or noisy ERP signal) and the estimated ERP signal using the wavelet denoising approach.

type of noise, the way the noise interacts with the signal and the number of available channels. To understand better the process of denoising we present an example based on the single trial ERP analysis. The ERP is the electric activity of the brain due to a stimulation. The measured responses can be considered as a combination of the brain activity due to stimulation plus the brain activity not related to the stimulation. The ERP is usually considered as transient - like smooth waveforms which are dominated by low frequencies [172]. A common approach to denoise the single trial ERP is to construct a filter and filter out the unwanted contribution of the on-going background activity of the brain. Digital filters can be used for this purpose. However, this approach presents two major drawbacks. First, the spectrum of the ERP must be completely known and secondly, the spectrum of the EP and the noise are usually overlapped. In this case the Wiener filter can be used. Using the Wiener filter the covariance matrix of the EP and the noise must be known a priori. The covariance of the noise can be estimated from EEG segments before the stimulation. However, the estimation of ERP covariance is a difficult task. Now, the problem is to proposed an accurate model for the covariance matrix of ERP. In our method we propose a simple and elegant structure for the covariance matrix of the clean signal through the prior distribution. In Fig. (5.1) a trial of EEG signal is shown. This EEG signal was obtained during the presence of a stimulus and can be also called the noisy ERP signal, since contains the EEG activity due to the stimulus plus the EEG activity unrelated to stimulus. The goal of denoising an ERP signal is to remove the irrelevant EEG activity and to recover the ERP signal. In Fig. (5.1), the recovered ERP signal using the wavelet denoising approach is shown.

Two general approaches are followed in biomedical signal denoising. The first is the model-based approach, where a model is used to explain the data. The model is fit to the data and the model parameters are estimated. The wavelet denoising [114, 115] and the linear model [116, 117] belong to this approach. The second approach is data driven and to this approach the Independent Component Analysis (ICA) [118] and the Principal Component Analysis (PCA) [120] belong.

In this chapter the model-based approach is adopted and the linear model is used due to its simplicity and its analytical expressions. The linear model finds many applications in biomedical signal processing, since, it has been used in the analysis of fMRI data [26] and in the estimation of Event Related Potentials (ERPs) [116]. When the linear model is used in a problem we face two problems. The first is related to the design matrix which is used and the second to the use of "best" parameters of the linear model. In most cases the design matrix is determined by the problem under discussion. Finding the optimal parameters values is related with the estimation framework. Two general schemes can be applied. The classical inference framework and the Bayesian inference framework [40]. In our approach we adopt the Bayesian framework since we can use prior knowledge in the estimation procedure through the prior distribution. In the Bayesian framework the most valuable quantity is the posterior distribution. In some cases the posterior distribution cannot be evaluated analytically and approximation techniques can be used such as the Variational Bayesian (VB)[45, 42], the Empirical Bayes (EB), the Laplace Approximation [45] and the Markov Chain Monte Carlo (MCMC). However, in the Laplace approximation the Gaussian assumption is based on the large data limit and the obtained posterior is poorly represented for small datasets, besides that we need many operations to compute the derivatives of the Hessian [42]. Similarly, in the MCMC methods the number of samples required for accurate estimates is infeasible large [42]. In addition, the absence of a global measure to ascertain whether the Markov Chain has reached equilibrium is a problem [42]. On the contrast the VB methodology is an efficient computational method since it results in closed form solutions and a universally accepted criterion exists to stop the process, which is the convergence of the variational bound. There exist similarities between the VB and EB methodologies, but the EB methodology results from a ML estimation procedure [45, 121].

In this chapter we present a method for the recovery of a biomedical signal which is observed in noise. The model we use is the additive one. To obtain a meaningful solution we need to impose some restrictions about the signal smoothness. The smoothness property is often used in biomedical signal processing, for example in [50] for fMRI data analysis, in [117] for ERP estimation and in [122] for the detrending of Heart Rate Variability (HRV) signal. In the Bayesian framework this property can be embedded through the use of a capable prior distribution [123, 124]. For the noise two cases are considered: the white Gaussian and the colored Gaussian noise. To estimate the various quantities of this model the VB methodology is used. The innovation of our work is related to the way we estimate the smoothness of the signal and the statistics of the noise. The proposed method provides with simultaneous estimation of the signal smoothness and the noise statistics within the same estimation framework. This feature avoids the visual tuning of the smoothness parameter as it is proposed in [122]. Also, for the smoothness parameter we obtain a posterior distribution in contradiction to [117, 50] where point estimates for the smoothness parameter are provided through the GCV criterion. Finally, the transformation of the resulting equations into the Fourier Domain, using the Discrete Fourier Transform (DFT), provides with efficient computational algorithms. What makes our approach different from others [117, 50, 122] is that all model parameters are estimated simultaneously into the same estimation framework. This is described for two cases: (a) white Gaussian noise and (b) colored Gaussian noise.

The chapter is organized as follows. First, the Bayesian model is described. Second, the VB methodology is applied to obtain the posterior distributions of the model. Next, the proposed algorithms are applied in simulated and real datasets. In the simulated datasets, first the numerical simulations and the evaluation metrics are described. After that, a comparison of the proposed algorithms with the Generalized Cross ὐ Validation (GCV) approach and the wavelet denoising is performed. In the real datasets, we applied the proposed algorithms in two cases: ERP estimation in EEG signal and removal of the trend in HRV signals, and compare to the wavelet denoising. Finally, the results are discussed along with the future work.

5.2 Methodology

We consider a signal $s(k)$ corrupted by additive Gaussian noise $n(k)$. The raw signal (observations) can be expressed as:

$$y(k) = s(k) + n(k), \tag{5.1}$$

where k is the index sample (or time), $k = 1, \cdots, N$, with N being the number of samples. Eq. 5.1 can be written in vector notation:

$$\mathbf{y} = \mathbf{s} + \mathbf{n}, \tag{5.2}$$

where $\mathbf{y} = [y(1), y(2), \cdots, y(N)]$, $\mathbf{s} = [s(1), s(2), \cdots, s(N)]$ and $\mathbf{n} = [n(1), n(2), \cdots, n(N)]$. The signal $\mathbf{s}$ is uncorrelated to the noise. The problem is to estimate the signal $\mathbf{s}$ given the observations $\mathbf{y}$. The ML solution in this problem is meaningless because the ML estimator corresponds to the observations. To obtain a meaningful solution regularization is required. The constraint is chosen ad hoc or it is based on some a priori information. In our study, the signal is constrained to be smooth. This means that we expect neighborhood samples of the signal to have similar values, i.e. the signal $\mathbf{s}$ has high correlation. This property is useful when we study biomedical signals. Since we use a Bayesian approach, the smoothness property must be introduced to our model through the prior distribution. For this reason we choose the smoothness prior [123]. However, the use of smoothness prior introduces a new parameter (see below) into our model, which enforces the use of a hierarchical Bayesian model to deal with it. We mention here that the term smoothness refers to temporal smoothness. In the next sections we will present the application of the VB methodology in the cases of white Gaussian and Colored Gaussian noise.

5.2.1 White Gaussian Noise

The smoothness prior over the signal $\mathbf{s}$ is given as:

$$p(\mathbf{s}|\alpha) \propto \Big(\frac{\alpha}{2\pi}\Big)^{N/2} \exp\Big\{-\frac{\alpha}{2}\mathbf{s}^T\mathbf{L}^T\mathbf{L}\mathbf{s}\Big\}. \tag{5.3}$$

This prior has been used to estimate the trend in HRV and fMRI time series [50, 122] and to estimate the ERP signal [117]. The matrix $\mathbf{L}$ is a discrete approximation of the d-th derivative operator. The noise is assumed to be white Gaussian, i.e.

$$p(\mathbf{n}|\lambda) = \Big(\frac{\lambda}{2\pi}\Big)^{N/2} \exp\Big\{-\frac{\lambda}{2}\mathbf{n}^T\mathbf{n}\Big\}. \tag{5.4}$$

where λ is the precision of the noise (inverse variance). Due to this assumption, the likelihood of the observations, given the signal $\mathbf{s}$ and the noise precision λ, is:

$$p(\mathbf{y}|\mathbf{s},\lambda) = \Big(\frac{\lambda}{2\pi}\Big)^{N/2} \exp\Big\{-\frac{\lambda}{2}(\mathbf{y}-\mathbf{s})^T(\mathbf{y}-\mathbf{s})\Big\}. \tag{5.5}$$

Finally, we assume that α and λ are Gamma variables:

$$\begin{aligned} p(\alpha) &= \Gamma(\alpha; b_\alpha, c_\alpha), & (5.6)\\ p(\lambda) &= \Gamma(\lambda; b_\lambda, c_\lambda), & (5.7)\end{aligned}$$

where

$$\Gamma(x;b,c) = \frac{1}{\Gamma(c)}\frac{x^{(c-1)}}{b^c}\exp\Big\{-\frac{x}{b}\Big\}. \tag{5.8}$$

The choice of Gamma distribution is based on the fact that the Gaussian and Gamma distributions are conjugates [125]. We observe that the prior over the signal is not completely known but depends on the parameter α, which is unknown and must be estimated. The same happens with the parameter λ. Since we assume that the signal is uncorrelated to the noise, the join prior of our model can be written as:

$$p(\mathbf{s},\alpha,\lambda) = p(\mathbf{s}|\alpha)p(\alpha)p(\lambda). \tag{5.9}$$

In the case, where the parameters α and λ are known, an estimate of the signal can be obtained as:

$$\hat{\mathbf{s}} = \lambda(\alpha\mathbf{L}^T\mathbf{L} + \lambda\mathbf{I})^{-1}\mathbf{y}, \tag{5.10}$$

where $\mathbf{I}$ is the identity matrix. The above estimator is the MAP estimator of the proposed model. The use of the MAP estimator assumes that the values of the parameters α and λ are known. However, in our problem these parameters are unknown and must be estimated using the observations. There exist several methods which address the estimation of those parameters: the evidence based approach [33, 152], the integration method [126], and the ensemble learning or VB methodology [42, 148, 152]. The above methods are based on a Bayesian treatment of the problem. Approaches outside the Bayesian framework can be also used such as the generalized cross validation criterion [50, 124].

In our study to perform inference about the signal $\mathbf{s}$ and the parameters α and λ we use the VB methodology (see Chapter 2). We approximate the true posterior with the factorized distribution:

$$q(\mathbf{s}, \alpha, \lambda|\mathbf{y}) = q(\mathbf{s})q(\alpha)q(\lambda). \tag{5.11}$$

Applying the VB methodology the Equations (5.12)-(5.19) are obtained: The posterior over the signal $\mathbf{s}$ is a Gaussian distribution with mean and covariance given by:

$$\begin{aligned} \hat{\mathbf{s}} &= \hat{\lambda}\mathbf{C}_{\mathbf{s}}\mathbf{y}, & (5.12)\\ \mathbf{C}_{\mathbf{s}} &= (\hat{\alpha}\mathbf{L}^T\mathbf{L} + \hat{\lambda}\mathbf{I})^{-1}. & (5.13)\end{aligned}$$

The posterior over the parameter λ is a Gamma distribution with parameters:

$$\begin{aligned} \frac{1}{b_{\lambda}^{'}} &= \frac{1}{2}\Big(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{s} + trace(\mathbf{C}_{\mathbf{s}} + \hat{\mathbf{s}}\hat{\mathbf{s}}^T), & (5.14)\\ c_{\lambda}^{'} &= \frac{N}{2} + c_{\lambda}, & (5.15)\\ \hat{\lambda} &= b_{\lambda}^{'}c_{\lambda}^{'}. & (5.16)\end{aligned}$$

The posterior over the parameter α is a Gamma distribution with parameters:

$$\begin{aligned} \frac{1}{b_{\alpha}^{'}} &= \frac{1}{2}\Big(trace(\mathbf{L}^T\mathbf{L}(\mathbf{C}_{\mathbf{s}} + \hat{\mathbf{s}}\hat{\mathbf{s}}^T)), & (5.17)\\ c_{\alpha}^{'} &= \frac{N}{2} + c_{\alpha}, & (5.18)\\ \hat{\alpha} &= b_{\alpha}^{'}c_{\alpha}^{'}. & (5.19)\end{aligned}$$

The algorithm consists of the iterative application of the Equations (5.12)-(5.19) until the convergence of the variational bound. This algorithm is called VarWhite. The variational bound is given by:

$$\begin{aligned} F(q, \mathbf{s}, \alpha, \lambda) = \Big\langle \log p(\mathbf{y}|\mathbf{s}, \lambda)\Big\rangle - KL(q(\mathbf{s})||p(\mathbf{s})))\\ -KL(q(\lambda)||p(\lambda))) - KL(q(\alpha)||p(\alpha))), \qquad (5.20)\end{aligned}$$

where:

$$\begin{aligned} KL(q(\mathbf{s})||p(\mathbf{s}))) &= \frac{N}{2}\log\hat{\alpha} - \frac{1}{2}\log|\mathbf{C}_{\mathbf{s}}| + \frac{1}{2}\Big(\hat{\alpha}\mathbf{L}^T\mathbf{L}\mathbf{C}_{\mathbf{s}}\Big) + \frac{\hat{\alpha}}{2}\hat{\mathbf{s}}^T\mathbf{L}^T\mathbf{L}\hat{\mathbf{s}}, & (5.21)\\ KL(q(\lambda)||p(\lambda))) &= (c_{\lambda}^{'} - 1)\Psi(c_{\lambda}^{'}) - \log b_{\lambda}^{'} - c_{\lambda}^{'} + \log\Gamma(c_{\lambda}^{'}) + \log\Gamma(c_{\lambda}) + \\ &\quad c_{\lambda}\log b_{\lambda} - (c_{\lambda} - 1)(\Psi(c_{\lambda}^{'}) + \log b_{\lambda}^{'}) + \frac{b_{\lambda}^{'}c_{\lambda}^{'}}{b_{\lambda}}, & (5.22)\\ KL(q(\alpha)||p(\alpha))) &= (c_{\alpha}^{'} - 1)\Psi(c_{\alpha}^{'}) - \log b_{\alpha}^{'} - c_{\alpha}^{'} + \log\Gamma(c_{\alpha}^{'}) + \log\Gamma(c_{\alpha}) + \\ &\quad c_{\alpha}\log b_{\alpha} - (c_{\alpha} - 1)(\Psi(c_{\alpha}^{'}) + \log b_{\alpha}^{'}) + \frac{b_{\alpha}^{'}c_{\alpha}^{'}}{b_{\alpha}}, & (5.23)\\ \Big\langle \log p(\mathbf{y}|\mathbf{s}, \lambda)\Big\rangle &= \frac{N}{2}\log\hat{\lambda} - \frac{\hat{\lambda}}{2}\Big(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{s} + trace(\mathbf{C}_{\mathbf{s}} + \hat{\mathbf{s}}\hat{\mathbf{s}}^T)\Big). & (5.24)\end{aligned}$$

The calculation of the KL divergence for various distributions is explained in [45].

5.2.2 Colored Gaussian Noise

In this section the previous model is extended in the case of colored Gaussian noise, i.e. a stationary process which follows the Gaussian distribution with full covariance. This extension makes the proposed model more robust since it includes the previous model as a special case and it can be used in cases where the noise is described as colored Gaussian. For example in ERP estimation the background EEG is better modeled by a colored Gaussian distribution [117]. The observation model is given by Eq. (5.2), the prior of the clean signal by Eq. (5.3) and the hyperprior for α by Eq. (5.6). We make the same assumptions about these parameters as in the case of the white noise. The noise is assumed to be colored Gaussian with zero mean and covariance matrix $\mathbf{C_n}$. The parameters in this case are the signal $\mathbf{s}$, the parameter α and the inverse covariance of the noise, $\mathbf{R} = \mathbf{C_n}^{-1}$. For the covariance of the noise we use as prior the Wishart density [125, 127, 128]

$$p(\mathbf{R}) = W(r_p, \mathbf{B}_p). \tag{5.25}$$

The prior over the parameters can be written as:

$$p(\mathbf{s}, \alpha, \mathbf{R}) = p(\mathbf{s}|\alpha)p(\alpha)p(\mathbf{R}). \tag{5.26}$$

We can apply the VB methodology as in the case of the white Gaussian noise. The posterior is approximated by $q(\mathbf{s}, \alpha, \mathbf{R}|\mathbf{y}) = q(\mathbf{s}|\alpha)q(\alpha)q(\mathbf{R})$. The approximate posteriors in this case are:

$$\begin{aligned} q(\mathbf{s}) &= \mathcal{N}(\hat{\mathbf{s}}, \mathbf{C_s}), && (5.27)\\ q(\alpha) &= \Gamma(\alpha; b'_\alpha, c'_\alpha), && (5.28)\\ q(\mathbf{R}) &= W(r, \mathbf{B}), && (5.29)\\ & && (5.30) \end{aligned}$$

where:

$$\begin{aligned} \mathbf{C_s} &= (\hat{\alpha}\mathbf{L}^T\mathbf{L} + \hat{\mathbf{R}})^{-1}, && (5.31)\\ \hat{\mathbf{s}} &= \mathbf{C_s}\hat{\mathbf{R}}\mathbf{y}, && (5.32)\\ \frac{1}{b'_\alpha} &= \frac{1}{2}\Big(trace\Big(\mathbf{L}^T\mathbf{L}\Big\langle \mathbf{ss}^T\Big\rangle\Big)\Big), && (5.33)\\ c'_\alpha &= \frac{N}{2} + c_\alpha, && (5.34)\\ r &= r_p + 1, && (5.35)\\ \mathbf{B} &= \mathbf{B}_p + \Big((\mathbf{y} - <\mathbf{s}>)(\mathbf{y} - <\mathbf{s}>)^T\Big) + \mathbf{C_s}. && (5.36) \end{aligned}$$

The required moments are easily evaluated as:

$$\begin{aligned} <\mathbf{s}> &= \hat{\mathbf{s}}, & (5.37)\\ \left\langle \mathbf{s}\mathbf{s}^T \right\rangle &= \mathbf{C}_{\mathbf{s}} + \hat{\mathbf{s}}\hat{\mathbf{s}}^T, & (5.38)\\ \hat{\alpha} &= b^{'}_{\alpha} c^{'}_{\alpha}, & (5.39)\\ \hat{\mathbf{R}} &= r\mathbf{B}^{-1}. & (5.40) \end{aligned}$$

The algorithm consists of the iterative application of Equations (5.31) - (5.36) until convergence of the variational bound. The algorithm is called VarColored. The variational bound is given by:

$$\begin{aligned} F(q, \mathbf{s}, \alpha, \lambda) = \Big\langle \log p(\mathbf{y}|\mathbf{s}, \lambda) \Big\rangle - KL(q(\mathbf{s})||p(\mathbf{s}))) \\ -KL(q(\mathbf{R})||p(\mathbf{R}))) - KL(q(\alpha)||p(\alpha))), \end{aligned} \qquad (5.41)$$

In the above equation the KL divergence for the signal $\mathbf{s}$ and the parameter α are the same as in the white noise case. For the other quantities we have:

$$\begin{aligned} KL(q(\mathbf{R})||p(\mathbf{R})) &= \frac{r-N-1}{2} L(r, \mathbf{B}) - \frac{r_p - N - 1}{2} L(r_p, \mathbf{B}_p) - \frac{rN}{2} \\ &\quad + \frac{r}{2} trace\Big\{ \mathbf{B}_p \mathbf{B} \Big\} + \log \frac{Z(r_p, \mathbf{B}_p)}{Z(r, \mathbf{B})}, & (5.42)\\ \Big\langle p(\mathbf{y}|\mathbf{s}, \mathbf{R}) \Big\rangle &= \frac{1}{2} L(r, \mathbf{B}) - \frac{1}{2} \Big\{ \hat{\mathbf{R}} \Big((\mathbf{y} - <\mathbf{s}>)(\mathbf{y} - <\mathbf{s}>)^T + \mathbf{C}_{\mathbf{s}} \Big) \Big\} . & (5.43) \end{aligned}$$

The terms $Z(r, \mathbf{B})$ and $Z(r_p, \mathbf{B}_p)$ and are the normalized quantities of the posterior and prior distribution of , which are Wishart distributions. The terms $L(r, \mathbf{B})$ and $L(r_p, \mathbf{B}_p)$ are given as:

$$\begin{aligned} L(r, \mathbf{B}) &= \int \log |\mathbf{R}| W(r, \mathbf{B}) d\mathbf{R}, & (5.44)\\ L(r_p, \mathbf{B}_p) &= \int \log |\mathbf{R}| W(r_p, \mathbf{B}_p) d\mathbf{R}, & (5.45) \end{aligned}$$

which can be seen as the expectation of the quantity $\log |\mathbf{R}|$ with respect to the posterior and the prior distributions of the noise precision matrix $\mathbf{R}$.

5.2.3 Stationarity assumptions

The stationarity of the proposed estimator for the signal $\mathbf{s}$ is depended on its prior and more specifically on the structure of the matrix $\mathbf{L}$. Because the signal is considered finite and includes N samples, the stationarity depends on the conditions at the beginning and at the end of the signal [123]. Assuming that the signal vanishes outside its domain i.e.

$y(-1) = y(N+1) = 0$, the matrix $\mathbf{L}$ for $d = 2$ becomes:

$$\mathbf{L} = \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & 0 & 0 & -1 & 2 \end{bmatrix} \tag{5.46}$$

This choice of such matrix gives a non stationary prior for the signal $\mathbf{s}$, because the matrix $\mathbf{L}^T\mathbf{L}$ takes the following form:

$$\mathbf{L}^T\mathbf{L} = \begin{bmatrix} 5 & -4 & 1 & 0 & 0 & \cdots & 0 & 0 \\ -4 & 6 & -4 & 1 & 0 & \cdots & 0 & 0 \\ 1 & -4 & 6 & -4 & 1 & \cdots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \cdots & 0 & 0 & 1 & -4 & 6 & -4 \\ 0 & 0 & \cdots & 0 & 0 & 1 & -4 & 5 \end{bmatrix}. \tag{5.47}$$

This type of prior can be used when the signal vanishes at the boundaries. When the signal is extended periodically outside its domain, i.e. $y(-1) = y(N)$ and $y(N+1) = y(1)$, the matrix $\mathbf{L}$ is:

$$\mathbf{L} = \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & -1 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -1 & \cdots & 0 & 0 & -1 & 2 \end{bmatrix} \tag{5.48}$$

and the matrix $\mathbf{L}^T\mathbf{L}$ is:

$$\mathbf{L}^T\mathbf{L} = \begin{bmatrix} 6 & -4 & 1 & 0 & 0 & \cdots & 1 & -4 \\ -4 & 6 & -4 & 1 & 0 & \cdots & 0 & 1 \\ 1 & -4 & 6 & -4 & 1 & \cdots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \cdots & 0 & 0 & 1 & -4 & 6 & -4 \\ -4 & 1 & \cdots & 0 & 0 & 1 & -4 & 6 \end{bmatrix} \tag{5.49}$$

which is a circulant one. In this case we can assume that the prior of the signal is stationary, due to the asymptotic equivalence between the circulant and Toeplitz matrices [129]. In the case we assume the stationary prior for the signal $\mathbf{s}$, we can write the equations for both algorithms, VarWhite and VarColored, in the Fourier domain to reduce the computational complexity. Another factor that affects the stationarity of the estimator of $\mathbf{s}$ is the statistical properties of the noise. If the noise is assumed to be non stationary then the estimator of $\mathbf{s}$ is non stationary, even if we assume stationary prior for the signal. This is due to the iterative nature of the algorithm. Assuming non stationary noise with non particular structure of the covariance N^2 parameters for the noise, when we have N data samples, need to be estimated. In our case, to reduce the number of parameters, we assume that the noise is stationary.

5.2.4 Equations in the frequency domain

Assuming stationary prior for the signal $\mathbf{s}$ and stationary noise $\mathbf{n}$ the VarWhite and VarColored algorithms can be written in the Fourier Domain. The resulting equations for the VarWhite algorithm are:

$$
\begin{aligned}
S(f) &= \frac{\hat{\lambda} Y(f)}{\hat{\lambda} + \hat{\alpha}|L(f)|^2}, f = 1, \cdots, N && (5.50)\\
P_s(f) &= \frac{1}{\hat{\lambda} + \hat{\alpha}|L(f)|^2}, f = 1, \cdots, N && (5.51)\\
\frac{1}{b'_\alpha} &= \frac{1}{2}\Big\{\sum_{f=1}^{N}\Big(|L(f)|^2 P_s(f) + \frac{1}{N}|L(f)|^2|S(f)|^2\Big)\Big\} + \frac{1}{b_\alpha} && (5.52)\\
c'_\alpha &= \frac{N}{2} + c_\alpha && (5.53)\\
\hat{\alpha} &= b'_\alpha c'_\alpha && (5.54)\\
\frac{1}{b'_\alpha} &= \frac{1}{2}\Big\{\frac{1}{N}\sum_{f=1}^{N}|Y(f)|^2 - \frac{2}{N}\sum_{f=1}^{N} Y^*(f)S(f) \\
&\quad + \sum_{f=1}^{N}(P_s(f) + \frac{1}{N}|S(f)|^2)\Big\} + \frac{1}{b_\lambda} && (5.55)\\
c'_\lambda &= \frac{N}{2} + c_\lambda && (5.56)\\
\hat{\lambda} &= b'_\lambda c'_\lambda, && (5.57)
\end{aligned}
$$

where $Y(f)$ and $S(f)$ are the DFT (Discrete Fourier Transform) coefficients of the vectors $\mathbf{y}$ and $\hat{\mathbf{s}}$, and $P_s(f)$ are the eigenvalues of the covariance matrix $\mathbf{C_s}$. The algorithm in the Fourier Domain consists of Eqs. (5.50)-(5.57). This algorithm is called VarWhiteFFT.

The equations for the VarColored algorithm are:

$$\begin{aligned}
S(f) &= \frac{R(f)Y(f)}{R(f)+\hat{\alpha}|L(f)|^2}, f=1,\cdots,N && (5.58)\\
P_s(f) &= \frac{1}{R(f)+\hat{\alpha}|L(f)|^2}, f=1,\cdots,N && (5.59)\\
\frac{1}{b'_\alpha} &= \frac{1}{2}\Big\{\sum_{f=1}^{N}\Big(|L(f)|^2P_s(f)+\frac{1}{N}|L(f)|^2|S(f)|^2\Big)\Big\}+\frac{1}{b_\alpha} && (5.60)\\
c'_\alpha &= \frac{N}{2}+c_\alpha && (5.61)\\
\hat{\alpha} &= b'_\alpha c'_\alpha && (5.62)\\
B(f) &= B_p(f)+|Y(f)-S(f)|^2/N+P_s(f), f=1,\cdots,N && (5.63)\\
r' &= r_p+1 && (5.64)\\
R(f) &= \frac{1}{B(f)}, f=1,\cdots,N && (5.65)
\end{aligned}$$

where $R(f)$, $B(f)$ and $B_p(f)$ are the eigenvalues of the matrices $\mathbf{R}$, $\mathbf{B}$ and $\mathbf{B}_p$, respectively. The algorithm in the Fourier domain consists of Eqs. (5.58)-(5.65), and it is called VarColoredFFT. In Eqs. (5.36) and (5.63) the second term is an approximation for the cross correlation matrix between the vectors $\mathbf{y}$ and $\hat{\mathbf{s}}$. However, since the quantity $(\mathbf{y}-\mathbf{s})^T(\mathbf{y}-\mathbf{s})$ is a rank one approximation, this makes it a very unstable term. On the other hand the term $\frac{|Y(f)-S(f)|^2}{N}$ is a good approximation for the cross correlation sequence from which we construct the cross correlation matrix. Thus, in the time domain the quantity $(\mathbf{y}-\mathbf{s})^T(\mathbf{y}-\mathbf{s})$ is replaced by a Toeplitz matrix, which is constructed using the cross correlation sequence. This sequence can be obtained by the inverse Fourier transform of $\frac{|Y(f)-S(f)|^2}{N}$.

5.3 Experimental results

5.3.1 Experiments using simulated signals

The electrophysiologic signal is constructed as a superposition of two Guassian components. Random fluctuation is introduced on the peaks position to simulate the latency variability. The matrix $\mathbf{L}$ describes the smoothness of the signal and it is an approximation of the d^{th} derivative. We test the proposed algorithms using as values of d = 2,4 and 6 (low d corresponds to smooth signal and high d in a "spiky" signal). The value of d determines the extent of the smoothness. Unlike the other quantities of the proposed model, d is difficult to be addressed in a theoretical basis. In the literature the choice of d is left to the user [117].

We compare the VarWhiteFFT algorithm with the generalized cross validation (GCV) criterion [124] and the wavelet denoising approach. From all the wavelet transforms, the Discrete Wavelet Transform (DWT) is the most widely used. However, the DWT presents a serious drawback in the estimation of ERP [114, 130], since it is shift invariant. To overcome this problem the Stationary Wavelet Transform (SWT) can be utilized. We use the biorthogonal mother wavelet (bior4.4) since we are interested in the morphology and the latency of the peaks [114]. The EEG signal is decomposed into five levels and soft thresholding is used. The thresholding rule is the 'sqtwolog' according to the wavelet toolbox of Matlab and level dependent estimation of level noise is applied. In the GCV approach the signal of interest is estimated as:

$$\hat{\mathbf{s}}_{GCV} = \Big(\frac{\alpha}{\lambda}\mathbf{L}^T\mathbf{L} + \mathbf{I}\Big)^{-1}\mathbf{y} = S(\phi)\mathbf{y}, \tag{5.66}$$

where $\phi = \frac{\alpha}{\lambda}$. To estimate the signal the parameter ϕ is needed. This is accomplished by minimizing the GCV criterion with respect to the parameter ϕ . The GCV criterion is given by the equation:

$$GCV(\phi) = \frac{\frac{1}{N}||\mathbf{y} - \mathbf{s}||^2}{(\frac{1}{N}trace(\mathbf{I} - S(\phi)))^2}. \tag{5.67}$$

To quantify the performance we calculate the SNR enhancement as:

$$SNRout = 10\log\frac{||\mathbf{s}||^2}{||\mathbf{s} - \hat{\mathbf{s}}||^2}, \tag{5.68}$$

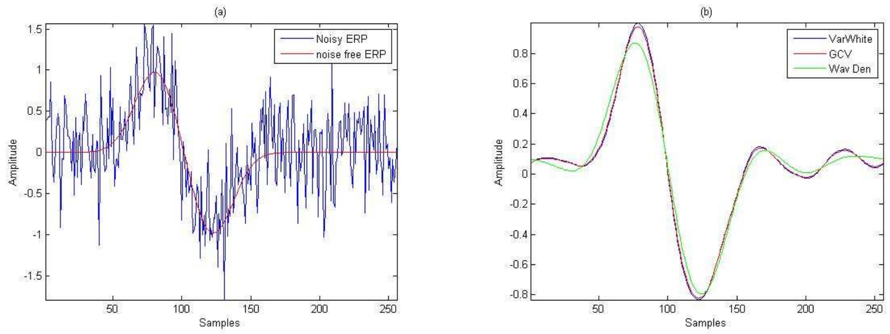
where $\mathbf{s}$ is the trial and $\hat{\mathbf{s}}$ is the corresponding estimate. To simulate M trials we generate M ERP waveforms and M realizations of noise for each SNR level. The simulated noisy trials are obtained by adding the noise to the ERP trials.

We apply the VarWhiteFFT algorithm in noisy ERP where the noise is white gaussian. The VarWhiteFFT algorithm is initialized using noninformative priors, $b_\alpha = b_\lambda = 10^6$ and $c_\alpha = c_\lambda = 10^{-6}$. The SNRout for SNR = 5, 3, 1, 0 dB using the VarWhiteFFT algorithm, the wavelet denoising and the GCV approach is calculated. SNR = 0 dB corresponds to realistic situations [131]. For GCV and VarWhiteFFT we use d = 2,4,6 and the obtained results are shown in Table 1. The best results are obtained when the VarWhiteFFT is used. A simulated ERP, a noisy ERP and the estimates are shown in Fig. 1 for SNR = 0 dB. The GCV and the VarWhiteFFT estimates present larger oscillations in the range of samples 150-256 than the wavelet denoising. The GCV and the VarWhiteFFT estimate better the simulated ERP in the range of samples 50-150. This is due to the fact that GCV and VarWhiteFFT assume that the ERP is stationary, in contrast to the wavelet denoising which assumes that the ERP is a non stationary signal.

In the colored noise case the VarColoredFFT algorithm is used and it is initialized using the noninformative prior for the parameter α , i.e. $b_\alpha = 10^6$ and $c_\alpha = 10^{-6}$. For the prior of the noise we set $r_p = 0$ and $\mathbf{B}_p = 0$ giving the improper prior $p(\mathbf{R}) \propto |\mathbf{R}|^{-\frac{N+1}{2}}$. We test the VarColoredFFT for input SNR = 5, 3, 1, 0 dB using low and high pass colored Gaussian noise. To create low pass noise an AR model of order 4 with AR coefficients [1, 1.5084, -0.1584, -0.3109, -0.0510] is used, while the AR coefficients for the high pass noise are [1, -1.5084, 0.1584, 0.3109, 0.0510]. In Table 5.2 we observe that the wavelet denoising approach presents better results compared to VarColoredFFT in the case of low

Table 5.1: $SNRout$ for different $SNRin$ in white gaussian noise case

	d=2		d=4		d=6		
SNR	VAR	GCV	VAR	GCV	VAR	GCV	WAV
5	18.3572	18.2973	17.9238	18.3428	15.5130	17.8851	14.1441
3	16.3498	16.2325	15.9253	16.1731	13.5032	15.5807	13.4763
1	14.6461	14.3728	14.1853	14.0988	11.5614	13.3881	12.8611
0	13.8054	13.4427	13.1249	13.1329	10.5247	12.4918	12.3807



(a)

(b)

Figure 5.2: (a) Simulated and noisy ERP and (b) Estimation using the GCV, VarWhiteFFT and Wavelet denoising approaches

Table 5.2: $SNRout$ for different $SNRin$ in colored low pass noise

SNR	VarColoredFFT ($d = 2$)	VarColoredFFT ($d = 4$)	VarColoredFFT ($d = 6$)	WAV
5	1.7643	13.5873	9.8299	13.9089
3	1.5652	12.1169	8.2163	12.7115
1	1.8186	10.5174	6.4513	11.6976
0	1.5985	9.3108	5.4439	10.5134

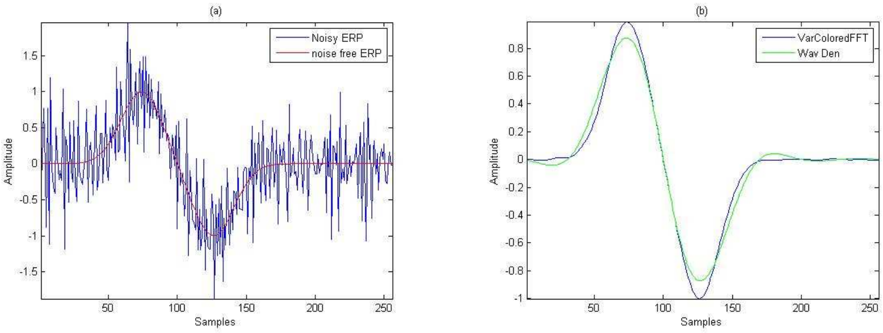
Table 5.3: $SNRout$ for different $SNRin$ in colored high pass noise

SNR	VarColoredFFT ($d = 2$)	VarColoredFFT ($d = 4$)	VarColoredFFT ($d = 6$)	WAV
5	0.6734	40.2934	36.2906	18.4265
3	0.7746	38.8252	34.8002	17.9421
1	0.8722	37.6405	33.2309	17.9271
0	1.0071	36.3522	32.4469	18.4125

pass noise. The best results for the VarColoredFFT are obtained for $d = 4$. In Table 5.3 the results in the case of high pass colored noise are shown. The VarColoredFFT approach presents much better results than the wavelet denoising. In Fig. 2 a trial of high pass noisy ERP along with the estimates obtained from the VarColoredFFT and the wavelet denoising are shown. In the next sections, results using real datasets are provided. However, in these sections we do not use the GCV approach since the GCV criterion is derived under the white noise assumption, while the noise in the real dataset is colored.

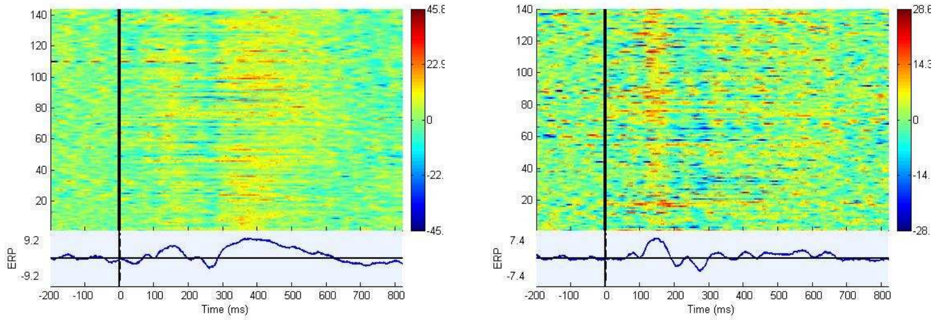
5.3.2 Application to Event Related Potential (ERP) estimation

We have used EEG data recorded during a go/nogo visual categorization task using natural photographs. This dataset has been used to study brain dynamics in [132]. Subjects



(a) (b)

Figure 5.3: (a) Simulated and noisy ERP and (b) Estimation using the VarColoredFFT and Wavelet denoising approaches



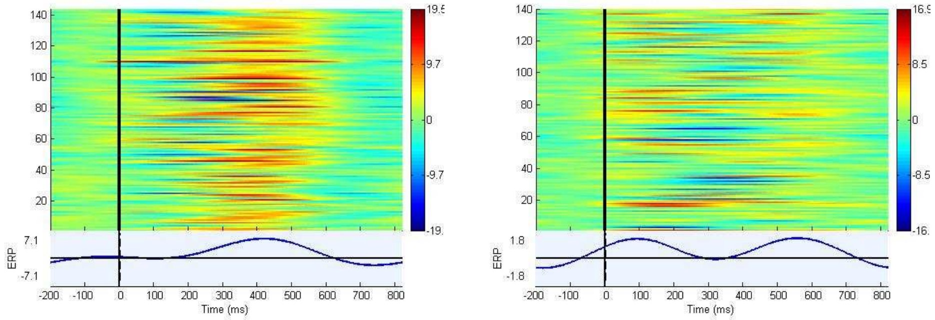
(a) (b)

Figure 5.4: Raw trials in the case of (a) target and (b) non target.

were presented with pictures which either contained or did not contain animal images. In the presence of an animal in the picture a button was pressed by the subject. The data have been processed using Independent Component Analysis (ICA) to remove muscle activity, eye blink etc. [132]. In our study the ICA - preprocessed data, derived from the Pz channel, are used. The dataset includes 14 subjects and consists of 4276 ERPs, where 2150 belong to the target case while the rest 2126 belong to the non target case.

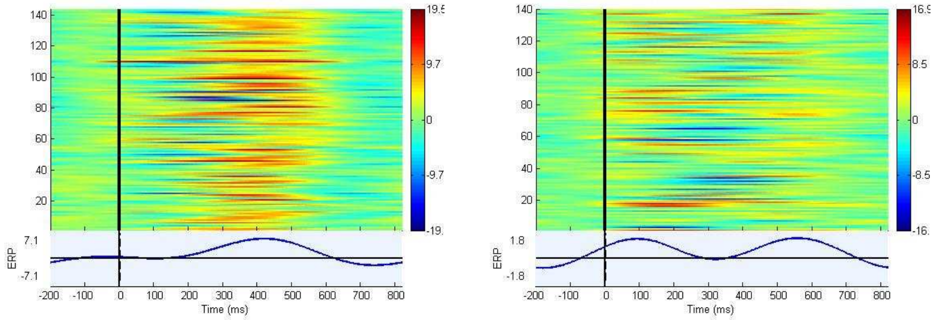
The dataset has been processed using the wavelet denoising and the VarColoredFFT approach with $d = 4$. The VarColoredFFT algorithm is initialized using $b_\alpha = 10^6$ and $c_\alpha = 10^6$ while for the covariance matrix of the noise improper prior is used. These values for the parameters of the prior of the smoothness parameter have provided the best results. In Figs. (5.4) - (5.6) ERP images of raw and processed trials from one subject for the target and the non target case are shown. The ERP image is a visualization tool which gives the ability to see the evolution of ERP in a trial-by-trial basis [119]. Besides the ERP image, we show the mean of trials from each method. It is obvious that both approaches produce a cleaner ERP signal than the raw ERP. Also, it is clear that features of ERP, such as the latency and the amplitude of latency, are more identifiable in the denoised data.

Besides the visual comparison of the two methods, we also present results for the classification of an ERP into target and non-target cases. The input of the classifier is a dataset of features. The features are extracted from raw ERPs and denoised ERPs. For the denoising of raw ERPs we have used the proposed approach and the wavelet analysis. We note here that the dataset from each subject has been processed separately. The classification procedure is based on two characteristics of the ERP: the latency and the amplitude of the P300 wave. These features are extracted and used as input to a quadratic discriminant classifier. The latency is taken as the maximum amplitude in a pre-specified window. In our case this window is 300-600ms from the onset of the stimulus. The dataset have been processed using 10 times fold cross validation. The mean classification rates were 63.4857 +/- 6.3096, 67.2993 +/- 7.3435 and 67.3457 +/- 7.5096 using the raw ERPs and the denoised ERPs extracted using the proposed approach and wavelet denoising.



(a) (b)

Figure 5.5: Estimated trials in the case of (a) target and (b) non target using the Var-ColoredFFT approach.



(a) (b)

Figure 5.6: Estimated trials in the case of (a) target and (b) non target using the Wavelet denoising approach.

Table 5.4: Classification rates of ERPs in target and non target cases for each subject.

	Raw EEG (%)	VarColored (%)	Wavelet Denoising (%)
Subject 1	68,25	72,54	74,91
Subject 2	65,30	70,64	71,84
Subject 3	72,37	76,43	77,12
Subject 4	64,82	66,33	68,21
Subject 5	57,40	62,78	53,51
Subject 6	64,33	69,21	67,42
Subject 7	54,73	52,19	58,41
Subject 8	61,20	69,07	65,32
Subject 9	76,54	73,71	75,40
Subject 10	64,82	76,26	72,13
Subject 11	58,58	55,67	57,82
Subject 12	61,02	68,14	68,27
Subject 13	65,22	69,65	73,39
Subject 14	54,22	59,57	59,09
Mean classification rate	63,48	67,30	67,35

We observe an increase on the classification rate 4% using the proposed approach and wavelet denoising. While the changes in the classification rate of the two approaches, VarColoredFTT and wavelet denoising, cannot be considered statistical significant in the whole dataset, the two approaches present different behavior as it is shown in Table 5.4.

5.3.3 Application to HRV time series analysis

The proposed approaches are employed for detrending HRV time series. HRV is used as a quantitative marker of the autonomous nervous system activity. The HRV time series contain components, which are related to slow linear or more complex trends. These effects can cause distortion on the subsequent processing such as time - frequency or spectral analysis. Our application has been inspired by the work proposed in [122]. In this work the authors used the smoothness prior to estimate the trend in the HRV time series and subtract it from them. This results to a detrended HRV time series. However, the value of the parameter α was based on visual analysis of the time series. This is a serious drawback of the proposed algorithm. To estimate the parameter α the generalized cross validation criterion can be used. However, the GCV criterion assumes that the noise is white Gaussian, something that is not always true. In this chapter we use the VarColored algorithm to estimate the trend in a HRV time series.

The HRV time series have been extracted from the MIT/BIH Arrhythmia Database

[133]. All records are utilized. The HRV time series have been processed using the VarColoredFFT algorithm and the wavelet denoising. After trend removal, an analysis is performed in the detrended HRV time series using time domain measures and spectral analysis. Before the detrending an impulse rejection filter is applied as described in [134]. The wavelet db3 and level equal to 6 are used [134]. Also, the thresholding rule is the 'sqtwolog' according to the wavelet toolbox of Matlab and level dependent estimation of level noise is applied. Finally, soft thresholding is used. The VarColoredFFT algorithm is initialized by using $b_\alpha = 10^6$ and $c_\alpha = 10^6$ for the smoothness parameter while for the covariance matrix of the noise improper prior is used.

The time domain measures used in our study are: the standard deviation of all RR intervals (SDNN) and the root mean square of differences of successive RR intervals (RMSSD) [135]. In Tables 5.5 and 5.6 the calculated measures are provided for the original HRV time series and the detrended HRV time series using the VarColoredFFT algorithm and the wavelet denoising approach. It is obvious that both approaches produce a significant change in the value of SDNN of original HRV time series. However, the VarColoredFFT approach tends to leave unchanged the RMSSD measure compared to the wavelet denoising approach. The detrending using the VarColoredFFT approach has a strong effect on the SDNN measure and only a small effect in the RMSSD measure which describes the short term RR variability. In contrast, the wavelet denoising has strong effect on both measures. To show that the two approaches, the proposed approach and the wavelet denoising, present statistically different results in terms of RMSSD and SDNN measures we perform t-tests [136]. The first t-test is related to the SDNN measure. Comparing the difference between the two approaches we have found that at 95% confidence level, the confidence interval is [4.0085, 16.3014], which does not include the zero value, and thus, the observed difference is statistically different with respect to the SDNN measure. The second t-test is related to the RMSDD measure. At 95% confidence level, the confidence interval is [4.8093, 20.6733], which indicated that the observed difference is statistical significant in term of RMSSD measure. Finally, a visual example of two records from the database (records 100 and 200) is shown in Fig. (5.7). In this figure, the HRV time series of the record and the estimated trends using the VarColoredFFT algorithm and the wavelet denoising approach are shown. It is obvious that both approaches are able to extract efficiently the trend.

5.4 Discussion

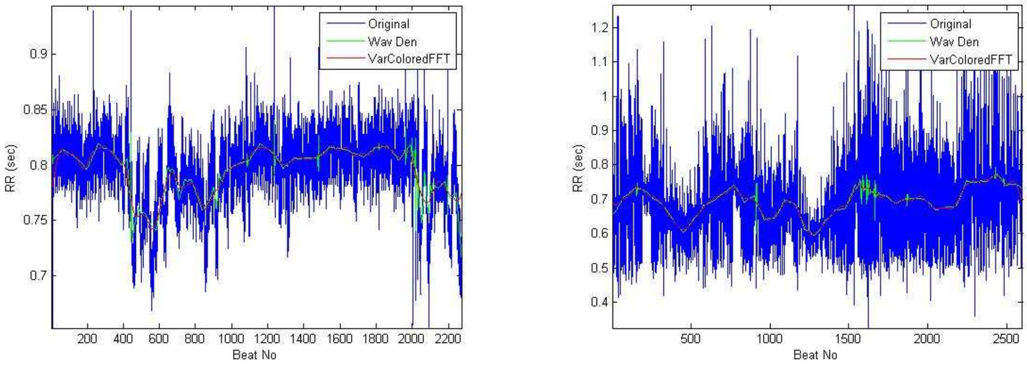
In this chapter we propose a new approach to estimate a biomedical signal contaminated by gaussian noise, either white or colored. More specifically we explore the smoothness of a biomedical signal. To incorporate the smoothness property in a Bayesian framework the smoothness prior is used. The proposed approach has been applied to estimate ERP potentials observed in noise and for detrending HRV time series. The results indicate the usefulness of the proposed approaches.

Table 5.5: RMSSD measure for all records of the MIT/BIH Arrythmia database.

Record	Raw time series (msec)	VarFFTColored (msec)	Wavelet (msec)
100	30.5676	30.5651	29.2378
101	39.0460	39.0418	38.5431
102	35.9438	35.9433	34.4788
103	32.5482	32.5422	32.1019
104	51.6978	51.6974	51.3571
105	22.5530	22.5502	21.4233
106	434.7516	434.7488	417.4801
107	30.1195	30.1188	29.4943
108	83.3694	83.3623	78.5698
109	36.8545	36.8516	29.0798
111	34.2006	34.1967	34.1119
112	17.5951	17.5915	17.3631
113	92.6397	92.6365	91.8428
114	109.2839	109.2817	76.0988
115	72.2478	72.2429	72.0577
116	18.6094	18.6078	17.4881
117	34.7143	34.7065	34.5804
118	71.5364	71.5298	40.1229
119	299.1581	299.1572	297.8358
121	19.5878	19.5732	19.5341
122	19.0312	19.0145	19.0124
123	103.4247	103.4227	103.0032
124	56.3651	56.3598	47.4654
200	260.7571	260.7564	260.4236
201	365.6437	365.6367	342.6606
202	208.2037	208.1961	138.8792
203	265.6832	265.6814	265.6476
205	26.1472	26.1372	17.7047
207	217.3939	217.3676	60.1048
208	188.9911	188.9900	188.9573
209	56.1245	56.1165	43.7349
210	159.5481	159.5471	157.4235
212	26.2510	26.2462	25.9468
213	25.3431	25.3428	24.4440
214	145.6556	145.6534	105.4569
215	43.9616	43.9607	35.6157
217	71.6366	71.6345	64.6255
219	220.9104	220.9058	220.3093
220	48.7048	48.6998	35.8222
221	272.4356	272.4347	272.4343
222	244.7885	244.7832	230.2498
223	86.8727	86.8694	67.3661
228	313.4371	313.4368	301.6098
230	28.5564	28.5368	28.2740
231	84.3748	84.2361	57.5353
232	143.9505	143.9482	62.2999
233	217.5260	217.5258	217.5076
234	17.8362	17.8252	17.3143
Mean value	114.3037	114.2961	101.5548

Table 5.6: SDNN measure for all records of the MIT/BIH Arrythmia database.

Record	Raw time series (msec)	VarFFTColored (msec)	Wavelet (msec)
100	36.8123	30.6589	29.1539
101	66.2710	43.2582	43.1299
102	29.0400	28.5746	27.9368
103	45.8076	40.4292	39.5033
104	36.2572	36.1162	35.8876
105	33.9255	21.7845	21.7154
106	260.9793	246.9869	235.2391
107	27.3528	27.0460	26.6920
108	98.5885	69.4819	67.2474
109	35.7987	30.5343	26.9596
111	37.7138	29.3101	29.6689
112	20.6286	13.0407	13.4732
113	94.3495	89.0375	89.0424
114	114.2124	80.8855	59.8151
115	85.8584	81.1964	80.4174
116	22.6307	14.0928	13.8859
117	39.6789	34.0083	31.4574
118	72.4890	51.0061	32.3252
119	176.5624	175.4097	174.8878
121	81.6542	27.3789	28.1441
122	39.8915	26.7308	27.6931
123	116.3484	113.7811	110.8904
124	77.1935	48.4703	43.4958
200	150.1990	144.5702	144.0939
201	347.5576	273.6207	240.5611
202	283.5881	147.8009	110.2574
203	198.9309	194.6069	194.4783
205	51.0154	32.6703	15.5988
207	290.0354	167.4336	50.3120
208	118.6269	114.5910	114.7254
209	77.4221	58.8256	39.0359
210	109.5262	108.4588	107.6870
212	40.2068	33.1682	33.5011
213	18.2943	16.2163	15.8691
214	102.4272	99.7495	82.6180
215	35.2039	34.6407	31.6939
217	58.1727	53.4981	49.6686
219	168.7913	155.5955	151.4202
220	59.6947	45.8786	34.4981
221	182.3230	178.9202	179.2007
222	214.3801	198.5178	181.2579
223	63.6052	53.9069	44.7834
228	177.9545	176.1317	170.4996
230	86.0102	46.9280	47.1409
231	312.8586	166.8755	126.9590
232	134.5248	132.4360	57.5120
233	124.1067	123.6997	123.6493
234	27.3740	20.7194	15.5568
Mean value	105.8932	86.2225	76.0675



(a) (b)

Figure 5.7: (a) Record 100 and (b) Record 200.

The Bayesian approach provides with the ability to use the prior knowledge of our problem through the prior distribution. Sometimes the prior distribution is not completely known but it comes in a parameterized form, one such example is the smoothness prior. These parameters of the prior distribution can be estimated within the Bayesian framework and they are called hyperparameters. To estimate the hyperparameters the Variational Bayesian Methodology is used. However, besides the VB approach, the hyperparameters can be estimated using the Empirical Bayes or the GCV criterion. The major difference between the EB and the VB is in the way that they estimate the hyperparameters. The EB approach is based on the ML estimation while the VB approach is based on the Bayesian estimation. This means that the EB provides with point estimates for the hyperparameters, while the VB approach provides with a posterior distribution for the hyperparameters. This means that the EB approach does not take into account the variability of the hyperparameters.

Another approach for the estimation of the hyperparameters is the GCV framework. We have observed in our experiments that in the case of the white gaussian noise the proposed approach and the GCV method result in similar performance. However, there exist some differences. In the Bayesian approach we know the assumptions, which does not happen in the GCV approach. Also, when we use multiple regularization parameters, we can handle them very easily in the Bayesian approach. However, there is not a widely used methodology where the GCV criterion is used to handle multiple regularization parameters. Finally, the GCV criterion has been developed under the assumption that the error follows a gaussian distribution. This means that we must expect low performance when the error follows a different distribution other than the gaussian. To conclude, the Bayesian approach provides with a structured way to estimate the regularization parameters, especially, in the case of multiple parameters.

The proposed algorithms can be applied in the wavelet domain. However, this implies that high correlation between the wavelet coefficients must exist. There is no evidence that something like that happens. The proposed algorithms assume global smoothness of the signal. Assuming local smoothness may be more appropriate for the wavelet coefficients and could be examined in future study.

In simulated experiments we compared the proposed algorithms, VarWhiteFFT and VarColoredFFT, with wavelet denoising. We observe that the VarWhiteFFT outperforms the wavelet denoising in terms of the SNR enhancement. In the case of colored noise we distinguish two cases: low pass and high pass noise. In the low pass case the wavelet denoising provides better results than the VarColoredFFT. In the case of high pass noise the VarColoredFFT presents better results compared to wavelet denoising. In the ERP estimation we showed that both methods present similar results. Besides the application of the proposed approaches in the ERP data, we use them for detrending of the HRV time series. In the HRV detrending we have shown that the use of VarColoredFFT to estimate the trend provides with the ability to remove the VLF components of the time series. Also, the method described in [17] for the detrending of the HRV time series can present computational problems since large matrices must be inverted [33]. The VarColoredFFT algorithm avoids this problem since the Fourier Domain is used.

Extensions of the proposed approaches in the Bayesian framework are straightforward. These extensions will help to explore the smoothness of a signal under different conditions. For example, in the case where we assume that the noise is not Gaussian but has impulsive nature, we can use a Gaussian mixture model with 2 components to model the impulsive nature of the noise [137, 138]. This will lead to a robust smoothing algorithm. The robustness in that case is given in terms of outliers rejection in the estimation procedure. Also, one obvious extension of the proposed approach is to use multiple parameters . This results to a non stationary prior [139] for the signal and we can explore the local smoothness of it. Finally, a third extension can be a combination of the above: the study of local smoothness of the signal in impulsive noise environment.

5.5 Conclusions

In this chapter we present a model for the estimation of a biomedical signal when this signal is contaminated by noise. More specifically we address the estimation of the smoothness of a biomedical signal. The smoothness property contains the highly correlated components of the signal. This property can be incorporated into a Bayesian framework by using the smoothness prior. This prior leads to the use of hierarchical modeling of the problem under discussion. To deal with the estimation of hyperparameters the VB methodology has been used. The VB methodology provides with closed form solutions and a convergence criterion to stop the process. The proposed approaches have been applied to the estimation of ERP and to the detrending of the HRV time series.

Chapter 6

Bayesian Methods for fMRI Time Series Analysis using a Non-Stationary Noise Model

6.1 Introduction

In this chapter, a Bayesian framework is presented for the analysis of fMRI data. The Bayesian framework is not new in fMRI data analysis. Many works have been published in this area. These works addressed several issues in the fMRI data analysis. In [30] the authors use the Bayesian framework to estimate the parameters of the GLM. However, in their analysis they use noninformative prior over the parameters of the GLM. This type of prior is used since there is no prior knowledge about the parameters. In [31] the authors are concentrated mostly to the estimation of the noise, which is modeled using an AR (autoregressive) model, rather than to the estimation of the parameters of the GLM. In [32] a Bayesian approach is presented which determines the design matrix in a flexible (automatic) way. To do that they assume sparsity over the parameters of GLM. The sparsity has been modeled by an hierarchical prior which is called Automatic Relevance Determination (ARD) [33]. However, in the estimation of the hyperparameters they use an ML (Maximum Likelihood) principle. This approach does not take into account the variability of the hyperparameters. To address it a full Bayesian approach must be used [34]. All the above works assume that the noise is temporal stationary and it is modeled using an AR model or a Gaussian distribution with zero mean and variance σ^2. However, the fMRI time series contains temporal non-stationarities which can be caused by subject movements, neurophysiological processes, or inaccuracies of the model [35, 36, 37], which in our study are described by the noise. The work presented in [35] is based on the weighted least squares (WLS) estimator, where the weighting matrix contains the non-stationarities of the noise. However, this matrix is calculated outside the estimation procedure. A Bayesian extension of the above work in presented in [37], but the estimation of the weighting matrix is confusing since it does not fit to the iterative nature of the proposed algorithm. Finally, Bayesian approaches using the spatial domain are presented in [38, 39], but, these approaches assume temporal stationarity.

In this chapter two algorithms are presented for the statistical analysis of the fMRI time series. The first algorithm is based on a voxel-by-voxel analysis of the data and it is based on the Generalized Linear Model (GLM). From the other, the second algorithm process all the voxels simultaneously and uses a spatio-temporal version of the GLM. Both algorithms estimate the variance of the noise across the images and the voxels and use a flexible design matrix to model the drift, as described in [32]. The use of the Bayesian approach is twofold in our study. First, to introduce any prior knowledge about the problem and second, to determine automatically the design matrix of the experiment. These two goals can be achieved through the choice of the prior distribution. The objective in a Bayesian approach is to obtain the posterior distribution and to make inference about the parameters of the GLM. However, this is not an easy task as multiple integrations are involved, which are intractable, and approximate approaches must be used. For this reason in this study the Variational Bayesian (VB) Methodology is adopted to make inference. The main advantage of the VB methodology is the closed form solutions that we obtain, as well as a criterion to assume convergence. The use of an extended design matrix allows the simultaneous estimation of the drift with the magnitude of the BOLD response and the spatial characteristics of the noise. This allows to better understand how the detection of activated regions of the brain depends on both the drift and the noise. The performance of the proposed algorithms (under the assumption of different noise models) is compared with the weighted least squares (WLS) method. Results using simulated and real data indicate the superiority of the proposed approach compared to the WLS method taking into account the complex noise structure of the fMRI time series.

In next sections the proposed algorithms for the analysis of the fMRI time series are presented. Also, results for both algorithms and the WLS method are given in the results section using simulated and real fMRI data. Finally, a discussion of the obtained results is presented in the discussion section.

6.2 Methodology

6.2.1 Voxel-by-Voxel analysis

The first algorithm performs a voxel-by-voxel analysis in the sense of treating each voxel independently on the others and is described by:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \tag{6.1}$$

where $\mathbf{y}$ is the fMRI time series (or voxel), $\mathbf{X}$ is the design matrix, $\mathbf{w}$ is the vector of regression coefficients and $\mathbf{e}$ is the vector of noise term.

The sparsity is a very helpful property since the processing is faster and simpler in a sparse representation where few coefficients reveal the information we are looking for.

Hence, sparse priors help us to determine the model order in an automatic way and reduce the complexity of the model. In our study the sparsity of the parameters is explored, hence a natural choice for the prior distribution is the ARD prior [43]. More specifically, the parameter vector $\mathbf{w}$ is treated as a random variable with Gaussian prior of zero mean and variance a_i^{-1} for each element in the vector $\mathbf{w}$:

$$p(\mathbf{w}|\mathbf{a}) = \prod_{i=1}^{p} N(0, a_i^{-1}). \tag{6.2}$$

where p is the length of the vector $\mathbf{w}$.

The noise in the fMRI data consists mainly of two components: a slow time - varying component, known as drift, and a high pass component, which in most cases is usually modeled by a white Gaussian distribution with zero mean and variance σ^2. The drift can be removed by high pass filtering or by introducing low frequency drift terms into the linear model. In our approach we adopt the second approach since it provides us with an estimation of the drift simultaneously with the effect of the BOLD response. Thus, the noise term $\mathbf{e}$ in Eq. (6.1) is related with the high pass component. From now the term "noise" is used to describe the high pass component. Traditionally, the noise is modeled as a stationary process. To overcome this restriction we assume that the noise is described by a non stationary model with time - varying variance. Also, we assume that the overall variance in a particular voxel is affected by the variance of each image in a multiplicative fashion. This means that the noise is modeled as a Gaussian distribution with zero mean and precision matrix (inverse covariance) $\beta\mathbf{V}$, where β is the overall variance of a voxel and the matrix $\mathbf{V}$ contains the scaling parameters v_n, i.e. $p(\mathbf{e}) = N(0, (\beta\mathbf{V})^{-1})$. The scaling parameters v_n, $n = 1, \cdots, N$ describe the variance of n-th image which is unknown. In this study two algorithms for the estimation of the parameters $\mathbf{w}$ and the scaling parameters v_n, are proposed. The main difference is found on the noise model. The first algorithm is based on a temporal model for the noise, while the second algorithm is based on a spatio - temporal model.

The overall precision (inverse variance) β of the noise follows a Gamma distribution:

$$p(\beta) = Gamma(\beta; b, c). \tag{6.3}$$

Also, each scaling parameter v_n follows a Gamma distribution. This means that the distribution for the diagonal matrix $\mathbf{V}$ is given by:

$$p(\mathbf{V}) = \prod_{n=1}^{N} Gamma(v_n; b_v, c_v). \tag{6.4}$$

In the above equations the Gamma distribution for a random variable x is given by:

$$Gamma(x; b, c) = \frac{1}{\Gamma(c)} \frac{x^{(c-1)}}{b^c} \exp\Big\{ -\frac{x}{b} \Big\}. \tag{6.5}$$

where b and c is the scale and the shape of the Gamma distribution, respectively. We use the Gamma distribution for the noise components for two reasons: First, this distribution is conjugate to the Gaussian distribution, which helps us in the derivation of closed form solutions, and second it places the positivity restriction on the overall variance and the scaling parameters. Each parameter a_i, which controls the prior distribution of the parameters $\mathbf{w}$, follows a Gamma distribution, so the overall prior over all a_i is a product of Gamma distributions given by:

$$p(\mathbf{a}) = \prod_{i=1}^{p} Gamma(a_i; b_a, c_a). \tag{6.6}$$

The likelihood of the data is given by:

$$\begin{aligned} p(\mathbf{y}|\mathbf{w}, \beta, \mathbf{V}) &= \frac{\mid \beta\mathbf{V} \mid^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}} \cdot \\ &\quad \exp\Big\{ -\frac{\beta}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T \mathbf{V}(\mathbf{y} - \mathbf{X}\mathbf{w})\Big\} &(6.7) \\ &= \frac{\beta^{\frac{N}{2}} \prod_{n=1}^{N} v_n^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}} \cdot \\ &\quad \exp\Big\{ -\frac{\beta}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T \mathbf{V}(\mathbf{y} - \mathbf{X}\mathbf{w})\Big\}. &(6.8) \end{aligned}$$

The prior over the parameters $\{\mathbf{w}, \mathbf{a}, \mathbf{V}, \beta\}$ is given by:

$$\begin{aligned} p(\mathbf{w}, \mathbf{a}, \mathbf{V}, \beta) &= p(\mathbf{w}|\mathbf{a})p(\mathbf{a})p(\mathbf{V})p(\beta) &(6.9) \\ &= p(\mathbf{w}|\mathbf{a}) \prod_{i=1}^{p} p(a_i) \prod_{n=1}^{N} p(v_n)p(\beta). &(6.10) \end{aligned}$$

To apply the VB methodology we need to define an approximate posterior based on one factorization over the parameters $\{\mathbf{w}, \mathbf{a}, \mathbf{V}, \beta\}$. In our study we choose the following factorization:

$$q(\mathbf{w}, \mathbf{a}, \mathbf{V}, \beta) = q(\mathbf{w}|\mathbf{a}) \prod_{i=1}^{p} q(a_i) \prod_{n=1}^{N} q(v_n)q(\beta). \tag{6.11}$$

Applying the VB methodology, and taking into account the above factorization, the following posteriors are obtained:

$$\begin{aligned} q(\mathbf{w}) &= N(\hat{\mathbf{w}}, \mathbf{C}_{\mathbf{w}}), &(6.12) \\ q(\beta) &= Gamma(\beta; b', c'), &(6.13) \\ q(\mathbf{a}) &= \prod_{i=1}^{p} Gamma(a_i; b_{a_i}^{'}, c_{a_i}^{'}), &(6.14) \\ p(\mathbf{V}) &= \prod_{n=1}^{N} Gamma(v_n; b_{v_n}^{'}, c_{v_n}^{'}), &(6.15) \end{aligned}$$

where

$$\mathbf{C}_{\mathbf{w}} = (\hat{\beta}\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X} + \hat{\mathbf{A}})^{-1}, \tag{6.16}$$

$$\hat{\mathbf{w}} = (\hat{\beta}\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X} + \hat{\mathbf{A}})^{-1}\hat{\beta}\mathbf{X}^T\hat{\mathbf{V}}\mathbf{y}, \tag{6.17}$$

$$\frac{1}{b'_{a_i}} = \frac{1}{2}(\hat{w_i}^2 + \mathbf{C}_{\mathbf{w}}(i,i)) + \frac{1}{b_a}, \tag{6.18}$$

$$c'_{a_i} = \frac{1}{2} + c_a, \tag{6.19}$$

$$\hat{a}_i = b'_{a_i}c'_{a_i}, \tag{6.20}$$

$$\begin{aligned} \frac{1}{b'_{\beta}} &= \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T\hat{\mathbf{V}}(\mathbf{y} - \mathbf{X}\mathbf{w}) \\ &\quad + tr(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X}\mathbf{C}_{\mathbf{w}}) + \frac{1}{b}, \end{aligned} \tag{6.21}$$

$$c'_{\beta} = \frac{N}{2} + c, \tag{6.22}$$

$$\hat{\beta} = b'_{\beta}c'_{\beta}, \tag{6.23}$$

$$\frac{1}{b_{v_n}} = \frac{\hat{\beta}}{2}(y_n - \mathbf{X}_n\mathbf{w})^2 + tr(\hat{\beta}\mathbf{C}_{\mathbf{w}}\mathbf{X}_n^T\mathbf{X}_n) + \frac{1}{b_v}, \tag{6.24}$$

$$c_{v_n} = \frac{1}{2} + c_v, \tag{6.25}$$

$$\hat{v_n} = b_{v_n}c_{v_n}. \tag{6.26}$$

In the above equations the matrix $\hat{\mathbf{A}}$ is a diagonal matrix with the mean of parameters a_i in its main diagonal and the matrix $\hat{\mathbf{V}}$ is also a diagonal matrix with the mean of the scaling parameters v_n in its main diagonal. The algorithm consists from the iterative application of Eqs. (6.16) - (6.26). This algorithm is called STNS (Sparse Temporal Non Stationary).

6.2.2 Simultaneously analysis of all voxels

In the above algorithm each voxel has been processed independently on the others. An extension of the above approach is to treat simultaneously all the voxels from one slice. This means that a spatio - temporal model must be used. We note here that the term "spatial" refers mainly to the noise model. Collecting all the voxels in one matrix, and using the fact that the design matrix is the same across the voxels, the fMRI dataset can be described by the following spatio - temporal linear model:

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \mathbf{E}, \tag{6.27}$$

where $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_T]$ is a NxT matrix containing all the voxels, $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_T]$ is a NxT matrix containing the noise, $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_T]$ is a pxT matrix containing the regression parameters of all voxels and $\mathbf{X}$ is the Nxp design matrix. The number N is the length of each voxel, while the number T is the number of voxels. Also, the dataset has been derived using N images.

The regression parameters are independent between the voxels. The probability distribution in that case is given by:

$$p(\mathbf{W}) = \prod_{t=1}^{T} p(\mathbf{w}_t). \tag{6.28}$$

Each regression parameter in a voxel is independent from the others *a priori*. This assumption is included in the proposed model through the prior distribution, which is called the ARD prior and is given as:

$$p(\mathbf{w}_t|\mathbf{a}_t) = \prod_{k=1}^{p} p(w_{tk}|a_{tk}) = \prod_{k=1}^{p} N(0, a_{tk}^{-1}). \tag{6.29}$$

A Gamma distribution is used for each parameter a_{tk}:

$$p(\mathbf{a}_t) = \prod_{k=1}^{p} \Gamma(a_{tk}; b_{tk}, c_{tk}). \tag{6.30}$$

where $\mathbf{a}_t = [a_{t1}, a_{t2}, \cdots, a_{tp}]$ is a vector containing the hyperparameters of the ARD prior at the t-th voxel. We assume a matrix Gaussian distribution for the noise given as:

$$p(\mathbf{E}) = N(0, \mathbf{V}^{-1}, \mathbf{B}^{-1}). \tag{6.31}$$

The matrix $\mathbf{V}$ is a NxN diagonal precision matrix and each element in the main diagonal describes the precision (inverse variance) in each image (slice of fMRI volume image). The matrix $\mathbf{B}$ is a TxT diagonal precision matrix and each diagonal element describes the precision in each voxel. The distribution of the noise for the t-th voxel is a Gaussian distribution given as:

$$p(\mathbf{e}_t) = N(0, (\beta_t \mathbf{V})^{-1}). \tag{6.32}$$

Also, in the proposed model the precision component of each image $\{v_1, v_2, \cdots, v_N\}$ and the precision component of each voxel $\{\beta_1, \beta_2, \cdots, \beta_T\}$ must be estimated, this means that we must place a prior distribution over each precision component. The prior distribution that is often used for a precision component is the Gamma distribution [45]. So, the prior over each precision component for each voxel is given as:

$$p(\beta_t) = \Gamma(\beta_t; b_{\beta_t}, c_{\beta_t}),\ t = 1, \cdots, T, \tag{6.33}$$

and for each image precision component:

$$p(v_n) = \Gamma(v_n; b_{v_n}, c_{v_n}),\ n = 1, \cdots, N. \tag{6.34}$$

The prior of all model parameters becomes:

$$\begin{aligned} p(\mathbf{W}, \{v_n\}_{n=1}^{N}, \{\beta_t\}_{t=1}^{T}, \{\mathbf{a}_t\}_{t=1}^{T}) = &\prod_{t=1}^{T} p(\mathbf{w}_t|\mathbf{a}_t) p(\mathbf{a}_t) \\ &\prod_{t=1}^{T} p(\beta_t) \prod_{n=1}^{N} p(v_n). \end{aligned} \tag{6.35}$$

Each voxel is independent from the others given the parameters $\{\mathbf{X}, \mathbf{W}, \mathbf{V}, \mathbf{B}\}$, so the likelihood of the observations $\mathbf{Y}$ can be written as:

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \mathbf{V}, \mathbf{B}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{X}, \mathbf{w}_t, \beta_t, \mathbf{V}). \tag{6.36}$$

Using the following factorization of the posterior:

$$\begin{aligned} q(\mathbf{W}, \{\beta_t\}_{t=1}^{T}, \{v_n\}_{n=1}^{N}, \{\mathbf{a}_t\}_{t=1}^{T}|\mathbf{Y}) = \prod_{t=1}^{T} q(\mathbf{w}_t|\mathbf{a}_t) q(\mathbf{a}_t) \cdot \\ \prod_{n=1}^{N} q(v_n) \cdot \prod_{t=1}^{T} q(\beta_t), \end{aligned} \tag{6.37}$$

and applying the VB methodology we obtain the posterior distributions:

$$\begin{aligned} q(\mathbf{w}_t) &= N(\hat{\mathbf{w}}_t, C_{\mathbf{w}_t}), t = 1, \cdots, T, & (6.38)\\ q(\mathbf{a}_t) &= \prod_{p=1}^{P} \Gamma(a_{tp}; b_{tp}^{'}, c_{tp}^{'}), t = 1, \cdots, T, & (6.39)\\ q(\beta_t) &= \Gamma(\beta_t; b_{\beta_t}^{'}, c_{\beta_t}^{'}), t = 1, \cdots, T, & (6.40)\\ q(v_n) &= \Gamma(v_n; b_{v_n}^{'}, c_{v_n}^{'}), n = 1, \cdots, N, & (6.41) \end{aligned}$$

where

$$\begin{aligned} \mathbf{C}_{\mathbf{w}_t} &= (\hat{\beta}_t \mathbf{X}^T \hat{\mathbf{V}} \mathbf{X} + \hat{\mathbf{A}}_t)^{-1}, & (6.42)\\ \hat{\mathbf{w}}_t &= (\hat{\beta}_t \mathbf{X}^T \hat{\mathbf{V}} \mathbf{X} + \hat{\mathbf{A}}_t)^{-1} \hat{\beta}_t \mathbf{X}^T \hat{\mathbf{V}} \mathbf{y}_t, & (6.43)\\ \frac{1}{b_{tp}^{'}} &= \frac{1}{2}(\hat{w}_{tp}^2 + \mathbf{C}_{\mathbf{w}_t}(p,p)) + \frac{1}{b_{tp}}, & (6.44)\\ c_{tp}^{'} &= \frac{1}{2} + c_{tp}, & (6.45)\\ \hat{a}_{tp} &= b_{tp} c_{tp}, & (6.46)\\ \frac{1}{b_{\beta_t}^{'}} &= \frac{1}{2}(\mathbf{y}_t - \mathbf{X}\mathbf{w}_t)^T \hat{\mathbf{V}} (\mathbf{y}_t - \mathbf{X}\mathbf{w}_t) \\ &\quad + tr(\mathbf{X}^T \hat{\mathbf{V}} \mathbf{X} \mathbf{C}_{\mathbf{w}_t}) + \frac{1}{b_{\beta_t}}, & (6.47)\\ c_{\beta_t}^{'} &= \frac{N}{2} + c_{\beta_t}, & (6.48)\\ \hat{\beta}_t &= b_{\beta_t}^{'} c_{\beta_t}^{'}, & (6.49)\\ \frac{1}{b_{v_n}^{'}} &= \frac{1}{2}(\mathbf{y}_n^T \hat{\mathbf{B}} \mathbf{y}_n - 2\mathbf{y}_n^T \hat{\mathbf{B}} \hat{\mathbf{W}} \mathbf{x}_n + \mathbf{x}_n^T G \mathbf{x}_n) + \frac{1}{b_{v_n}}, & (6.50)\\ c_{v_n}^{'} &= \frac{T}{2} + c_{v_n}, & (6.51)\\ \hat{v}_n &= b_{v_n}^{'} c_{v_n}^{'}. & (6.52) \end{aligned}$$

In the above equations the matrices $\hat{\mathbf{A}}_t$, $t = 1, \cdots, T$ are pxp diagonal matrices having the parameters $\hat{a}_{t1}, \hat{a}_{t2}, \cdots, \hat{a}_{tp}$ in the main diagonal. The matrix $\hat{\mathbf{B}}$ is a TxT diagonal matrix containing in the main diagonal the mean of the precision components for each voxel, $\hat{\beta}_t,\ t = 1, \cdots, T$, and the matrix $\hat{\mathbf{V}}$ is a NxN diagonal matrix containing in the main diagonal the mean of the precision components for each image, $\hat{v}_n,\ n = 1, \cdots, N$. The quantity G is calculated as follows:

$$G = \sum_{t=1}^{T} \beta_t (\mathbf{C}_{\mathbf{w}_t} + \hat{\mathbf{w}}_t \hat{\mathbf{w}}_t^T). \tag{6.53}$$

Also, in the above equations $\mathbf{y}_t$ describes the t-th voxel (t-th column of the data matrix $\mathbf{Y}$), while $\mathbf{y}_n$ describes the n-th image (n-th row of the data matrix $\mathbf{Y}$). The vector $\mathbf{x}_n$ is the n-th row of the design matrix $\mathbf{X}$. The algorithm consists of the iterative application of Eqs. (6.42)-(6.53). First, the Eqs. (6.42)-(6.49) and (6.53) are applied over all voxels to obtain the estimates of the regression parameters and the precision component for each voxel. Also, in this step the quantity G is calculated. Then, Eqs. (6.50)-(6.52) are applied to estimate the precision component of each image. This algorithm is called SSTNS (Sparse Spatio - Temporal Non Stationary).

6.2.3 Construction of the design matrix

The construction of the design matrix is crucial for the statistical analysis of fMRI data. The design matrix usually contains regressors related to the experiment plus the mean constant. In a block related experiment, which we study in this chapter, the design matrix has one regressor for the BOLD response plus the mean constant. This is the minimum number of regressors that the design matrix must contain to obtain an accurate analysis of the fMRI data. However, we can use an extended design matrix containing regressors related to other components of the fMRI time series than activation like drift terms [32] and movement effects [38]. In our study we adopt the idea of extended design matrix.

The drift in fMRI time series is described by polynomial [46], spline [50], wavelet [51, 52] and Gaussian basis functions [32]. In our approach we use Gaussian basis functions to model the drift. Using wavelets or splines a different configuration of the GLM is needed, which is out of the scope of this work. To remove the drift from the fMRI time series, we can estimate it and then substract it from the fMRI time series to perform the estimation of the GLM parameters. However, this approach does not give us any understanding on the effect of drift removal on the estimation of the GLM parameters. To overcome this inconsistency we can include the drift into the GLM model through the use of an extended design matrix. The extended design matrix contains regressors to model the drift of the fMRI time series. According to the above observations the extended design matrix has the form:

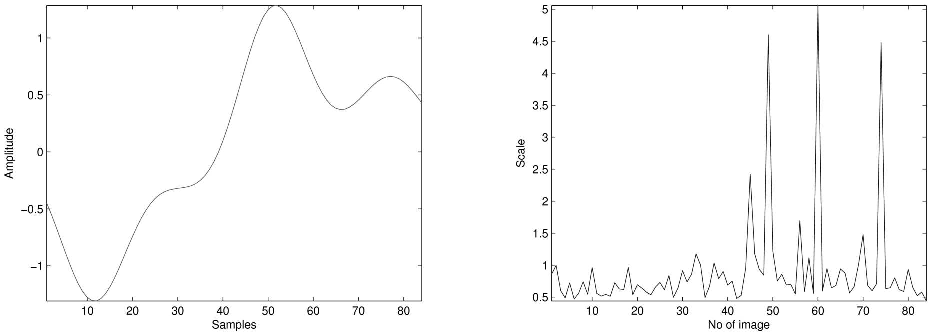
$$\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_N\ \mathbf{s}\ \mathbf{1}],$$

where $\mathbf{x}_n,\ n = 1, \cdots, N$ are the regressors obtained from the Gaussian basis functions, in the same way as described in [32], $\mathbf{s}$ is the BOLD response and $\mathbf{1}$ is a vector with 1s.

6.3 Experimental results

The proposed algorithms are compared with the WLS approach using two statistics: the conventional t - test and the PPM. The two proposed algorithms and the WLS approach are applied on simulated and real fMRI data. For the WLS approach the design matrix has two regressors: one for the BOLD response and one for the mean. For the initialization of the proposed algorithms we set the scale and the shape of each Gamma distribution to 10^6 and 10^{-6}, respectively. The free energy is used to assume convergence in the VB algorithms. The proposed algorithms are terminated when the relative change in free energy drops below 0.02.

6.3.1 Experiments with simulated fMRI data



(a) The drift signal used for the creation of the simulated data.

(b) The main diagonal (scaling parameters) of the noise covariance matrix, without considering the overall variance, used for the creation of the simulated data.

Figure 6.1: The drift signal and the main diagonal of the noise covariance matrix used for the creation of the simulated data.

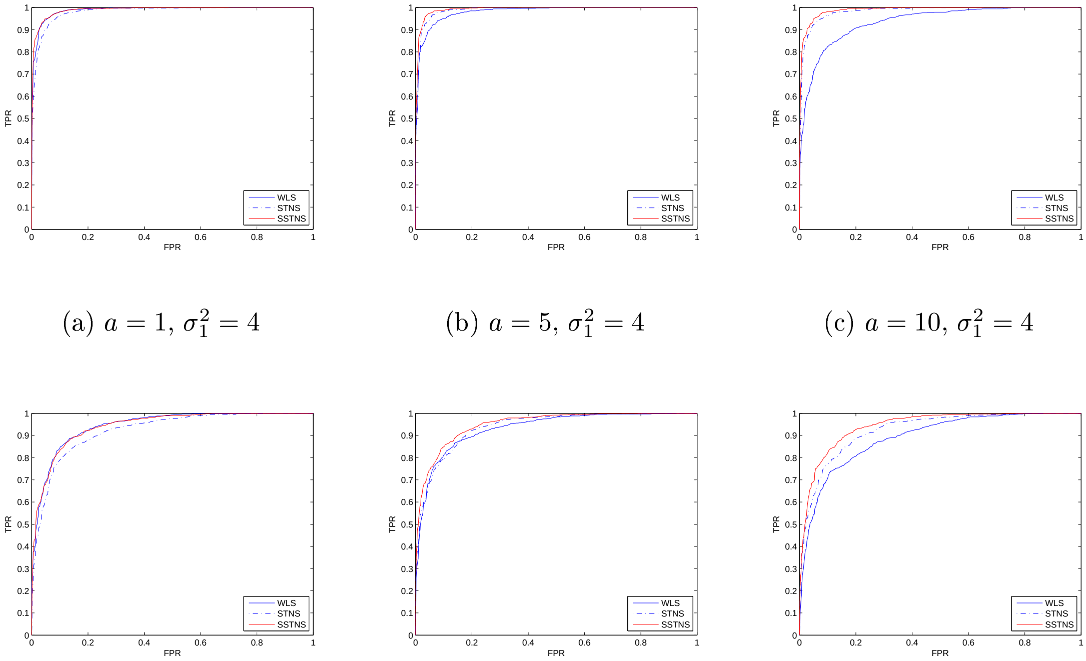
The model used to create the simulated fMRI time series is described by $\mathbf{y} = \alpha\mathbf{s}+\mathbf{K}\mathbf{b} + \mathbf{e}$. The fMRI time series have been modeled as the BOLD response plus a constant mean plus a drift term plus the noise to simulate the activated voxels, while for the non activated voxels the BOLD response is absent. The noise comes from a Gaussian distribution with zero mean and covariance $\sigma_1^2\mathbf{V}_1$, where $\mathbf{V}_1$ is a diagonal matrix (shown in Fig. 6.1(b)) and simulates the matrix of scaling parameters and σ_1^2 simulates the overall variance of the voxel. The matrix $\mathbf{K}$ is a design matrix used to create the simulated time series. The size of $\mathbf{K}$ is $N \times 2$ and it has two regressors, one for the BOLD response and one the mean constant. The vector $\mathbf{b}$ is the vector of simulated coefficients and has size 2×1. The first element of the vector $\mathbf{b}$ is responsible for the BOLD response and takes two values, zero for the non activated voxels and one for the activated. The second element of the vector $\mathbf{b}$ is responsible for the mean constant and in our simulated experiments is equal to 100. The drift $\mathbf{s}$ (see Fig. 6.1(a)) is extracted from real fMRI data. The parameter α controls the amplitude of the drift in the simulated time series. We compare the proposed algorithms with the WLS approach for different values of the parameter α and the overall variance σ_1^2. For each pair values of parameters α and σ_1^2 we create 2000 fMRI time series, 1000 of them correspond to activated voxels, while the other 1000 correspond to the non activated voxels.

The detection performance of the two proposed algorithms is compared to the one of WLS approach in terms of t-statistic. The comparison is performed using the receiver operatic characteristic (ROC) analysis. ROC analysis reflects the ability of the processing method to detect most of the real activations while minimizing the detections of false activations. In ROC analysis, two values must be computed the true positive ratio (TPR) and the false positive ratio (FPR). The ROC curve is a plot of TPR versus FPR under different threshold ratio. In Fig. 7.3 the ROC curves of the three methods for various pair values of parameters α and σ_1^2 are shown. To produce these ROC curves the t-statistic is used. We observe that when the drift is not very obvious inside the fMRI times series (small α) the WLS approach and the SSTNS algorithm present similar behavior, and both result to superior performance compared to STNS algorithm. As the drift tends to become more obvious inside the fMRI time series we observe that the performance of WLS deteriorates and the STNS algorithm results better performance than the WLS. Finally, in all cases we observe that the SSTNS algorithm results into better performance than the other two methods. In Table 6.1 the area under each ROC curve is presented, which verifies the aforementioned visual inspected results.

In Table 6.2 the mean value of the estimated overall variance for each approach for the 2000 Monte Carlo simulations is presented. It is shown how the drift affects the estimation procedure of the WLS. The WLS method does not take into account the presence of the drift and this results to a deteriorated estimation of the overall variance which at the end affects the calculation of t-statistic. When the drift is small inside the fMRI time series the best results are obtained by the WLS method, while as the drift becomes larger the SSTNS algorithm leads to the best results. Also, we can observe that the SSTNS algorithm is the most stable as it does not present strong fluctuations for different α and σ_1^2 values.

Table 6.1: Area under curve for STNS, SSTNS and WLS

		WLS	STNS	SSTNS
$\sigma_1^2 = 4$	$\alpha = 1$	0.9876	0.9812	0.9892
	$\alpha = 5$	0.9803	0.9882	0.9917
	$\alpha = 10$	0.9356	0.9832	0.9883
$\sigma_1^2 = 9$	$\alpha = 1$	0.9438	0.9220	0.9421
	$\alpha = 5$	0.9324	0.9392	0.9503
	$\alpha = 10$	0.8904	0.9239	0.9423



(a) $a = 1$, $\sigma_1^2 = 4$

(b) $a = 5$, $\sigma_1^2 = 4$

(c) $a = 10$, $\sigma_1^2 = 4$

(d) $a = 1$, $\sigma_1^2 = 9$

(e) $a = 5$, $\sigma_1^2 = 9$

(f) $a = 10$, $\sigma_1^2 = 9$

Figure 6.2: ROC curves, using various values of the parameters α and σ_1^2, for STNS, SSTNS and WLS approaches in the case of simulated data

Table 6.2: Estimated overall variances

		WLS	STNS	SSTNS
$\sigma_1^2 = 4$	$\alpha = 1$	4.2239	2.1676	5.5617
	$\alpha = 5$	18.5887	3.9322	5.5771
	$\alpha = 10$	63.7429	6.2939	5.5239
$\sigma_1^2 = 9$	$\alpha = 1$	8.7925	4.9365	9.6499
	$\alpha = 5$	23.1096	6.6947	9.6312
	$\alpha = 10$	68.2391	10.7487	9.6408

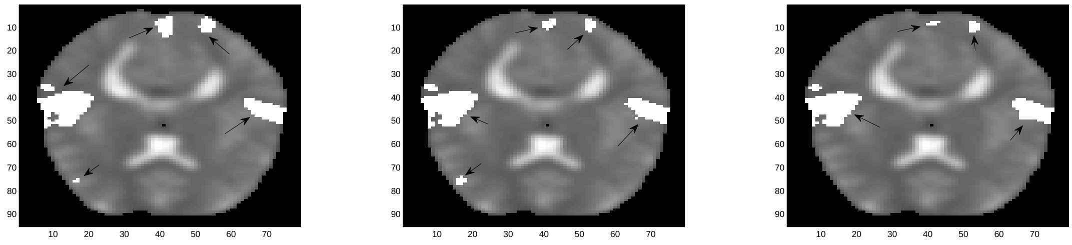
The aforementioned results show how the drift affects the estimation procedure when the GLM is used with the non stationary noise model. The STNS algorithm is based solely in temporal information while a spatio - temporal extension of it is the SSTNS algorithm. Comparing these two methods we observe that the SSTNS is superior to STNS. This is something expected since the SSTNS algorithm uses more information for the estimation of the scaling parameters. Using the STNS method we try to estimate $2N+3$ parameters from N observations. This is a very difficult problem and constraints must be imposed. In our study the constraints are introduced into our model through the prior distribution. The STNS method process one voxel at a time, while the SSTNS processes all voxels simultaneously.

6.3.2 Experiment with real fMRI data

The proposed algorithms are validated on a block design real fMRI data. This fMRI experiment was designed for auditory processing task on a healthy volunteer. It consisted of 96 acquisitions. The acquisitions were made in blocks of 6, giving 16 blocks of 42sec duration. The condition for successive blocks alternated between rest and auditory stimulation, starting with rest. Auditory stimulation was performed with bi-syllabic words presented binaurally at a rate of 60 words per minute. The functional data starts at acquisition 16. Due to T1 effects the first two blocks were discarded. The whole brain BOLD/EPI images were acquired on a modified 2T Siemens MAGNETOM Vision system. Each acquisition consisted of 64 slices (6x64x64, 3mm x 3mm x 3mm voxels). Acquisition lasted 6.05sec, with the scan to scan repetition time set to 7sec. After preprocessing, functional images consisted of 68 slices (79x95x68, 2mm x 2mm x 2mm voxles). The data have been downloaded from [55].

The PPMs enable Bayesian inference about specific effect in neuroimaging and are images of the probability that an activation exceeds some specified threshold [54]. In the PPMs two thresholds must be defined. The first and most important is the γ (see Eq. (3.57)) and it is the effect size threshold. This defines what we mean by the term "activation". The second threshold defines the probability the voxel has to exceed in order to be displayed. This threshold is called the probability threshold [54].

The PPMs of slice 30 for each method are shown in Fig. 6.3. These PPMs were derived for effect size threshold equal to 0.5 and probability threshold equal to $1-\frac{1}{N}$. The WLS is a technique based on classical inference and the use of PPM for it may be a little confusing. However, it is well known that under Gaussian errors and assuming uniformative prior we can obtain an estimator identical to WLS based on the Bayesian framework [30]. Thus, in our experiments we use this fact to obtain a posterior distribution for the WLS. This posterior distribution is a Gaussian distribution with mean $(\mathbf{X}^T\mathbf{V}_{\mathbf{wls}}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}_{\mathbf{wls}}\mathbf{y}$ and covariance $\sigma^2_{wls}\mathbf{X}^T\mathbf{V}_{\mathbf{wls}}\mathbf{X}$. We see that all approaches detect the activation on the acoustic cortex. However, we see that the WLS and the STNS methods produce activation in regions not related to the experiment. In Fig. 6.4 the scaling parameters of each image, estimated by the SSTNS and by the residuals of the LS approach, are depicted. We observe that when the stimulus starts or ends, there exists an increase of the scaling parameter in these images. This observation has been also reported in [35, 38]. This effect can arise by the presence of motion artefacts or by the true properties of the hemodynamic response that are not captured by the design matrix. This problem can be addressed by using an appropriate noise model, such as the one presented in this chapter, or by integrating the temporal derivatives of hrf (hemodynamic response function) in the design matrix. We observe that the proposed algorithms take into account this inconsistency of the GLM.



(a) WLS (b) STNS (c) SSTNS

Figure 6.3: PPMs for $\gamma = 0.5$ and $p_T = 1 - \frac{1}{N}$. The activated regions are shown in white color with an arrow. The background image describes the mean brain activation.

6.4 Discussion

We have proposed two algorithms for the detection of activated regions of the brain using a modified GLM. Inside the fMRI time series exists many components such as the BOLD response, the drift and high frequency noise. In our study, the BOLD response and the drift were modeled through the use of an extended design matrix. This matrix contains additional regressors to model the drift. The noise was modeled by a non stationary model. After the construction of the linear model inference for the regression parameters and the noise is carried out. For this reason the VB methodology is used.

The two algorithms have been applied to simulated and real fMRI data and compared

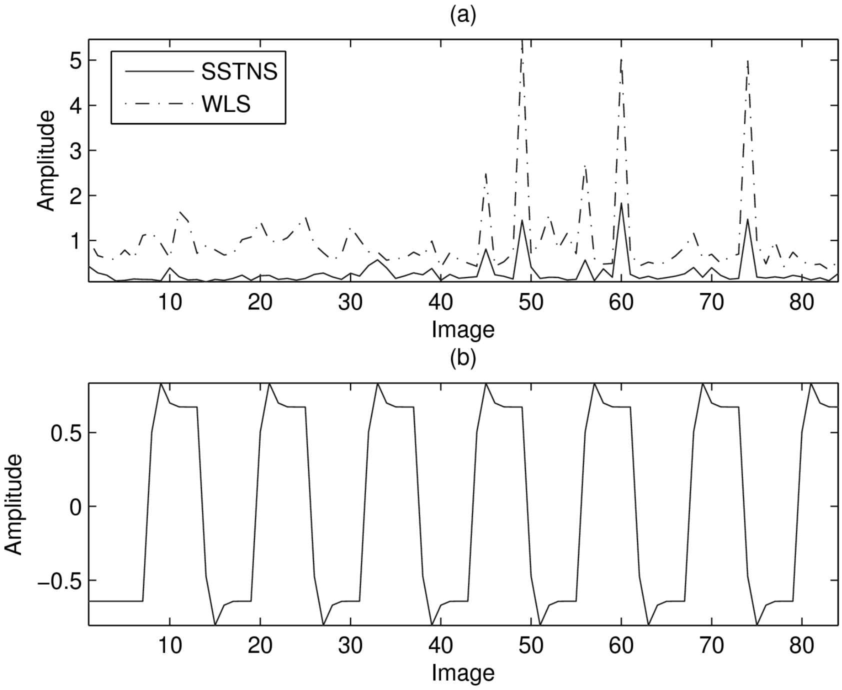


Figure 6.4: (a) Estimated scaling parameter for each image (time instant) using the SSTNS and WLS approaches, and (b) BOLD response.

to the WLS approach. The results have shown the superiority of the proposed algorithms when the drift is present in the fMRI time series. The two algorithms have simultaneously estimated the drift, the BOLD response and the noise, in contrast to the WLS approach [35] or the Bayesian extension of it [37] where a high pass filter has been applied to remove the drift.

A critical assumption of the proposed algorithms is the noise model. In our study the noise model consists from two components interacting in a multiplicative way. This means that voxels with high overall variance will present a larger increase of an artefact than the voxels with low overall variance. An alternative to the multiplicative model is the additive noise model, where the two components interact additively. In [35] the two models are studied in the context of the fMRI analysis. The main conclusion of [35] is that in most cases the noise components can be well modeled by the multiplicative noise model.

An obvious way to remove the drift from the fMRI time series is to apply a high pass filter. However, this approach suffers from the following limitations: First, the cut off frequency of the filter must be known *a priori*, Second the filter is the same for all voxels, and third the drift in that case is assumed to be a stationary signal. In our study the drift is assumed to be non stationary and the width of the Gaussian basis functions is the only easily tuned parameter. The width must be greater than the duration of one block of activation. The aim of this restriction is to avoid modeling incorrectly the BOLD response as drift term. In the future we intend to study the additive model as the noise process in the analysis of fMRI data. Also, more complicated priors over the regression parameters can be used to take into account the spatial characteristics. Finally, modeling the drift using splines [50] or the wavelet domain [51] are possible extensions of this work.

In addition, the proposed algorithms can be modified in order to be applied in real - time for the detection of activated regions. This modification can be done in two ways: The first approach is similar to the one proposed in [56, 57] for the GLM. In this case an incremental form of the equations of the proposed algorithms could be obtained. However, in that case the properties of the VB algorithm, such as the convergence, are lost. The second approach is based on an online version of the VB framework [58]. In general, the real-time fMRI data analysis is a very consuming procedure, particularly due to the preprocessing steps of MR images. To reduce the time of fMRI processing parallel computing or a computer - cluster can be employed as reported in [56, 57]. To conclude, the proposed algorithms can be modified for real-time fMRI time series but it is expected that this modification would deteriorate the detection performance. Finally, concerning the computational requirements of the proposed algorithms, the most time consuming operation is the calculation of the covariance matrix of the regression parameters, since it involves the calculation of an inverse matrix. All the other operations are related to multiplications and additions.

Another modification concerns the use of non linear forward model. More specifically, the proposed algorithms are based on the linear forward model. However, easily can be extended to the non linear forward model in the same spirit as have been done in [59]. Chappell et al [59] use the VB framework to estimate the parameters of the non - linear forward model. However, to obtain a useful algorithm at the end, they resort to linearize the likehood term through Taylor expansion. This fact results in no VB algorithm, which the main consequence is that the guarantee of convergence can no longer apply [59]. Also, it is questionable how good is the Taylor approximation for each problem under study.

6.5 Conclusions

The analysis of the fMRI data using the GLM is based on two steps. The first step is related to the estimation of the parameters of the GLM, while the second step is related to the detection of activated regions based on the previous step. In this chapter two methods for the estimation of parameters $\mathbf{w}$ of the GLM are proposed: one based on temporal formulation, and one based on a spatio - temporal formulation of the problem. Also, other components of the fMRI time series such as the drift, are incorporated into the estimation procedure through an extended design matrix. These two algorithms are applied in simulated and real fMRI data, and compared to the WLS algorithm. The simulated experiments have shown that the proposed methods outperform the WLS when the drift is present inside the fMRI time series. While, the experiments based on real fMRI data have shown that the proposed methods can be used in cases where the properties of the true hemodynamic response is not modeled correctly from the design matrix. As we observe incorrect modeling of the timing of the hemodynamic response results in an increase of variance in the particular image.

Chapter 7

A sparse and spatially constrained generative regression model for fMRI data analysis

7.1 Introduction

In this chapter we present an advanced Bayesian framework for the analysis of functional Magnetic Resonance Imaging (fMRI) data that simultaneously employs both spatial and sparse properties. The basic building block of our method is the general linear model (GML) that constitutes a well-known probabilistic approach. By treating regression coefficients as random variables, we can apply an enhanced Gibbs distribution function that captures spatial constrains and at the same time allows sparse representation of fMRI time series. The proposed scheme is described as a maximum a posteriori (MAP) approach, where the known Expectation Maximization (EM) algorithm is applied offering closed form update equations for the model parameters. We have demonstrated that our method produces improved performance and functional activation detection in both simulated data and real applications.

A significant drawback of the basic GLM approach is that spatial and temporal properties of fMRI data are not taken into account. However, the fMRI data are biologically generated by structures that involve spatial properties, since adjacent voxels tend to have similar activation level [86]. Moreover, the produced activation maps contain many small activation islands and so there is a need for spatial regularization. Another desirable property is to handle temporal correlations derived from neural, physiological and physical sources [38] and have a mechanism that can automatically address the model order. The latter is a very important issue in many model based applications including regression. If the order of the regression model is too large it may overfit the observations and does not generalize well, while if it is too small it might miss trends in the data [87].

Within the literature there are several methods that include either spatial correlations, or sparse properties into the estimation procedure, but only a few of them have investigated the simultaneous incorporation of both features. Spatial characteristics of fMRI have been proposed through the use of Markov Random Fields (MRF) priors [88, 89], mixture models [90], autoregressive (AR) spatial models [31, 39], or a Laplacian affinity matrix [38]. On the other hand, sparse models for fMRI data analysis have been developed, covering sparseness over regression coefficients of GLM [32, 37, 52], over the coefficients of spatio-temporal AR models [39], the weights on the space domain of images [92], through the use of elastic nets [93], or by converting the estimation problem into a linear programming problem [91]. Training of the above methods is performed by either Markov Chain Monte Carlo (MCMC), or Variational Bayes (VB) framework. A more compact methodology has been been presented in [180] that address both spatial and sparse capabilities in an hierarchical framework. In particular, the image of the regression coefficients is first decomposed using wavelets, and then a sparse prior is applied over the wavelet coefficients. An alternative approach has been presented in [95] where the regression coefficients are indirectly spatially smoothed using an Ising prior over their indicator variables. Finally, a recent work is described in [96], that applies a multivariate Laplacian prior over coefficients written as a scale mixture following by a spatial distribution on an auxiliary variable, that allows a spatio-temporal smoothing of data.

In this chapter we propose an advanced Bayesian framework that simultaneously employs both spatial and sparse properties in a more systematic way. The contribution of this chapter is two-fold. First, we provide directly the regression coefficients with the desired two properties by considering an enhanced prior distribution. Additionally, we manage to establish an efficient EM-based framework with closed-form update equations for the model parameters that facilitates the learning procedure.

More specifically, the general-purpose GLM is used for fMRI time series modeling. The key aspect of our method is the enhanced exploitation of the Markov Random Fields [97, 98] by using an effective Gibbs potential function. Traditionally, Gibbs distribution is used for addressing only spatial correlations. In our study we present a modification of the potential function that, apart from spatial, it is able to simultaneously impose sparseness based on the Relevance Vector Machine (RVM)[87]. A maximum a posteriori expectation maximization algorithm (MAP-EM) [205] is applied next to train this model and estimate its parameters. This is very efficient since it leads to update rules in closed form during the M-step and improves data fitting. The performance of the proposed methodology is quantitatively and qualitatively evaluated using a variety of simulated and real datasets. Comparison has been made using the typical maximum likelihood (ML) and the spatially variant GLM methods without sparseness. We also present some visualizable examples of the performance of our approach on real applications of block design and event related cases.

In section 7.2 we briefly describe the basic GLM framework and show how we can introduce a Gibbs prior so as to allow spatial correlations. The proposed simultaneous sparse spatial regression model is then presented in section 7.3 together with the MAP-based learning procedure. In section 7.4 a view of the proposed model in the spirit of EM algorithm is provided. To assess the performance of the proposed methodology we present in section 7.5 numerical experiments with artificial and real fMRI datasets. Finally, in section 7.6 we give conclusions and suggestions for future research.

7.2 A spatially variant linear regression model

7.2.1 Background

Suppose we are given a set of N time-series $Y = \{\mathbf{y}_1 \ldots, \mathbf{y}_N\}$, where each observation $\mathbf{y}_n$ is a sequence of M values over time, i.e. $\mathbf{y}_n = \{y_{nm}\}_{m=1}^{M}$. The application of the Generalized Linear Model (GLM) assumes that the fMRI time series $\mathbf{y}_n$ are described with the following manner:

$$\mathbf{y}_n = \mathbf{\Phi}\mathbf{w}_n + \mathbf{e}_n \ , \tag{7.1}$$

where $\mathbf{\Phi}$ is the design matrix of size $M \times D$ and $\mathbf{w}_n$ is the vector of the D regression coefficients which are unknown and must be estimated. The last term $\mathbf{e}_n$ is a M-dimensional vector determining the model error. In most cases temporal correlations exist over the fMRI time series that arise from neural, physiological and physical sources, and unmodeled neuronal activity [24, 25]. In order to model them we can apply an auto-regressive (AR) process [31, 39]. Note that long range correlations can be additionally included by using an appropriate extension of the design matrix [24, 32]. According to an AR process of order p, the error term $\mathbf{e}_n$ can be written as:

$$\mathbf{e}_n = \mathbf{E}_n \xi_n + \varepsilon_n \tag{7.2}$$

where $\mathbf{E}_n$ is an $M \times p$ matrix containing past error samples, ξ_n is the vector of the p AR coefficients and ε_n is an i.i.d. M-length zero mean Gaussian vector with a precision (inverse variance) λ_n, i.e. $\varepsilon_n \sim \mathcal{N}(0, \lambda_n^{-1}\mathbf{I})$. Alternatively we can consider the next formulation:

$$\Xi_n \mathbf{e}_n = \varepsilon_n \tag{7.3}$$

where Ξ_n is a $M \times M$ upper diagonal matrix containing the AR coefficients. From this scheme we obtain the distribution of error as $\mathbf{e}_n \sim \mathcal{N}(0, (\lambda_n \Xi_n^T \Xi_n)^{-1})$. Both versions of the AR model will help us to write the likelihood in a more convenient way.

The design matrix $\mathbf{\Phi}$ contains some explanatory variables (or effects) that describe various experimental factors. Its construction is crucial for the statistical analysis of fMRI data. The number of regressors (columns of the design matrix) depends on the experiment and on the problem formulation in order to address several factors of the fMRI time series such as long range correlations and movement effects [32, 38]. During the experimental study we have considered various cases related to the design matrix.

In fMRI data analysis the goal is to find the involvement of experimental factors in the generation process of time series through the estimation of coefficients $\mathbf{w}_n$. Following Eq. 7.1 and since $\mathbf{\Phi w}_n$ is deterministic we can model the sequence $\mathbf{y}_n$ with a normal distribution

$$p(\mathbf{y}_n|\mathbf{w}_n, \lambda_n, \xi_n) = \mathcal{N}(\mathbf{\Phi w}_n, (\lambda_n \Xi_n^T \Xi_n)^{-1}) \ . \tag{7.4}$$

Thus, the problem can be viewed as a maximum likelihood (ML) estimation problem for the model parameters $\Theta = \{\mathbf{w}_n, \lambda_n, \xi_n\}_{n=1}^N$. The log-likelihood function can be written in two equivalent forms using Eqs. (7.2) and (7.3), respectively:

$$\begin{align} L_{ML}(\Theta) &= \sum_{n=1}^{N} \log p(\mathbf{y}_n|\mathbf{w}_n, \lambda_n, \xi_n) = \sum_{n=1}^{N} \left\{ \frac{M}{2} \log \lambda_n - \frac{\lambda_n}{2} \|\Xi_n(\mathbf{y}_n - \mathbf{\Phi w}_n)\|^2 \right\} \ , \tag{7.5} \\ L_{ML}(\Theta) &= \sum_{n=1}^{N} \log p(\mathbf{y}_n|\mathbf{w}_n, \lambda_n, \xi_n) = \sum_{n=1}^{N} \left\{ \frac{M}{2} \log \lambda_n - \frac{\lambda_n}{2} \|\mathbf{y}_n - \mathbf{\Phi w}_n - \mathbf{E}_n \xi_n\|^2 \right\} \tag{7.6} \end{align}$$

The maximization procedure leads to the following rules that are iteratively applied[1]:

$$\begin{align} \hat{\mathbf{w}}_n &= (\mathbf{\Phi}^T \Xi_n^T \Xi_n \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \Xi_n^T \Xi_n \mathbf{y}_n \ , \tag{7.7} \\ \hat{\lambda}_n &= \frac{M}{\|\Xi_n(\mathbf{y}_n - \mathbf{\Phi} \hat{\mathbf{w}}_n)\|^2} \ , \tag{7.8} \\ \hat{\xi}_n &= (\mathbf{E}_n^T \mathbf{E}_n)^{-1} \mathbf{E}_n (\mathbf{y}_n - \mathbf{\Phi} \hat{\mathbf{w}}_n) \ . \tag{7.9} \end{align}$$

7.2.2 GLM with MRF-based spatial prior

The GLM framework does not support inference about the spatial aspects of functional anatomy. A common technique dealing with this subject is by performing a preprocessing step with a Gaussian filter to smooth the fMRI signal [25]. However, this may cause the construction of overestimated activated maps with a loss of local information. Another difficulty is the selection of the Gaussian window size that may deteriorate the performance.

The Bayesian formulation offers a natural platform for automatically incorporating spatial properties. This can be accomplished through the use of a Gibbs prior distribution over the voxel coefficients. Introducing of such prior constrains the local characteristics of the voxels and the brain response based on the notion of Markov random field (MRF) [97, 98, 100]. It must be noted that Gibbs spatial priors have been successfully applied to the task of image segmentation, see for example [101, 102, 103].

[1] we apply the Eq. 7.5 for the regression coefficients $\mathbf{w}_n$ and the Eq. 7.6 for the AR coefficients ξ_n

The Gibbs density function for the n-th voxel takes the following form:

$$p(\mathbf{w}_n|\beta_n) = Z(\beta_n)\exp\{-\frac{1}{2}V_{N_n}(\mathbf{w}_n)\} \ . \tag{7.10}$$

The function V denotes the clique potential function within the neighborhood N_n of n-th voxel. In our case we have selected the next potential

$$V_{N_n}(\mathbf{w}_n) = \beta_n \sum_{k\in N_n} \|\mathbf{w}_n - \mathbf{w}_k\|^2 \ , \tag{7.11}$$

where β_n is the regularization parameter. The neighborhood N_n is the set of voxels that are horizontally, vertically or diagonally adjacent to the voxel n, having a cardinality $|N_n|$. Finally, the first term Z of Eq. 7.10 is the normalization factor and can be written as $Z(\beta_n) \propto \beta_n^{|N_n|/2}$. In addition, a Gamma prior is imposed on the regularization parameter β_n as well as the noise precision λ_n of the form

$$\begin{aligned} p(\beta_n) &= Gamma(\beta_n|b_\beta, c_\beta) \propto \beta_n^{c_\beta - 1} e^{-b_\beta \beta_n} \ , & (7.12)\\ p(\lambda_n) &= Gamma(\lambda_n|b_\lambda, c_\lambda) \propto \lambda_n^{c_\lambda - 1} e^{-b_\lambda \lambda_n} \ . & (7.13)\end{aligned}$$

The estimation problem is now formulated as a maximum a posteriori (MAP) framework, in the sense of maximizing the posterior density of model parameters $\Theta = \{\mathbf{w}_n, \beta_n, \lambda_n, \xi_n\}_{n=1}^N$. The MAP log-likelihood function is given by:

$$\begin{aligned} L_{MAP}(\Theta) &= \sum_{n=1}^{N} \Big\{ \log p(\mathbf{y}_n|\mathbf{w}_n, \lambda_n, \xi_n) + \log\{p(\mathbf{w}_n|\beta_n)p(\beta_n)p(\lambda_n)\}\Big\} \\ &= \sum_{n=1}^{N} \Big\{ \frac{M}{2}\log\lambda_n - -\frac{\lambda_n}{2}\|\Xi_n(\mathbf{y}_n - \boldsymbol{\Phi}\mathbf{w}_n)\|^2 + \frac{|N_n|}{2}\log\beta_n - \\ &\quad \frac{\beta_n}{2}\sum_{k\in N_n}\|\mathbf{w}_n - \mathbf{w}_k\|^2 + G(\beta_n) + G(\lambda_n)\Big\} \ . \end{aligned} \tag{7.14}$$

where function $G()$ has the following form[2]

$$G(x) = c_x \log x - b_x x \ . \tag{7.15}$$

By taking the partial derivatives of function L_{MAP} with respect to model parameters the next updated rules are obtained

$$\begin{aligned} \hat{\mathbf{w}}_n &= (\lambda_n \boldsymbol{\Phi}^T \Xi_n^T \Xi_n \boldsymbol{\Phi} + B_n)^{-1}(\lambda_n \boldsymbol{\Phi}^T \Xi_n^T \Xi_n \mathbf{y} + BW_n) \ , & (7.16)\\ \hat{\beta}_n &= \frac{|N_n| + 2c_\beta}{\sum_{k\in N_n}\|\hat{\mathbf{w}}_n - \hat{\mathbf{w}}_k\|^2 + 2b_\beta} \ , & (7.17)\\ \hat{\lambda}_n &= \frac{M + 2c_\lambda}{\|\Xi_n(\mathbf{y}_n - \boldsymbol{\Phi}\hat{\mathbf{w}}_n)\|^2 + 2b_\lambda} \ , & (7.18)\end{aligned}$$

[2] We follow the methodology described in [87] where the maximization is made over a logarithmic scale using that $p(\log x) = xp(x)$.

where $B_n = \sum_{k \in N_n} (\beta_n + \beta_k)\mathbf{I}$ and $BW_n = \sum_{k \in N_n} (\beta_n + \beta_k)\mathbf{w}_k$ that correspond to the effect of neighbors of n-th voxel to the computation of its regression coefficients. Note that the update equation for the AR coefficients ξ_n is the same as in the ML case (Eq. 7.9). The above learning scheme can be incorporated in an Expectation-Maximization (EM) framework [205]. In particular, during the E-step the expectation of the hidden variables $(\mathbf{w}_n)$ are computed (Eq. 7.16) and use them next for updating the model parameters β_n, ξ_n and λ_n during the M-step (Eqs. 7.17, 7.9 and 7.18, respectively). This spatially variant regression model will be referred next as SVGLM.

7.3 Simultaneous Sparse and Spatial GLM

A desired property of the linear regression model is to offer an automatic mechanism that will zero out the coefficients which are not significant and maintain only large coefficients that are considered significant according to the model. Moreover, an important issue when using the regression model is how to define its order D. The appropriate value of D depends on the shape of data to be fitted, that is models of smaller order lead to underfitting, while large values of D may lead to overfitting. It is well known that both cases may lead to serious deterioration of the fitting performance. The problem can be tackled using the Bayesian regularization method that has been successfully employed in the Relevance Vector Machine (RVM) model [87].

In order to simultaneously capture both spatial and sparse properties, the Gibbs distribution function needs to be reformulated. This can be accomplished by using the following Gibbs density function

$$p(\mathbf{w}_n|\beta_n, z_n, \alpha_n) = Z(\beta_n, z_n, \alpha_n) \exp\Big(-\frac{1}{2}\Big\{ V_{N_n}^{(1)}(\mathbf{w}_n) + V_{N_n}^{(2)}(\mathbf{w}_n) \Big\}\Big) \ . \tag{7.19}$$

The first term in the exponential part of the above function is the sparse term used for describing local relationships of the n-th voxel coefficients. This can be expressed as

$$V_{N_n}^{(1)}(\mathbf{w}_n) = \mathbf{w}_n^T \mathbf{A}_n \mathbf{w}_n \ , \tag{7.20}$$

where $\mathbf{A}_n$ is a diagonal matrix containing the D elements of the hyperparameter vector $\alpha_n = (\alpha_{n1}, \ldots, \alpha_{nD})^T$. In addition, a Gamma prior is imposed on the hyperparameters α_{nd}

$$p(\alpha_n) = \prod_{d=1}^{D} Gamma(\alpha_{nd}|b_\alpha, c_\alpha) \propto \prod_{d=1}^{D} \alpha_{nd}^{c_\alpha - 1} e^{-b_\alpha \alpha_{nd}} \ . \tag{7.21}$$

In this way, a two-stage hierarchical prior is achieved which is actually a Student-t distribution with heavy tails [87]. Sparsity is obtained since this scheme enforces most α_{nd} to be large, thus the corresponding coefficients w_{nd} are set zero and finally eliminated.

The second term of the exponential part of the proposed Gibbs function (Eq. 7.19) captures the spatial correlation and is responsible for the clique potential of the n-th voxel:

$$V_{N_n}^{(2)}(\mathbf{w}_n) = \beta_n \sum_{k \in N_n} z_{nk} \|\mathbf{w}_n - \mathbf{w}_k\|^2 \ . \tag{7.22}$$

In comparison with the potential function of the SVGLM method (Eq. 7.10), this formulation provides a variation in the neighbors' contribution to the calculation of the clique potential value, as reflected by the parameters z_{nk}. Experiments have shown that the introduction of such weights can increase the flexibility of spatial modeling and can be proved advantageous in cases around the borders of activation regions (edges). It must be noted that in the literature there are other model-based methods that embody the same desired property around edges, see for example [104]. However, they are not possible to offer closed form update rules such as in our case. Finally, the first term Z of Eq. 7.19 acts as a normalization factor and can be expressed as:

$$Z(\beta_n, z_n, \alpha_n) \propto \beta_n^{|N_n|/2} \prod_{k \in N_n} z_{nk}^{1/2} \prod_{d=1}^{D} \alpha_{nd}^{1/2} \ . \tag{7.23}$$

We also assume that the regularization parameter β_n, the noise precision λ_n and the weights z_{nk} are variables following the Gamma distribution. Based on the above formulation, the data analysis problem can be treated as a maximum a posteriori (MAP) approach for the set of regression model variables $\Theta = \{\mathbf{w}_n, \beta_n, \xi_n, \lambda_n, z_n, \alpha_n\}_{n=1}^{N}$. The MAP log-likelihood function can be given as:

$$\begin{aligned} L_{MAP}(\Theta) &= \sum_{n=1}^{N} \Big\{ \log p(\mathbf{y}_n | \mathbf{w}_n, \xi_n, \lambda_n) + \log\{p(\mathbf{w}_n|\beta_n, z_n, \alpha_n) p(\beta_n) p(\lambda_n) p(z_n) p(\alpha_n)\} \Big\} \\ &= \sum_{n=1}^{N} \Big\{ \frac{M}{2} \log \lambda_n - \frac{\lambda_n}{2} \|\Xi_n(\mathbf{y}_n - \mathbf{\Phi}\mathbf{w}_n)\|^2 - \frac{1}{2}\mathbf{w}_n^T \mathbf{A}_n \mathbf{w}_n - \\ &\quad \frac{\beta_n}{2} \sum_{k \in N_n} z_{nk} \|\mathbf{w}_n - \mathbf{w}_k\|^2 + \frac{|N_n|}{2} \log \beta_n + \frac{1}{2} \sum_{k \in N_n} \log z_{nk} + \\ &\quad \frac{1}{2} \sum_{d=1}^{D} \log \alpha_{nd} + G(\beta_n) + G(\lambda_n) + \sum_{k \in N_n} G(z_{nk}) + \sum_{d=1}^{D} G(\alpha_{nd}) \Big\} \ . \end{aligned} \tag{7.24}$$

Setting the partial derivatives with respect to regression coefficients equal to zero the following closed form update rule is obtained

$$\hat{\mathbf{w}}_n = (\lambda_n \mathbf{\Phi}^T \Xi_n^T \Xi_n \mathbf{\Phi} + BZ_n + \mathbf{A}_n)^{-1} (\lambda_n \mathbf{\Phi}^T \Xi_n^T \Xi_n \mathbf{y}_n + BZW_n) \ , \tag{7.25}$$

where the matrices BZ_n and BZW_n are

$$BZ_n = \sum_{k \in N_n} (\beta_n z_{nk} + \beta_k z_{kn}) \mathbf{I} \ , \text{ and } BZW_n = \sum_{k \in N_n} (\beta_n z_{nk} + \beta_k z_{kn}) \mathbf{w}_k \ . \tag{7.26}$$

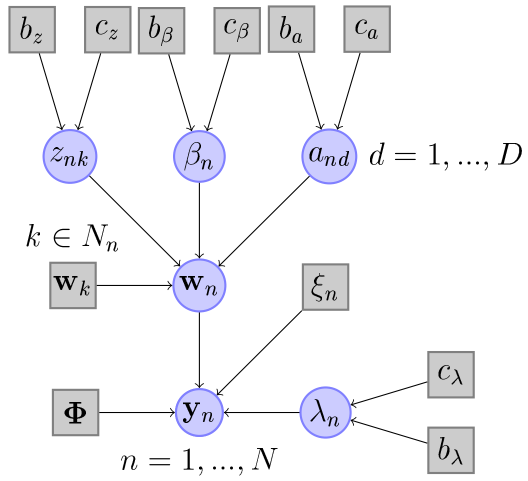


Figure 7.1: Graphical representation of the proposed model.

For the rest three model variables $\{\beta_n, z_n, \alpha_n\}$ we also produce update equations

$$\hat{\beta}_n = \frac{|N_n| + 2c_\beta}{\sum_{k\in N_n} z_{nk}\|\hat{\mathbf{w}}_n - \hat{\mathbf{w}}_k\|^2 + 2b_\beta}, \tag{7.27}$$

$$\hat{z}_{nk} = \frac{1 + 2c_z}{\hat{\beta}_n\|\hat{\mathbf{w}}_n - \hat{\mathbf{w}}_k\|^2 + 2b_z}, \tag{7.28}$$

$$\hat{\alpha}_{nd} = \frac{1 + 2c_a}{\hat{w}_{nd}^2 + 2b_a}, \tag{7.29}$$

while the AR coefficients ξ_n and the noise precision λ_n have the same form as previously defined (Eq. 7.9 and Eq. 7.18, respectively).

Again, the whole procedure can be incorporated in an EM framework by treating the regression coefficients as hidden variables. In this way, their expectation is computed in the E-step governed by Eq. 7.25, while the maximization of the complete-data MAP log-likelihood function is performed during the M-step giving update equations for model parameters (Eqs. 7.27-7.29). The above scheme is iteratively applied until the convergence of the MAP function. We call this method SSGLM. Following Eq. 7.24 it is easy to see that when $a_{nd} = 0$ the proposed method is reduced to the previously described SVGLM (setting also $z_{nk} = 1$) keeping only the spatial component. On the opposite case, when $\beta_n = 0$ or $z_{nk} = 0$ it maintains only the sparse part and becomes equivalent to the RVM-based sparse regression modeling [87]. A graphical representation of the proposed method is presented in Fig. 7.1. In the Appendix A we present an EM-based alternative description of the above model where we obtain the marginal distribution of the observations $\mathbf{y}_n$ by integrating out the regression coefficients $\mathbf{w}_n$ and treating them as hidden variables.

7.4 EM-based model estimation framework

In the previous analysis the regression coefficients $\{\mathbf{w}_n\}_{n=1}^{N}$ have been treated as model parameters. However, following the same strategy as the RVM methodology [87], we can integrated them out and obtain a reduced model with less parameters $\Theta = \{\beta_n, z_{nk}, \alpha_{nd}, \lambda_n, \xi_n\}_{n=1}^{N}$. The marginal log-likelihood for each voxel n can be obtained by the following integration

$$\log p(\mathbf{y}_n|\beta_n, z_{nk}, \alpha_{nd}, \xi_n) = \log \int p(\mathbf{y}_n|\mathbf{w}_n, \lambda_n, \xi_n) p(\mathbf{w}_n|\beta_n, z_{nk}, \alpha_{nd}) d\mathbf{w}_n \ . \tag{7.30}$$

Since both densities are known (Eqs. 7.4 and 7.19), we can easily found the marginal log-likelihood

$$\begin{aligned}\log p(\mathbf{y}_n|\beta_n, z_{nk}, \alpha_{nd}, \xi_n) \ \propto \ & \frac{|N_n|}{2}\log\beta_n - \frac{1}{2}\sum_{k\in N_n}\log z_{nk} - \frac{1}{2}\sum_{d=1}^{D}\log\alpha_{nd} - \\ & \frac{1}{2}\log|\mathbf{S}_n| + \frac{M}{2}\log\lambda_n - \frac{1}{2}\{\mathbf{m}_n\mathbf{S}_n\mathbf{m}_n - \lambda_n\mathbf{y}_n^T\Xi_n^T\Xi_n\mathbf{y}_n - \\ & \beta_n\sum_{k\in N_n} z_{nk}\mathbf{w}_k^T\mathbf{w}_k\} \ ,\end{aligned} \tag{7.31}$$

where

$$\begin{aligned}\mathbf{m}_n &= (\lambda_n\mathbf{\Phi}^T\Xi_n^T\Xi_n\mathbf{\Phi} + BZ_n + \mathbf{A}_n)^{-1}(\lambda_n\mathbf{\Phi}^T\Xi_n^T\Xi_n\mathbf{y}_n + BZW_n) \ , & (7.32)\\ \mathbf{S}_n &= (\lambda_n\mathbf{\Phi}^T\Xi_n^T\Xi_n\mathbf{\Phi} + BZ_n + \mathbf{A}_n)^{-1} \ . & (7.33)\end{aligned}$$

Therefore, the MAP log-likelihood function is written as

$$l_{map}(\Theta) = \sum_{n=1}^{N}\left\{\log p(\mathbf{y}_n|\beta_n, z_n, \alpha_n, \xi_n) + G(\beta_n) + G(\lambda_n) + \sum_{k\in N_n} G(z_{nk}) + \sum_{d=1}^{D} G(\alpha_{nd})\right\} \ . \tag{7.34}$$

The maximization of the above function can be done either directly by taking the partial derivatives and find the updated rules, or by following the EM-MAP framework, that we adopt here. According to the EM algorithm, the regression coefficients $\mathbf{w}_n$ are treated as hidden variables where their expectation is calculated at the E-step (Eq. 7.25). Equation 7.24 describes the complete data MAP log-likelihood function. During the M-step the expectation of this function is maximized, where the expectation is made with respect to the posterior distribution of regression coefficients $\mathbf{w}_n$. Notice here that this posterior can be considered as Gaussian with mean $\mathbf{m}_n$ and covariance $\mathbf{S}_n$. By setting the partial derivatives with respect to model parameters equal to zero, the next update rules are obtained

$$\hat{\beta}_n = \frac{|N_n| + 2c_\beta}{\sum_{k\in N_n} \hat{z}_{nk}\mathcal{E}_{\mathbf{w}_n|\mathbf{y}_n,\theta_n}\left\{\|\mathbf{w}_n - \mathbf{w}_k\|^2\right\} + 2c_\beta}\ , \tag{7.35}$$

$$\hat{z}_{nk} = \frac{1 + 2c_z}{\hat{\beta}_n\mathcal{E}_{\mathbf{w}_n|\mathbf{y}_n,\theta_n}\left\{\|\mathbf{w}_n - \mathbf{w}_k\|^2\right\} + 2b_z}\ , \tag{7.36}$$

$$\hat{\alpha}_{nd} = \frac{1 + 2c_a}{\mathcal{E}_{\mathbf{w}_n|\mathbf{y}_n,\theta_n}\left\{w_{nd}^2\right\} + 2b_a}\ , \tag{7.37}$$

$$\hat{\lambda}_n = \frac{M + 2c_\lambda}{\mathcal{E}_{\mathbf{w}_n|\mathbf{y}_n,\theta_n}\left\{\|\Xi_n(\mathbf{y}_n - \mathbf{\Phi}\mathbf{w}_n)\|^2\right\} + 2b_\lambda}\ , \tag{7.38}$$

$$\hat{\xi}_n = (\mathbf{E}_n^T\mathbf{E}_n)^{-1}\mathbf{E}_n\mathcal{E}_{\mathbf{w}_n|\mathbf{y}_n,\theta_n}\left\{\|\Xi_n(\mathbf{y}_n - \mathbf{\Phi}\mathbf{w}_n)\|^2\right\}\ , \tag{7.39}$$

which are iteratively applied. The above expectations can be easily calculated using that

$$\mathcal{E}_{\mathbf{w}_n|\mathbf{y}_n,\theta_n}\left\{\mathbf{w}_n\right\} = \mathbf{m}_n\ , \tag{7.40}$$

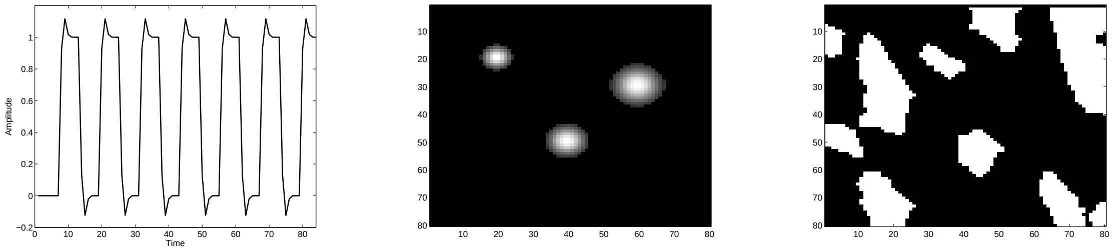
$$\mathcal{E}_{\mathbf{w}_n|\mathbf{y}_n,\theta_n}\left\{\mathbf{w}\mathbf{w}_n^T\right\} = \mathbf{S}_n + \mathbf{m}_n\mathbf{m}_n^T\ . \tag{7.41}$$

7.5 Experimental results

We have tested the proposed method (SSGLM) using various simulated and real datasets. Comparison has been made with two versions of the spatially variant GLM: the simplest one (SVGLM) as described at section 7.2.2 and those (SVGLM-2) obtained by ignoring the sparse term of the Gibbs distribution function of the SSGLM method. The difference of both versions is found on the enforcement of parameters z_{nk} in the case of SVGLM-2, in an attempt to provide a weighting scheme for the clique potential function. The aim of this study is to evaluate the usefulness of these parameters. All methods are initialized with the same strategy. First, the ML estimates of the regression coefficients $\mathbf{w}_n$ are obtained (Eq. 7.7) and then are used for initializing the rest model parameters ξ_n, λ_n, β_n, z_{kn} and a_{np}, according to Eqs. 7.9, 7.18 and 7.27-7.29, respectively. It must be noted that in the case of SSGLM method, since there is a dependency between parameters β_n and z_{kn}, we use the Eq. 7.17 (instead of the Eq. 7.27). During the experiments all Gamma parameters were set equal to 0.5, except for the sparse responsible parameters $\{b_\alpha, c_\alpha\}$ that were set as $b_\alpha = c_\alpha = 10^{-8}$ making them non-informative, as suggested in [87].

7.5.1 Experiments with simulated data

The simulated datasets used in our experiments were created according to the following generation mechanism. We used a design matrix with two pre-specified regressors. The



(a) (b) (c)

Figure 7.2: Simulated data generation features: (a) Bold signal, (b) random and (c) circular shaped image of activated areas.

first one was responsible for the BOLD signal ($\mathbf{s}$) of length $M = 84$ and has been derived by a real experiment found on the SPM package shown in Fig. 7.2(a), while the second one being a constant of ones. Then, we constructed an image with the activated areas where the pixel intensities correspond to the value of the first coefficient (w_{n1}). In our study we have used two such simulated images of size 80×80 with two different shapes of activation: circular (Fig. 7.2(b)) and random[3] (Fig. 7.2(c)). The second coefficient w_{n2} had a constant value equal to 100. The time series data ($\mathbf{y}_n$) were finally calculated by using the generative equation of GLM (Eq. 7.1) with an additive Gaussian noise of various signal-to-noise-ratio (SNR) levels. The noise was constructed according to an AR model of order $p = 3$ whose coefficients ξ_n took the values $\xi_n = (1\ -0.8\ 0.6\ -0.4)^T$, that were the same used in [31]. Finally, the SNR value was calculated as follows:

$$SNR = 10 \log \frac{\mathbf{s}^T \mathbf{s}}{M(1/\lambda_n)} \tag{7.42}$$

where $\mathbf{s}$ is the BOLD signal (Fig.7.2(a)).

Two evaluation criteria were used during the experiments.

- The Area Under Curve (AUC) of the Receiver Operating Curve (ROC) based on t-statistic calculations. ROC curves were generated by considering a voxel to be active if its effect size is greater than a predefined threshold. In our experiments the above threshold varied from the minimum to the maximum value of the t-statistic as calculated by each method. ROC analysis reflects the ability of the method to detect the real activations, while minimizing the detections of false activations.

- The normalized mean square error (NMSE), between the estimated ($\hat{w}_{n1}$) and the

[3] It has been created by sampling from an MRF model using a Gibbs sampler and has been obtained from [102]

Table 7.1: Comparative results for simulated data in various noisy environments.

circular-shaped areas						
	AUC			*NMSE*		
SNR	SSGLM	SVGLM-2	SVGLM	SSGLM	SVGLM-2	SVGLM
0	0.9989	0.9990	0.9990	0.0680	0.2431	0.2537
-2	0.9985	0.9990	0.9990	0.0952	0.2858	0.3015
-4	0.9987	0.9989	0.9989	0.1497	0.3312	0.3542
-6	0.9972	0.9985	0.9983	0.2218	0.3730	0.4054
-8	0.9934	0.9982	0.9976	0.3027	0.4081	0.4517
-10	0.9818	0.9974	0.9962	0.4114	0.4394	0.4961

random-shaped areas						
	AUC			*NMSE*		
SNR	SSGLM	SVGLM-2	SVGLM	SSGLM	SVGLM-2	SVGLM
0	0.9845	0.9813	0.9864	0.2015	0.2145	0.2211
-2	0.9841	0.9712	0.9737	0.2247	0.2421	0.2712
-4	0.9824	0.9627	0.9641	0.2547	0.2871	0.3190
-6	0.9777	0.9460	0.9445	0.3027	0.3431	0.3724
-8	0.9715	0.9257	0.9248	0.3631	0.4076	0.4329
-10	0.9594	0.9075	0.9005	0.4323	0.4814	0.4904

true (ω_{n1}) coefficients responsible for the BOLD signal which are known:

$$NMSE = \frac{\sum_{n=1}^{N}(\hat{w}_{n1} - \omega_{n1})^2}{\sum_{n=1}^{N}\omega_{n1}^2} . \tag{7.43}$$

NMSE measures the quality of the curve fitting procedure.

For every noise realization (SNR value), we performed 50 different runs of each comparative method and the mean values of AUC and NMSE measurements were calculated. Moreover, during the experiments with simulated data we have used a design matrix ($\boldsymbol{\Phi}$) with four columns ($D = 4$): one for the BOLD signal, two others for the time and the dispersion derivatives, and a last column with ones for the constant term.

We present in Table 7.1 the comparative results in terms of the above two criteria for several SNR values. As it is obvious, the proposed method improves the quality of fitting process (NMSE quantity), as well as the activation detection ability (AUC quantity). This is more apparent in random-shaped areas and in lower values of examined SNR values. An example of the obtained ROC curves by three methods is displayed in Fig. 7.3, giving the ability of the SSGLM to detect larger real activations (sensitivity) and simultaneously reduce the detection of false positive activations (specificity). However, its calculated AUC values are slightly worst than those of its spatially constrained peers during experiments with circular regions that has much smoother borders. Between the two

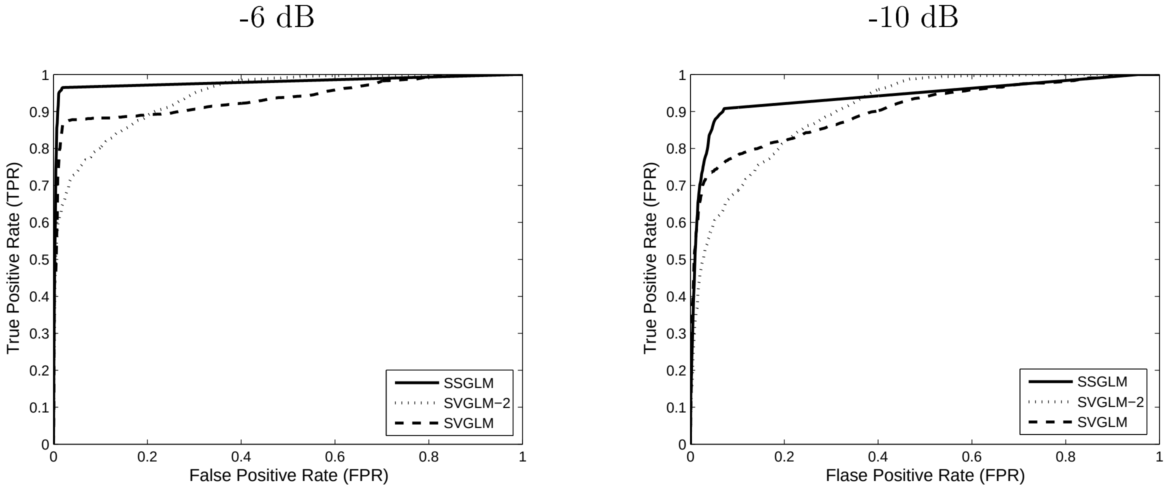
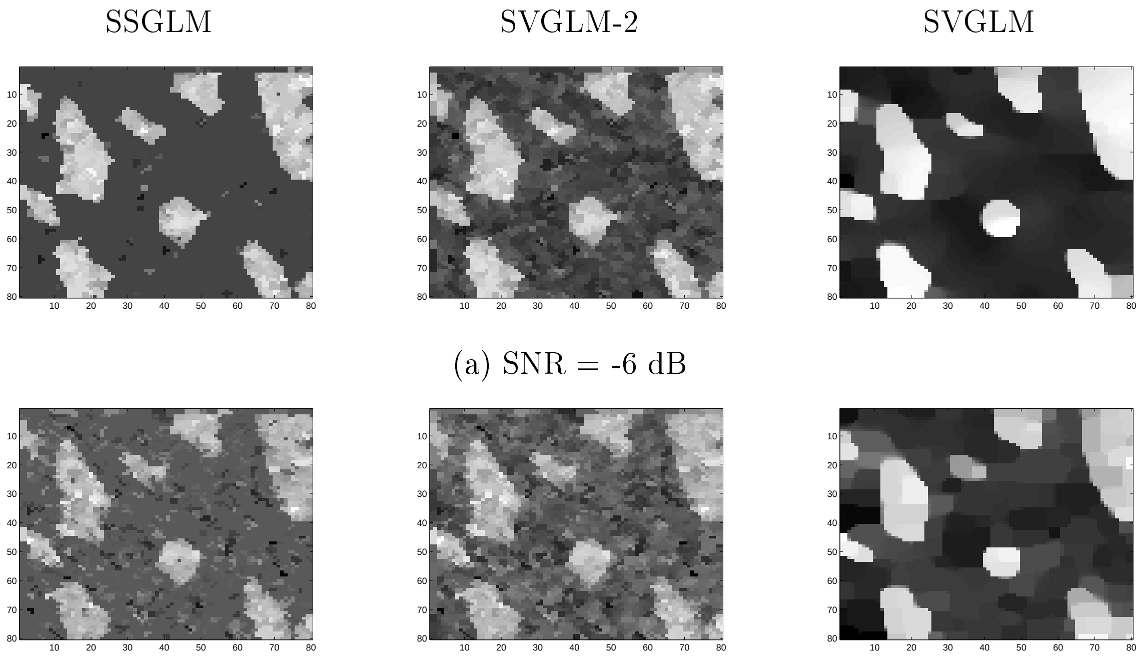


Figure 7.3: Example of ROC curves created by the estimates of SSGLM, SVGLM-2 and SVGLM methods for two different SNR values.

spatially-constrained methods, the SVGLM-2 had better performance in all cases. The introduction of the parameters z_{nk} in the potential function of the SVGLM-2 (Eq. 7.22) gives better detection capability and fitting accuracy. In addition it manages to improve the effect of over-smoothing at the boundaries of activation regions that happens with the simple SVGML method. Finally, the proposed approach SSGLM not only supports this property, but also achieves enhanced activation detection capabilities by making the distinguish between activated and non-activated areas more significant. This behavior is shown in Figure 7.4 that presents the produced BOLD contrast images of three comparative methods when studying the simulated data with random-shaped regions for two different SNR values.

7.5.2 Experiments with real fMRI data

The proposed approach was also evaluated in a variety of real applications. For any selected dataset we followed the standard preprocessing steps of the SPM package, i.e. realignment, segmentation, and spatial normalization, without performing the spatial smoothing step. Data are then scaled by means of their mean value, as described in [180], and finally were high pass filtered using a set of discrete cosine basis functions. Our method (SSGLM) was compared with the spatially variant (SVGLM), as well as with the maximum-likelihood (ML) approach. In the latter case (ML), time-series are initially spatially smoothed. During all experiments we have chosen an AR model of order $p = 3$ as was suggested in [31].



(a) SNR = -6 dB

(a) SNR = -10 dB

Figure 7.4: The estimated BOLD contrast by three comparative methods in simulated data with random shaped activated areas.

Block design fMRI data

At first we have studied a real block design fMRI dataset[4] designed for auditory processing task on a healthy volunteer. Its functional images consisted of $\mathcal{M} = 68$ slices ($79 \times 95 \times 68$, $2mm \times 2mm \times 2mm$ voxels). Experiments were made with the slice 29 of this dataset. We have used two regressors ($D = 2$) for the design matrix, one for the BOLD response and another having constant values of ones for modeling the mean brain activity. Figure (7.5) illustrates the images of the BOLD contrast produced by the three comparative approaches. All methods show maximum BOLD signals at the expected areas of the auditory cortex. However the SSGLM approach is significantly biased to these areas and suppress more efficiently the weights of the rest of the brain giving a cleaner activation pattern.

In order to perform a more comprehensive study, further experiments are made on this slice. In particular, we find it useful to visually inspect the resulting activation maps obtained by the t-test. In Fig. 7.6 the SPMs of each method are shown, calculated without (Fig. 7.6(a)), or with setting a threshold[5] (Fig. 7.6(b)) on t-values. It is interesting to observe that both maps produced by SSGLM are very similar achieving less sensitivity

[4] It was downloaded from the SPM web page http://www.fil.ion.ucl.ac.uk/spm/

[5] The significance level was set to 0.05, which gives a threshold value $t_0 = 1.66$.

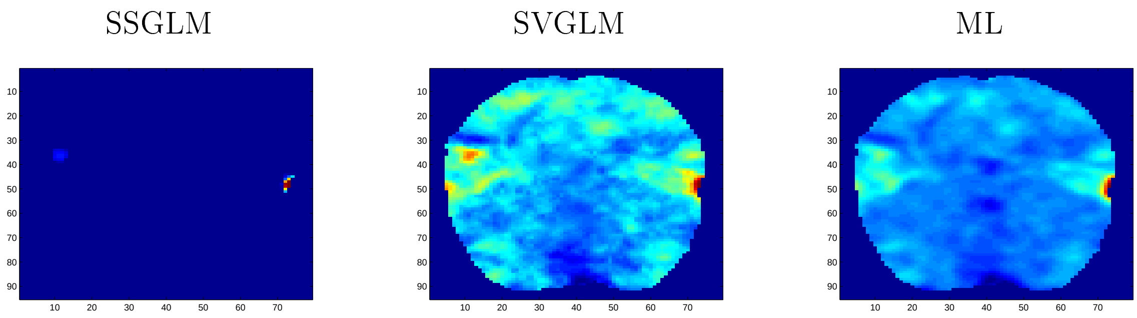
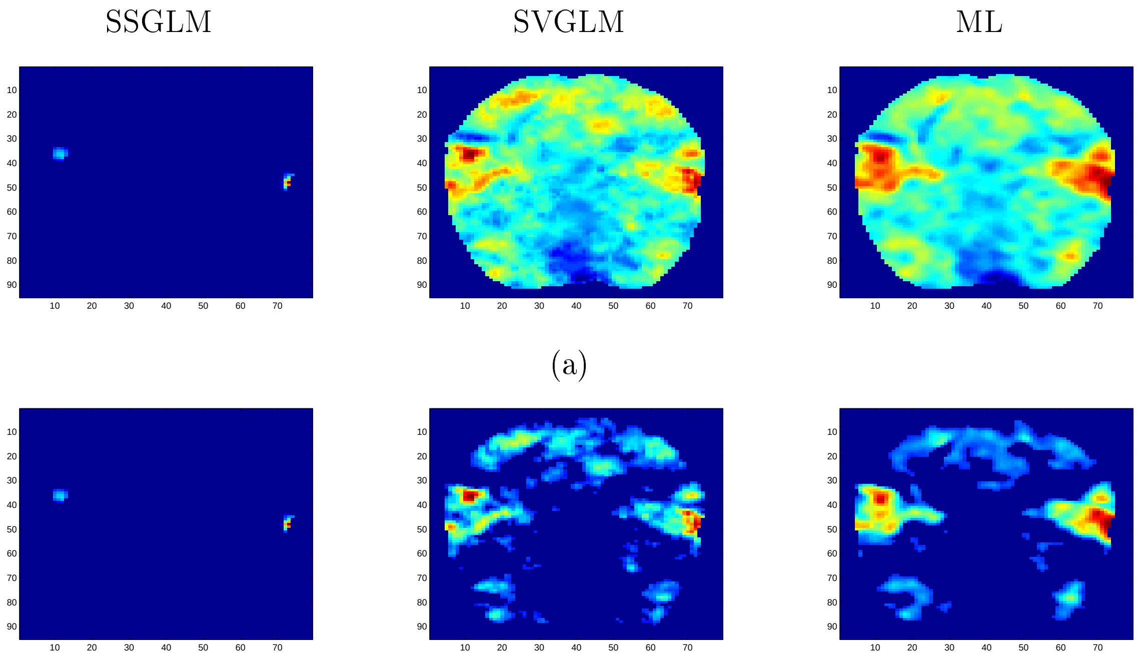


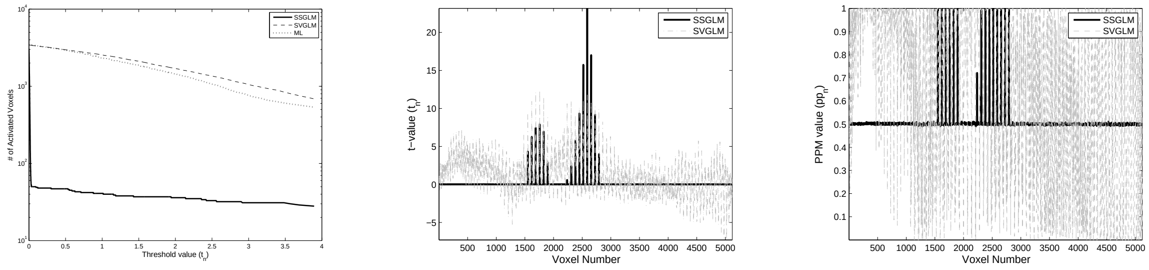
Figure 7.5: BOLD contrast images estimated by three methods in real block design experiment.



(a)

(b)

Figure 7.6: Statistical parametric maps (SPMs) of comparative methods based on t-values (a) without and (b) with setting a threshold value.



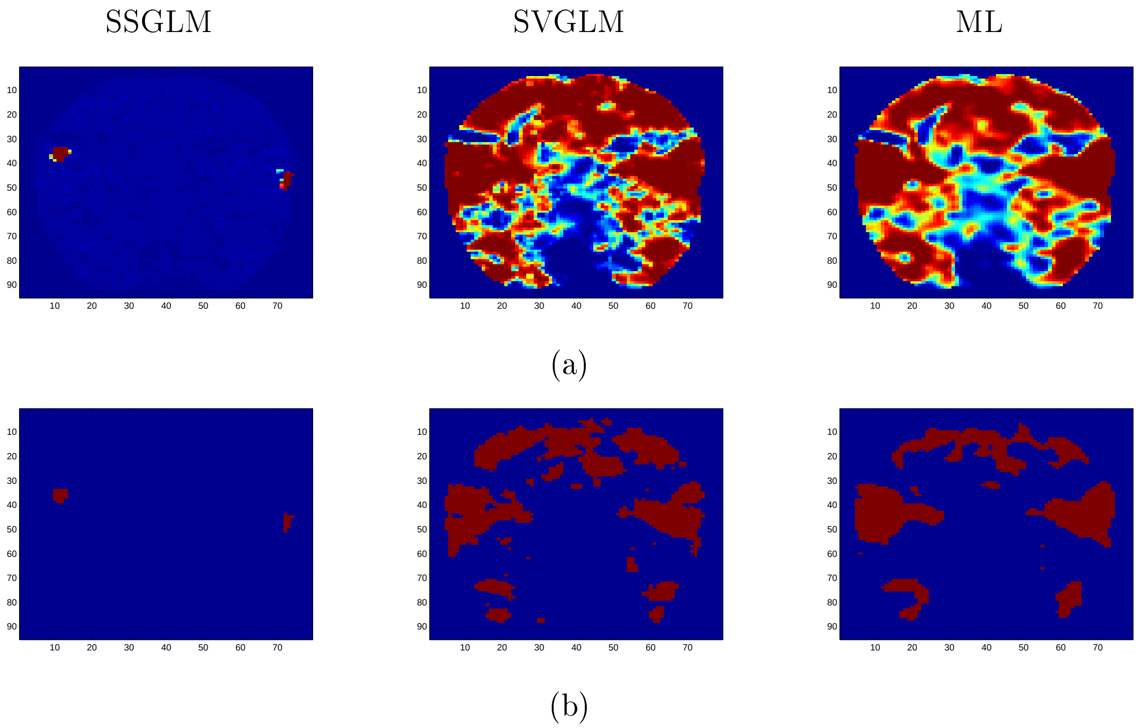
(a) (b) (c)

Figure 7.7: (a) Plots (in logarithmic scale) of the estimated number of activated voxels in terms of threshold value used for producing the SPMs. (b) Plots of the t-values of SPMs and (c) plots of the posterior values of PPMs as computed by the SSGLM (thick line) and the SVGLM method (thin line).

to the threshold value. This is more apparent in Fig. 7.7(a) where we plot the estimated size (number of voxels) of activated areas by each method in terms of the threshold value (obtained with a varying significance level from 0.0001 to 0.5). The same observation is made by plotting the calculated t-values of the SSGLM and the SVGLM methods in Fig. 7.7 (b). As it is obvious the distinction between the activated and non activated areas becomes much cleaner in the SSGLM plot. The above observation is very important from a clinical perspective, since in the standard fMRI analysis methods the activation boundary varies significantly with the smoothing and the statistical threshold used. This dependence complicates clinical decisions based on fMRI results [106]. This problem is alleviated by our methodology which does not require smoothing and produces results that are very insensitive to the threshold choice.

Similar observations are obtained by studying the posterior probability maps (PPM) of brain activity, displayed in Fig. 7.8 without (a) or with setting a threshold[6] (b). Again the SSGLM method produced much smoother and cleaner areas, while the rest methods showed almost similar results. Finally, in Fig.7.7(c) we plot the calculated PPM values of the SSGLM and the SVGLM methods. The distinguish between activated and non activated areas is more obvious in our method. Moreover, by comparing all different activation maps which can be obtained (Figs. 7.5, 7.6, 7.8) it is interesting to observe that the proposed method maintains similar behavior regarding the same estimated activation areas.

[6] We have used the threshold value $pp_0 = 1 - 1/M$ as suggested in [54]



(a)

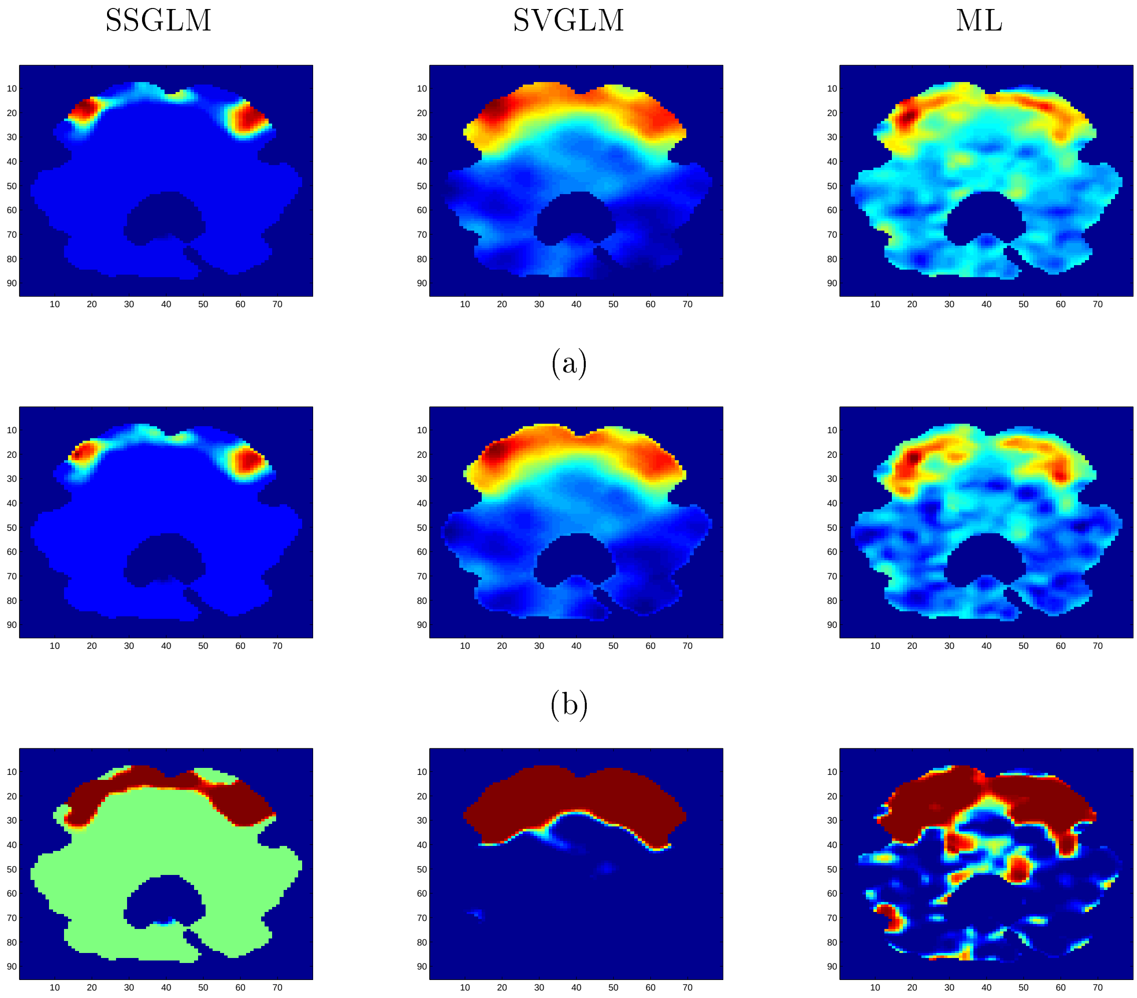
(b)

Figure 7.8: Posterior Probability Maps (PPMs) of three comparative methods (a) without and (b) with setting a threshold value.

Event related design fMRI data

Additional experiments were made considering event related cases. At first we have used a public available dataset obtained from the SPM web page designed for face recognition using grayscale images of faces, where we have selected the slice 18 for study. The contrast vector was set as $\mathbf{c} = [1, 1, 1, 1, 0]$ that describes the response to the presentation of a face image. We have used a design matrix that consists of five ($D = 5$) regressors related to 4 types of events. In particular, the first four regressors indicate the presence of a face and have been convolved with a "canonical" HRF, while the last one is the constant term. Figure 7.9 presents the produced maps of (a) the BOLD signal, (b) the SPMs (b) and (c) the PPMs. As it is obvious, all methods show large responses in the occipital lobe. However, the SSGLM method produces more localized and less dispersed activation areas.

In the second event related experiment we analyzed fMRI data consisted of images acquired from a motor event related paradigm available at the affiliated University Hospital of Ioannina. During this experiment patients with RLS (restless legs syndrome) performed random and spontaneous limb movements evoked by sensory leg uneasiness. These movements were used to create the indicator vector in our modeling that was convolved next with the hemodynamic response function (HRF) in order to provide the BOLD signal. The design matrix had four columns ($D = 4$): the BOLD signal, its time and dispersion

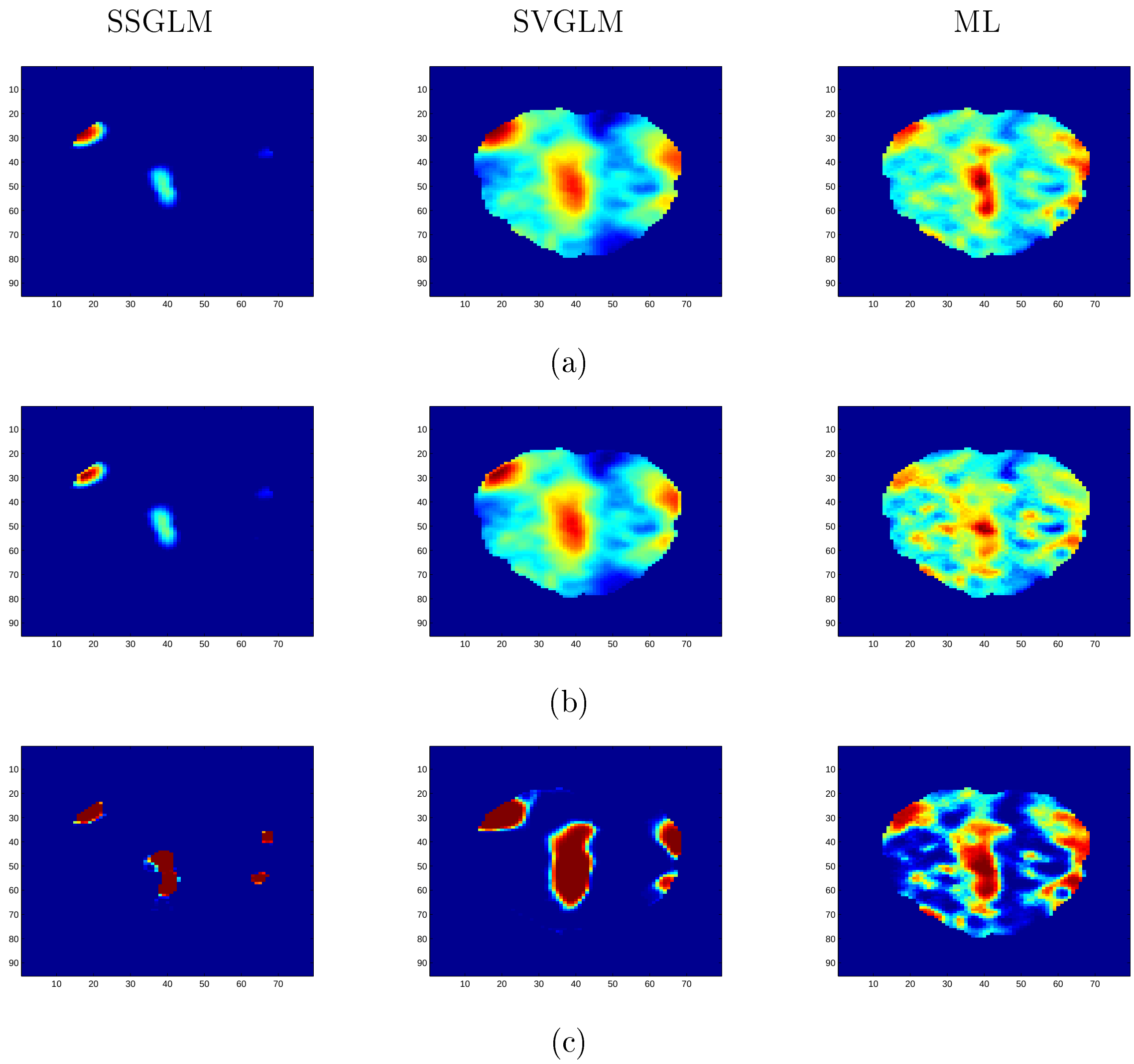


(a)

(b)

(c)

Figure 7.9: Images of (a) the BOLD contrasts, (b) the SPMs and (c) the PPMs estimated by three methods in a real event related experiment.



(a)

(b)

(c)

Figure 7.10: Created images of the (a) BOLD contrasts, (b) SPMs and (c) PPMs produced by three methods in the real motor event related experiment.

derivatives, and a constant.

In our study we examined the main effect (leg movement vs rest) which means that the contrast vector has the form $\mathbf{c} = [1, 0, 0, 0]$. Maps of the estimated BOLD contrasts, the SPMs and the PPMs, are shown in Fig. 7.10 for the slice 54 of the dataset. All the methods revealed similar brain activated regions related with motor function such as: a) the supplementary motor area b) the primary motor areas (precentral gyrus) and c) the superior parietal lobe. In comparison with the other approaches, the SSGLM method provides contrast maps where even a simple visual inspection reveals localized maxima in good agreement with the current knowledge of the locations and extent of motor circuitry. The other approaches need further thresholding in order to detect activated areas.

7.6 Conclusions

In fMRI data analysis, the spatial extension of the hemodynamic response in a neighborhood of voxels introduces a significant weakness for the detection process of the activated areas. Moreover, the presence of temporal correlations deteriorates the performance. In this work we present an advanced method to tackle these two problems by efficiently incorporating both spatial correlations and sparse properties. This is done by using a powerful prior over the regression coefficients based on Markov Random Fields (MRFs) modeling and Relevance Vector Machines (RVMs). Training of the proposed model is achieved through a maximum a posteriori (MAP) framework that allows the EM algorithm to be effectively used for estimating the model parameters providing update rules in closed form. Experiments on artificial and real datasets have demonstrated the ability of the method to improve the detection performance and robustness, especially in noisy environments, and to enhance the estimation accuracy. Our method showed a reduced sensitivity to the threshold value of the produced statistical map without needing to make multiple comparisons. Our future research study is focused to three directions: a) to examine the appropriateness of other types of sparse priors [107], b) to try alternative potential functions of the Gibbs distribution and c) to assume a Student-t distribution instead of Gaussian for modeling the excitation noise aiming to achieve more robust statistical inference and handle more efficiently outlying observations [208].

Chapter 8

Clustering fMRI time-series by using a mixture of regression models with spatial and sparse properties

8.1 Introduction

Classification and clustering methods can be used for the determination of fMRI time series into activated or non activated. However, the classification methods meet an obstacle: they require a training set. This is not an easy task since the fMRI response depends on many experimental factors such the image acquisition parameters, paradigm design, subject and region of the brain activated. On the other hand setting the problem as a clustering problem seems to be more natural. In the literature they are plenty of works in this direction [182, 184, 185, 186, 187, 188, 189, 190, 191]. Clustering is the process to divide a set of samples into groups (called clusters) so that samples from the same cluster to be similar each other while samples belong to different clusters to be dissimilar. It is a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics [192, 45].

Two major classes of clustering methods are the distance - based methods and the model - based methods [192]. The first category assumes a weak structure of the data, while the second category assumes a compact and informative structure. Distance-based methods, such as the well-known k-means algorithm, usually require the number of clusters to be known a priori. However, model-based methods can incorporate prior knowledge more naturally into the clustering approach which can help us to estimate the number of clusters.

Probabilistic mixture modeling is a well established model-based approach for clustering that offers many advantages. One such advantage is that it provides a natural platform to evaluate the quality of the clustering solution [45]. Clustering time-series is a special case of clustering in which the available data have one or both of the following two features: first they are of very large dimension and second they are not of equal length and thus conventional clustering methods cannot straightforwardly be applied. In such cases it is natural to initially fit the available data with a parametric model and then to cluster based on that model. Through the literature there are different types of models that have been used for time series clustering [193]. Among them, Hidden Markov Models [194], polynomial and spline regression models [195, 196], mixtures of ARMA models [197, 198] and mixtures of Gaussian processes [199] are commonly used models. The main drawback of these methods is that they do not automatically address the problem of model order selection, which is very important in regression. If the order of the regressor model is too large, it overfits the observations and does not generalize well. On the other hand if it is too small, it might miss trends in the data.

fMRI belongs to the spatiotemporal class of data that capture both spatial and temporal properties of data [200]. Clustering such kind of data must consider how to group voxels into spatial regions where voxels exhibit similar temporal behavior. But this is not a problem rather a challenge. In such cases it is important to measure both the temporal characteristics of the grouped voxels and simultaneously to accurately classify voxels in groups of similar temporal behavior. Thus, for this type of data, determining class membership, apart from the distance between the coefficients of the model, it is also beneficial to use spatial constraints. Such constraints must capture our prior knowledge that adjacent voxels most likely belong to the same class and have the same label.

From the fMRI data analysis perspective, in the literature many works have been presented about the clustering of fMRI time series. In most of them the clustering procedure is made using raw data or features that are extracted from the fMRI signals [182, 184, 185, 186, 187, 188, 189, 190, 191]. Mixture models have been recently to the task of clustering [191, 90]. In [202] a mixture of General Linear Regression models (GLMs) is used that takes into account the spatial correlation of voxels using a spatial prior based on the distances between voxels and cluster centers. Recently, in [203] a mixture of linear regression models is used, where spatial correlations among the time series is achieved through Potts models over the hidden variables of the mixture model.

In this chapter we proposed a new probabilistic mixture modeling approach for clustering fMRI time series based on linear regression models where each cluster is described as a linear regression model. The innovation of the proposed method is found on three issues. First, present a sparse representation of every cluster regression model through the use of an appropriate sparse prior over the regression coefficients [87]. Enforcing sparsity is a fundamental machine learning regularization principle and has been used to tackle several problems, such as feature selection. The key idea behind use of sparse priors is that we can obtain more flexible inference methods by employing models having initially many degrees of freedom than can be uniquely adapted to given data. In particular, in sparse Bayesian regression a heavy tail prior is imposed to the coefficients of the regressor. During training such prior will zero out the coefficients that are not significant and maintain only a few large coefficients that are considered significant based on the training data. Spatial constraints of data have been also incorporated to the mixture model through the notion of Markov Random Field (MRF). This is done by considering the class labels parameter of each voxel as random variables that follows a Gibbs distribution so as to achieve similar behavior in every voxel neighborhood. Special care is given during the optimization procedure in order to meet the constraints of those parameters.

To avoid sensitivity of the design matrix to the choice of kernel matrix, we have used a kernel composite design matrix constructed as linear combination of Gaussian kernel matrices with different scaling parameter. Each kernel matrix has each own weight that is unknown and must be estimated. During the learning process the constraints of these kernel weights are also taken into account. The clustering procedure is formulated as a Maximum A Posteriori (MAP) estimation problem where the Expectation - Maximization (EM) algorithm constitutes a powerful framework for solving it. At the end of the training phase, we select the cluster that is more similar to the BOLD signal according to the Pearson correlation measure. An incremental strategy for building the mixture model is also presented. The advantage of doing that is twofold: First, it makes the EM-based learning procedure independent on initialization of model parameters. At the second level it allows us to introduce a stopping criterion of the repeating splitting process based on the correlation measurement. Intuitively, this can be seen as a model order selection for the complexity of the mixture model. As experiments with artificial and real fMRI dataset have shown, the proposed method offers very promising results with an excellent behavior in difficult and noisy environments.

8.2 The mixture of linear regression models

Let $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$ be a set of N fMRI time series of equal length T, where each element $\mathbf{y}_n$ is a sequence of data points measured at T successive time instances x_l, i.e. $\mathbf{y}_n = \{y_{nl}\}_{l=1,\cdots,T}$. The linear regression model follows the next functional description

$$\mathbf{y}_n = \mathbf{X}\mathbf{w} + \mathbf{e} \tag{8.1}$$

where $\mathbf{w}$ is the vector of M unknown regression coefficients, while $\mathbf{e}$ is the noise term that is assumed to be zero mean Gaussian with variance σ^2, i.e. $\mathbf{e} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Finally, $\mathbf{X}$ is the M-order design matrix of size $T \times M$ where its construction plays an important role for the data analysis. A typical design matrix scheme is by using the Vandermonde or B-splines matrix dealing with polynomial or splines models, respectively [**?**]. However a more powerful strategy is to assume a kernel design matrix using an appropriate kernel basis function over time instances $\{x_l\}_{l=1}^{T}$, such as the Gaussian kernel which is the most commonly used

$$K^{\lambda}(x_l, x_k) = \exp(-\frac{(x_l - x_k)^2}{2\lambda}) \ .$$

However, selecting the proper value of the scalar parameter λ is a significant issue since it depends on the amount of local variations of the data.

According to this model, the conditional probability density of the sequence $\mathbf{y}_n$ given the set of model parameters $\theta = \{\mathbf{w}, \sigma^2\}$ is also of Gaussian form

$$p(\mathbf{y}_n|\theta) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}) \ .$$

In this study we consider the problem of clustering the set of time series Y into a set of K clusters, in such a way that each cluster to contain similar time series, i.e. to have been generated from the same linear regression model. Mixture modeling provides a natural and powerful platform of establishing the clustering procedure based on linear regression models. It is described with the following probability density:

$$f(\mathbf{y}_n|\Theta) = \sum_{j=1}^{K} \pi_j p(\mathbf{y}_n|\theta_j) \ , \tag{8.2}$$

where π_j are the weights (prior probabilities) of every cluster that satisfy the constraints: $\pi_j \geq 0$ and $\sum_{j=1}^{K} \pi_j = 1$. Following this scheme, each sequence $\mathbf{y}_n$ is generated by first selecting a cluster (or source) j according to probabilities π_j and then performing a sampling based on the corresponding j-th linear regression model with parameters, $\theta_j = \{\mathbf{w}_j, \sigma_j^2\}$, as described by the normal density function $p(\mathbf{y}_n|\theta_j) = \mathcal{N}(\mathbf{X}\mathbf{w}_j, \sigma_j^2\mathbf{I})$.

Based on the above formulation, the clustering problem can be transformed into an estimation problem for the model parameters by maximizing the data log-likelihood function

$$L(\Theta) = \sum_{n=1}^{N} \log\{\sum_{j=1}^{K} \pi_j p(\mathbf{y}_n|\theta_j)\}. \tag{8.3}$$

The EM algorithm [205] constitutes an efficient method for applying to such ML estimation problem. It consists of two main steps which are applied iteratively. The E-step where the current posterior probabilities probabilities of time series to belong to each cluster are calculated:

$$z_{nj} = p(j|\mathbf{y}_n, \Theta) = \frac{\pi_j p(\mathbf{y}_n|\theta_j)}{f(\mathbf{y}_n|\Theta)} \ , \tag{8.4}$$

and the M-step where the maximization of the expected complete log-likelihood (Q-function) is performed with respect to model parameters,

$$Q(\Theta|\Theta^{(t)}) = \sum_{n=1}^{N} \sum_{j=1}^{K} z_{nj}\{\log \pi_j - \frac{1}{2} T \log \sigma_j^2 - \frac{\|\mathbf{y}_n - \mathbf{X}\mathbf{w}_j\|^2}{2\sigma_j^2}\} \tag{8.5}$$

The maximization leads to the following update rules:

$$\begin{aligned}
\pi_j &= \frac{\sum_{n=1}^{N} z_{nj}}{N}, & (8.6)\\
\mathbf{w}_j &= (\sum_{n=1}^{N} z_{nj}\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \sum_{n=1}^{N}(z_{nj}\mathbf{y}_n), & (8.7)\\
\sigma_j^2 &= \frac{\sum_{n=1}^{N} z_{nj}\|\mathbf{y}_n - \mathbf{X}\mathbf{w}_j\|^2}{T\sum_{n=1}^{N} z_{nj}}. & (8.8)
\end{aligned}$$

After the convergence of the EM algorithm, the association of N observations with the K clusters is done following the rule of the maximum posterior probability values.

8.3 Regression mixture modeling with spatial and sparse properties

The above structure of the linear regression mixture model for clustering fMRI data has some limitations and is not capable of handling some important characteristics arisen from the nature of the observations. In particular, the fmri data are structures that involve spatial properties, since adjacent voxels tend to have similar activity behavior [86]. Another desirable property is to handle temporal correlations derived from neural, physiological and physical sources [38] and have a mechanism that can automatically address the model order. Bayesian framework allows the incorporation of all these features through the use of appropriate prior distributions over the model parameters that act as useful constraints.

In order to capture spatial properties we can consider that the probabilities π_{nj} of each fMRI sequence $\mathbf{y}_n$ to belong to the j-th cluster are additional model parameters that satisfy the constraints $\pi_{nj} \geq 0$ and $\sum_{j=1}^{K} \pi_{nj} = 1$. The mixture model is now modified as

$$f(\mathbf{y}_n|\Theta) = \sum_{j=1}^{K} \pi_{nj} p(\mathbf{y}_n|\theta_j). \tag{8.9}$$

where the total set of parameters are $\Theta = \{\{\pi_{nj}\}_{n=1}^{N}, \theta_j\}_{j=1}^{K}$. We can handle the local characteristics of the voxels using the Markov Random Fields (MRF) since they have successfully applied to computer vision applications, such as the task of image segmentation [201]. In particular, we can assume the Gibbs prior distribution [98, 100] over the set of voxel labels $\Pi = \{\pi_n\}_{n=1}^{N}$ having a density function

$$p(\Pi) = \frac{1}{Z}\exp\{-\sum_{n=1}^{N} V_{N_n}(\Pi)\} \ . \tag{8.10}$$

The function $V_{N_n}(\Pi)$ denotes the clique potential function of the labels of the n-th time series vectors, where in our study it takes the following form:

$$V_{N_n}(\Pi) = \sum_{m \in N_n} \sum_{j=1}^{K} \beta_j (\pi_{nj} - \pi_{mj})^2. \tag{8.11}$$

The neighborhood N_n around the n-th voxel is the set of eight (8) voxels that are horizontally, diagonally or vertically adjacent. We also assume that every cluster has its own regularization parameter β_j providing us with a way to enforce different degree of smoothness at each cluster. Finally, the term Z is the normalizing factor that is analogous to $Z \propto \prod_{j=1}^{K} \beta_j^N$.

An important role in using a regression model is how to estimate its order M. This affects the vector of the regression coefficients $\mathbf{w}_j$. The appropriate value of M depends on the shape of data to be fitted, where models of small order may lead to underfitting while large values of M may become responsible for data overfitting. As a results this phenomenon deteriorates significantly the clustering performance. A solution to this problem can be given using the Bayesian regularization framework that penalizes models of large order [87]. In particular, we can initially assume large value of order M and impose a heavy tailed prior distribution $p(\mathbf{w}_j)$ over the regression coefficients. After training only a part of them will become active while most of them will be zero out.

The sparsity of the regression coefficients $\mathbf{w}_j$ can be achieved in an hierarchical way by considering first a zero-mean Gaussian distribution over them

$$p(\mathbf{w}_j|\boldsymbol{\alpha}_j) = N(\mathbf{w}_j|\mathbf{0}, \mathbf{A}_j^{-1}) = \prod_{l=1}^{M} N(w_{jl}|0, \alpha_{jl}^{-1}) \ , \tag{8.12}$$

where A_j is a diagonal matrix containing the M components of the hyperparameter vector $\boldsymbol{\alpha}_j = (a_{j1}, \ldots, a_{jM})$. At a second level, a Gamma prior distribution is imposed on hyperparameters α_{jl}

$$p(\alpha_j) = \prod_{l=1}^{M} \Gamma(\alpha_{jl}|b, c) \propto \prod_{l=1}^{M} \alpha_{jl}^{b-1} \exp^{-c\alpha_{jl}} \ . \tag{8.13}$$

The above two-stage hierarchical sparse prior is actually the Student's-t distribution enforcing most of the values α_{jl} to be large and thus eliminating the effect of the corresponding coefficients w_{jl} by setting to zero. In such way the regression model order for every cluster is automatically selected and overfitting is avoided.

As mentioned before, the construction of the design matrix $\mathbf{X}$ is a crucial part of the regression model. In our case we have considered that each cluster has its own design matrix $\mathbf{X}_j$ written as a mixture of kernel matrices[206, 204]

$$\mathbf{X}_j = \sum_{s=1}^{S} u_{js} \mathbf{X}_s$$

where $\mathbf{X}_s$ is the kernel matrices with scalar parameter λ_s. The weights u_{js} satisfy the constraints $u_{js} \geq 0$ and $\sum_{s=1}^{S} u_{js} = 1$. This scheme performs an inference from a pool of S kernel functions which are combined into a composite space. Every candidate kernel matrix $\mathbf{X}_s$ has its own scale parameter λ_s value. The parameters u_{js} must be estimated in order to obtain the weighted scheme of the kernel combination that better suits to every cluster.

From the above analysis, the clustering procedure becomes a Maximum-A-Posteriori (MAP) estimation problem, where the log-likelihood of the model (Eq. 8.3) is augmented with two penalty terms: a) one that corresponds to the prior for the labels Π (spatial constraints) and b) another one that corresponds to the sparse prior for the regression coefficients $\mathbf{w}_j$ (sparse constraints)

$$\mathcal{L}_{MAP}(\Theta) = \sum_{n=1}^{N} \log\{\sum_{j=1}^{K} \pi_{nj} p(\mathbf{y}_n|\theta_j)\} + \log p(\Pi) + \sum_{j=1}^{K} \left\{ \log p(\mathbf{w}_j|\boldsymbol{\alpha}_j) + \log p(\boldsymbol{\alpha}_j) \right\} \ . \tag{8.14}$$

The application of the EM algorithm to the MAP estimation problem requires the conditional expectation values z_{nj} of the hidden variables to be computed during the E-step

$$z_{nj} = P(j|\mathbf{y}_n, \Theta) = \frac{\pi_{nj} p(\mathbf{y}_n|\theta_j)}{f(\mathbf{y}_n|\Theta)} \ . \tag{8.15}$$

At the M-step, the maximization of the the expected value of the MAP log-likelihood of the complete data is performed:

$$\begin{aligned} Q(\Theta|\Theta^{(t)}) \quad = \quad & \sum_{n=1}^{N}\sum_{j=1}^{K} z_{nj}\{\log \pi_{nj} - \frac{1}{2} T \log \sigma_j^2 - \frac{\|\mathbf{y}_n - \mathbf{X}_j \mathbf{w}_j\|^2}{\sigma_j^2}\} \\ & - \log \beta_j - \beta_j \sum_{m \in \mathcal{N}_n} (\pi_{nj} - \pi_{mj})^2 - \sum_{j=1}^{K} \frac{1}{2} \mathbf{w}_j^T \mathbf{A}_j \mathbf{w}_j + \\ & \sum_{l=1}^{M} \{(b-1) \log \alpha_{jl} - c\alpha_{jl}\} \ . \end{aligned} \tag{8.16}$$

By setting the partial derivatives of the above Q function with respect to label parameters π_{nj} equal to zero, we obtain the following quadratic equation:

$$\pi_{nj}^2 - < \pi_{nj} > \pi_{nj} - \frac{1}{2\beta_j |\mathcal{N}_n|} z_{nj} = 0 \ , \tag{8.17}$$

where $< \pi_{nj} >$ is the mean value of the j-th cluster's probabilities of the spatial neighbors of the n-th voxel, i.e. $< \pi_{nj} > = \frac{1}{|\mathcal{N}_n|} \sum_{m \in \mathcal{N}_n} \pi_{mj}$. The above quadratic expression has two roots, where we select only the root with the positive sign since it yields the constraint $\pi_{ij} \geq 0$:

$$\pi_{nj} = \frac{< \pi_{nj} > + \sqrt{< \pi_{nj}^2 > + \frac{2}{|\mathcal{N}_n|} z_{nj}}}{2} \ . \tag{8.18}$$

However, these values do not satisfy the constraints $0 \leq \pi_{nj} \leq 1$ and $\sum_{j=1}^{K} \pi_{nj} = 1$, and there is a need to project them on their constraint convex hull. For this purpose an efficient convex quadratic programming method is used as presented in [201].

For the rest model parameters $\theta_j = \{\mathbf{w}_j, \boldsymbol{\alpha}_j, \sigma_j^2\}$ the update rules can be easily obtained as

$$\mathbf{w}_j = \Big[\Big(\sum_{n=1}^{N} z_{nj}\Big)\frac{1}{\sigma_j^2}\mathbf{X}_j^T\mathbf{X}_j + \mathbf{A}_j\Big]^{-1}\frac{1}{\sigma_j^2}\mathbf{X}_j^T\Big(\sum_{n=1}^{N} z_{nj}\mathbf{y}_n\Big) \tag{8.19}$$

$$\alpha_{jl} = \frac{1+2c}{w_{jl}^2 + 2b} \tag{8.20}$$

$$\sigma_j^2 = \frac{\sum_{n=1}^{N} z_{nj}\|\mathbf{y}_n - \mathbf{X}_j\mathbf{w}_j\|^2}{T\sum_{n=1}^{N} z_{nj}}. \tag{8.21}$$

Finally, the parameters u_{jr} of the kernel composite design matrix are obtained after solving the following optimization problem:

$$\max_{\mathbf{u}_j}\Big\{\mathbf{u}_j^T\mathbf{K}_{\mathbf{w}}^{j\;T}\mathbf{K}_{\mathbf{w}}^{j}\mathbf{u}_j - 2\mathbf{u}_j^T\mathbf{K}_{\mathbf{w}}^{j\;T}\frac{\sum_{n=1}^{N} z_{nj}\mathbf{y}_n}{\sum_{n=1}^{N} z_{nj}}\Big\}, \text{ s.t. } \sum_{s=1}^{S} u_{js} = 1 \text{ and } u_{js} \geq 0 \ .$$

The matrix $\mathbf{K}_{\mathbf{w}}^{j}$ is derived by rearranging the terms in the linear regression model as describe below:

$$\mathbf{X}_j\mathbf{w}_j = \Big(\sum_{s=1}^{S} u_{js}\mathbf{X}_s\Big)\mathbf{w}_j = \mathbf{K}_{\mathbf{w}}^{j}\mathbf{u}_j. \tag{8.22}$$

At the end of the learning process the activation map of the brain is constructed. In particular, we initially select the cluster h that best match with the BOLD regressor (which is known before the data analysis) among the K mixture components. This is done following the Pearson correlation measurement between each cluster's estimated mean value $(\mathbf{X}_j\mathbf{w}_j)$ and the BOLD regressor, which is in fact the cosine similarity. The set of voxels that belong to this cluster h draw the brain activation region, while all the rest voxels from different clusters assign the non-activation region.

8.3.1 Incremental learning

A drawback of the EM algorithm is its sensitivity to the initialization of the model parameters due to its local nature. Improper initialization may lead to poor local maxima of the log-likelihood that sequentially affects the quality of the clustering solution. A solution is to test several initial values and select the one set of values that reach the maximum log-likelihood function value after running one-step of the EM algorithm. However, other more advanced methods have been recently presented about incrementally building Gaussian mixture models [207, 209, 210]. We have adopted such scheme in our approach and have developed a framework that iteratively adds a new component to the mixture by performing a component split procedure.

Initially, we start with a model having one component that comes from a single linear regression model. Let now assume that we have already constructed a mixture f_k of k linear regression components

$$f_k(y_i|\Theta_k) = \sum_{j=1}^{k} \pi_j p(y_i|\theta_j) \ . \tag{8.23}$$

The component j^* which is more similar to the BOLD regressor is then selected for splitting and a new component $k+1$ is generated. For initializing its parameters we perform the following steps:

- Among the time series that currently belong to the selected for splitting cluster j^*, find a small percentage of the worst fitted cases and calculate their mean value $\overline{y_*}$.

- Fit a regression model to the this mean sequence $\overline{y_*}$ and obtain initial values for the new component's regression coefficients w_{k+1}, regularization parameter $\alpha_{k+1,l} = 1/w^2_{k+1,l}$ and noise variance σ^2_{k+1}. The kernel weights of the design matrix $\mathbf{X}_{k+1}$ are equivalent, i.e. $u_{k+1,s} = 1/S$.

- The label parameters are initialized as $\pi_{n,k+1} = \{\pi_{n,j^*}\}^{new} = \frac{\{\pi_{n,j^*}\}^{old}}{2}$

Subsequently, the EM algorithm can be applied for estimating the parameters Θ_{k+1} of the new mixture model.

The splitting procedure is responsible for adding one linear regression component at a time. Intuitively thinking, it can be seen as a pruning mechanism that is repeated until found the cluster that best describes the BOLD effect in terms of its curve representation and also its homogeneous appearance. For terminating the procedure we have used the criterion of the percentage of the correlation increase between two successive steps. When this percentage becomes very small the incremental training process is terminated. In this case the mixture increment from Θ_k to Θ_{k+1} does not offer any significant improvement to the correlation criterion, and thus the best found cluster from the previous step is the final solution.

8.4 Experimental results

The proposed method have been tested using simulated and real fMRI data. We have compared the proposed mixture model with spatial and spatial properties (SSRM), using both the incremental (iSSRM) and the regular version, with the ML regression mixture (MLRM) approach. The MLRM approach is similar to the SSRM. The only difference is that the estimation of regression coefficients of each component is based on the ML principle (i.e. we have not used the sparse prior over the regression coefficients). The matrices $\mathbf{X}_s, s = 1, \cdots, S$, where $S = 10$, were created by using Gaussian kernels for different values of the width parameter λ_s varying from 0.1 to 2 with step 0.2. In all experiments, first we applied the incremental version of the algorithm to determine among others the number of clusters and then we applied the SSRM and the MLRM algorithms using this number.

To initialize all the algorithms, expect the iSSRM, the following procedure is adopted. First, we select randomly K time series, one for each cluster, from the dataset. Then, the ML learning rule is applied in each regression model to estimate the regression coefficients. After that the parameters a_{jl} can be estimated as $\alpha_{jl} = \frac{1}{w_{jl}^2}$, where j is the cluster and l the corresponding regressor. The mixing probabilities π_{nj} are initially set to $\frac{1}{K}$ and the parameters u_{js} are initially set to $\frac{1}{S}$. Finally, two steps of the EM algorithm were executed to improve the estimation of model parameters and to evaluate the loglikelihood. This approach is applied for one hundred different trials and the solution with the maximum log-likelihood value is selected for initializing the parameters of EM algorithm.

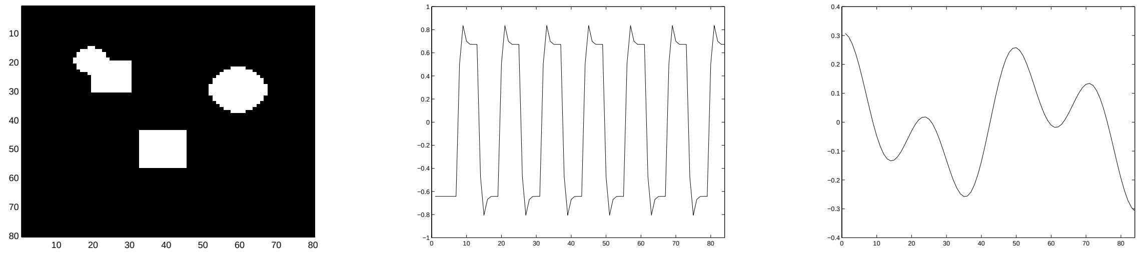
In experiments, the design matrix of each cluster $\mathbf{X}_j$ has two components, one component which comes out from the combination of kernel matrices and one component which is common in all cluster and it is the BOLD regressor, $\mathbf{X}_j = [\mathbf{X}_j^{(F)} \; \mathbf{b}^T]$. This slightly different design matrix from that described previously does not change at all the proposed model. It must be only taken into account when the estimation of parameters u_{js} is performed. In that case the term $w_p\mathbf{b}$ must be removed for the observations $\mathbf{y}_n$ where w_p denotes the corresponding weight to the BOLD regressor.

8.4.1 Experiments using simulated fMRI data

In experiments with simulated fMRI data we create 3-D dataset of time series from a linear regression model where the design matrix was known as well as the regression coefficients. In these time series we have added white gaussian noise of various SNR levels. The SNR is defined between the BOLD regressor and the white gaussian noise component of the model. The spatial correlation between the time series is achieved through the regression coefficients. The regression coefficients of the BOLD regressor have a spatial pattern which is drawn in Fig. 8.1a. The BOLD regressor, which is used to model the neural activity, is shown in Fig. 8.1b. Also, in the time series we have added a slow varying component to model the drift in the fMRI time series (Fig. 8.1c).

To quantify the performance and measure the quality of the clustering, we have used two criteria:

- the Performance (success rate), which is the percentage of correctly classified time series and quantifies the ability of the method to assign each time series to the correct cluster.



(a) (b) (c)

Figure 8.1: (a) Spatial pattern of time series (black : non - activation, white: activation), (b) BOLD regressor and (c) Drift component

- the normalized mutual information (NMI), which is an information theoretic measure based on the mutual information of the true labeling (Ω) and the clustering (C) normalized by their entropies:

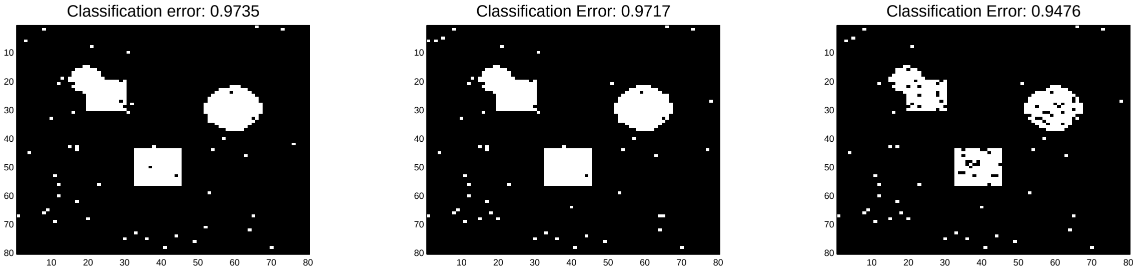
$$NMI(\Omega, C) = \frac{I(\Omega, C)}{(H(\Omega) + H(C))/2}. \tag{8.24}$$

We have compared the iSSRM and the SSRM algorithms with the MLRM algorithm. The SSRM and the MLRM algorithm assumed that the number of cluster is known. To define the number of clusters we use first the iSSRM algorithm which provides us with the estimated number of clusters, then this number is used to the others algorithms. The results are shown in Table (8.1), we can see that the iSSRM algorithm present better performance from the other algorithms, in terms of the classification error and the mutual information. Comparing the iSSRM and the SSRM algorithms we can see that the incremental version provides better results from the SSRM. Since the difference of these two algorithms is on the initialization strategy of the EM algorithm, we can concluded that the initialization is responsible for the difference in the results. Finally, in Fig. 8.2 we shown an example of the clustering in the case of -8 dB. We provide the activation of each method as well as the classification error of them. It is obvious the ability of the iSSRM and SSRM algorithms to fill the holes that are observed in the MLRM algorithm.

To shown the usefulness of the proposed approach in the construction of design matrix we compare the iSSRM algorithm with a version of SSRM with out using a combination of matrices but only one of them (we called this method sRM). In our experiments the extended design matrix $\mathbf{X}_j$ was constructed as a combination of 10 design matrices ($R = 10$) (based on the idea of kernels) $\mathbf{F}_r, r = 1, \cdots, 10$. In Table 8.2 we shown the results for iSSRM and the sRM algorithms. The sRM algorithm has been run 10 times with a different design matrix $\mathbf{F}_r$ and we have choose the best results. Again, the iSSRM algorithm presents better performance in terms of classification error and mutual information.

Table 8.1: Comparative results for simulated data in various noisy environments.

	Performance			*NMI*		
SNR	iSSRM	SSRM	MLRM	iSSRM	SSRM	MLRM
0	0.9989	1.0000	1.0000	0.9937	1.0000	1.0000
-2	0.9993	0.9999	0.9999	0.9956	0.9993	0.9989
-4	0.9983	0.9983	0.9967	0.9864	0.9862	0.9751
-6	0.9901	0.9688	0.9827	0.9380	0.8945	0.8973
-8	0.9713	0.9049	0.9466	0.8513	0.7237	0.7468
-10	0.9456	0.8296	0.8862	0.7515	0.5458	0.5565
-12	0.8384	0.8961	0.8075	0.5621	0.6310	0.3705
-14	0.8035	0.7870	0.6571	0.4648	0.4138	0.1436

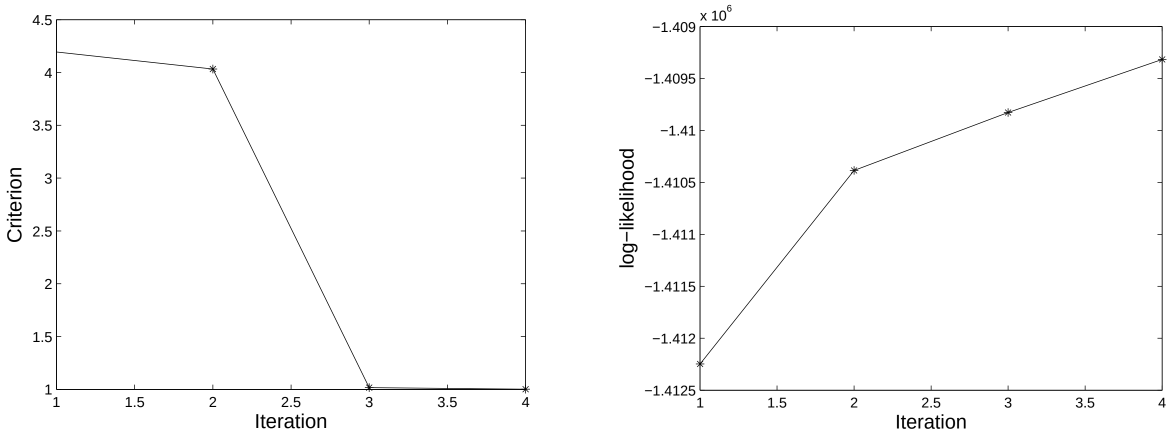


iSSRM SSRM MLRM

Figure 8.2: Activation patterns using (a) iSSRM, (b) SSRM and (c) MLRM

Table 8.2: Comparative results for simulated data in various noisy environments.

	Performance		*NMI*	
SNR	iSSRM	sRM (λ)	iSSRM	sRM (λ)
0	1.0000	0.9405 (0.1)	1.0000	0.7926 (0.1)
-2	0.9998	0.9386 (0.1)	0.9986	0.7872 (0.1)
-4	0.9979	0.9152 (0.1)	0.9836	0.7144 (0.1)
-6	0.9897	0.7677 (0.9)	0.9366	0.4488 (0.9)
-8	0.9730	0.7129 (1.7)	0.8578	0.3658 (1.7)
-10	0.9507	0.6592 (1.7)	0.7687	0.2870 (1.7)
-12	0.8589	0.6793 (1.3)	0.5980	0.2949 (1.3)
-14	0.7804	0.6475 (0.5)	0.4307	0.2366 (0.5)



(a) (b)

Figure 8.3: Termination criterion (SNR=-14dB).

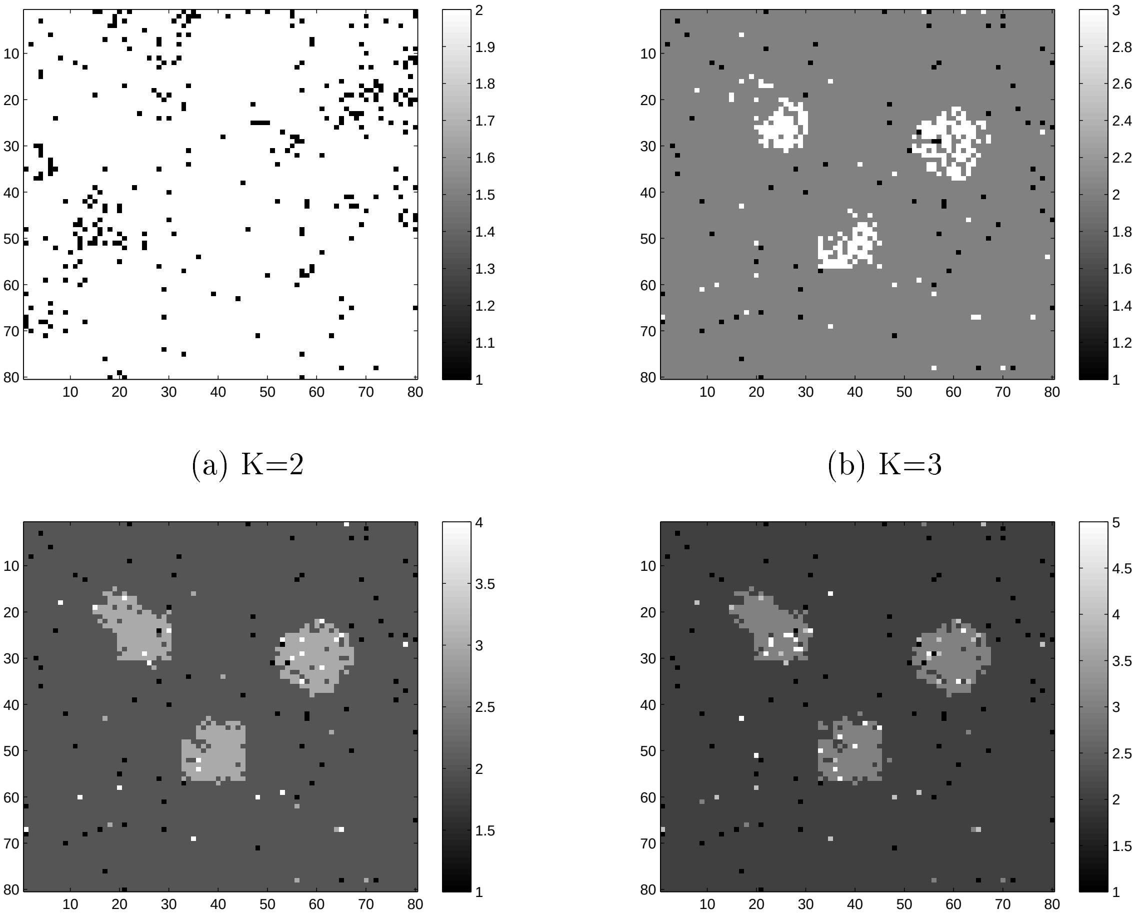
In Fig. 8.3 we provide the termination criterion with respect to the iteration of the incremental algorithm. Also, in the same figure the log-likelihood is given. These results have been obtained from a dataset where the SNR was -14 dB. We can observe that at every step of the incremental algorihm the log-likelihood is always increased while the termination criterion converges after few iterations. This is more obvious in Fig. 8.4 where images of the clustering procedure for the same dataset are provided. We see that the increase of number clusters from $K = 3$ to $K = 4$ does not provide any new information.

8.4.2 Experiments using real fMRI data

We have applied the iSSRM, SSRM and the MLRM algorithms using real fMRI data concerns block design and event related experiments. In both datasets, we followed the standard preprocessing steps of the SPM package, i.e. realignment, segmentation, normalization and spatial smoothing steps. Data are then scaled by using the global mean value of all time series as a factor. Finally, each time series was then high pass filtered using a set of discrete cosine basis functions. At first we have studied a real block design fMRI dataset[1] designed for auditory processing task on a healthy volunteer. Its functional images consisted of $\mathcal{M} = 68$ slices ($79 \times 95 \times 68$, $2mm \times 2mm \times 2mm$ voxels). Experiments were made with the slice 29 of this dataset. We have applied the iSSRM algorithm, which provides us with the number of clusters. After that, the SSRJM and the MLRM algorithms have been applied.

The results of clustering are shown in Fig. 8.5. The images show the position of clusters inside the brain. Also, an activation map is provided, which is produces by

[1] It was downloaded from the SPM web page http://www.fil.ion.ucl.ac.uk/spm/



(a) K=2

(b) K=3

(c) K=4

(d) K=5

Figure 8.4: Clustering results (SNR=-14 dB).

adopting an analysis based on the GLM. More specifically in our case the activation map produces by using the spatial method described in [213]. We can observe that there is one cluster that coincides with the activation region which is the auditory cortex. Comparing the clustering results with the activation map we can observe that there is great similarity between the cluster responsible for activation and the strength of the regression coefficient.

In Fig. 8.6 we depict the center of each cluster with the BOLD response, also we provide the correlation coefficients of each center cluster with the BOLD response. This is shown for all the methods. We see that the iSSRM algorithm provides us with a cluster center which is more correlated to the BOLD response than the others algorithms.

In event related experiments we analyzed fMRI data consisted of images acquired from a motor event related paradigm available at the affiliated University Hospital of Ioannina. During this experiment patients with RLS (restless legs syndrome) performed random and spontaneous limb movements evoked by sensory leg uneasiness. These movements were used to create the indicator vector in our modeling that was convolved next with the hemodynamic response function (HRF) in order to provide the BOLD signal. In Fig. 8.7 we show the clustering results together with the activation map. Similar observations, just like the auditory experiment, can be done here. All methods provides us with the cluster related to the primary motor region and supplementary motor areas of the brain. We can observe the similarity between the cluster of activation and the activation map. Also, in Fig. 8.8 we depict the center of each cluster with the BOLD response, in addition the correlation coefficients of each center cluster with the BOLD response are provided.

8.5 Conclusions

In this chapter, we proposed a probabilistic mixture modeling approach for the clustering of fMRI time series. More, specifically a mixture of linear regression models with sparse and spatial properties is presented. Sparse priors are placed on the weights of each linear regression model helping us to deal with problem of model order selection. Also, spatial priors are used on the mixing coefficients to take into account the spatial correlation between the voxels. This is achieved by using a Gibbs distribution. Furthermore, to avoid sensitivity of the design matrix to the choice of kernel matrix, we have used a kernel composite design matrix constructed as linear combination of Gaussian kernel matrices with different scaling parameter. Our future research study is focused to three directions: a) to examine the appropriateness of other types of sparse priors [107], b) to try alternative potential functions of the Gibbs distribution and c) to try different approaches for learning the design matrix [211].

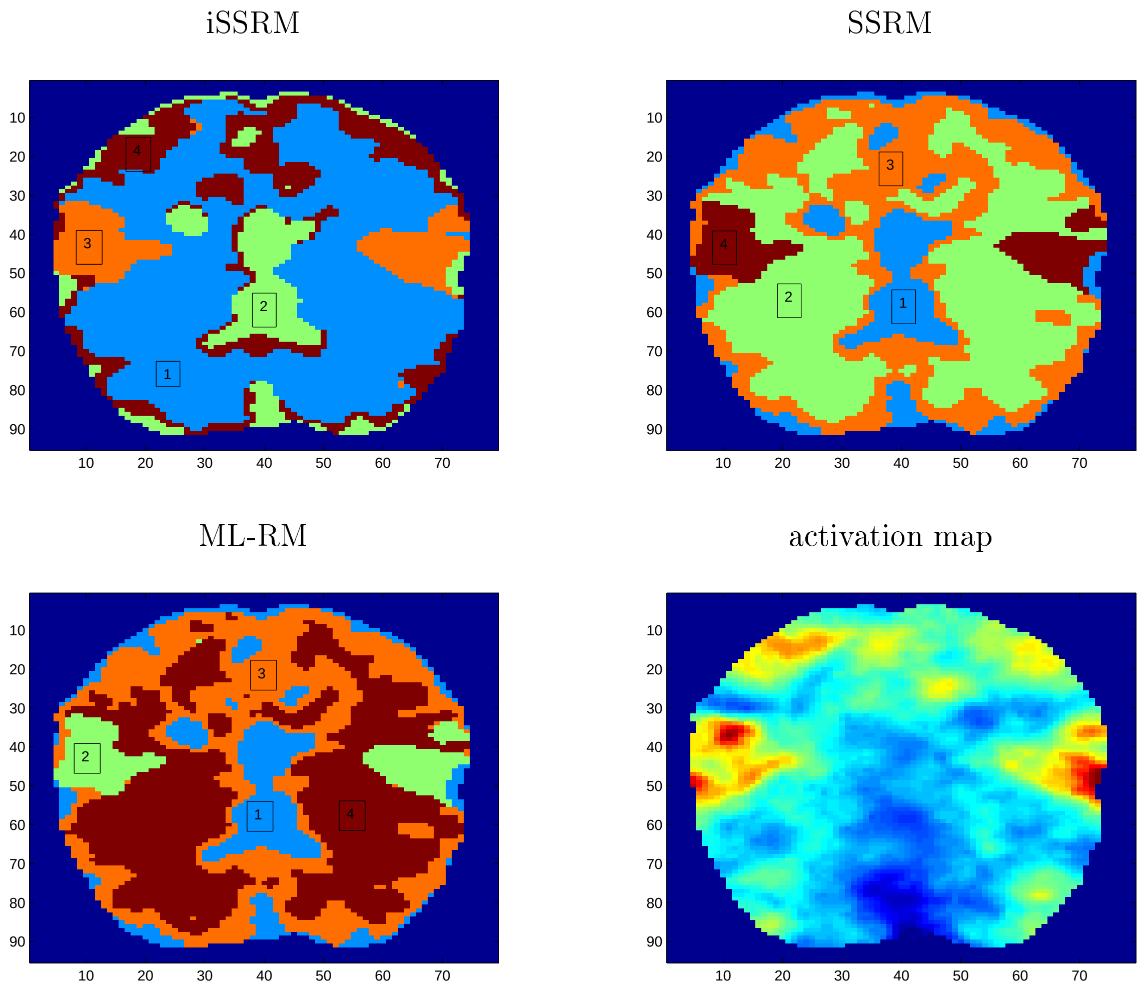
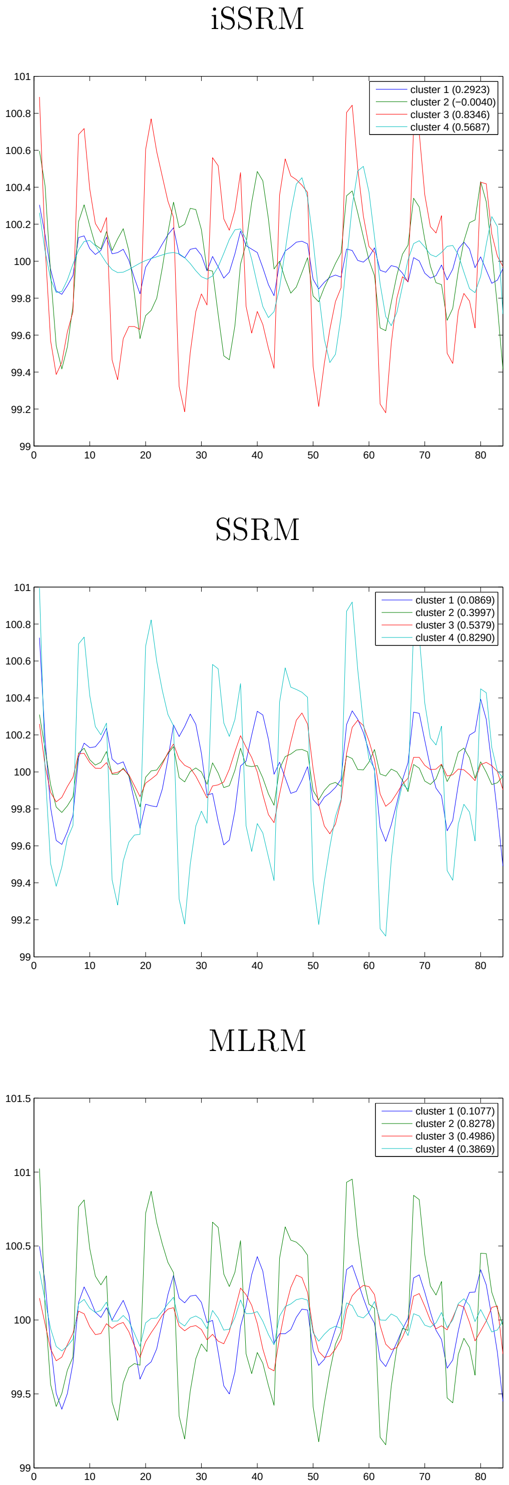


Figure 8.5: Images of voxels clustering (auditory experiment).



iSSRM

SSRM

MLRM

Figure 8.6: Center clusters (auditory experiment).

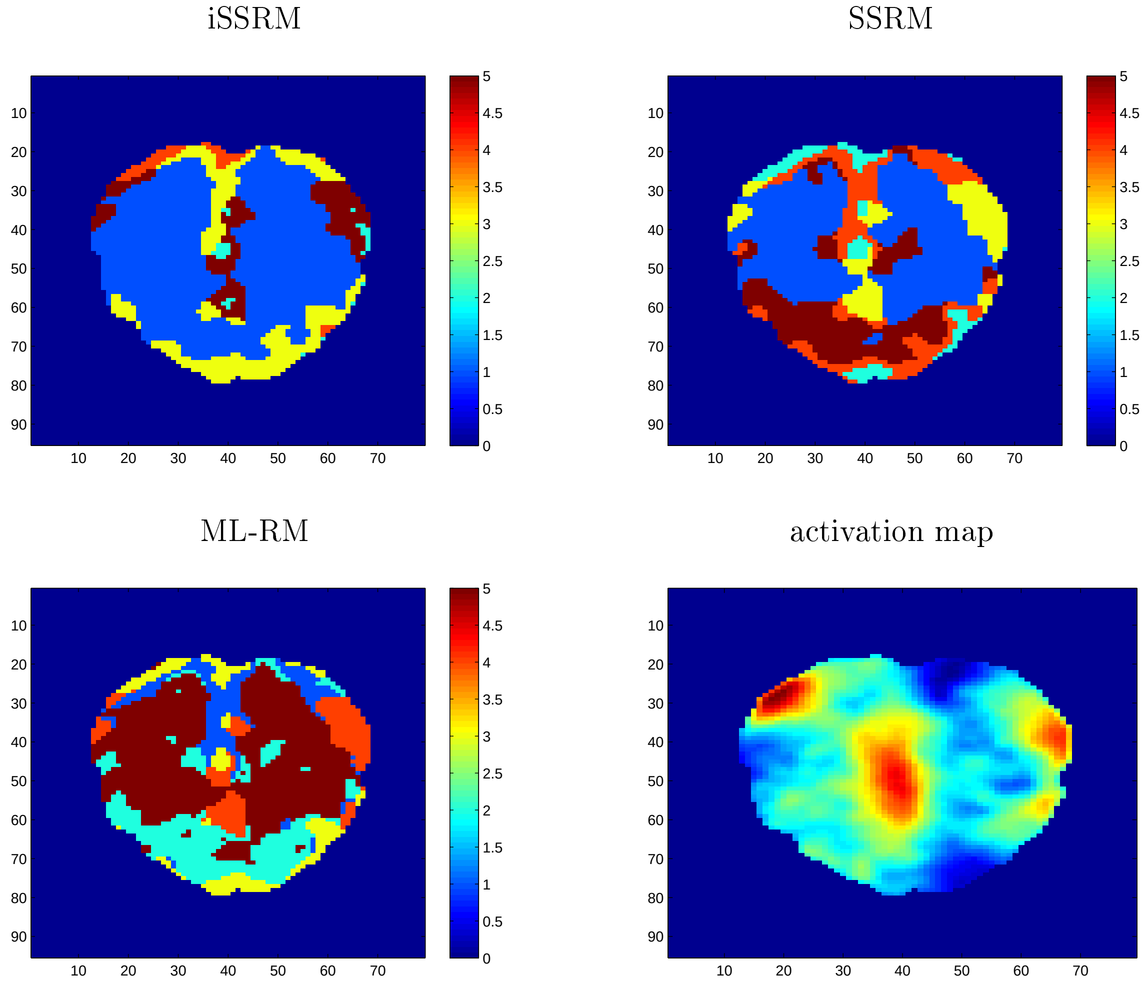
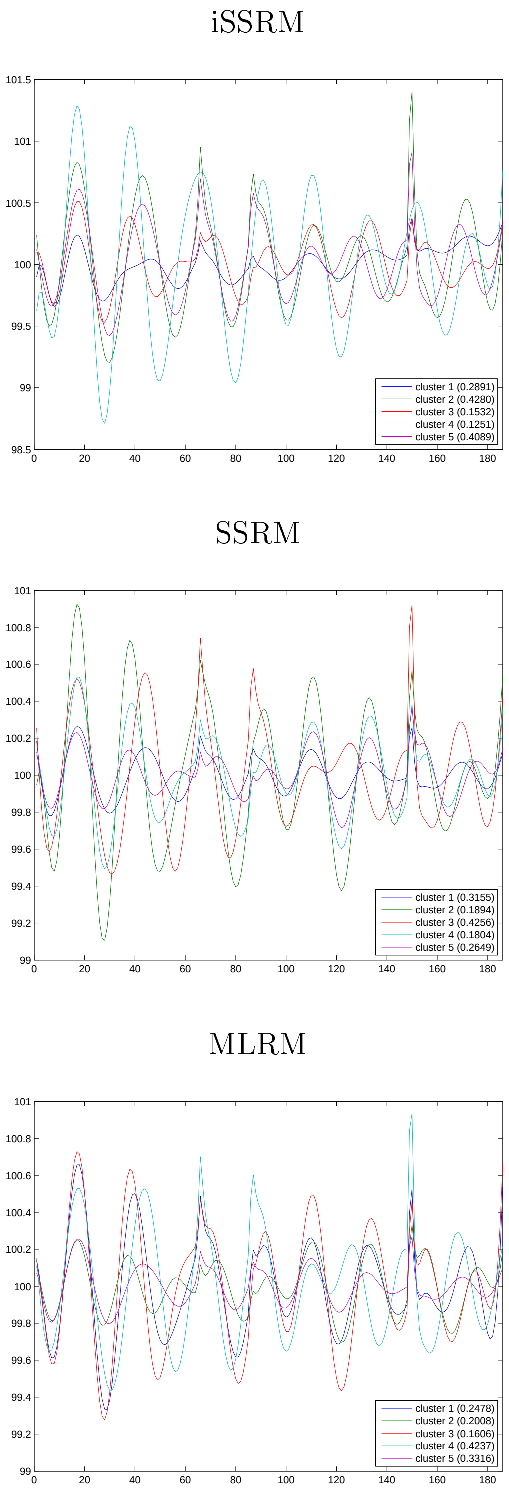


Figure 8.7: Images of voxels clustering (motor experiment).



iSSRM

SSRM

MLRM

Figure 8.8: Center clusters (motor experiment).

Chapter 9

Conclusions

In this thesis we have studied the linear regression model and the time varying autoregressive model and its applications on problems of biomedical signal processing. The time varying autoregressive model was proposed for the enhancement of epileptic EEG spikes, while variations of the linear regression model were proposed for the analysis of fMRI time series, the estimation of ERPs and the drift removal from HRV time series. More, specifically a model, based on the smoothness prior, was proposed for the estimation of a signal inside a noisy environment. For the estimation procedure we adopted the Variational Bayesian Methodology and the proposed model was used to find the ERPs and the drift inside the HRV time series. Next, two algorithms, using the linear regression model, were proposed for the analysis of fMRI time series. In these algorithms, we have focused on issues concerning the noise model. The noise decomposed into two components, one component originates for the time series, while the other originates from the images. Furthermore, the linear regression model with sparse and spatial properties was used for the analysis of fMRI time series. To include these properties into the model an enhanced version of the Gibbs distribution was used. Finally, we proposed a clustering technique for fMRI time series. An extended version of mixture modeling, based on the linear regression model and the Gibbs distribution, was proposed for the clustering. Also, an incremental algorithm was derived based on the above mixture models.

In chapter 4, the time varying autoregressive model was used for the enhancement of epileptic spikes. This model was represented in the form of a state-space model, and then the Kalman Filter was used to estimate the autoregressive coefficients. The results were indicated that the proposed method is able to enhance the epileptic spikes in terms of SNR. Also, when the proposed method is used as a preprocessing step into a detection procedure, is able to reduce the false alarms while keep at acceptable level the loose of epileptic spikes.

In chapter 5, we proposed a method for the recovery of biomedical signal from a noisy environment. More specifically, we assumed that the signal of interest was smooth. This assumption guided us to proposed the smoothness prior for the signal. The noise was studied in two cases: white gaussian noise and colored Gaussian noise. To estimate the various model parameters we adopt a probabilistic approach based on the Variational Bayesian Methodology. This approach uses an approximate posterior, instead of the true, helping us to obtained closed form solutions. The results had shown the usefulness of the proposed method comparing to the wavelet denoising approach and the generalized cross validation criterion. The proposed method was applied to estimate the ERPs from the EEG signal and the drift inside the HRV time series.

In chapter 6, two methods were proposed to find the activation of brain using fMRI time series. More specifically, the linear regression model was used and the variance of the noise was decomposed into two components, one across time - series and the other across images. Again, the Variational Bayesian Methodology was used for the estimation procedure of various model parameters. The results shown the ability of the proposed methods to find accurately the brain activation.

In chapter 7, we extended our study in the analysis of fMRI time series by introducing spatial properties into the linear regression model. More specifically, an enhance prior distribution, based on Gibbs distribution, was proposed. This prior includes simultaneously sparse and spatial properties into the linear regression model. Experiments were performed using real and simulated data. The results indicated the superiority of the proposed method.

In chapter 8, a clustering method, based on mixture modeling, was proposed for the analysis of fMRI time series. The proposed mixture model uses spatial over the mixing probabilities to take into account the correlation between adjacent fMRI time series. Also, the mixture components are based on linear regression models which help us to model better the time series and confront the large dimension of time series. Furthermore, an incremental algorithm based on mixture modeling was proposed for the clustering of fMRI time series. The incremental algorithm help us to confront the problem of ill-balanced data observe in fMRI time series.

In future work, it would be interesting to study the enhancement of epileptic spikes using multichannel recordings. This will help us to include spatial information into the model. Also, a method, based on the EM algorithm, where the model parameters will be estimated from the data it would be useful, especially for automatic monitoring of EEG signal. In the signal estimation method of chapter 5, we could extended the model by using a non stationary smoothness prior or by adopting other noise distributions such as the Student's t - distribution. However, these extensions to be computationally efficient will be needed to resort into approximation techniques to estimate the posterior covariance.

In chapter 6, the drift was modeled by using Gaussian basis functions with fixed the scale parameter. An extension of this approach is to use other basis functions for the drift. In addition, we could use a learning procedure for the design matrix as that described in chapter 8 to avoid the need to make assumptions about the scale parameter. Also, a similar learning procedure should be examined in conjunction with the linear regression model of chapter 7. Finally, in the model of chapter 8 it would be useful to examine other distribution component than the gaussian in the clustering of time series.

In the analysis of fMRI time series, it would be interesting to construct generative models that include more preprocessing steps into the same framework. This will help to better understand how the various components and properties of fMRI time series interact each other. One crucial aspect in the analysis of fMRI time series is the estimation of HRF during the analysis of time series and how this affects the subsequent analysis of data. Finally, the use of multiple imaging techniques, to overcome the limitations of each method, is very appealing. For example, EEG and fMRI data can be collected simultaneously. Merging these two techniques we hope to get the best of both worlds. In this direction new models will be constructed to explain the observations.

Furthermore, the use of proposed models is not restricted only to problems that were described in this thesis. For example, the clustering method that was described in chapter 8 can be easily applied in the analysis of other biomedical signals such as the clustering of ERPs. Besides the classification results that we obtain due to the clustering, we also obtain time series which corresponds to the means of each cluster. These time series can be used as regressors into a linear regression model and we could applied a similar procedure, as that of the statistical analysis of fMRI data, to obtain statistical brain maps based on ERPs.

Bibliography

[1] S. Cerutti, A. Goldberger, and Y. Yamamoto, “Recent advances in heart rate variability signal processing and interpretation,” *IEEE Transactions on Biomedical Engineering*, vol. 53, pp. 1–3, January 2006.

[2] M. D. Rienzo, G. Mancia, G. Parati, A. Pedotti, and A. Zanchetti, *Frontiers of Blood Pressure and Heart Rate Analysis.* New York, NY, USA: IOS Press, 1997.

[3] M. Akay, *Nonlinear Biomedical Signal Processing, Volume 2, Dynamic Analysis and Modeling.* Wiley-IEEE Press, 2000.

[4] Electrophysiology, Task Force of the European Society of Cardiology the North American Society of Pacing, “Heart Rate Variability : Standards of Measurement, Physiological Interpretation, and Clinical Use,” *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[5] U. R. Acharya, K. P. Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, “Heart Rate Variability: A Review,” *Medical and Biological Engineering and Computing*, vol. 44, pp. 1031–1051, December 2006.

[6] J. Mateo and P. Laguna, “Analysis of heart rate variability in the presence of ectopic beats using the heart timing signal,” *IEEE Transactions of Biomedical Engineering*, vol. 50, pp. 334–343, March 2003.

[7] K. Solem, P. Laguna, and L. Sornmo, “An efficient method for handling ectopic beats using the heart timing signal,” *IEEE Transactions of Biomedical Engineering*, vol. 53, pp. 13–20, January 2006.

[8] J. Malmivuo and R. Plonsey. *Bioelectromagnetism - Principles and Applications of Bioelectric and Biomagnetic Fields.* Oxford University Press, New York, 1995.

[9] E. Niedermeyer and E. F. Lopes da Silva. *Electroencephalography: Basic Principles, Clinical Applications and Related Fields. 3rd ed.* Baltimore: Williams and Wilkins, 1993.

[10] M. Palus. Nonlinear in normal human EEG: cycles, temporal asymmetry, nonstationarity and randomness, not chaos. *Biol. Cybern.*, vol. 75, pp. 389–396, 1996.

[11] G. Pfurtscheller. Graphical display and statistical evaluation of event-related desynchronization (ERD). *Electroenceph. Clin. Neurophysiol.*, vo. 43, pp. 757–760, 1977.

[12] G. Pfurtscheller. Spatiotemporal analysis of alpha frequency components with the erd technique. *Brain Topogr.*, vol. 2, pp. 3–8, 1989.

[13] G. Pfurtscheller and A. Aranibar. Event-related cortical desynchronization detected by power measurements of scalp EEG. *Electroenceph. Clin. Neurophysiol.*, vol. 42, pp. 817–826, 1977.

[14] G. Pfurtscheller and F. L. D. Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, vol. 110, pp. 1842–1857, 1999.

[15] J. Pijn, J. V. Neerven, A. Noest, and F. L. da Silva. Chaos or noise in EEG signals dependence on state and brain site. *Electroencephalogr. Clin. Neurophysiol.*, vol. 79, pp. 371-ủ381, 1991.

[16] C. Krause, H. Lang, M. Laine, S. Helle, M. Kuusisto, and P. B. Event related desynchronization evoked by auditory stimuli. *Brain Topogr.*, vol. 7, pp. 107–112, 1994.

[17] P. Lahteenmaki, C. Krauseb, L. Sillanmaki, T. Salmia, and A. Lang. Event-related alpha synchronization/desynchronization in a memory-search task in adolescent survivors of childhood cancer. *Clinical Neurophysiology*, vol. 110, pp. 2064–2073, 1999.

[18] V. Goel, A. Brambrink, A. Baykal, R. Koehler, D. Hanley, and N. Thakor. Dominant frequency analysis of EEG reveals brain's response during injury and recovery. *IEEE Trans. Biomed. Eng.*, vol. 43, pp. 1083–1092, 1996.

[19] N. Thakor and S. Tong. Advances in quantitative electroencephalogram analysis methods. *Annual Review of Biomedical Engineering*, vol. 6, pp. 453–495, 2004.

[20] A. Tzallas, P. Karvelis, C. Katsis, D. Fotiadis, S. Giannopoulos, and S. Konitsiotis. A method for classification of transient events in EEG recordings: application to epilepsy diagnosis. *Methods Inf. Med.*, vol. 45,pp. 610–621, 2006.

[21] S. Wilson and R. Emerson. Spike detection: a review and comparison of algorithms. *Clin. Neurophysiol.*, vol. 113,pp. 1873–1881, 2002.

[22] F. Mormann, R. Andrzejak, C. Elger, and K. Lenhnertz. Seizure prediction: the long and the winding road. *Brain*, vol. 130, pp. 314–333, 2007.

[23] E. Waterhouse. New horizons in ambulatory electroencephalography. *IEEE Eng. Med. Biol. Mag.*, vol. 22, pp. 74–80, 2003.

[24] P. Jezzard, P. Matthews, and S. Smith, *Functional MRI : An Introduction to Methods.* Oxford University Press, USA, 2001.

[25] R. Frackowiak, J. Ashburner, W. Penny, S. Zeki, K. Friston, C. Frith, R. Dolan, and C. Price, *Human Brain Function, Second Edition.* Elsevier Science, USA, 2004.

[26] K. Friston, "Analysis of fMRI time series revisited," *Neuroimage*, vol. 2, pp. 45–53, 1995.

[27] M. Mohamed, F. Abou-Chadi, and B. Ouda, "Analysis of fMRI data using classical and bayesian approaches: A comparative study," *IFMBE Proceedings, World Congress on Medical Physics and Biomedical Engineering 2006*, vol. 14, pp. 924–931, 2006.

[28] K. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, "Classical and bayesian inference in neuroimaging: Theory," *NeuroImage*, vol. 16, pp. 465–483, June 2002.

[29] K. Friston, D. Glaser, R. Henson, S. Kiebel, C. Phillips, and J. Ashburner, "Classical and bayesian inference in neuroimaging: Applications," *NeuroImage*, vol. 16, pp. 484–512, June 2002.

[30] J. Kershaw, B. Ardekani, and I. Kanno, "Application of bayesian inference to fMRI data analysis," *Medical Imaging, IEEE Transactions on*, vol. 18, pp. 1138–1153, Dec 1999.

[31] W. Penny, S. Kiebel, and K. Friston, "Variational bayesian inference for fMRI time series," *NeuroImage*, vol. 19, pp. 727–741, July 2003.

[32] H. Luo and S. Puthusserypady, "A sparse bayesian method for determination of flexible design matrix for fMRI data analysis," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 52, pp. 2699–2706, Dec. 2005.

[33] D. MacKay, "Probable networks and plausible predictions - a review of practical bayesians methods for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, pp. 469–505, 1995.

[34] C. Bishop and M. Tipping, “Variational relevance vector machines,” *Proc. 16th Conf. Uncertainty in Artificial Intelligence*, pp. 46–53, 2000.

[35] J. Diedrichsen and R. Shadmehr, “Detecting and adjusting for artifacts in fMRI time series data,” *NeuroImage*, vol. 27, no. 3, pp. 624 – 634, 2005.

[36] C. Long, E. Brown, C. Triantafyllou, I. Aharon, L. Wald, and V. Solo, “Nonstationary noise estimation in functional MRI,” *NeuroImage*, vol. 28, no. 4, pp. 890 – 903, 2005.

[37] H. Luo and S. Puthusserypady, “fMRI data analysis with nonstationary noise models: A bayesian approach,” *IEEE Transactions on Biomedical Engineering*, vol. 54, pp. 1621–1630, Sept. 2007.

[38] W. Penny, N. Trujillo-Barreto, and K. Friston, “Bayesian fMRI time series analysis with spatial priors,” *NeuroImage*, vol. 24, pp. 350–362, Jan. 2005.

[39] M. Woolrich, M. Jenkinson, J. Brady, and S. Smith, “Fully bayesian spatio-temporal modeling of fMRI data,” *IEEE Transactions on Medical Imaging*, vol. 23, pp. 213–231, Feb. 2004.

[40] S. Kay, *Fundamentals of statistical signal processing.* Englewood Cliffs, NJ: Prentice Hall, 1993.

[41] T. Jaakola, *Variational methods for inference and learning in graphical models.* PhD thesis, Mass.Inst.Technol., Campribge, MA, 1997.

[42] M. Beal, *Variational Algorithms for Approximate Bayesian Inference.* PhD thesis, Gatsby Computational Neuroscience Unit, Univ. College London, London, U.K., 2003.

[43] D. J. MacKay, “Bayesian interpolation,” *Neural Computation*, vol. 4, pp. 415–447, 1992.

[44] B. Charlin and T. Louis, *Bayes and Empirical Bayes Methods for Data Analysis.* New York, NY: CRC Press, 2000.

[45] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer, October 2007.

[46] O. Friman, M. Borga, P. Lundberg, and H. Knutsson, “Detection and detrending in fMRI data analysis,” *Neuroimage*, vol. 22, pp. 645–655, 2004.

[47] R. Baumgartner, L. Ryner, W. Richter, R. Summers, M. Jarmasz, and R. Somorjai, “Comparison of two exploratory data analysis methods for fMRI: fuzzy clustering vs.

principal component analysis," *Magnetic Resonance Imaging*, vol. 18, no. 1, pp. 89 – 94, 2000.

[48] A. Meyer-Baese, A. Wismueller, and O. Lange, "Comparison of two exploratory data analysis methods for fMRI: unsupervised clustering versus independent component analysis," *IEEE Transactions on Information Technology in Biomedicine,*, vol. 8, pp. 387 –398, Sept. 2004.

[49] M. J. Mckeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fMRI data by blind separation into independent spatial components," *Human Brain Mapping*, vol. 6, no. 3, pp. 160–188, 1998.

[50] J. Carew, G. Wahba, X. Xie, E. Nordheim, and M. Meyerand, "Optimal spline smoothing of fMRI time series by generalized cross-validation," *Neuroimage*, vol. 18, pp. 950–961, Apr 2003.

[51] F. Meyer, "Wavelet-based estimation of a semiparametric generalized linear model of fMRI time-series," *IEEE Transactions on Medical Imaging*, vol. 22, p. 315ΰ322, March 2003.

[52] H. Luo and S. Puthusserypady, "Analysis of fMRI data with drift: Modified general linear model and bayesian estimator," *IEEE Transactions on Biomedical Engineering*, vol. 55, pp. 1504 – 1511, May 2008.

[53] K. Friston and W. Penny, "Posterior probability maps and SPMs," *NeuroImage*, vol. 19, July 2003.

[54] K. J. Friston, *Statistical Parametric Mapping: The Analysis of Functional Brain Images.* Academic Press, 2006.

[55] K. Friston, "Statistical parametric mapping." `http://www.fil.ion.ucl.ac.uk/spm/`, 2009.

[56] E. Bagarinao, K. Matsuo, T. Nakai, and S. Sato, "Estimation of general linear model coefficients for real-time application," *NeuroImage*, vol. 19, no. 2, pp. 422 – 429, 2003.

[57] E. Bagarinao, T. Nakai, and Y. Tanaka, "Real-time functional MRI: Development and emerging applications," *Magnetic Resonance in Medical Science*, vol. 5, no. 3, pp. 157–165, 2006.

[58] A. Honkela and H. Valpola, "On-line variational bayesian learning," in *In Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003*, pp. 803–808, 2003.

[59] M. Chappell, A. Groves, B. Whitcher, and M. Woolrich, “Variational bayesian inference for a nonlinear forward model,” *Signal Processing, IEEE Transactions on*, vol. 57, pp. 223–236, Jan. 2009.

[60] E. Niedermeyer, F. D. Silva, Electroencephalography, Basic Principles, Clinical Applications, and Related Fields, Williams and Wilkins, Baltimore, 1999.

[61] L. Tarrasenko, Y. U. Khan, M. R. G. Holt, Identification of inter-ictal spikes in the EEG using neural networks analysis, *IEE Proc. - Sci. Meas. Technol.*, vol. 145, pp. 270–278, 1998.

[62] N. Acir, I. Oztura, M. Kuntalp, B. Baklan, C. Guzelis, Automatic Detection of Epilepticform Events in EEG by a Three-Stage Procedure Based on Artificial Neural Networks, *IEEE Transactions on Biomedical Engineering*, vol. 52, pp. 30–40, 2005.

[63] C. C. C. Pang, A. Upton, C. Shine, M. Kamath, A Comparison of Algorithms for Detection of Spikes in the Electroencephalogram, *IEEE Transactions on Biomedical Engineering* vol. 50, pp. 521–525, 2003.

[64] H. S. Park, Y. Lee, N. Kim, D. Lee, S. Kim, Detection of Epileptiform Activities in the EEG Using Neural Network and Expert System, *Med. Info.*, vol. 9, pp. 1255–1259, 1998.

[65] J. Gotman, P. Gloor, Automatic Recognition of inter-ictal epileptic activity in the human scalp EEG recordings, *Electroencephal Clin Neurophysiol*, vol. 41, pp. 513–529, 1976.

[66] J. Gotman, Automatic Recognition of epileptic seizures in the EEG, *Electroencephal Clin Neurophysiol*, vol. 54, pp. 530–540, 1982.

[67] P. Ktonas, J. Glover, L. Webster, R. Antonathanasap, W. V. Leeuwen, C. V. Veelen, W. Vliegenthart, Automatic detection of epileptogenic sharp EEG transients, *Electroencephal Clin Neurophysiol*, pp. 38 – 58,1984.

[68] T. Kalayci, O. Ozdamar, Wavelet preprocessing for automated Neural Network Detection of EEG spikes, *IEEE Eng. in Med. and Biology*, pp. 160–166, 1995.

[69] F. H. L. D. Silva, A. Dijk, H. Smits, Detection of nonstationarities in EEG’s using autoregressive model - An Application to EEG’s of epileptics, *CEAN - Computerized EEG analysis*, Stuttgart, Germany, 1975.

[70] A. J. Gabor, M. Seyal, Automated interictal EEG spike detection using artificial neural networks, *Electroencephal Clin Neurophysiol*, vol. 83, pp. 271–280, 1992.

[71] A. J. Gabor, Seizure detection using a self - organizing neural network: validation and comparison with other detection strategies, *Electroencephal Clin Neurophysiol*, vol. 107, pp. 27–32,1998.

[72] W. R. Webber, B. Litt, K. Wilson, R. P. Lesser, Practical detection of epileptiform discharges (EDs) in the EEG using an artificial neural network: a comparison of raw and parameterized EEG data, *Electroencephal Clin Neurophysiol*, vol. 91, pp. 194–204, 1994.

[73] O. Ozdamar, T. Kalayci, Detection of spikes with artificial neural networks using raw EEG, *Electroencephal Clin Neurophysiol*, vol. 31, pp. 122–142, 1998.

[74] C. J. James, R. D. Jones, P. J. B. abd G. J. Caroll, Detection of epileptiform discharges in the EEG by a hybrid system comprising mimetic, self - organized artificial neural network and fuzzy logic stages, *Electroencephal Clin Neurophysiol*, vol. 110, pp. 2049–2063, 1999.

[75] C. W. Ko, H. W. Chung, Automatic spike detection via artificial neural network using raw EEG data effects of data preparation and implications in the limitations of online recognition, *Clin Neurophysiol*, vol. 111, pp. 477–481, 2000.

[76] G. Hellmann, Multifold features determine linear equation for automatic spike detection applying neural network interictal ECoG, *Clin Neurophysiol*, vol. 110, pp. 887–894, 1999.

[77] R. Benlamri, M. Batouche, S. Rami, C. Bouanaka, An automated system for analysis and interpretation of epileptiform activity in the EEG, *Comput. Biol. Med.*, vol. 27, pp. 129–139, 1997.

[78] A. Dingle, R. Jones, G. Caroll, W. Fright, A multistage system to detect epileptiform activity in the EEG, *IEEE Transactions on Biomedical Engineering*, vol. 40, pp. 1260–1268, 1993.

[79] B. L. Davey, W. R. Fright, G. J. Caroll, R. D. Jones, Expert system approach to detection of epileptiform activity in the EEG, *Med Biol Eng Comput*, vol. 27, pp. 365–370, 1989.

[80] C. James, M. Hagan, R. Jones, P. Bones, G. Carroll, Multireference adaptive noise canceling applied to the EEG, *IEEE Transactions on Biomedical Engineering*, vol. 44, pp. 775–779, 1997.

[81] V. Parsa, P. Parker, Multireference adaptive noise cancellation applied to somatosensory evoked potentials, *IEEE Transactions on Biomedical Engineering*, vol. 41, pp. 792–800, 1994.

[82] P. Sadasivan, D. N. Dutt, ANC schemes for the enhancement of EEG signals in the presence of EOG artifacts, *Computers and Biomedical Research*, vol. 29, pp. 27–40, 1996.

[83] W. D. Penny, S. J. Roberts, Dymanic models for nonstationary signal segmentation, *Computer and Biomedical Research*, vol. 32, pp. 483–502, 1999.

[84] M. Arnold, W. H. R. Miltner, H. Witte, R. Bauer, C. Braun, Adaptive AR Modeling of Nonstationary Time Series by Means of Kalman Filtering, *IEEE Transactions on Biomedical Engineering*, vol. 45, pp. 553–562, 1998.

[85] H. H. Jasper, The Ten - Twenty Electrode System of the International Federation, *Clinical Neurophysiology*, vol. 10, pp. 371–375, 1958.

[86] L. Harrison, W. Penny, J. Daunizeau, and K. Friston, "Diffusion-based spatial priors for functional magnetic resonance images," *NeuroImage*, vol. 41, no. 2, pp. 408–423, 2008.

[87] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[88] X. Descombes, F. Kruggel, and von D.Y. Cramon, "fMRI signal restoration using a spatio-temporal Markov Random Field preserving transitions," *NeuroImage*, vol. 8, pp. 340–349, 1998.

[89] C. Gossl, D. P. Auer, and L. Fahrmeir, "Bayesian spatiotemporal inference in functional magnetic resonance imaging," *Biometrics*, vol. 57, pp. 554–562, 2001.

[90] N. Hartvig and J. Jensen, "Spatial mixture modeling of fMRI data," *Human Brain Mapping*, vol. 11, no. 4, 2000.

[91] Y. Li, P. Namburi, Z. Yu, C. Guan, J. Feng, and Z. Gu, "Voxel selection in fMRI data analysis based on sparse representation," *IEEE Transactions on Biomedical Engineering*, vol. 56, pp. 2439 –2451, Oct. 2009.

[92] A. Lukic, M. Wernick, D. Tzikas, X. Chen, A. Likas, N. Galatsanos, Y. Yang, F. Zhao, and S. Strother, "Bayesian kernel methods for analysis of functional neuroimages," *IEEE Trans. on Medical Imaging*, vol. 26, no. 12, pp. 1613–1622, 2007.

[93] M. Carroll, G. Cecchi, I. Rish, R. Garg, and A. R. Rao, "Prediction and interpretation of distributed neural activity with sparse models," *NeuroImage*, vol. 44, pp. 112–122, 2009.

[94] G. Flandin and W. Penny, “Bayesian fMRI data analysis with sparse spatial basis function priors,” *NeuroImage*, vol. 34, pp. 1108–1125, 2007.

[95] M. Smith and L. Fahrmeir, “Spatial bayesian variable selection with application to functional magnetic resonance imaging,” *Journal of the American Statistical Assosiation*, vol. 102, no. 478, pp. 417–431, 2007.

[96] M. A. van Gerven, B. Cseke, F. P. de Lange, and T. Heskes, “Efficient bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior,” *NeuroImage*, vol. 50, no. 1, pp. 150 – 161, 2010.

[97] J. Besag, “Spatial interaction and the statistical analysis of lattice systems (with discussion),” *Journal of Royal Statistical Society*, vol. 36, no. 2, pp. 192–326, 1975.

[98] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[99] A. Dempster, L. A., and R. D., “Maximum likelihood from incomplete data via the em algorithm,” *Journal of Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[100] S. Z. Li, *Markov Random Field Modeling in Image Analysis.* Springer Publishing Company, Incorporated, 2009.

[101] P. J. Green, “Bayesian Reconstructions from Emission Tomography Data Using a Modified EM Algorithm,” *IEEE Trans. on Medical Imaging*, vol. 9, no. 1, pp. 84–93, 1990.

[102] Y. Zhang, M. Brady, and S. Smith, “Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm,” *IEEE Trans. on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[103] K. Blekas, A. Likas, N. P. Galatsanos, and I. E. Lagaris, “A Spatially-Constrained Mixture Model for Image Segmentation,” *IEEE Trans. on Neural Networks* , vol. 16, pp. 494–498, 2005.

[104] D. Geman and C. Yang, “Nonlinear image recovery with half-quadratic regularization ,” *IEEE Trans. on Image Processing* , vol. 4, pp. 932–946, 1995.

[105] The FIL Methods Group, *SPM8 Manual.* Functional Imaging Laboratory Wellcome Trust Centre for Neuroimaging Institute of Neurology, 2009.

[106] E. Vlieger, C. Majoie, S. Leenstra2, and G. den Heeten, “Functional magnetic resonance imaging for neurosurgical planning in neurooncology,” *European Radiology*, vol. 17, no. 7, pp. 1143–1153, 2004.

[107] M. Seeger, “Bayesian Inference and Optimal Design for the Sparse Linear Model,” *Journal of Machine Learning Research*, vol. 9, pp. 759–813, 2008.

[108] M. Lindquist, “The Statistical Analysis of fMRI Data,” *Statistical Science*, vol. 23, pp. 439–464, no 4, 2008.

[109] N. Logothetis, J. Pauls, M. Augath, T. Trinath and A. Oeltermann, “Neurophysiological investigation of the basis of the fMRI signal,” *Nature*, vol. 412, pp. 150–157, July 2001.

[110] W. J. Tompkins, *Biomedical digital signal processing : C-language examples and laboratory experiments for the IBM PC.* Englewood Cliffs, N.J.: Prentice Hall, 1993.

[111] M. Akay, *Biomedical signal processing.* San Diego: Academic Press, 1994.

[112] V. P. Oikonomou, A. T. Tzallas, and D. I. Fotiadis, “A kalman filter based methodology for EEG spike enhancement,” *Computer Methods and Programs in Biomedicine*, vol. 85, no. 2, pp. 101–108, 2007.

[113] S. V. Vaseghi, *Advanced digital signal processing and noise reduction.* Chichester, West Sussex, England ; Hoboken: Wiley, 3rd ed., 2006.

[114] A. P. Bradley and W. J. Wilson, “On wavelet analysis of auditory evoked potentials,” *Clinical Neurophysiology*, vol. 115, no. 5, pp. 1114–1128, 2004.

[115] Z. S. Wang, A. Maier, D. A. Leopold, N. K. Logothetis, and H. L. Liang, “Single-trial evoked potential estimation using wavelets,” *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 463–473, 2007.

[116] P. A. Karjalainen, J. P. Kaipio, A. S. Koistinen, and M. Vauhkonen, “Subspace regularization method for the single-trial estimation of evoked potentials,” *IEEE Transactions on Biomedical Engineering*, vol. 46, no. 7, pp. 849–860, 1999.

[117] G. Sparacino, S. Milani, E. Arslan, and C. Cobelli, “A bayesian approach to estimate evoked potentials,” *Computer Methods and Programs in Biomedicine*, vol. 68, no. 3, pp. 233–248, 2002.

[118] T. P. Jung, S. Makeig, M. J. McKeown, A. J. Bell, T. W. Lee, and T. J. Sejnowski, “Imaging brain dynamics using independent component analysis,” *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1107–1122, 2001.

[119] T.-P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne and T.J. Sejnowski. “Analysis and Visualization of Single-Trial Event-Related Potentials,” *Human Brain Mapping*, vol. 14,pp. 166–185, 2001.

[120] J. Kayser and C. E. Tenke, "Optimizing PCA methodology for ERP component identification and measurement: theoretical rationale and empirical evaluation," *Clinical Neurophysiology*, vol. 114, no. 12, pp. 2307–2325, 2003.

[121] C. P. Robert, *The Bayesian choice : from decision-theoretic foundations to computational implementation.* Springer texts in statistics, New York: Springer, 2nd ed., 2001.

[122] M. P. Tarvainen, P. O. Ranta-aho, and P. A. Karjalainen, "An advanced detrending method with application to hrv analysis," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 2, pp. 172–175, 2002.

[123] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems.* New York: Springer, 2005.

[124] J. O. Ramsay and B. W. Silverman, *Functional data analysis.* Springer series in statistics, New York ; Berlin: Springer, 2nd ed., 2005.

[125] G. E. P. Box and G. C. Tiao, *Bayesian inference in statistical analysis.* New York: Wiley, Wiley classics library ed., 1992.

[126] P. M. Williams, "Bayesian regularization and pruning using a laplace prior," *Neural Computation*, vol. 7, no. 1, pp. 117–143, 1995.

[127] H. Uhlig, "On singular wishart and singular multivariate beta-distributions," *Annals of Statistics*, vol. 22, no. 1, pp. 395–405, 1994.

[128] J. A. Diaz-Garcia, R. G. Jaimez, and K. V. Mardia, "Wishart and pseudo-wishart distributions and some applications to shape theory," *Journal of Multivariate Analysis*, vol. 63, no. 1, pp. 73–87, 1997.

[129] T. K. Moon and W. C. Stirling, *Mathematical methods and algorithms for signal processing.* Upper Saddle River, NJ: Prentice Hall, 2000.

[130] Fatourechi, "A wavelet-based approach for the extraction of event related potentials from EEG," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 2, pp. ii–737–40 vol.2, 2004.

[131] R. Q. Quiroga and H. Garcia, "Single-trial event-related potentials with wavelet denoising," *Clinical Neurophysiology*, vol. 114, no. 2, pp. 376–390, 2003.

[132] A. Delorme, S. Makeig, M. Fabre-Thorpe, and T. Sejnowski, "From single-trial EEG to brain area dynamics," *Neurocomputing*, vol. 44, pp. 1057–1064, 2002.

[133] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet - components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–E220, 2000.

[134] R. A. Thuraisingham, "Preprocessing RR interval time series for heart rate variability analysis and estimates of standard deviation of RR intervals," *Computer Methods and Programs in Biomedicine*, vol. 83, no. 1, pp. 78–82, 2006.

[135] A. J. Camm, M. Malik, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, R. E. Kleiger, F. Lombardi, A. Malliani, A. J. Moss, J. N. Rottman, G. Schmidt, P. J. Schwartz, and D. Singer, "Heart rate variability - standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[136] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining.* Boston: Pearson Addison Wesley, 2006.

[137] W. D. Penny, J. Kilner, and F. Blankenburg, "Robust bayesian general linear models," *Neuroimage*, vol. 36, no. 3, pp. 661–671, 2007.

[138] M. E. Tipping and N. D. Lawrence, "Variational inference for student-t models: Robust bayesian interpolation and generalised component analysis," *Neurocomputing*, vol. 69, no. 1-3, pp. 123–141, 2005.

[139] G. K. Chantas, N. P. Galatsanos, and A. C. Likas, "Bayesian restoration using a new nonstationary edge-preserving image prior," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 2987–2997, 2006.

[140] T. Budinger and H.F VanBrocklin. Positron-emission tomography (PET). In *The Biomedical Engineering Handbook, Second Edition. 2 Volume Set*, Electrical Engineering Handbook. CRC Press, 1999.

[141] A. Cohen. Biomedical signals. In *The Biomedical Engineering Handbook, Second Edition. 2 Volume Set*, Electrical Engineering Handbook. CRC Press, 1999.

[142] I. Cunningham and P. Judy. Computed tomography. In *The Biomedical Engineering Handbook, Second Edition. 2 Volume Set*, Electrical Engineering Handbook. CRC Press, 1999.

[143] J. Pauly, A. Macovski, Steven S. Conolly, and J. Schenck. Magnetic resonance imaging. In *The Biomedical Engineering Handbook, Second Edition. 2 Volume Set*, Electrical Engineering Handbook. CRC Press, 1999.

[144] N. Thakor, B. Gramatikov, and D. Sherman. Wavelet (time-scale) analysis in biomedical signal processing. In *The Biomedical Engineering Handbook, Second Edition. 2 Volume Set*, Electrical Engineering Handbook. CRC Press, 1999.

[145] S.L. Bressler and M. Ding. Event-Related Potentials. In *Wiley Encyclopedia of Biomedical Engineering, 6-Volume Set*, John Wiley and Sons, Inc. 2006.

[146] D. Iyer and G. Zouridakis. Single-trial evoked potential estimation: Comparison between independent component analysis and wavelet denoising. *Clinical Neurophysiology*, vol. 118, no. 3, pp. 495–504, 2007.

[147] A. Hyvarinen and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, vol. 13, pp. 411–430, 2000.

[148] J.W. Miskin and D.J.C. MacKay. Ensemble learning for blind source separation. *Independent Component Analysis: Principles and Practice*, Cambridge University Press, 2000.

[149] R.A. Choudrey and S.J. Roberts. Variational mixture of Bayesian independent component analyzers. *Neural Computation*, vol. 15, pp. 213–252, 2003.

[150] H. Attias. Independent Factor Analysis. *Neural Computation*, vol. 11, pp. 803–851, 1999.

[151] S. Mallat, *A Wavelet Tour of Signal Processing, 3rd ed.* Academic Press, 2008.

[152] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[153] K.H. Choo, J.C. Tong, L. Zhang, “Recent applications of Hidden Markov Models in computational biology”. *Genomics Proteomics Bioinformatics*, vol. 2, pp. 84–96,2004.

[154] D. Husmeier, “Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models”. *Bioinformatics*, vol. 21, pp. 166-172,2005.

[155] L.R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition”, *Proc. IEEE*,vol. 77, pp. 257–286, 1989.

[156] O. Cappe, E. Moulines and T. Ryden, *Inference in Hidden Markov Models.* Springer Science+Business Media, 2005.

[157] D.P. Wipf and B.D. Rao. Sparse bayesian learning for basis selection. *IEEE Transactions on Signal Processing*, vol. 52, pp. 2153–2164, August 2004.

[158] G. Deng, "Iterative Learning Algorithms for Linear Gaussian Observation Models", *IEEE Transactions on Signal Processing*, vol. 52, pp. 2286–2297, August 2004.

[159] S. Kay, *Modern Spectral Estimation: Theory and Application.* Englewood Cliffs, NJ: Prentice Hall, 1988.

[160] R. Kalman, "A new approach to linear filtering and prediction problems", *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[161] H. Rauch "Solutions to the linear smoothing problem", *IEEE Transactions on Automatic Control*, vol. 8, pp. 371–372, 1963.

[162] B. Anderson and J. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, N.J., 1979.

[163] R. Brown, *Introduction to random signal analysis and Kalman Filtering*, John Wiley and Sons Inc., 1983.

[164] M. Grewal and A. Andrews. *Kalman Filtering: Theory and Practice using Matlab, Second Edition.* John Wiley and Sons Inc., 2001.

[165] S. Haykin. *Kalman Filtering and Neural Networks.* John Wiley and Sons Inc., 2001.

[166] M. Tarvainen, J. Hiltunen, P. Ranta-aho, and P. Karjalainen, "Estimation of nonstationary EEG with kalman smoother approach: An application to event related synchronization (ERS)", *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 516–524, March 2004.

[167] O. Sayadi and M. Shamsollahi,"ECG denoising and compression using a modified extended kalman filter structure", *Signal Processing*, vol. 88, pp. 2114–2121, 2008.

[168] M. Khan and D. Dutt. An expectation maximization algorithm based kalman smoother approach for event - related desynchronization (ERD) estimation from EEG. *IEEE Transactions on Biomedical Engineering*, 54(7):1191 – 1198, July 2007.

[169] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm", *Journal of Time Series Analysis*, vol. 3, pp. 253–264, 1982.

[170] V. Oikonomou, A. Tzallas, and D. Fotiadis, "A kalman filter based methodology for EEG spike enhancement", *Computer Methods and Programs in Biomedicine* , vol. 85, pp. 101–108, 2007.

[171] R. Kazemi, A. Farsi, M. Ghaed, and M. K. Ghartemani, “Detection and extraction of periodic noises in audio and biomedical signals using kalman filter”, *Signal Processing*, vol. 88, pp. 2114 – 2121, 2008.

[172] S. Georgiadis, P. Ranta-aho, M. Tarvainen, and P. Karjalainen, “Single - trial dynamical estimation of event - related potentials: A kalman filter-based approach”, *IEEE Transactions on Biomedical Engineering*, vol. 52, pp. 1397– 1406, 2005.

[173] M. Aboy, O. Marquez, J. McNames, R. Hornero, T. Trong, and B. Goldstein, “Adaptive modeling and spectral estimation of nonstationary biomedical signals based on kalman filtering”, *IEEE Transactions on Biomedical Engineering*, vol. 52, pp. 1485–1489, August 2005.

[174] F. La Foresta, N. Mammone and F.C. Morabito, “PCA-ICA for automatic identification of critical events in continuous coma-EEG monitoring”, *Biomedical Signal Processing and Control*,vol. 4, pp. 229 – 235, 2009.

[175] A.R. Webb. *Statistical Pattern Recognition, 2nd Edition.* John Wiley & Sons, October 2002.

[176] S.J. Luck, *An Introduction to the Event-Related Potential Technique.* MIT Press, 2005.

[177] W.R. Gilks, S. Richardson and D.J. Spiegelhalter, *Markov chain Monte Carlo in practice.* Chapman & Hall/CRC, 1996.

[178] J.J.K. O Ruanaidh and W.J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing* . Springer, New York, 1996.

[179] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning.* MIT Press, 2006.

[180] G. Flandin and W. Penny, “Bayesian fMRI data analysis with sparse spatial basis function priors,” *NeuroImage*, vol. 34, pp. 1108–1125, 2007.

[181] K. J. Worsley, C. H. Liao, J. Aston, V. Petre, G. H. Duncan, F. Morales, and A. C. Evans, “A general statistical analysis for fMRI data,” *NeuroImage*, vol. 15, no. 1, pp. 1 – 15, 2002.

[182] R. Baumgartner, C. Windischberger, and E. Moser, “Quantification in functional magnetic resonance imaging: Fuzzy clustering vs. correlation analysis,” *Magnetic Resonance Imaging*, vol. 16, no. 2, pp. 115 – 125, 1998.

[183] O. Friman, J. Cedefamn, P. Lundberg, M. Borga, and H. Knutsson, "Detection of neural activity in functional MRI using canonical correlation analysis," *Magnetic Resonance in Medicine*, vol. 45, pp. 323ὐ–330, 2001.

[184] C. Goutte, P. Toft, E. Rostrup, F. . Nielsen, and L. K. Hansen, "On clustering fMRI time series," *NeuroImage*, vol. 9, no. 3, pp. 298 – 310, 1999.

[185] A. Wismüller, O. Lange, D. R. Dersch, G. L. Leinsinger, K. Hahn, B. Pütz, and D. Auer, "Cluster analysis of biomedical image time-series," *Int. J. Comput. Vision*, vol. 46, no. 2, pp. 103–128, 2002.

[186] F. G. Meyer and J. Chinrungrueng, "Spatiotemporal clustering of fMRI time series in the spectral domain," *Medical Image Analysis*, vol. 9, no. 1, pp. 51 – 68, 2005.

[187] A. Mezer, Y. Yovel, O. Pasternak, T. Gorfine, and Y. Assaf, "Cluster analysis of resting-state fMRI time series," *NeuroImage*, vol. 45, no. 4, pp. 1117 – 1125, 2009.

[188] C. Windischberger, M. Barth, C. Lamm, L. Schroeder, H. Bauer, R. C. Gur, and E. Moser, "Fuzzy cluster analysis of high-field functional MRI data," *Artificial Intelligence in Medicine*, vol. 29, no. 3, pp. 203 – 223, 2003.

[189] A. Meyer-Base, A. Saalbach, O. Lange, and A. Wismóller, "Unsupervised clustering of fMRI and MRI time series," *Biomedical Signal Processing and Control*, vol. 2, no. 4, pp. 295 – 310, 2007.

[190] A. Wismuller, A. Meyer-Base, O. Lange, D. Auer, M. F. Reiser, and D. Sumners, "Model-free functional MRI analysis based on unsupervised clustering," *Journal of Biomedical Informatics*, vol. 37, no. 1, pp. 10 – 18, 2004.

[191] L. He and I. R. Greenshields, "An mrf spatial fuzzy clustering method for fMRI SPMs," *Biomedical Signal Processing and Control*, vol. 3, no. 4, pp. 327 – 333, 2008.

[192] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264 – 323, 1999.

[193] T. W. Liao, "Clustering of time series data–a survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857 – 1874, 2005.

[194] P. Smyth in *Advances in Neural Information Processing Systems*, pp. 648ὐ–654, 1997.

[195] S. J. Gaffney and P. Smyth, "Curve clustering with random effects regression mixtures," in *Proc. of the Ninth Intern. Workshop on Artificial Intelligence and Statistics (C. M. Bishop and B. J. Frey, eds.)*, 2003.

[196] D. Chudova, S. Gaffney, E. Mjolsness, and P. Smyth, "Mixture models for translation-invariant clustering of sets of multi-dimensional curves," in *Proc. of the Ninth ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*, pp. 79ὐ–88, 2003.

[197] Y. Xiong and D. Y. Yeung, "Mixtures of arma models for model-based time series clustering," in *IEEE International Conference on Data Mining (ICDM)*, pp. 717ὐ–720, 2002.

[198] Y. Xiong and D. Y. Yeung, "Time series clustering with arma mixtures," *Pattern Recognition*, vol. 37, no. 8, pp. 1675 – 1689, 2004.

[199] J. Shi and B. Wang, "Curve prediction and clustering with mixtures of gaussian process functional regression models," *Statistics and Computing*, vol. 18, pp. 267ὐ–283, 2008.

[200] K. Blekas, C. Nikou, N. P. Galatsanos, and N. V. Tsekos, "A regression mixture model with spatial constraints for clustering spatiotemporal data," *International Journal*, vol. 17, pp. 1023–1041, 2008.

[201] K. Blekas, A. Likas, N. P. Galatsanos, and I. E. Lagaris, "A Spatially-Constrained Mixture Model for Image Segmentation," *IEEE Trans. on Neural Networks* , vol. 16, pp. 494–498, 2005.

[202] W. Penny and K. Friston, "Mixtures of general linear models for functional neuroimaging," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 504 –514, April 2003.

[203] J. Xia, F. Liang, and Y. M. Wang, "On clustering fMRI using potts and mixture regression models," in *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4795–4798, 2009.

[204] M. Girolami and S. Rogers, "Hierarchic bayesian models for kernel learning," in *ICML '05: Proceedings of the 22nd international conference on Machine learning*, (New York, NY, USA), pp. 241–248, ACM, 2005.

[205] A. Dempster, L. A., and R. D., "Maximum likelihood from incomplete data via the em algorithm," *Journal of Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[206] S. Gunn and J. Kandola, "Structural modelling with sparse kernels," *Machine Learning*, vol. 48, pp. 137–163, 2002.

[207] J. Li and A. Barron, "Mixture density estimation," in *Advances in Neural Information Processing Systems*, vol. 12, pp. 279–285, The MIT Press, 2000.

[208] K. Lange, R. Little and J. Taylor, "Robust statistical modeling using the t distribution," *Journal of American Statistical Association*, vol. 84, pp. 881ὐ-896, 1989.

[209] N. Ueda, R. Nakano, Z. Ghahramani, and G. Hinton, "SMEM algorithm for mixture models," *Neural Computation*, vol. 12, no. 9, pp. 2109–2128, 2000.

[210] N. Vlassis and A. Likas, "A greedy EM algorithm for Gaussian mixture learning," *Neural Processing Letters*, vol. 15, pp. 77–87, 2001.

[211] D. Tzikas, A. Likas, and N. Galatsanos, "Sparse bayesian modeling with adaptive kernel learning," *IEEE Transactions on Neural Networks*, vol. 20, pp. 926–937,2009

[212] N. Metropolis A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, "Equations of State Calculations by Fast Computing Machines,"*Journal of Chemical Physics*, vol. 21, pp. 1087ὐ-1092, 1953

[213] V. Oikonomou and K. Blekas, "A sparse spatial linear regression model for fMRI data analysis," in *SETN 2010*, pp. 203–213, 2010.

[214] S. Mallat, *A wavelet tour of signal processing: The Sparse Way.* Academic Press, 2009.

[215] J. Li and Z.J. Wang and S.J. Palmer and M.J. McKeown, "Dynamic Bayesian network modeling of fMRI: A comparison of group-analysis methods," *NeuroImage*, vol. 41, pp. 398–407, 2008.

[216] J.C. Rajapakse and J. Zhou, "Learning effective brain connectivity with dynamic Bayesian networks," *NeuroImage*, vol. 37, pp. 749–760, 2007.

Author's Publications

Journal Publications

- V.P. Oikonomou, A.T. Tzallas and D.I. Fotiadis, A Kalman Filter based Methodology for EEG Spike Enhancement, Computer Methods and Programs in Biomedicine, 85 (2007), pp. 101-108.

- V.P. Oikonomou, E.E. Tripoliti and D.I. Fotiadis, Bayesian Methods for fMRI Time-Series Analysis Using a Nonstationary Model for the Noise, accepted in IEEE Transactions on Information Technology in Biomedicine.

- V.P. Oikonomou and D.I. Fotiadis, Biomedical Signal Denoising using the Variational Bayesian Approach with Applications to ERP Estimation and HRV Analysis, submitted to Computer Methods and Programs in Biomedicine (under revision)

- V.P. Oikonomou, K. Blekas and L. Astrakas, A sparse and spatially constrained generative regression model for fMRI data analysis, submitted to IEEE Transactions on Biomedical Engineering

- V.P. Oikonomou and K. Blekas, Clustering fMRI time-series by using a mixture of regression models with spatial and sparse properties, to be prepared for submission

Conference Publications

- V.P. Oikonomou, M.G. Tsipouras and D.I. Fotiadis, Knowledge-Based Systems for Arrhythmia Detection and Classification, in Proceedings of the 6th International Workshop on Scattering Theory and Biomedical Technology 18-21 September 2003, Tsepelovo, Greece.

- M.G. Tsipouras, V.P. Oikonomou, D.I. Fotiadis, L.K. Michalis, D. Sideris, Classification of Atrial Tachyarrhythmias in Electrocardiograms Using Time Frequency Analysis, in Proceedings of Computer in Cardiology 19-22 September 2004, Chicago, USA

- V.P. Oikonomou and D.I. Fotiadis, A Bayesian PCA Approach to Fetal ECG extraction, in Proceedings of 3rd European Medical and Biological Engineering Conference, Prague, Czech, 2005.

- A.T. Tzallas, V.P. Oikonomou and D.I. Fotiadis, Epileptic Spike Detection Using a Kalman Filter based Approach, Proceedings of the 28th EMBS Annual International Conference, New York City, USA, Aug. 30-Sep. 3, 2006.
- V.P. Oikonomou and D.I. Fotiadis, A Bayesian Approach for Biomedical Signal Denoising, Proceedings of the 5th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB) , Ioannina, Greece, 26-28 October 2006
- V.P. Oikonomou and D. I. Fotiadis, A Bayesian Approach for the Estimation of AR Coefficients from Noisy Biomedical Data, Proceedings of the 29th EMBS Annual International Conference, Lyon, France, Aug. 23-26, 2007.
- V.P. Oikonomou, E.E. Tripoliti and D.I. Fotiadis, A sparse variational bayesian approach for fMRI data analysis, 8th International Conference on Bioinformatics and Bioengineering (BIBE 2008), Athens, Greece, October 8-10, 2008.
- V.P. Oikonomou, E.E. Tripoliti and D.I. Fotiadis, A sparse linear model for the analysis of fMRI data with non stationary noise, 4th International IEEE EMBS Conference on Neural Engineering, 2009.
- V.P. Oikonomou and D.I. Fotiadis, A Template - based Approach for the estimation of Event Related Potentials using the Bayesian Linear Model, 16th International Conference on Digital Signal Processing (DSP 2009).
- V.P. Oikonomou, E.E. Tripoliti and D.I. Fotiadis, A Bayesian Spatio - Temporal Approach for the Analysis of fMRI Data with Non - Stationary Noise, 31st Annual International IEEE EMBS Conference, September, 2-6, 2009, Hilton Minneapolis, Minnesota, USA
- V.P. Oikonomou and D.I. Fotiadis, A Detrending Technique for HRV signals based on the bayesian framework and a non ủ stationary model for the non - trend component, 9th International Workshop on Mathematical Methods in Scattering Theory and Biomedical Engineering, 9-11 October 2009, Patras, Greece
- V.P. Oikonomou, K. Blekas, A sparse linear regression model for fMRI data analysis, 6th Hellenic Conference on Artificial Intelligence (SETN 2010), Athens, Greece

Chapter book:

- V.P. Oikonomou, A.T. Tzallas, S. Konitsiotis, D.G. Tsalikakis and D.I. Fotiadis, The use of Kalman Filter in Biomedical Signal Processing, Chapter book in: "Kalman Filter:Recent Advances and Applications".

Short Vita

Vangelis P. Oikonomou was born in Karditsa, Greece, in 1976. He received the diploma degree and the M.Sc. in computer science from the University of Ioannina, Greece, in 2000 and 2003, respectively. He is currently working toward the Ph.D. degree in the Dept. of Computer Science at the University of Ioannina. His research interests include bayesian methods and statistical signal processing and its applications on biomedical signals.