Mapping multivariate measures of brain response onto stimulus information during emotional face classification

Oliver G. B. Garrod

Department of Psychology
University of Glasgow

Presented for the degree of
Doctor of Philosophy
at the University of Glasgow
September 2009

Contents

List of Figures

Abstract

The relationship between feature processing and visual classification in the brain has been explored through a combination of reverse correlation methods (i.e."Bubbles" [22]) and electrophysiological measurements (EEG) taken during a facial emotion categorization task [63]. However, in the absence of any specific model of the brain response measurements, this and other [60] attempts to parametrically relate stimulus properties to measurements of brain activation are difficult to interpret. In this thesis I consider a blind data–driven model of brain response. Statistically independent model parameters are found to minimize the expectation of an objective likelihood function over time [55], and a novel combination of methods is proposed for separating the signal from the noise. The model's estimated signal parameters are then objectively rated by their ability to explain the subject's performance during a facial emotion classification task, and also by their ability to explain the stimulus features, as revealed in a Bubbles experiment.

Chapter 1

General Introduction

Interest in parametric models of the brain Electroencephalograph (EEG) in terms of some properties of input stimuli has, in recent years, grown over the standard hypothesis–driven paradigm in cognitive neuroscience [62][63][67][60]. The benefits of this approach are that the results are not biased by any fixed pre–conceived hypotheses, and that subtle distinctions between the influence of various stimulus properties can be observed from only a single experiment. Often, however, these approaches have failed to pay sufficient focus to the underlying model hidden within the EEG measurements themselves [60][63][67]. The parametric manipulation approach to empirical research is common in fields where the properties of the brain responses (e.g. single–cell recordings) are well understood, but are these methods still appropriate when the processes underlying the brain responses are not so well understood, such as the EEG response? Here it is proposed that, to be valid, these methods do not require true knowledge of the processes underlying the brain response, as long as some alternative model is provided where the parameters of the measured data are statistically independent — i.e a 'Blind Source Model' [17].

This approach to modelling the EEG data is tested on data generated from a 'Bubbles' experiment where no explicit model of the EEG responses was provided [63]. The purpose of the experiment was to explore the timing of facial feature processing in the brain during the classification of emotional expressions. The facial features themselves were not controlled parametrically, but rather their *availability* in the stimulus was controlled by the smooth random masking of locations in the face. The rest of this chapter provides a brief framework for framing visual classification problems, followed by a brief overview of the distinction between stimulus contrast–based experimental designs and parametrically–based experimental designs; and finally an overview of the cognitive neuroscience of facial emotion processing. The

following chapter then specifies and tests certain assumptions of the blind source model, before presenting the results of its application to the EEG data from [63]; and the final chapter relates both the classification task performance and the input stimulus to each of the independent parameters of the EEG model.

Visual Classification

The primary role of the brain during visual classification is clearly to assign to a particular visual input a particular category of the world, taken from a finite set of categories. However, to practically perform this role, it must first be able to encode the visual input, encode the set of possible categories, and, where appropriate, to plan and execute the behaviour associated with this category. Figure 1.1 illustrates this arrangement: a visual input, drawn from a category of the world, is encoded by the brain; this visual code is assigned by some inferential process to a category code; and this category code is transformed into a behaviour code. In an explicit visual classification task, perfect classification performance corresponds to maximum agreement between the executed behaviour and the behaviour expected by the task designer, while degrees of performance are proportional to degrees of agreement. Assuming that the brain makes no errors either in encoding the set of categories or in translating from its category assignment to the appropriate behaviour, and that the behaviour expected by the task designer is associated with the true category of the world, performance can be considered as the degree of agreement between the category assigned by the brain and the true category of the world.

The focus in this thesis will then be on the other two roles that the brain must perform — namely encoding the visual input and inferring from this representation the likeliest category of the world. More precisely, it will be about *when* these events might take place in the brain, and *what* features of a given stimulus does it require to do them. Although visual encoding is not itself part of the formal classification process, a visual code that extracts features from the input that vary most across categories will clearly enhance classification performance, so the visual encoding process is unlikely to be completely separated from the category encoding process in a functional sense (i.e. knowledge of the category set should optimise visual encoding towards that set). Previous EEG research has associated the visual encoding stage with a negative ERP effect at $170ms$ following stimulus onset [25][26] and the category selection stage with a positive ERP effect at $300ms$ following stimulus onset [25][26][43].

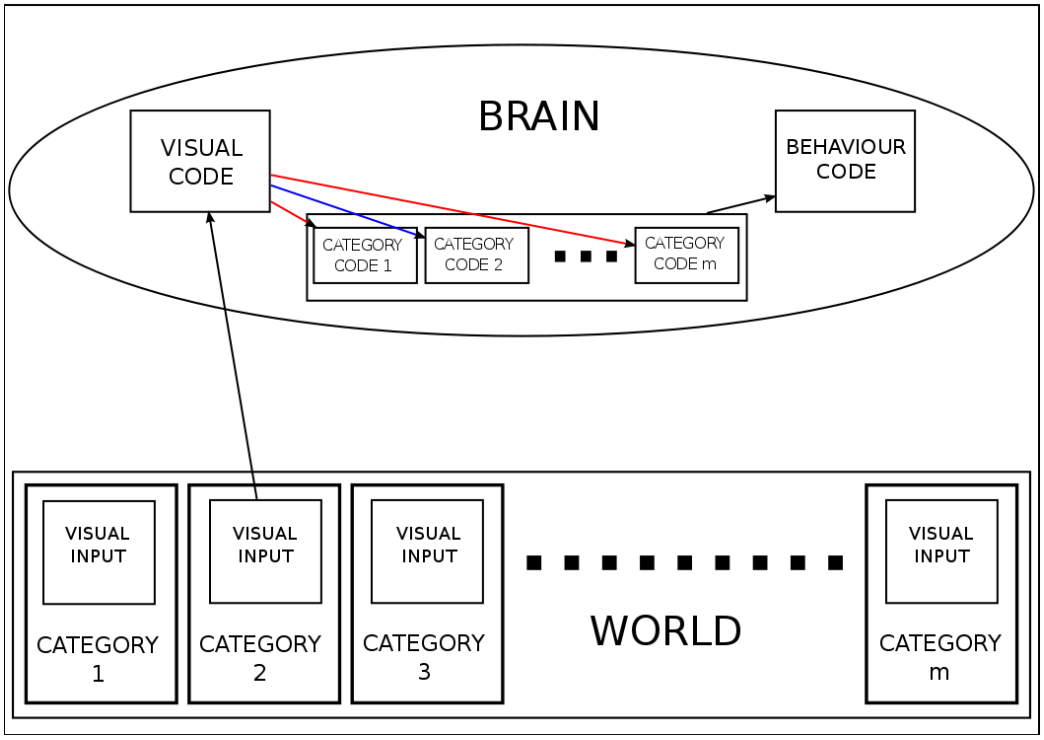


Figure 1.1: Illustration of the visual classification process. Black arrows represent directed causal relations; Red arrows represent incorrect inferences; Blue arrows represent correct inferences. Correct classification corresponds to agreement between the category of the world leading to visual input and the category code inferred by the brain.

Experimental design: Contrast Stimuli vs. Parametric Modulation

The history of empirical cognitive science has been shaped, in the most part, by the classic experimental paradigm (expressed in [59], for example). Within this paradigm, a certain model of brain function is put forth, together with corresponding inputs and outputs, and predictions are made on the output of this model given certain specific inputs. Such a model is said to be "falsifiable" if one can conceive of an input for which a given output would be inconsistent with the predictions of the model. While this is, and should be, seen as a necessary requirement for any scientific model to be taken seriously, it cannot be seen as a judgement of the model's quality. Many a trivial model would pass the test of falsifiability. The test of a model's quality is rather the practical test of how useful it is in explaining what we, or others, want to know. This usefulness may be in terms of its accuracy of prediction, or in terms of its wealth of prediction, or in terms of both. Either way, the process of model formation is key to the production of useful empirical predictions. How this process is performed is no trivial matter.

Let us consider a particular approach to model formation in empirical research — one where the empirical phase is designed, not to test a specific prediction of a previously fixed model, but rather to estimate the parameters of some general model form. This approach is common in psychophysical research (see e.g. [53]), where the phenomenon of interest is generally at quite a low level in the processing stream. For these kinds of low–level processes one can expect there to be less unknown variables influencing the measured outcome, so it is quite possible to reliably fit a complex non–linear model to the empirical data. The approach is less common, however, in higher-level cognitive research, where less is known about the relationship between stimulus input and measured output.

Linear Template Model

The linear template is a simple model of visual encoding for perceptual classification that has its roots in behavioural psychophysics (e.g. [13]). It relies upon the definition of a given set of outputs — let us denote this by S — and a given input space — let us denote this by the vector–valued variable $\mathbf{x}$. The linear template model then assumes that the behavioural output chosen by the observer when faced with a particular input is some function of the inner product of this input, $\mathbf{x}_i$, and an internal template, $\mathbf{t}$, under internal encoding noise, $\boldsymbol{\epsilon_c}$, and internal output noise, $\boldsymbol{\epsilon_b}$:

$$s = f[(\mathbf{x}_i + \boldsymbol{\epsilon_c})^{\mathbf{t}}\mathbf{t} + \boldsymbol{\epsilon_b}], s \in S, \tag{1.1}$$

where:

$$f(a) = \begin{cases} +1 & a \geq c \\ -1 & a < c \end{cases}$$

in the case of a two–category classification. Figure 1.2 displays this graphically for a given template in a two dimensional input space.

The empirical problem is to find an estimate of the observer's template based on a set of known inputs and observed behavioural outputs. One way to solve this problem is with the method of 'reverse correlation' [3][40][47]. If we have two inputs that are known to produce alternate responses, such as the two inputs in figure 1.2, then we know also that a dividing line in input space, defined by the unknown template, exists somewhere between these two points. Treating these points as the centres of two Gaussian distributions, we know that some covariance parameters exist such that a fixed proportion of randomly drawn samples from each distribution are expected to be 'mis–classified' by the observer as different from the original sample. Figure 1.3 displays this graphically for the same two points as in figure 1.2 and a given proportion of mis–classified points. Provided both the proportion of mis–classified points and the total number of points is sufficiently large, and the covariance matrix for each distribution is scalar (i.e. equal for all dimensions of input space) the template can be estimated empirically by:

$$\hat{\mathbf{t}} = k\left[(\bar{\mathbf{x}}_{ij} + \bar{\mathbf{x}}_{ii}) - (\bar{\mathbf{x}}_{ji} + \bar{\mathbf{x}}_{jj})\right], \tag{1.2}$$

where $\bar{\mathbf{x}}_{ij}$ is the average, mean–centred value of points generated by distribution i that are classed as distribution j and k is some scalar.

This particular situation can be created by taking a pair of such contrasting input stimuli and corrupting their feature space by additive zero–mean Gaussian noise. If enough noise samples are created, and the variance of the noise is set to allow a sufficient number of mis–classifications, then the experimenter must simply note the response of the observer under each noise sample and equation 1.2 can be applied to the known noise distribution. The result of such an estimate is known as the 'Classification Image'.

Bubbles

The experimental data that this thesis is based upon come from a related paradigm to reverse correlation, named "Bubbles". The original Bubbles experiment [22] consisted of one of two tasks for the observer — judging the expressiveness of a face (expressive or not expressive — 'EXNEX') or judging the gender of a face (male or female — 'GENDER'). The stimulus on any trial consisted of a masked photograph of a face displaying one of the

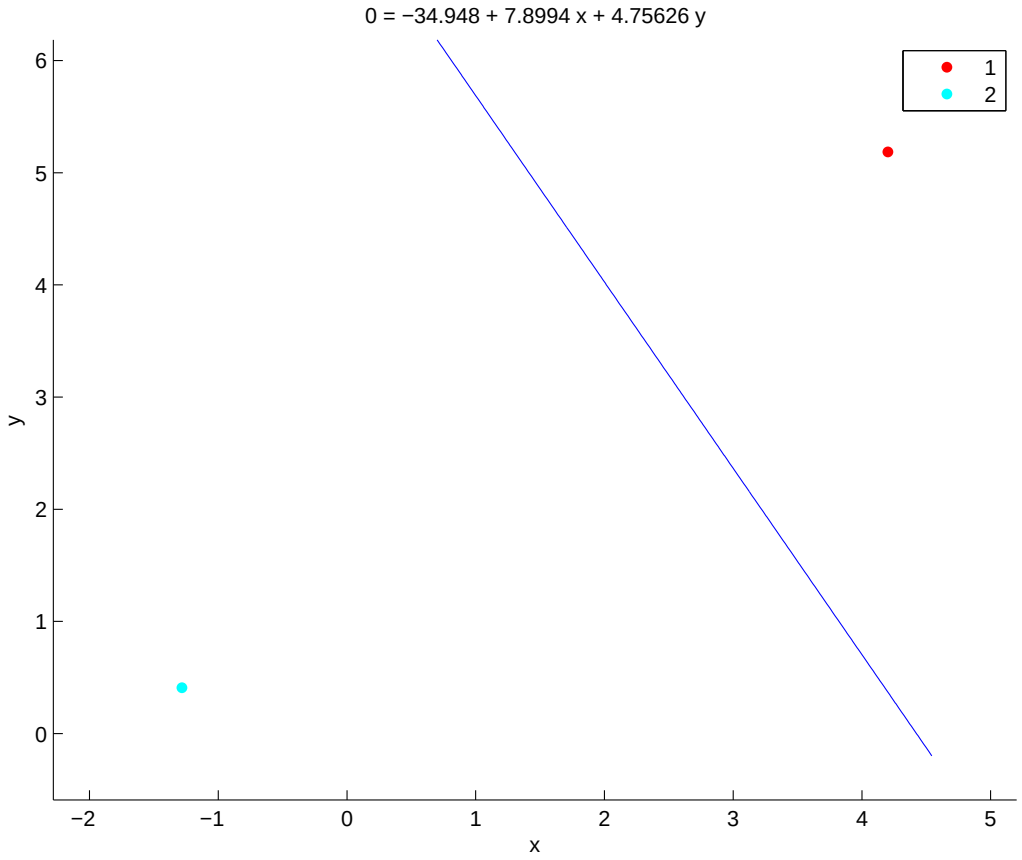


Figure 1.2: Illustration of two–category classification problem in two dimensions. X–axis — dimension 1; Y–axis — dimension 2. Two inputs, each from a different category, are displayed as the red and blue points. The blue line indicates the boundary in two–dimensional space, given by the equation above the graph, either side of which a given input should be categorized as one or the other class. The linear template is given by the unit vector orthogonal to this line.

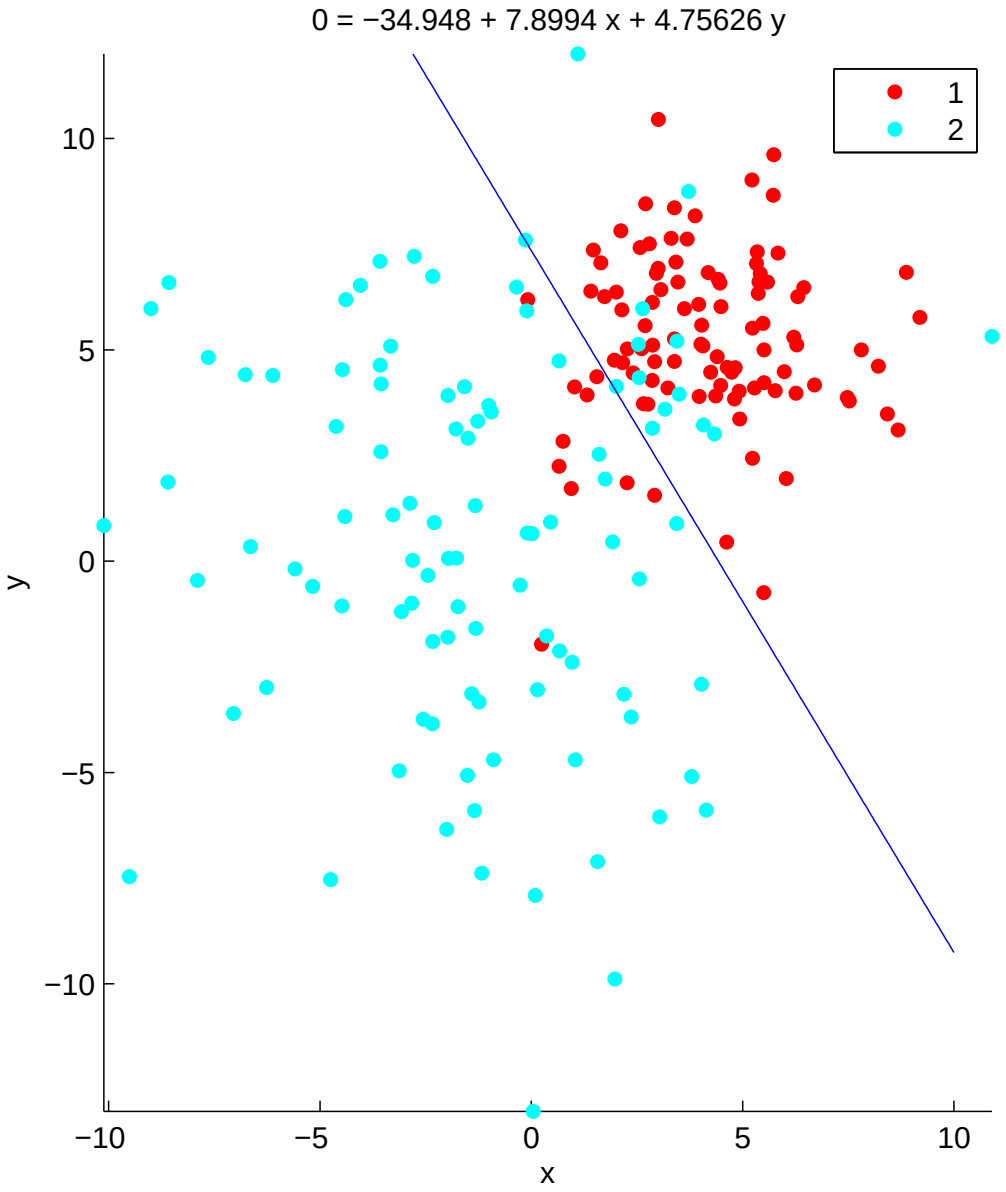


Figure 1.3: Illustration of the reverse–correlation process. For each of the two points in figure reffig:lte1 a random sample is generated from the Gaussian distribution centred at that point, leading to a subset of mis–classified samples for each distribution (red marks to the left of the line and blue marks to the right of the line). The linear template is estimated as the difference between the mean of the two sets of mis–classified samples.

two genders and one of the states of expressiveness. The mask completely covered the face in black, save for a set of Gaussian apertures ("bubbles") acting as holes through which the face was visible. Each Gaussian bubble had a diameter slightly larger than that of the iris. On each trial, the locations of these bubbles was randomly assigned, and the total number of bubbles per trial was determined by the performance of the subject in preceding trials. If the subject was performing poorly up to that point ($< 75\%$ accuracy), then the number of bubbles was increased and, if they were performing well ($> 75\%$ accuracy), the number was decreased.

The original Bubbles data were analysed, per subject, in relation to their performance. Each trial was categorized as either 'correct' or 'incorrect' and the retained mask for that trial was added to those of all other trials in the same category. The bubbles equivalent of the classification image was found by taking the proportion of each pixel's aggregate value in the 'correct' set to the grand aggregate value for this pixel over all trials. What the bubbles result provides is therefore slightly different to what the reverse correlation result provides. The bubble classification image as a function of pixel, ρ, is actually equivalent to a spatially smoothed (by the standard deviation of the Gaussian) estimate of the likelihood distribution for being 'correct', over the range of possible locations, or means, of the Gaussian aperture:

$$C_B(\rho) \simeq P(\text{`}correct'|\mu = \rho). \tag{1.3}$$

The principles underlying the Bubbles paradigm are therefore similar to those underlying the reverse correlation paradigm — an input space is chosen (in this case, the space of pixels); an output response is recorded (the accuracy of the subject); and the input space is corrupted somehow — but different in that the bubbles classification image does not provide an estimate of the linear template [48]. An extension of the bubbles paradigm to EEG signals [62], which shall be discussed in more detail later, can be viewed as an approximation of a linear template. Not to the stimulus features themselves, but to their availability.

Facial emotion

The human face is capable of communicating a wide range of non–verbal signals. Some of these signals can be considered as ancillary to ongoing verbal communication (see, e.g., [18] for analyses of so–called 'conversational expressions'), but a subset of them express information that cannot be expressed verbally. Perhaps the most widely studied of these expressions are the emotional facial expressions. Ekman & Friesen [19] reported 6 emotional

expressions that are recognized across a wide range of cultures — 'Happy', 'Fear', 'Disgust', 'Surprise', 'Anger', 'Sadness'. These are commonly referred to as the 6 'universal' emotional expressions.

Areas in the brain associated with general facial processing — the fusiform gyrus [38], the superior temporal sulcus [23] — have been implicated in the specific processing of facial emotion [49]. But the regions that are most often implicated are those associated with the affect that the emotion evokes, rather than face–specific areas. Expressions of negative affect — fear [1][2] and sadness [11] — are associated with amygdala activation. Happy is associated with activation in the cingulate cortex [57], and disgust is associated with activation in the insula [14]. In terms of EEG/MEG, specific effect of facial emotion have been reported in MEG [69] and EEG [63] recordings at $170ms$ following stimulus presentation. The EEG study cited is the study of Schyns et al. that is tested later.

Chapter 2

The signal model

Introduction

To be able to quantitatively relate the classification task that a subject is performing to the brain signals that we measure, we first need a quantitative model of the brain signals themselves. The Electroencephalogram (EEG) signal is a measurement of the electric field disturbance at a set of electrodes placed on the surface of the scalp. While this electric field disturbance is directly affected by electrophysiological activity in the brain, it does not, in itself, offer an unambiguous answer to the question of what the components of this electrophysiological activity are, or where they are taking place in the brain [5]. The purpose of the EEG signal model is to remove this ambiguity — to be able to record that, for any set of scalp measurements $m(t)$, the brain currently has some electrophysiological state $x(t)$, and that this particular state is the output of some known function β of the measurements. How this function is chosen is entirely up to us but, once chosen, it will determine both the state that is assigned to the brain and also the interpretation of that state in terms of the classification task at hand. What makes any particular model "good", or at least "better" than another, is therefore driven by two completely different kinds of criteria — one based upon what is already known about how the brain ought to function, and one based upon the task that the brain is currently performing. There would be little use in producing a model that could perfectly explain a single isolated task if the state the brain didn't also make sense for the brain itself, but at the same time there would be little to gain from the exercise if an attempt wasn't at least made to explain the classification task currently undertaken by the subject.

Returning to the task–oriented assessment of different models later, let us for now just consider the models in terms of the brain that they aim to

represent. For this, it is first worthwhile separating the models into three classes of approach:

1. "data–driven" or statistical approach,
2. "theory–driven" or *a priori* approach,
3. "naive" or non–model approach.

Of the three, the third is really less of a set and more of a singular entity. It refers to the case where, for whatever reason, the electrophysiological state $x(t)$ that is assigned to the brain at any particular point in time during the course of the task is simply the unaltered set of signals measured at the electrodes. The data–driven approach consists of those models that attempt to optimize some statistical properties of the brain electrophysiological state $x(t)$, without explicitly considering the kind of mechanism that might produce it. The theory–driven approach, on the other hand, consists of those models that do explicitly attempt this — usually by considering the electro–physiological basis of the measured brain response. This distinction is somewhat artificial, as instances of either type may also be interested to some extent in what the other type is primarily trying to achieve (even if this interest is not integral to the model solution, both can certainly be judged on what the other is trying to achieve) but the distinction is useful nonetheless, if only to make explicit the two different kinds of consideration that they represent.

Data–driven model

The goal of the data driven model is to explain the measured data by a set of unknown 'sources' that have no explicit physical relation to the set of electrodes from which the data is measured. In this type of model, although the sources are assumed to reflect the activity of electrophysiological generators at real physical locations in the brain, these physical assumptions are not directly used to estimate their activity, nor are their physical locations directly estimated. Each source is instead an abstract statistical concept — some unambiguous transformation of the measured electrical disturbance at the scalp that adheres to a set of defined statistical constraints.

The basic hypothesis is that the data measured at n electrode sites for any trial of the experiment and at any time point is generated from some unknown linear mixture of $m \leq n$ signal sources and $e = n - m$ noise sources, where each source is considered as a random variable drawn from some parametric distribution. The relation between sources and data could

be also be expressed by a non–linear function, but only the linear case is considered here. This can be written as:

$$\begin{aligned}\mathbf{m}_i(t) &= s_{i1}(t)\mathbf{a}_1 + s_{i2}(t)\mathbf{a}_2 + \cdots + s_{in}(t)\mathbf{a}_n + \mathbf{c} \\ &= \mathbf{A}\mathbf{s}_i(t) + \mathbf{c}, \end{aligned} \tag{2.1}$$

where $\mathbf{m}(t)_i$ is an $n \times 1$ vector of EEG measurements for trial i at time t, each entry $s_{ij}(t)$ of the $n \times 1$ vector $\mathbf{s}_i(t)$ is the value of some unknown signal or noise source j for trial i at time t, each $n \times 1$ vector $\mathbf{a}_j$ is column j of some unknown $n \times n$ invertible "mixing matrix" $\mathbf{A}$, and the $n \times 1$ vector $\mathbf{c}$ is an unknown additive constant. Without any loss of generality, the constant can be ignored if we assume that both the EEG data and the sources are centred at each time point. The model should hold for all trials that we measure, so we can write:

$$\mathbf{M}(t) = \mathbf{A}\mathbf{S}(t), \tag{2.2}$$

where column i of the $n \times k$ matrices $\mathbf{M}$(t) and $\mathbf{S}$(t) is $\mathbf{m}_i(t)$ and $\mathbf{s}_i(t)$ respectively.

If we define the "unmixing matrix" $\mathbf{W}$, such that $\mathbf{W^t A} = \mathbf{I}$, then the sources can be recovered from the measured data according to:

$$\mathbf{S}(t) = \mathbf{W^t M}(t) \tag{2.3}$$

if only the unmixing matrix is known. The problem is then to find an estimate of this unmixing matrix when the only part of the above model that is actually observed is the data. This problem is termed 'Blind Source Separation' (BSS). In order to solve this problem, some constraints must be imposed upon the statistical nature of the sources. Here we constrain the sources to be independent Gaussian random variables at each point in time t.

Independent Gaussian Sources

The benefit of assuming Gaussian sources is that their statistics go no higher than second order — Gaussian sources are independent if and only if their covariance matrix is diagonal. The drawback is that second order statistics alone do not provide enough constraint to solve for $\mathbf{A}$ [17] — an infinite number of unmixing matrices could equally produce Gaussian variables with diagonal covariance matrices from the observed data. Consequently, the BSS problem is more commonly solved under the constraint of non–Gaussian — in particular super–Gaussian (i.e. positively Kurtotic) — sources [17][28][37]. However, while non–Gaussian BSS of EEG data is effective at finding recording artifacts, meaningful EEG activity is more likely to be close to a Gaussian distribution [29]. Due to the above–mentioned ambiguity of second

order statistics, some additional constraints are necessary to separate Gaussian sources. If the sources are assumed to be sufficiently non–stationary over time (i.e. if their covariance matrix changes sufficiently over time) then these additional constraints are provided by the temporal sequence of the data [44][56][52].

To understand this, let us first consider the situation where the sources are active at only a single point in time and we have measurements at exactly this point. In this case, the covariance matrix of the data is given by:

$$\mathbf{\Sigma_m} = \frac{1}{k-1}\mathbf{M}\mathbf{M^t},$$

Substituting Equation 2.2 into the above equation, the data covariance can be expressed in terms of both the source covariance and the mixing matrix:

$$\begin{aligned}\mathbf{\Sigma_m} &= \mathbf{A}[\frac{1}{k-1}\mathbf{S}\mathbf{S^t}]\mathbf{A^t} \\ &= \mathbf{A}\mathbf{\Sigma_s}\mathbf{A^t},\end{aligned} \tag{2.4}$$

for which there is no unique solution for $\mathbf{A}$, even if the source covariance is known. To see why, imagine that $\mathbf{A}$ is the product of some orthogonal matrix $\mathbf{V}$ and some diagonal matrix $\mathbf{D^{\frac{1}{2}}}$, such that $\mathbf{A^t}\mathbf{A} = \mathbf{D}$. In this case Equation 2.4 can be rewritten as:

$$\mathbf{\Sigma_m}\mathbf{A} = \mathbf{A}\mathbf{\Sigma_s}\mathbf{D}, \tag{2.5}$$

which, because $\mathbf{\Sigma_s}\mathbf{D}$ is diagonal, is the eigenvalue equation for $\mathbf{\Sigma_m}$. As the eigenvalue equation is only unique up to permutation and scale of the eigenvectors, so the solution for $\mathbf{A}$ is also. The permutation ambiguity can be resolved by sorting the eigenvectors by eigenvalue, and the scaling ambiguity can be resolved by fixing the unknown $\mathbf{\Sigma_s}$. Referring to the individual source variances by σ_j^2 , each corresponding element of $\mathbf{D}$ is given by $d_{jj} = \sqrt{\lambda_j/\sigma_j^2}$, where λ_j is the j^{th} eigenvalue of $\mathbf{\Sigma_m}$, and each corresponding column of $\mathbf{A}$ is given by $\mathbf{a}_j = d_{jj}\mathbf{v}_j$. So, with a known source covariance, $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Sigma_s}^{-\frac{1}{2}}$ is a possible solution to equation 2.4, and the only solution where $\mathbf{A^t}\mathbf{A} = \mathbf{D}$. However, rearranging equation 2.5 slightly, we get:

$$\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V^t}\mathbf{\Sigma_m}\mathbf{V}\mathbf{\Lambda}^{-\frac{1}{2}} = \mathbf{I},$$

which we can multiply on both sides by any arbitrary orthogonal matrix $\tilde{\mathbf{V}}$ and its transpose without changing the equality. So, dropping the restriction that $\mathbf{A^t}\mathbf{A} = \mathbf{D}$, any solution for $\mathbf{A}$ of the form $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}\tilde{\mathbf{V}}\mathbf{\Sigma_s}^{-\frac{1}{2}}$ also satisfies

equation 2.4. $\mathbf{A}$ is therefore ambiguous, not just up to scale and permutation, but also up to a rotation of $\mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}$.

Let us now consider the situation where the sources are active at two points in time and the data are measured at exactly these two points. The data covariance and the source covariance at both times t_1 and t_2 are again related by $\mathbf{A}$:

$$\mathbf{\Sigma_m}(t_1) = \mathbf{A}\mathbf{\Sigma_s}(t_1)\mathbf{A^t},$$

and:

$$\mathbf{\Sigma_m}(t_2) = \mathbf{A}\mathbf{\Sigma_m}(t_2)\mathbf{A^t}.$$

Multiplying both equations by $\mathbf{W}$:

$$\mathbf{\Sigma_m}(t_1)\mathbf{W} = \mathbf{\Sigma_m}(t_2)\mathbf{W}\mathbf{\Lambda}, \tag{2.6}$$

where $\Lambda = \mathbf{\Sigma_s}(t_1)\mathbf{\Sigma_s}^{-1}(t_2)$ is a diagonal matrix. This is the unique[1] generalized eigenvalue equation for the two matrices $\mathbf{\Sigma_m}(t_1)$ and $\mathbf{\Sigma_m}(t_2)$, where the ratio $\sigma_j^2(t_1)/\sigma_j^2(t_2)$ is equal to a generalized eigenvalue of the two matrices and the columns of the unmixing matrix are their corresponding generalized eigenvectors ([52]). This solution for $\mathbf{W}$ is clearly unique only in the case where $\Lambda = \mathbf{\Sigma_s}(t_1)\mathbf{\Sigma_s}^{-1}(t_2) \neq c\mathbf{I}$. Hence the source covariance matrix must be different at the two times if they are to be separated.

From a practical point of view, the situation that we are attempting to model will probably involve source covariance matrices at more than two time points. It is of course possible that, over the time course of measurement, only two periods of source activity with distinguishable covariance exist, but there is no particular reason to expect this to be the case. This then leads us to consider the situation where the sources are active over a sequence of time, T, and the data are measured at each point, t, in this sequence. We now have, for all $t \in T$:

$$\mathbf{\Sigma_m}(t) = \mathbf{A}\mathbf{\Sigma_s}(t)\mathbf{A^t},$$

and we are looking for an inverse of matrix $\mathbf{A}$ such that, for all time $t \in T$, $\mathbf{\Sigma}(t)_\mathbf{s} = \mathbf{W^t}\mathbf{\Sigma}(t)_\mathbf{m}\mathbf{W}$ is a diagonal matrix. We saw earlier, in equations 2.5 and 2.6, that this process of matrix diagonalization by invertible square matrices is possible for a single symmetric matrix or jointly for two different symmetric matrices. In the first case the solution is only unique if we restrict it to be orthogonal, while in the second case it is unique without restriction. Does a solution exist for diagonalizing more than two symmetric matrices? The answer is: probably not exactly. However, even if no exact solution does

[1] Again, up to scale and permutation. The same considerations of unit length scaling of eigenvectors and sorting by eigenvalue apply as for the standard eigenvalue equation.

exist, Pham & Cardoso ([56]) show that the invertible square matrix which merely maximizes the joint diagonalization of the set of data covariance matrices, in the sense of Kullback–Liebler (KL) divergence, also minimizes the Gaussian mutual information between the sources. Imposing the additional constraints of multiple simultaneous covariance matrices therefore changes the problem from that of finding the unique unmixing matrix that transforms the observed data into a set of independent Gaussian sources to that of finding the unique unmixing matrix that transforms the observed data into a set of Gaussian sources that are as independent as possible.

Following Pham & Cardoso, let us define the mutual information between n random vectors $Y_1, \ldots, Y_n$ with joint density $f_{Y_1,\ldots,Y_n}$ and marginal densities $f_{Y_1}, \ldots, f_{Y_n}$ as:

$$I(Y_1, \ldots, Y_n) = -\mathrm{E}\left[\log \frac{\prod_{i=1}^{n} f_{Y_i}(Y_i)}{f_{Y_1,\ldots,Y_n}(Y_1, \ldots, Y_n)}\right], \tag{2.7}$$

which is equal to the KL divergence between the joint density and the product of the marginal densities. The (asymmetrical) KL divergence between two zero mean Gaussian distributions with covariance matrices $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$ is given by:

$$D\{\mathbf{\Sigma_1}|\mathbf{\Sigma_2}\} = \frac{1}{2}[\mathrm{tr}(\mathbf{\Sigma_2}^{-1}\mathbf{\Sigma_1}) - \log(|\mathbf{\Sigma_2}^{-1}\mathbf{\Sigma_1}|) - n], \tag{2.8}$$

where $|\mathbf{X}|$ and $\mathrm{tr}(\mathbf{X})$ are respectively the determinant and trace of square matrix $\mathbf{X}$. For any estimate of the unmixing matrix $\hat{\mathbf{W}}$, we have $\hat{\mathbf{S}}(t) = \hat{\mathbf{W}}^{\mathbf{t}}\mathbf{M}(t)$, with covariance matrix of the joint density:

$$\hat{\mathbf{\Sigma}}_{\mathbf{s}}(t) = \hat{\mathbf{W}}^{\mathbf{t}}\mathbf{\Sigma}_{\mathbf{m}}(t)\hat{\mathbf{W}}, \tag{2.9}$$

and covariance matrix of the product of the marginal densities:

$$\hat{\mathbf{\Sigma}}_{\mathbf{\Pi s}}(t) = \mathrm{diag}[\hat{\mathbf{\Sigma}}_{\mathbf{s}}(t)]. \tag{2.10}$$

Assuming that the observed data are temporally independent, the mutual information between estimated sources over the sequence T is simply the sum of their mutual information at each time $t \in T$. The solution for $\hat{\mathbf{W}}$ that minimizes the mutual information between sources for all time points is therefore that which minimizes:

$$\begin{aligned}\sum_{t=1}^{l_T} I[\hat{\mathbf{S}}_1(t), \ldots, \hat{\mathbf{S}}_n(t)] &= \sum_{t=1}^{l_T} D\{\hat{\mathbf{\Sigma}}_{\mathbf{s}}(t)|\hat{\mathbf{\Sigma}}_{\mathbf{\Pi s}}(t)\}, \\ &= \sum_{t=1}^{l_T} D\{\hat{\mathbf{W}}^{\mathbf{t}}\mathbf{\Sigma}_{\mathbf{m}}(t)\hat{\mathbf{W}}|\mathrm{diag}[\hat{\mathbf{W}}^{\mathbf{t}}\mathbf{\Sigma}_{\mathbf{m}}(t)\hat{\mathbf{W}}]\},\end{aligned} \tag{2.11}$$

where $\hat{\mathbf{S}}_j(t)$ is row j of matrix $\hat{\mathbf{S}}(t)$ ([56].

This terms is related to the likelihood function for the sources, which is a monotonically decreasing function of $D\{\hat{\mathbf{\Sigma}}_\mathbf{s}(t)|\mathbf{\Sigma}_\mathbf{s}(t)\}$. Using the fact that, for any positive–definite matrix $\mathbf{R}$ and any diagonal matrix $\mathbf{\Sigma}$,

$$D\{\mathbf{R}|\mathbf{\Sigma}\} = D\{\mathbf{R}|\text{diag}(\mathbf{R})\} + D\{\text{diag}(\mathbf{R})|\mathbf{\Sigma}\},$$

Pham & Cardoso show that this likelihood function can be written as:

$$\begin{aligned} L^* &= \sum_{t=1}^{l_T} D\{\hat{\mathbf{\Sigma}}_\mathbf{s}(t)|\text{diag}[\hat{\mathbf{\Sigma}}_\mathbf{s}(t)]\} + \sum_{t=1}^{l_T} D\{\text{diag}[\hat{\mathbf{\Sigma}}_\mathbf{s}(t)]|\mathbf{\Sigma}_\mathbf{s}(t)\}, \\ &= \sum_{t=1}^{l_T} I[\hat{\mathbf{S}}_1(t), \ldots, \hat{\mathbf{S}}_n(t)] + \sum_{t=1}^{l_T} D\{\text{diag}[\hat{\mathbf{W}}^\mathbf{t}\mathbf{\Sigma}_\mathbf{m}(t)\hat{\mathbf{W}}]|\mathbf{\Sigma}_\mathbf{s}(t)\}. \end{aligned} \tag{2.12}$$

The first term, as we have seen, is minimized when the sources are as independent as possible, while the second term is minimized when the (unknown) source covariance is set to that of the product of the marginal densities for any estimate of $\hat{\mathbf{W}}$. So, by maximizing statistical independence between sources, we have simultaneously minimized the left hand term of the negative likelihood function, and we can minimize the right hand hand term by simply using the marginal variance of our maximally independent estimate as the true source covariance matrix ([56]).

The problem faced by Pham & Cardoso is slightly different to the problem as stated here, however. They are dealing with the case where, for each time t, only a single observation of the data is available. Or, in terms of the model presented here, the experiment consists of only a single trial. But, clearly, the mutual information term given in Equation 2.11 relies upon our having an estimate of the data covariance at each time t. To overcome this problem, they used a kernel estimator for the data covariance (essentially a temporal smoothing window). For long sequences, this is fine, but the data we are dealing with cover only very small periods of time ($l_T \simeq 100$ time points). The benefit of our design, however, is that we have, for each time t, the sample covariance matrix of the data, estimated over $k \simeq 3000$ independent observations.

Minimization of the mutual information As stated above, our goal is to minimize the mutual information between the source signals with respect to an estimate of the unmixing matrix $\hat{\mathbf{W}}$. The gradient of the mutual information with respect to changes in $\hat{\mathbf{W}}$ is given by the square $n \times n$

matrix $\mathbf{G}$, where:

$$g_{ij} = \frac{1}{l_T}\sum_{t=1}^{l_T}\frac{\hat{\sigma}_i(t)\hat{\sigma}_j(t)}{\hat{\sigma}_i^2(t)} - \delta_{ij},$$
$$= \frac{1}{l_T}\sum_{t=1}^{l_T}\frac{[\hat{\mathbf{W}}^{\mathbf{t}}\mathbf{\Sigma_m}(t)\hat{\mathbf{W}}]_{ij}}{[\hat{\mathbf{W}}^{\mathbf{t}}\mathbf{\Sigma_m}(t)\hat{\mathbf{W}}]_{ii}} - \delta_{ij}. \tag{2.13}$$

A Jacobi–like algorithm, named 'Joint Diagonalization for the Block gaussian Likelihood', or JD–BGL for short, that uses $\mathbf{G}$ to iteratively update pairs of columns of $\hat{\mathbf{W}}$ is developed in [55]. If we further define the square $n \times n$ matrix $\mathbf{\Omega}$, where:

$$\omega_{ij} = \frac{1}{l_T}\sum_{t=1}^{l_T}\frac{\hat{\sigma}_j^2(t)}{\hat{\sigma}_i^2(t)} - \delta_{ij},$$
$$= \frac{1}{l_T}\sum_{t=1}^{l_T}\frac{[\hat{\mathbf{W}}^{\mathbf{t}}\mathbf{\Sigma_m}(t)\hat{\mathbf{W}}]_{jj}}{[\hat{\mathbf{W}}^{\mathbf{t}}\mathbf{\Sigma_m}(t)\hat{\mathbf{W}}]_{ii}} - \delta_{ij}, \tag{2.14}$$

then the likelihood of the sources will always increase if we change columns $\hat{\mathbf{w}}_i$ and $\hat{\mathbf{w}}_j$ according to:

$$\begin{bmatrix}\hat{\mathbf{w}}_i^{\mathbf{t}(\text{new})} \\ \hat{\mathbf{w}}_j^{\mathbf{t}(\text{new})}\end{bmatrix} = \begin{bmatrix}\hat{\mathbf{w}}_i^{\mathbf{t}} \\ \hat{\mathbf{w}}_j^{\mathbf{t}}\end{bmatrix} - \mathbf{T}_{ij}\begin{bmatrix}\hat{\mathbf{w}}_i^{\mathbf{t}} \\ \hat{\mathbf{w}}_j^{\mathbf{t}}\end{bmatrix}, \tag{2.15}$$

where:

$$\mathbf{T}_{ij} = \frac{2}{1+\sqrt{1-4h_{ij}h_{ji}}}\begin{bmatrix}0 & h_{ij} \\ h_{ji} & 0\end{bmatrix}, \tag{2.16}$$

and:

$$\begin{bmatrix}h_{ij} \\ h_{ji}\end{bmatrix} = \begin{bmatrix}\omega_{ij} & 1 \\ 1 & \omega_{ji}\end{bmatrix}^{-1}\begin{bmatrix}g_{ij} \\ g_{ji}\end{bmatrix}, \tag{2.17}$$

for each of the $n(n-1)/2$ pairs of columns. It will only fail to increase where $g_{ij} = g_{ji} = 0$ — i.e. where the gradient of the likelihood with respect to changes in $\hat{\mathbf{W}}$ is zero ([55]).

Simulations To get an idea of what it takes for sources to be deemed 'sufficiently stationary', simulations were run testing the performance of the JD–BGL algorithm for varying levels of stationarity in the source signals in the absence of noise. Performance under noise is considered in a later section. Here, the measure of 'stationarity', q_j, for any source $s_j(t)$ was considered as the variance over time of $\sigma_j^2(t)$. As the $\sigma_j^2(t)$ must always be positive, we

were careful to control its variance over time without changing its expected value. The approach taken here was to treat each $\sigma_j^2(t)$ as an independent gamma random variable over time:

$$\sigma_j^2(t) \sim \Gamma(\kappa, \theta), \tag{2.18}$$

The mean and variance of a gamma distribution are given by $\kappa\theta$ and $\kappa\theta^2$, so both parameters must be changed together if we want to change the variance while leaving the mean constant. To control for this, each level of stationarity, $q_j(p)$, was set under the joint constraints that $\kappa(p)\theta^2(p) = q_j(p)$ and $\kappa(p)\theta(p) = \kappa(p)\theta(p) = c$, where c is some positive scalar. These constraints are both met if $\theta(p) = \frac{1}{c}q_j(p)$, and $\kappa(p) = c^2\frac{1}{q_j(p)}$.

Two different conditions of non–stationarity were tested — homogeneous and non–homogeneous. In the homogeneous condition, separating performance was measured at 10 equally spaced levels of q_j, ranging from 0.01 to 1, with $q_j = q_i, \forall j, i$, and expected variance was set to $c = 10$. Figure 2.1 displays the gamma p.d.f. at each of these levels. In the non–homogeneous condition, each of 10 levels was set by controlling the difference between the maximum value of q_j and the average of all other values, whilst leaving the maximum value unchanged. A simple scheme was chosen so that $[q_1(p), \ldots, q_n(p)] = [0.01^{(ap)} \ldots, 1^{(ap)}]$, with a being some small positive constant. This also served to test separating performance when more than one source was close to being stationary. Figure 2.2 displays the varying distributions of q_j at different levels of non–homogeneous non–stationarity. For both conditions, 100 datasets were generated per level, each with $l_T = 100$ time points and $k = 1000$ observations per time point. Each dataset, level, and time point consisted of 6 randomly generated Gaussian sources, and the same randomly generated 6×6 mixing matrix (uniform, on the range $[0, 1]$) was used for all levels of a single dataset.

Performance was measured as the KL divergence between the matrix $\mathbf{H} = \hat{\mathbf{W}}^{\mathbf{t}}\mathbf{A}$ and the identity matrix, with perfect separation in the case where $\mathbf{H} = \mathbf{I}$. However, some steps had to be taken to overcome the ambiguity in permutation and scaling in $\hat{\mathbf{W}}^{\mathbf{t}}$ first. For the scaling problem, each row of $\hat{\mathbf{W}}^{\mathbf{t}}$ was scaled to unit length. Then, to construct the permutation matrix, $\mathbf{P}$, for each column j of the mixing matrix, the "best matching" row i in $\hat{\mathbf{W}}^{\mathbf{t}}$ was selected as the index of the maximum absolute value in $\hat{\mathbf{W}}^{\mathbf{t}}\mathbf{a}_j$ (as all rows are unit length, this is simply the cosine of their angles, multiplied by the length of the column). p_{ji} was then set to 1 and the process was repeated for all columns of $\mathbf{A}$. The rows of $\hat{\mathbf{W}}^{\mathbf{t}}$ were finally rescaled so that the diagonal elements of $\hat{\mathbf{W}}^{\mathbf{t}}\mathbf{A}$ were all equal to 1. Note that, when a single row of $\mathbf{W}^{\mathbf{t}}$ was the best match for more than one column of $\mathbf{A}$, then the

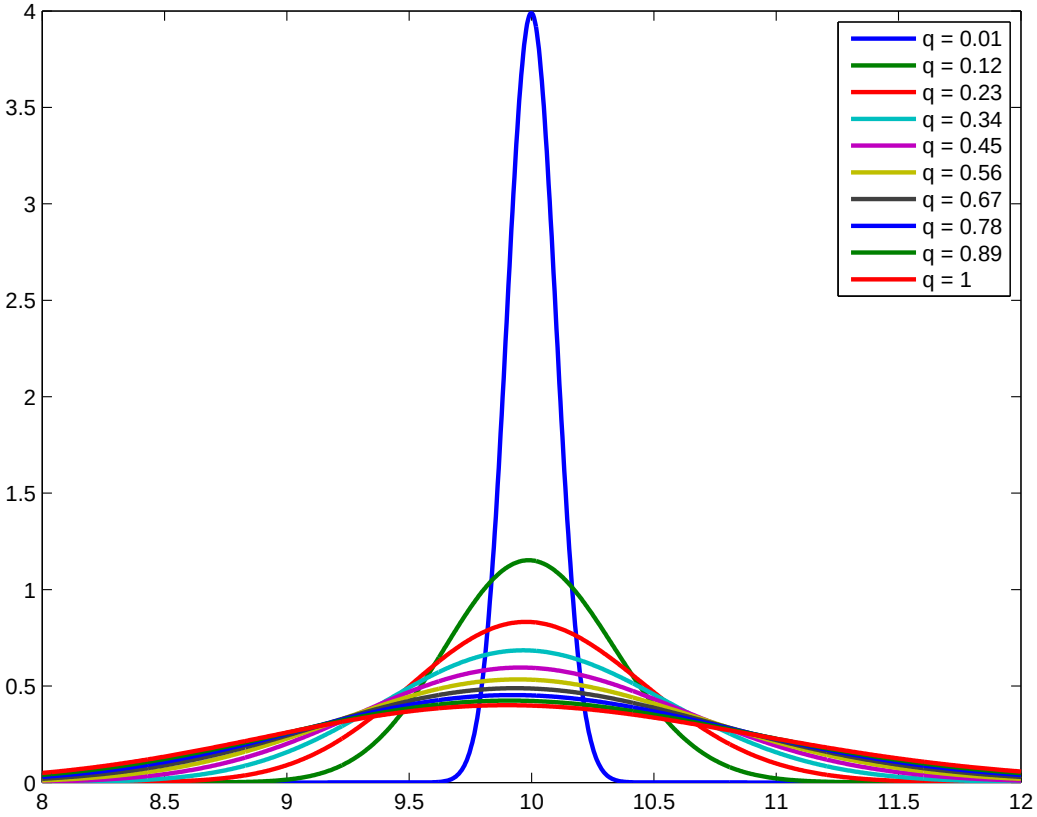


Figure 2.1: Source variance gamma densities for each of the 10 levels of stationarity.

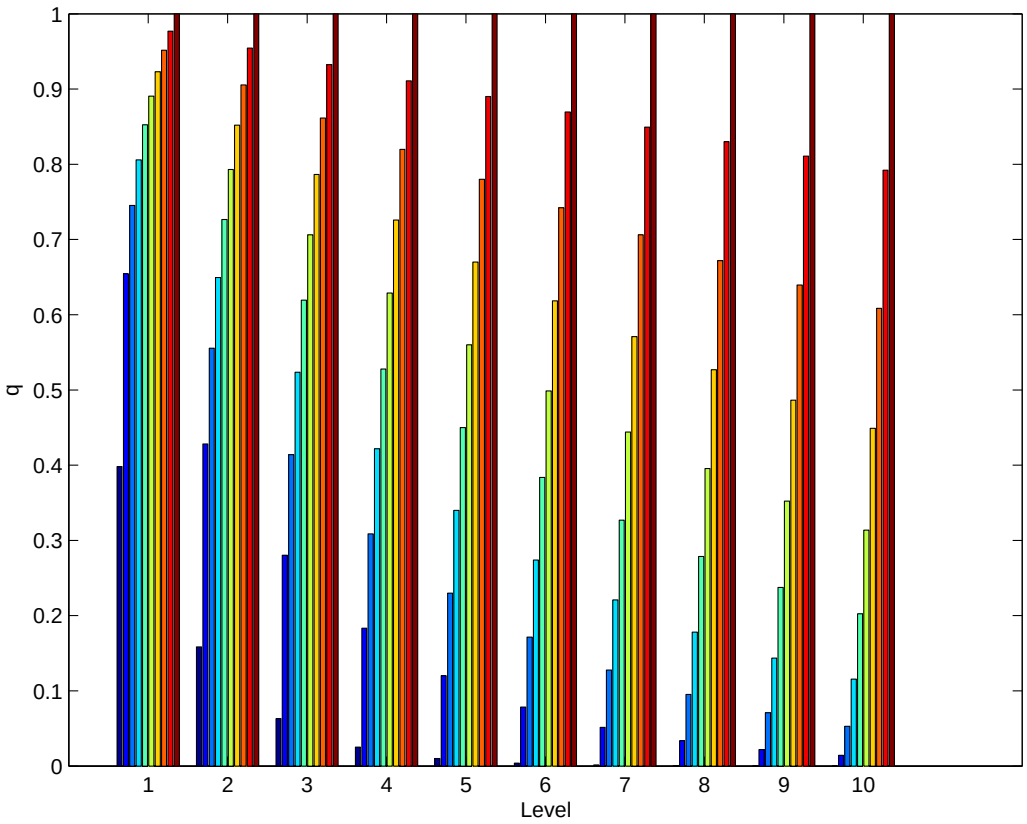


Figure 2.2: Distribution over sources of q_j for each of the 10 levels of non–homogeneous stationarity. Low levels of non–homogeneity have more sources close to stationary than not, high levels vice versa.

only solution was to assign it to the closest column (normalized by column length) and the second best matching row would then be set to the remaining columns (and the same check was carried out for the second best matching rows, etc...). However, as a result of this procedure, the absolute value of at least one off–diagonal element of **H** would greater than 1, and hence **H** would no longer be positive semi–definite — a requisite condition for using the KL divergence as a matrix distance metric [56]. The likeliest cause of this event was when two or more columns of the mixing matrix were too close together, so when this occurred the dataset was just rerun with a new mixing matrix.

Figure 2.3 shows performance as a function of homogeneous non–stationarity. A value of zero here indicates that the matrix **H** is equal to the identity matrix, and hence the sources are perfectly separated. These results clearly show that the JD–BGL algorithm reaches optimal performance quickly as the homogeneous non–stationarity increases, more or less plateauing at $q \simeq 0.25$. To confirm that the improvement in separating performance was indeed caused by changes in the variance, rather than changes in the kurtosis or skewness of the gamma distribution, the process was repeated at various fixed levels of κ (both skewness and kurtosis are functions of κ but not of θ.) Figure 2.4 shows separating performance as a function of homogeneous θ at 10 different levels of fixed κ, from which it is clear that changes in kurtosis and skewness were not the cause of the change in separating performance as a function of variance in figure 2.3.

Figure 2.5 shows separating performance at each of the 10 different levels of non–homogeneous stationarity. Due to the way that the levels were defined, we can see if there are any effect of the distribution of q_j. First, note that only the variance of the three sources with the lowest q_j at level 10 were below $q_j = 0.1$. Comparing the separating performance here at level 10 with separating performance at $q_j = 0.1$ for homogeneous stationarity, the two are almost identical, despite the fact that most of the sources are above that level in this case, while all sources were at that level in the homogeneous case. On the other hand, separating performance at level 4 is far superior to the corresponding performance for its minimum value $q_j = 0.016$ in the homogeneous case. So separating performance can be high in the non–homogeneous case, even in the presence of one or more sources that are close to stationarity, provided that the average non–stationary is sufficiently high amongst the rest of the sources.

Subspace Partitioning

As stated at the beginning of this section, we propose that the observed data are generated from some unknown mixture of independent Gaussian

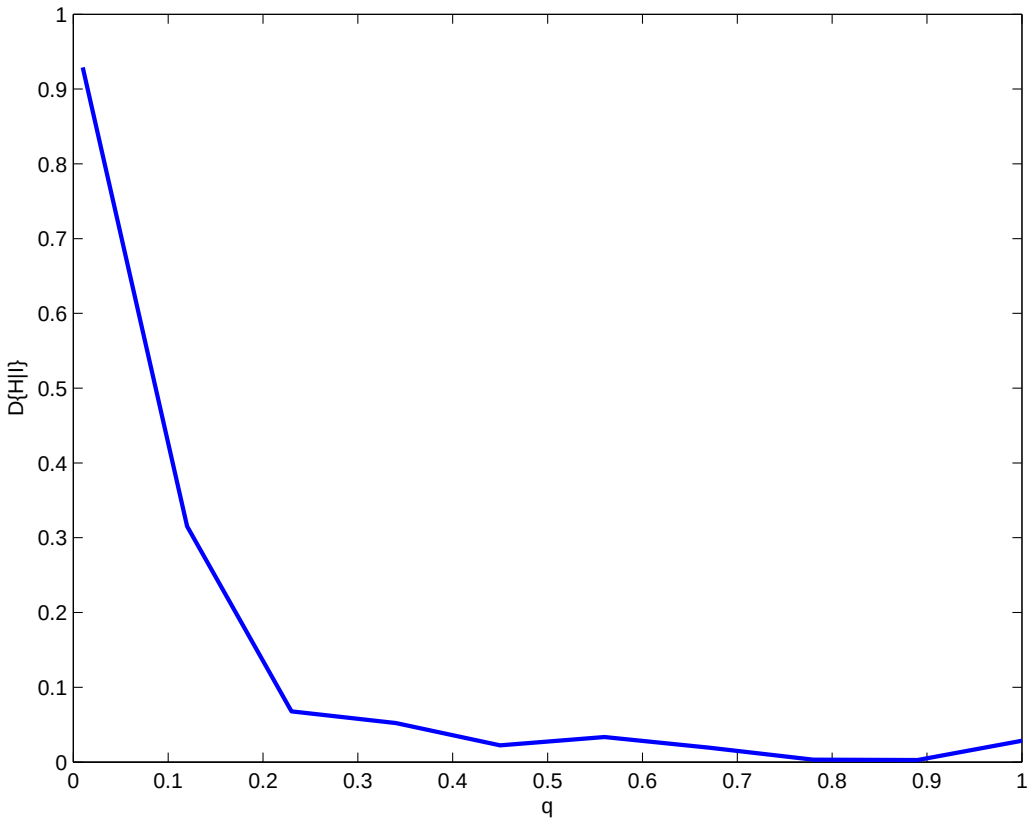


Figure 2.3: Separating performance as a function of homogeneous non–stationarity. Horizontal axis — level of non–stationarity, $q = \kappa\theta^2$. Vertical axis — separating performance, $Sep = D\{H|I\}$.

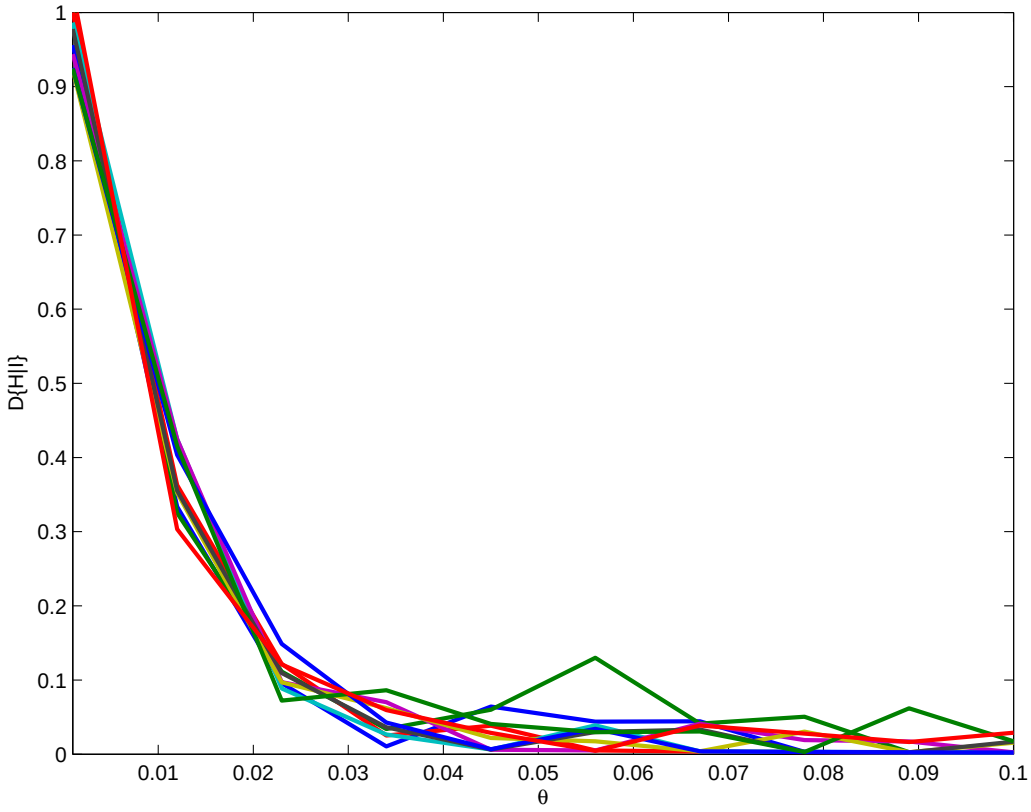


Figure 2.4: Separating performance at different homogeneous levels of k over range of θ. Individual curves — different levels of κ. Horizontal axis — θ. Vertical axis — separating performance, $Sep = D\{H|I\}$.

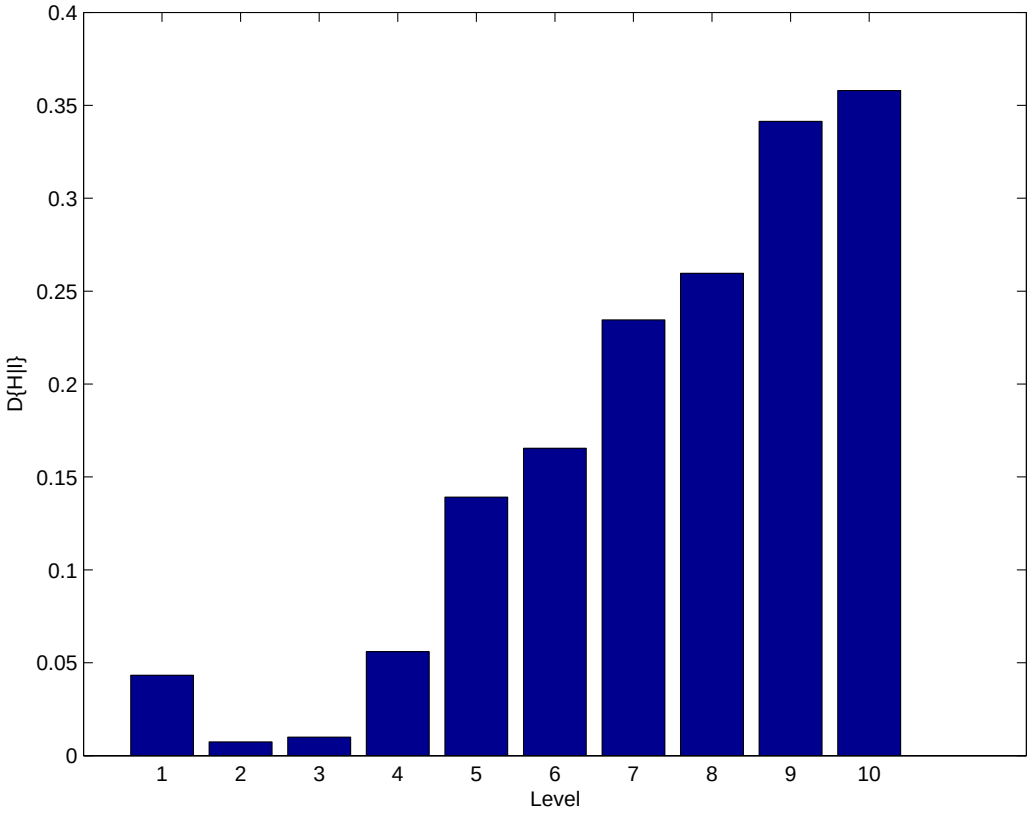


Figure 2.5: Separating performance at each of 10 different levels of non–homogeneous non–stationarity. Horizontal axis — level of non–homogeneity in source stationarity, p, corresponding to $[q_1, \ldots, q_n] = [0.01^{(ap)}, \ldots, 1^{(ap)}]$. Vertical axis — separating performance, $Sep = D\{H|I\}$

signal and noise sources. We have shown, following Pham & Cardoso ([56]), that the best estimate of these sources can be recovered by maximizing the joint diagonalization of the data covariance over time, hence minimizing the mutual information between sources. We have not, however, shown how the signal sources can be separated from the noise sources. Consider again equation 2.2 — this time in terms of our best estimates for $\mathbf{A}$ and $\mathbf{S}(t)$:

$$\mathbf{M}(t) = \hat{\mathbf{A}}\hat{\mathbf{S}}(t).$$

Let us now define the $n \times k$ matrix $\hat{\mathbf{M}}(t)$ to be an estimate of the observed data in terms of some subset, B, of the combined signal and noise sources, and $n \times k$ matrix $\mathbf{E}(t)$ to be the error in this estimate, given $\hat{\mathbf{A}}$, such that:

$$\begin{aligned} \mathbf{M}(t) &= \hat{\mathbf{M}}(t) + \mathbf{E}(t), \\ &= \hat{\mathbf{A}}^{(B)}\hat{\mathbf{S}}^{(B)}(t) + \hat{\mathbf{A}}^{(-B)}\hat{\mathbf{S}}^{(-B)}(t), \end{aligned} \tag{2.19}$$

where $\hat{\mathbf{A}}^{(B)}$ is an $n \times l_B$ matrix consisting of the columns of $\hat{\mathbf{A}}$ corresponding to the l_B elements of subset B, $\hat{\mathbf{A}}^{(-B)}$ is an $n \times n - l_B$ matrix consisting of the remaining columns, and $\hat{\mathbf{S}}^{(B)}(t)$ and $\hat{\mathbf{S}}^{(-B)}(t)$ are similarly defined $l_B \times k$ and $n - l_B \times k$ matrices. Now, because:

$$\begin{aligned} \begin{bmatrix} \hat{\mathbf{S}}^{(B)}(t) \\ \hat{\mathbf{S}}^{(-B)}(t) \end{bmatrix} &= \begin{bmatrix} \hat{\mathbf{W}}^{(B)\mathbf{t}} \\ \hat{\mathbf{W}}^{(-B)\mathbf{t}} \end{bmatrix} \mathbf{M}(t) \\ &= \begin{bmatrix} \hat{\mathbf{W}}^{(B)\mathbf{t}}\mathbf{M}(t) \\ \hat{\mathbf{W}}^{(-B)\mathbf{t}}\mathbf{M}(t) \end{bmatrix}, \end{aligned}$$

we can rewrite equation 2.19 as:

$$\mathbf{M}(t) = \hat{\mathbf{A}}^{(B)}\hat{\mathbf{W}}^{(B)\mathbf{t}}\mathbf{M}(t) + \mathbf{E}(t).$$

The expected error variance in the estimate for the set $\hat{B}$ given the mixing matrix $\hat{\mathbf{A}}$ is therefore given by:

$$\epsilon = \frac{1}{l_T}\sum_{t=1}^{l_T} \mathrm{tr}\left[[\mathbf{I} - \hat{\mathbf{A}}^{(\hat{B})}\hat{\mathbf{W}}^{(\hat{B})\mathbf{t}}]\mathbf{\Sigma_m}(t)[\mathbf{I} - \hat{\mathbf{A}}^{(\hat{B})}\hat{\mathbf{W}}^{(\hat{B})\mathbf{t}}]^{\mathbf{t}}\right], \tag{2.20}$$

On this basis, it might seem reasonable to simply find the set $\hat{B}$ that minimizes this error and call that set the signal set. However, as can be seen from equation 2.20, the error can be made to zero if $\hat{B}$ is set to some permutation of the full mixing matrix, as $\hat{\mathbf{A}}^{(\hat{B})}\hat{\mathbf{W}}^{(\hat{B})\mathbf{t}} = \mathbf{I}$ in this case. This makes sense, because our model is a model of both the signal and the noise, and we do not know how many signal sources there ought to be.

Known signal quantity If the number of signal sources, l_B, is known beforehand, the problem is theoretically solvable but becomes increasingly computationally expensive as the number of known signal sources and the number of observed channels increases (the number of unordered source subsets being given by $\binom{n}{l_B}$). The problem is simplified if we assume some natural order to the sources, such that the signal occupies the top l_B sources in the sorted list. Two slightly different, but related, ordering criteria are introduced here, both utilizing the estimated variance of the recovered sources. The first criterion, 'Maximum Expected Energy' (MEE), assumes that the expected variance in the signal sources is always higher than the expected variance in the noise sources. The MEE sorting criterion is easily applied to the estimated data, but should be sensitive to the relative energy in the noise. The second criterion, 'Maximum Non-stationary Energy' (MNsE), assumes that the noise is generated by a stationary process, but that its energy is otherwise unknown. The advantage of the MNsE criterion is that it is should distinguish equally under varying levels of expected noise variance, and it is no more difficult to implement than the MEE criterion. As the previous simulations demonstrate, the JD–BGL algorithm still provides good estimates of stationary source distributions, provided that a few of the sources are non–stationary (see fig 2.5). But it is unclear how sensitive it should be to the level of stationarity in the noise. Simulations, described in the next section, were run to test both criteria in the presence of both stationary and non–stationary noise.

Signal quantity estimation Such sorting criteria are only useful if the number of signal sources is known, however. And, as we do not know this, we must estimate it somehow. To solve this problem, first note that, in the absence of noise, the rank of the data covariance matrix at time t is limited by the number of signal sources ($r(t) \leq l_B$). If we define the 'effective rank', $rE(t)$, of the data covariance at any point in time to be the rank of the subspace spanned by the *currently active* signal sources, then the problem of estimating the number of signal sources can be reduced to that of estimating the maximum of $rE(t)$ over time. Let's call this the 'Maximum Effective Rank' (MER) principle. This makes the problem much easier, as methods already exist for estimating the rank of the signal subspace for a single covariance matrix. Note that in estimating the rank of the subspace spanned by the active signal sources we are not interested in estimating the true sources themselves. We already have an estimate of that; we simply do not know which of our estimated sources correspond to the signal. We are instead searching for some orthogonal basis within which all of the signal

sources lie.

The classic method for effective rank estimation is what shall be called here the 'Cumulative Energy Function' (CEF) method. The CEF is defined as the cumulative sum of the eigenvalues of the covariance matrix, sorted by decreasing magnitude:

$$\mathrm{CEF}(x) = \sum_{i=1}^{n} \lambda_i. \tag{2.21}$$

The basic CEF method is to chose the effective rank as the value rE that satisfies $\mathrm{CEF}(rE-1) < p\mathrm{CEF}(n) < \mathrm{CEF}(rE)$, where p is some number between 0 and 1. In other words, the effective rank of the covariance matrix is chosen as the rank of the minimal subspace required to generate some fixed proportion of the variance in the data. Like the MEE criterion introduced earlier, the CEF method relies upon the assumption that the noise variance is lower than the signal variance but, more crucially, it also relies upon the assumption that the noise variance is some known proportion of the signal variance.

An alternative method that uses the same variance–based sorting criterion as the basic CEF method, but that makes no assumptions about the magnitude of the noise variance, is 'row–wise cross–validation' (CV–RW) (e.g. [41]). The principal behind the CV–RW method is to find the rank of the maximal subspace required to *predict* any data point generated according to this covariance matrix. To find this subspace, a covariance matrix is estimated from a subset of the data where an observation (or a subset of observations) are left out of the sample. Let us refer to this data as $\mathbf{M}^{(-i)}$. This covariance matrix is subjected to an eigenvalue decomposition, and the 'left out' data, $\mathbf{M}^{(i)}$ are estimated at each level, f, of the CEF, according to:

$$\hat{\mathbf{M}}^{(i,f)} = \mathbf{V}^{(-i,f)}\mathbf{V}^{(-i,f)\mathbf{t}}\mathbf{M}^{(i)}, \tag{2.22}$$

where $\mathbf{V}^{(-i,f)}$ is the first f columns of the eigenvector matrix of the covariance matrix of $\mathbf{M}^{(-i)}$. The mean predictive squared error is then calculated at each level of the CEF over all left out sub–sets, and the effective rank is chosen as the value of f that minimizes this error. The problem with this approach, however, is that the left out data are not themselves independent of their estimate, meaning that the predictive error is naturally under–estimated ([12]).

A stricter cross–validation procedure is suggested in [12]. They name it 'Eigenvector cross–validation' (CV–EV) (named after a chemometrics company who uses the method in their software.) The CV–EV method works initially like the CV–RW method — a covariance matrix is estimated from a subset of the data and its eigenvalue decomposition is performed. Then, for

each level f of the CEF, the values at each variable, d, of the left out data is estimated individually from the left out data with only the other variables included:

$$\hat{\mathbf{M}}^{(i,d,f)} = \mathbf{V}^{(-i,d,f)}\mathbf{H}\mathbf{V}^{(-i,-d,f)\mathbf{t}}\mathbf{M}^{(i,-d)}, \tag{2.23}$$

where $\mathbf{H} = \left[\mathbf{V}^{(-i,-d,f)}\mathbf{V}^{(-i,-d,f)\mathbf{t}}\right]^{-1}$ and the superscripts again correspond to either included or left out variables or observations. With this approach, the predictive error is truly independent of the left out data, but it relies upon there being some correlation between the data variables ([12]). If there is no correlation between variables, then the entire space is effectively considered as noise. But also, if there is no correlation between variables at all time points, then we could not hope separate the sources anyway.

Simulations are described in the next section that test these three effective rank estimation methods on single covariance matrices. The MER principal as a means of separating the signal and noise subspaces of the non–stationary blind Gaussian source model are then tested, followed by tests of the two source sorting criteria — MEE and MNsE.

Simulations Simulation were first run to compare the various rank estimation methods for a single covariance matrix. Data were generated for 1000 datasets. Each dataset consisted of multivariate normal data of rank order $R = 5$ with $d = 10$ variables. To test the effect of different eigenvalue distributions on the rank estimation methods, in half of the models 5 eigenvalues were drawn randomly from a uniform distribution on the range $(0, 1]$ while in the other half they were drawn randomly from an exponential distribution with $\lambda = 1$. The remaining 5 were always set to zero, and 10 eigenvectors were created by performing a QR decomposition on a 10×10 matrix with entries drawn randomly from a uniform distribution on the range $(0, 1]$. A model covariance matrix was created, according to $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathbf{t}}$, and 1000 samples were drawn randomly from the corresponding zero–mean multi–variate normal distribution. Finally, zero–mean uncorrelated Gaussian noise was added at 5 different levels — ranging from 1 to 20 percent of the sample variance — providing $1000 \times 5 = 5000$ datasets in total.

Averaged over all models and noise levels, the CV–EV method was the most effective at estimating the true rank of the data (Acc = 74.76%), ahead of CEF (Acc = 43.64%) and CV–RW (Acc = 8.18%). Figure 2.6 displays the accuracy of the three rank estimation methods as a function of noise level. Not surprisingly, CEF outperformed the other two methods when the noise variance equalled exactly its programmed cut–off (90% of the total variance), but CV–EV performed well at most levels of noise. There was little effect of the two eigenvalue distributions. Figure 2.7 shows the probability dis-

tribution over various rank estimates for the three methods. CV–EK only ever underestimates the true rank of the data, while CV–RW systematically over–estimates. As expected, CEF over–estimates when the true noise variance is higher than the programmed threshold and under–estimates when it is lower. Taken as a whole, CV–EV is clearly the best method to use when the variance of the noise is unknown beforehand. The potential for under–estimation of the signal space is preferable to over–estimation and it appears fairly robust to high levels of noise.

Next, simulations testing the two components of the subspace partitioning problem — namely the MER principle for estimation of signal quantity and the MEE and MNsE sorting criteria for identification of a known number of signals — were run for the blind source model. Two factors were considered for each problem: the total expected noise variance, as a proportion of the expected variance for any signal; and the degree of stationarity in the noise, as a proportion of the stationarity in the signals. Stationarity here is defined as it was for the earlier simulations. The basic generative model for these data was different to that used in the previous simulations, as the number of signals was less than the number of observed variables. Consequently, the mixing matrix was no longer square. However, the variance of each signal was again treated as a gamma random variable over time, but the variance of this gamma distribution was held constant over levels of the noise factors and over the different signals. Noise was generated according to a multivariate normal distribution with zero mean and a non-diagonal covariance. To allow for control over its stationarity, a scalar covariance matrix was created first and then transformed into a new arbitrary basis by a random orthogonal matrix. As with the signal variance, the total noise variance over time was treated as a gamma random variable.

For the data testing MEE and MNsE for a known signal quantity, 10 levels of total expected noise variance, between 10% and 100% of the expected variance for any signal, were compared at two levels of relative stationarity (1% and 100% of the stationarity in the signal). 100 datasets were generated for each combination of the two factors, each with $l_T = 100$ time points and $k = 1000$ observations per time point. Each dataset, level, and time point consisted of 6 randomly generated Gaussian signal sources and a $12 \times k$ matrix of noise samples drawn from the multivariate normal distribution described above. The data were generated by first mixing the signal sources and then adding the noise to the result. The same randomly generated 12×6 mixing matrix was used to mix the signals for all levels of a single dataset. Due to the relatively high computational demands of the CV–EV algorithm, only 5 levels of total expected noise variance were compared for the data testing the MER principle for signal quantity estimation, again between 10% and

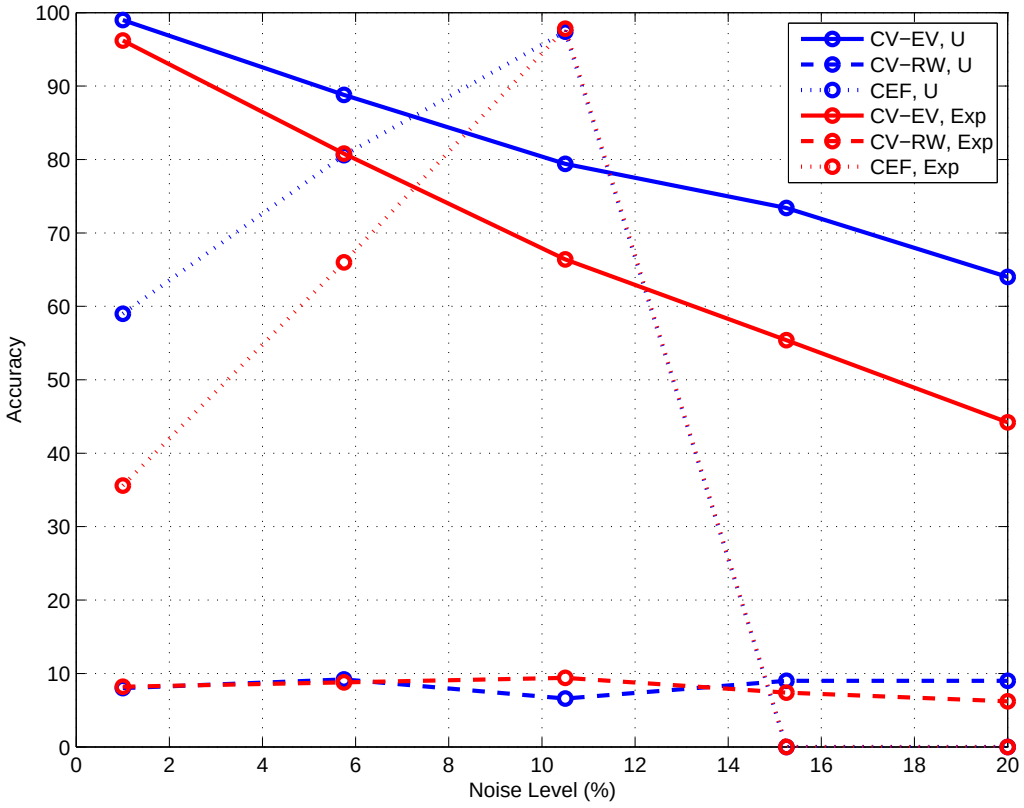


Figure 2.6: Accuracy of the rank estimate of a single covariance matrix as a function of noise variance for three estimation methods. Solid curves — 'Eigenvector' cross–validiation. Dashed curves — row–wise cross–validation. Dotted curved — cumulative energy function. Blue curves — uniformly distributed eigenvalues. Red curves — exponentially distributed eigenvalues. Horizontal axis — noise variance as a percentage of total variance. Vertical axis — rank estimation accuracy.

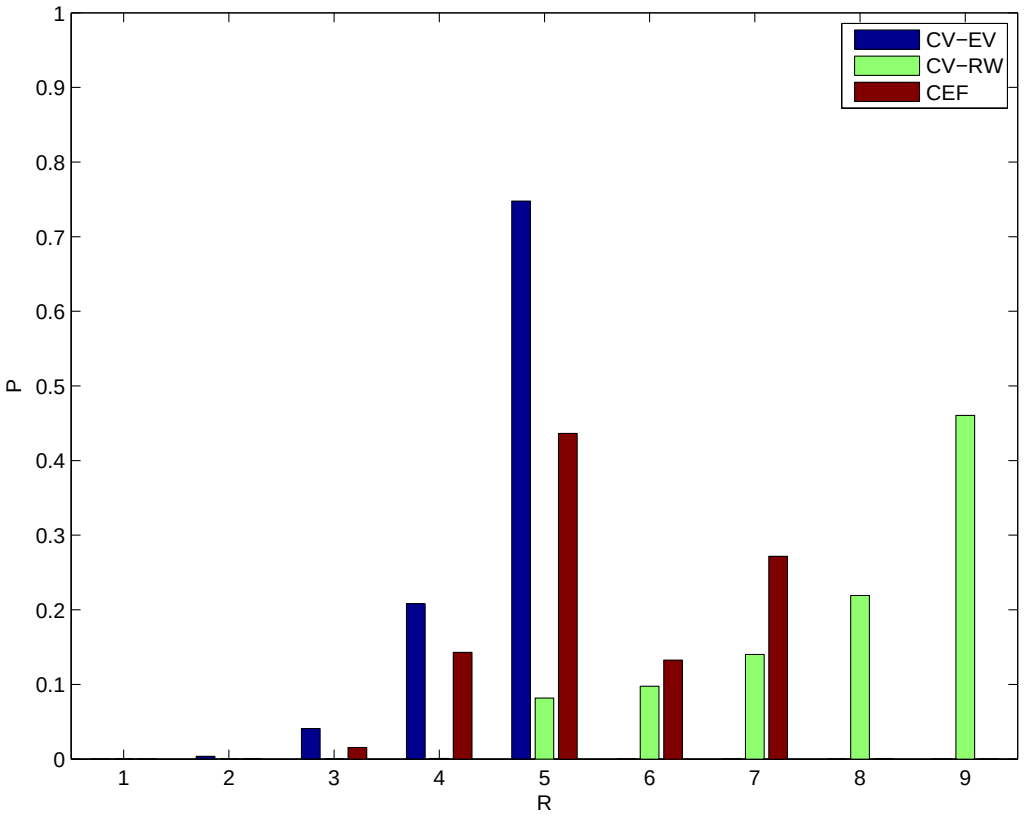


Figure 2.7: Probability of rank estimate of a single covariance matrix for three estimation methods. Blue bars — 'Eigenvector' cross–validation. Green bars — row–wise cross–validation. Red bars — cumulative energy function. Horizontal axis — Rank estimate. Vertical axis — Probability. True rank is 5.

100% of the expected variance for any signal. 50 datasets were generated for each combination of factors, each with $l_T = 10$ time points and $k = 50$ observations per time point. There were 3 sources, a $6 \times k$ matrix of noise samples, and the mixing matrix was 6×3.

Partitioning performance of both the MEE and MNsE criteria was measured with d–prime. D–prime is a measure of sensitivity, taking into account both the proportion of signal sources correctly recognized as such and the proportion of noise sources incorrectly recognized as signal sources. A value of d–prime around +2 indicates that about 90% of signal sources are correctly recognized, while only about 10% of noise sources are incorrectly attributed to the signal. A negative value of d–prime indicates that there are more mis–attributed noise sources than there are correctly assigned signal sources. Figure 2.8 shows the d–prime sensitivity for the two noise conditions (stationary noise, non–stationary noise) over the range of expected noise variance when sorting by the MEE criterion. The average sensitivity in both cases of stationary and non–stationary noise is quite good (d–prime $\simeq 2.5$). For stationary noise, performance drops slightly as the noise variance increases, but not by a great deal. Surprisingly, performance actually improves with increased noise variance when the noise is non–stationary and a significant negative bias appears when the noise variance is around 20% of the signal variance. Figure 2.9 displays the test results for the MNsE criterion. These results follow a similar pattern, but the negative bias towards the noise is present at all levels of noise variance when the noise is non–stationary.

To test the MER principle, estimates of the signal quantity for a each dataset were taken as the maximum effective rank estimate from the CV–EV method for any covariance matrix during the time course of the data. The bar graphs in figures 2.10a and 2.10b each display the average rank estimates for all 5 noise levels, for the stationary and non–stationary noise conditions respectively. These graphs show that the average effective rank estimate for any time point was always close to the signal quantity, but generally slightly lower. Whether the noise was stationary or not made little difference to the temporal distribution of these estimates, but the average estimate tended to be slightly higher in the case of non–stationary noise. This is reflected in the higher accuracy of the signal quantity estimation for non–stationary sources, shown in figure 2.11. Figure 2.11 also shows that the estimation of signal quantity in the non-stationary blind source model is more accurate than the pure estimation of effective rank for a single covariance matrix. This can be attributed to the fact that the effective rank estimator gets more independent 'chances' to guess the noise subspace, of which only the maximum is taken. As the CV–EV method tends to underestimate the rank of the signal subspace, we should expect performance to be superior when

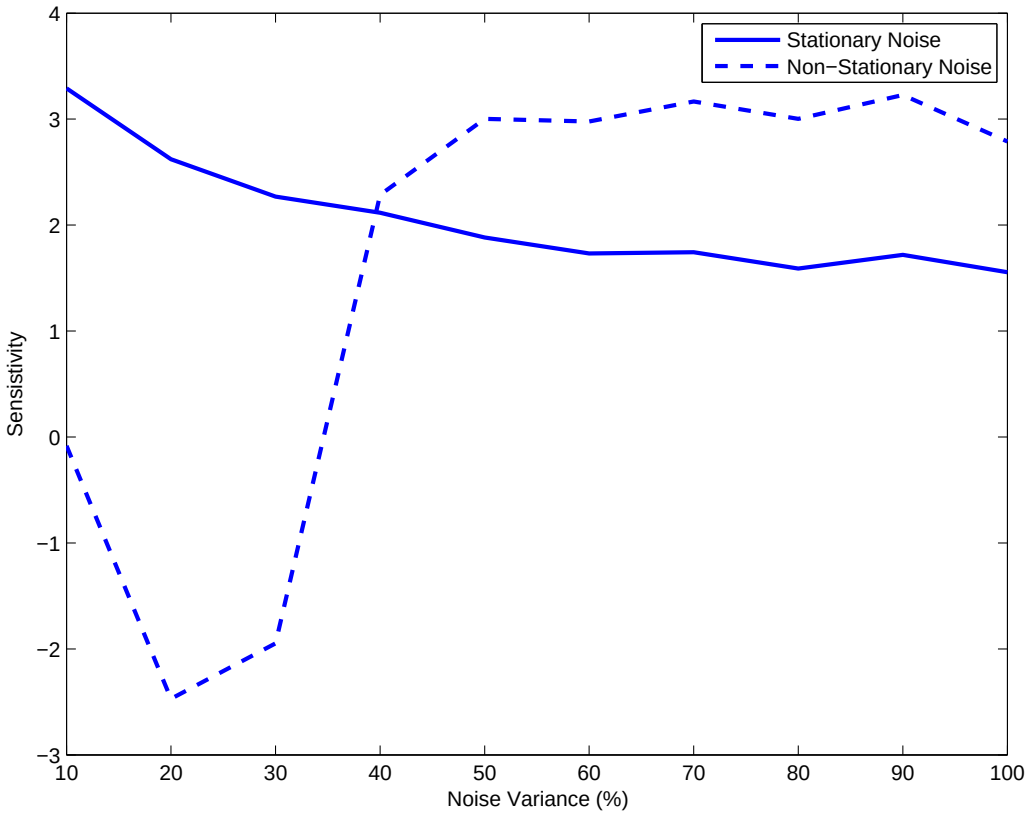


Figure 2.8: Sensitivity to signal sources as a function of noise variance when sorting by mean of estimated source variance (MEE). Solid curve — stationary noise. Dashed curve — non–stationary noise. Horizontal axis — noise variance as a percentage of signal variance. Vertical axis — Sensitivity (d–prime)

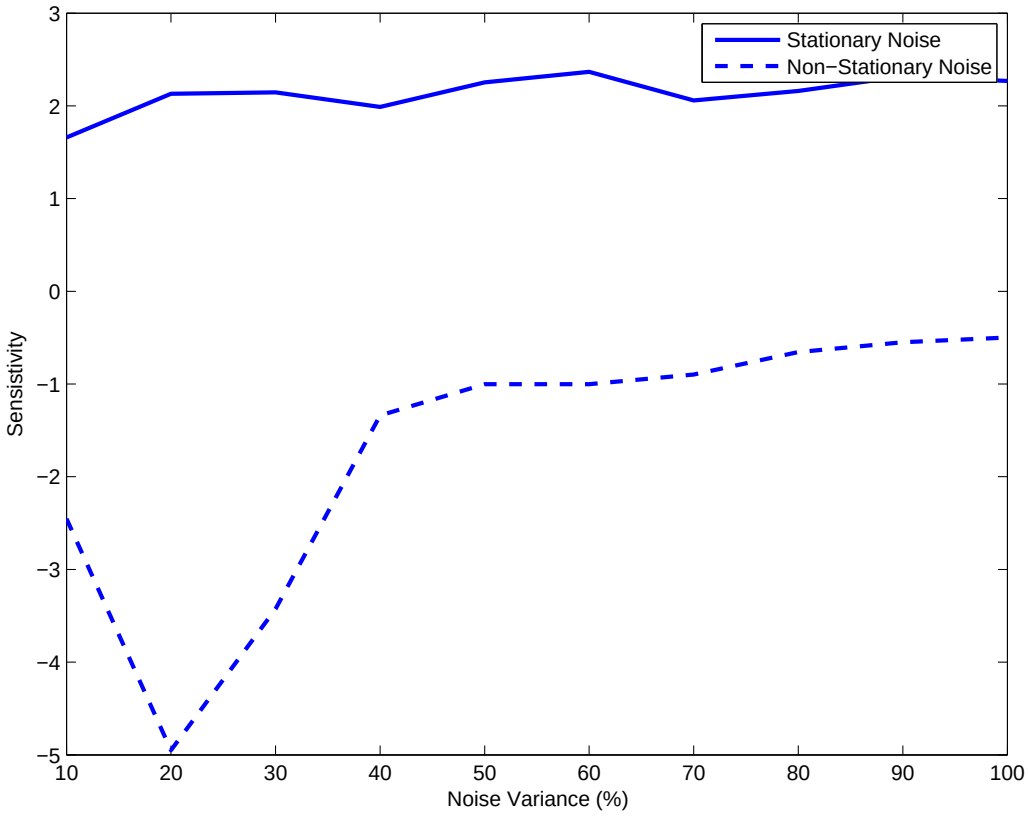
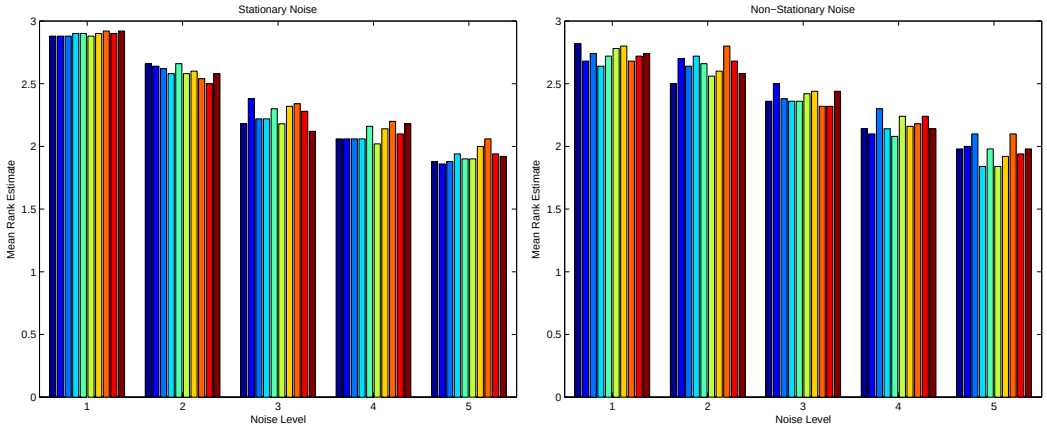


Figure 2.9: Sensitivity to signal sources as a function of noise variance when sorting by variance of estimated source variance (MNsE. Solid curve — stationary noise. Dashed curve — non–stationary noise. Horizontal axis — noise variance as a percentage of signal variance. Vertical axis — Sensitivity (d–prime)



(a) (b)

Figure 2.10: Mean rank estimate over time at each of 5 levels of (a): stationary, (b): non–stationary noise variance between 10% and 100% of signal variance. Horizontal axis — noise level. Vertical axis — mean rank estimate. Sub–bars — time points. Number of signal sources was 3.

the maximum effective rank out of multiple covariance matrices is used.

Taken together, the results of the above simulations suggest that, while the the presence of non–stationary noise does appear to improve the estimation of signal quantity, it also induces a large error in the estimation of signal identity. Non–stationary noise should therefore be minimized if possible.

EEG Data

EEG data were taken from a Bubbles experiment published by Schyns et al. in 2007 [63]. On each trial of the experiment, subjects were asked to classify a static Bubble–masked face image as displaying either one of the 6 universal emotions ['Happy' (H); 'Fear' (F); 'Surprise' (Su); 'Disgust' (D); 'Anger' (A); or 'Sadness' (Sa)], or 'Neutral' (N).

Some details on the authors' recording set–up (taken from the original paper): *EEG data were recorded on sintered Ag/AgCl electrodes mounted in a* 62*–electrode cap (Easy–Cap) at scalp positions including the standard 10–20 system positions along with intermediate positions and an additional row of low occipital electrodes. Vertical electro-oculogram (vEOG) was bipolarly registered above and below the dominant eye, and the horizontal electro-oculogram (hEOG) was registered at the outer canthi of both eyes. Electrode impedance was maintained below 10 kU throughout recording. Electrical activity was continuously sampled at* 1024 *Hz. Analysis epochs were generated*

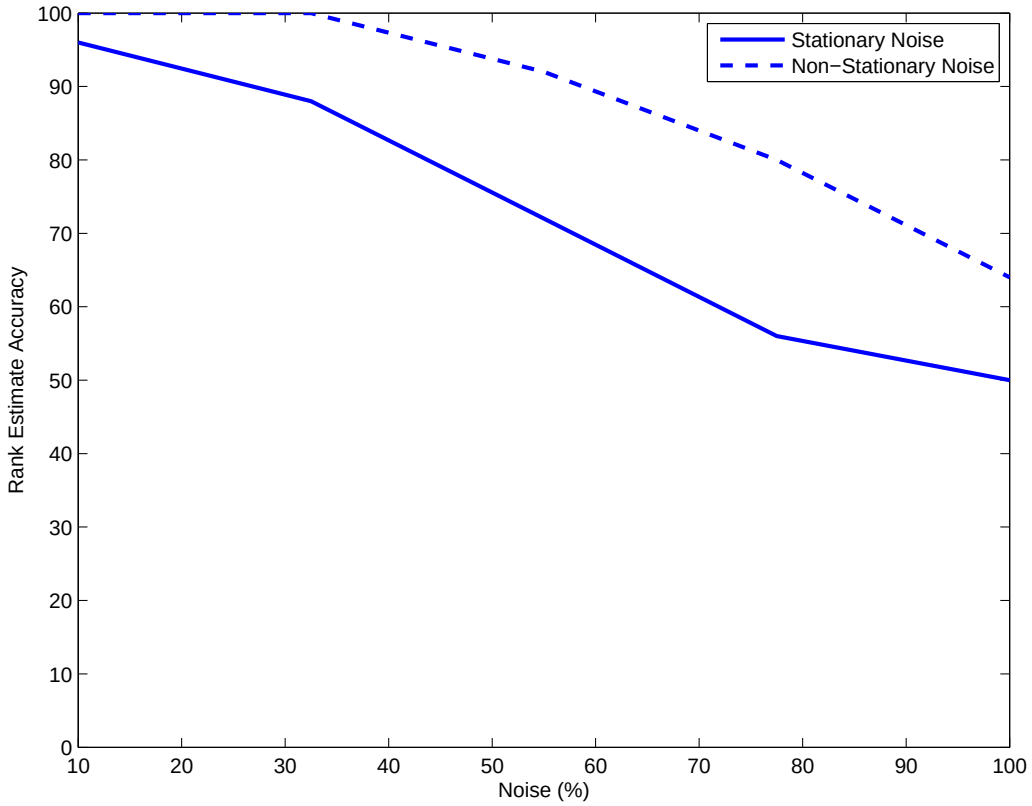


Figure 2.11: Accuracy of number of signal source, estimated as the maximum rank of a single covaraince matrix, plotted as a function of noise variance. Solid curve — stationary noise. Dashed curve — non–stationary noise. Horizontal axis — Noise variance as a percentage of signal variance. Vertical axis — Accuracy.

offline, beginning 500 ms prior to stimulus onset and lasting for 1500 *ms in total. EEG and EOG artefacts were rejected by using a* $[230mV; +30mV]$ *deviation threshold over* 200 *ms intervals on all electrodes. The EOG rejection procedure rejected rotations of the eyeball from* 0.9^o *inward to* 1.5^o *downward of visual angle - the stimulus spanned* $5.36^o \times 3.7^o$ *of visual angle on the screen.*

Pre–processing

Prior to running the JD–BGL algorithm on the EEG data, some additional preprocessing was performed:

1. Referencing
2. Baseline correction
3. Detrending

The most common method for referencing EEG data is the average reference. The argument for the average reference [51] relies upon two assumptions — the first quite a strong one, but the second less so. It first assumes that EEG potentials measured on the scalp are a result of the action of one or more current dipoles in a homogeneous spherical conducting medium. In this situation, the integral of the electrical potential over the surface of the sphere should be 0. Second, it assumes that EEG measurements are taken across the whole surface of this sphere. If both assumptions are correct then the potential should sum to 0 over all measurement channels and, if it does not, then the measurement channels are incorrectly referenced and they should be re–referenced in such a way that it does. This re–referencing corresponds to the removal from each channel of the mean measurement over all channels, and can be expressed as a linear transformation by the average reference operator, $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}^{\mathbf{t}}/e$.

The first problem with the second assumption is that the head is not actually spherical and the brain does not constitute a homogeneous conducting medium, but we can assume that these are close enough to being true to ignore them for now. The main problem with it is that EEG potentials are not measured over the full surface of the sphere. At best, they are measured over not much more than half of the sphere – the top half that makes up the scalp. If the potential over the top half of the sphere does not sum to 0, this does not necessarily mean that the electrodes are incorrectly referenced — it may simply mean that the potential over the bottom half of the sphere is opposite. Forcing the summed potential measurements to be 0 in this case

might actually distort the true picture, as a globally positive potential will be lowered and a globally negative potential will be raised, even though both of these potentials are valid for the top half of the sphere. This will, in turn, force the summed potential over the rest of the sphere to be neutral, which is certainly possible but unlikely to always be true.

The best solution to this problem is probably to use a local surface Laplacian filter to estimate the current densities at each region immediately below the scalp (e.g. [4]), but this approach was deemed beyond the scope of the thesis. An simpler alternative to using the average reference is to use a fixed EEG measurement channel, or set of channels, outwith the data array as reference. Any artefactual changes in potential over time caused by external factors, such as the amplifier for example, should be reflected equally in these electrodes and hence removed after referencing. They do, however, introduce their owns sources of noise and, perhaps more importantly, bias. Linked mastoid electrodes were used here as reference.

Although the fixed electrode reference should remove any any external artefactual potential changes that are also measured on the reference source, any internal potential differences between pairs of electrodes will probably not be captured by this source (nor by the average reference, for that matter) and hence remain in the data. One way to remove these is by baseline correction. Baseline correction assumes that the expected electrical potential measurement over a given time window should be zero for all channels, as long as that time window contains no experimental stimulus. Removing from each time point of the EEG measurement the average potential over the pre–stimulus interval to satisfy this assumption should hopefully remove any artefactual potential differences between pairs of electrodes. A final step is the removal of any linear trend from the EEG data at each electrode. This step is crucial if we want to minimize the risk of non–stationary noise in our data, as any two electrodes with opposing trends will tend to amplify the noise variance over time. Detrending was carried out under the assumption that any linear trend in the data is already present during the pre–stimulus time window. A regression line was fit to this window for each trial and electrode, and the full time course of activity for the electrode was shifted by the projected path of this line.

Algorithmic details

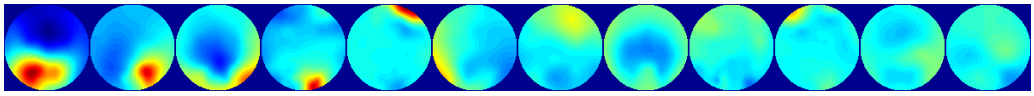
The JD–BGL algorithm was implemented in Matlab. The algorithm was terminated either upon convergence of the source estimate likelihood or after 500 iterations. In practise, the algorithm always converged within the 500 iteration limit. Joint diagonalization was performed independently per subject

and per expression (expression shown to the subject, rather than expression perceived by the subject) over the full time course of EEG measurement (from $-500ms$ to $1024ms$). All 58 channels of the 62 channel EEG set–up corresponding to scalp locations were included in the analysis (of the remaining 4, 2 are linked mastoid reference electrodes and 2 are EOG electrodes). Effective rank estimation was performed with the CV–EV algorithm. Due to the high number of trials at each time point, k–fold cross-validation was preferred over 'leave–one–out' cross-validation. 5 folds were used. Estimated sources were scaled to make each column of $\hat{\mathbf{A}}$ unit length and were then sorted by the MEE criterion. The scaled topographical distribution of a given source across the scalp surface is contained in its corresponding column of $\hat{\mathbf{A}}$. This distribution was interpolated over a physiscal scalp model containing the electrode coordinates for presentation purposes. The assumption of Gaussianity in the recovered sources was tested with the Jarque–Bera test [31].

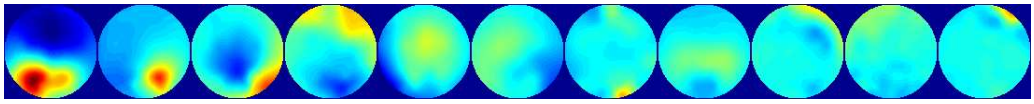
Results & Discussion

Results are discussed for all three subjects and two example expressions — 'H' and 'F'. The effective rank varied little over time, subject, or expression. The maximum effective ranks for subject 'LP' were 12 and 11 for expressions 'H' and 'F' respectively; for subject 'LF', 12 and 12; and for subject 'UM', 15 and 15. Figures 2.12a and 2.12b display topographic maps of the estimated signal columns of the mixing matrix for subject 'LP'. Figures 2.13a and 2.13b, and 2.14a and 2.14b display the equivalent results for subjects 'LF' and 'UM' respectively. The maps are weighted by the expected variance of their corresponding sources over time.

The top three or four source maps are remarkably consistent across subjects and expressions. Most contain a left and a right lateralized temporal–parietal source, corresponding roughly to electrodes 'P3' and 'P4', and a bi–lateral occipito–temporal source, corresponding roughly to electrodes 'P7' and 'P8'. The left temporal–parietal activation can be seen in the first source for subject 'LP' for both expressions, in the second source for subject 'LF' for both expressions, and in the first and third source for subject 'UM' for expression 'H' and 'F' respectively. The right temporal–parietal activation can be seen in the second source for subject 'LP' for both expressions, in the first source for subject 'LF' for both expressions, and in the second source for subject 'UM' for both expressions. The bi–lateral occipito–temporal source can be seen in the third source for subject 'LP' for both expressions, in the fourth source for subject 'LF' for expression 'F' (and possibly in the fifth source for expression 'H'), and in the third and first sources for subject 'UM' for ex-

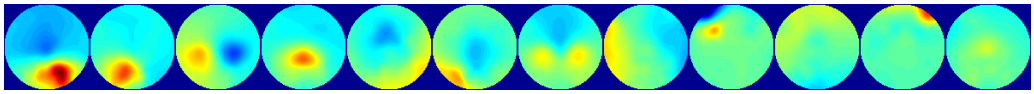


(a)

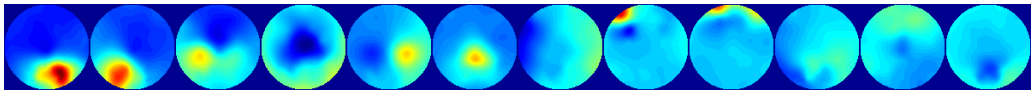


(b)

Figure 2.12: Mixing vectors for each of the signal sources for subject 'LP' and two expressions. Vectors are displayed as topologies over the scalp and sorted and weighted by decreasing expected variance. (a): Trials where the subject was shown a 'Happy' face. (b): Trials where subject was shown a 'Fear' face.

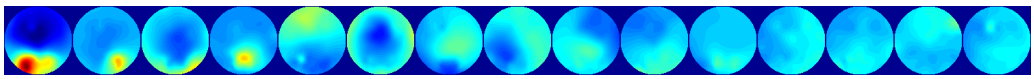


(a)

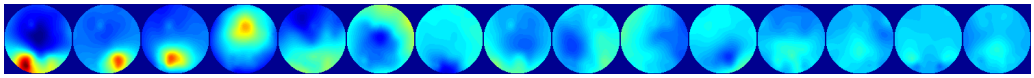


(b)

Figure 2.13: Mixing vectors for each of the signal sources for subject 'LF' and two expressions. Vectors are displayed as topologies over the scalp and sorted and weighted by decreasing expected variance. (a): Trials where the subject was shown a 'Happy' face. (b): Trials where subject was shown a 'Fear' face.



(a)



(b)

Figure 2.14: Mixing vectors for each of the signal sources for subject 'UM' and two expressions. Vectors are displayed as topologies over the scalp and sorted and weighted by decreasing expected variance. (a): Trials where the subject was shown a 'Happy' face. (b): Trials where subject was shown a 'Fear' face.

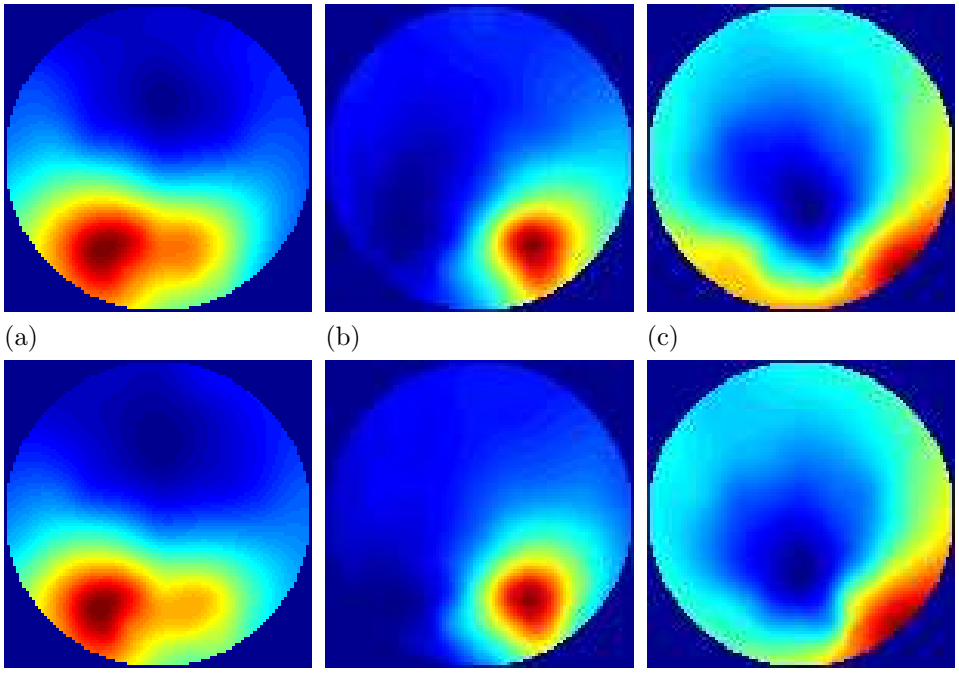


Figure 2.15: Mixing vectors for the three highlighted sources for subject 'LP' and two expressions. Vectors are displayed as topologies over the scalp. Top row: 'Happy' trials'. Bottom row: 'Fear' trials. (a), (d): Left temporal–parietal source. (b), (e): Right temporal–parietal source. (c), (f): Occipito–temporal source

pression 'H' and 'F' respectively. In all cases, although active bi–laterally, the occipito–temporal source is slightly biased towards either the right or left. Figures 2.15a – 2.15c and 2.15d – 2.15f display these source topologies for subject 'LP' and expressions 'H' and 'F' respectively, while Figures 2.16a – 2.16f and 2.17a – 2.17f display the same results for subjects 'LF' and 'UM' respectively.

The electrodes highlighted by these topologies are all commonly implicated in visual classification studies and, in the case of the occipto–temporal electrodes, in facial perception studies in particular. The occipito–temporal electrodes are the site of the famous 'N170' ERP negative bias towards face stimuli [10] at $170ms$ following stimulus onset. This is also the time window of general visual encoding processes reported previously. Recent studies have shown an association between the N170 ERP component and the location of the 'Fusiform Face Area' (FFA) in the context of a facial classification task [30][61][27][24][20]. The FFA is a region in the fusiform gyrus of the temporal

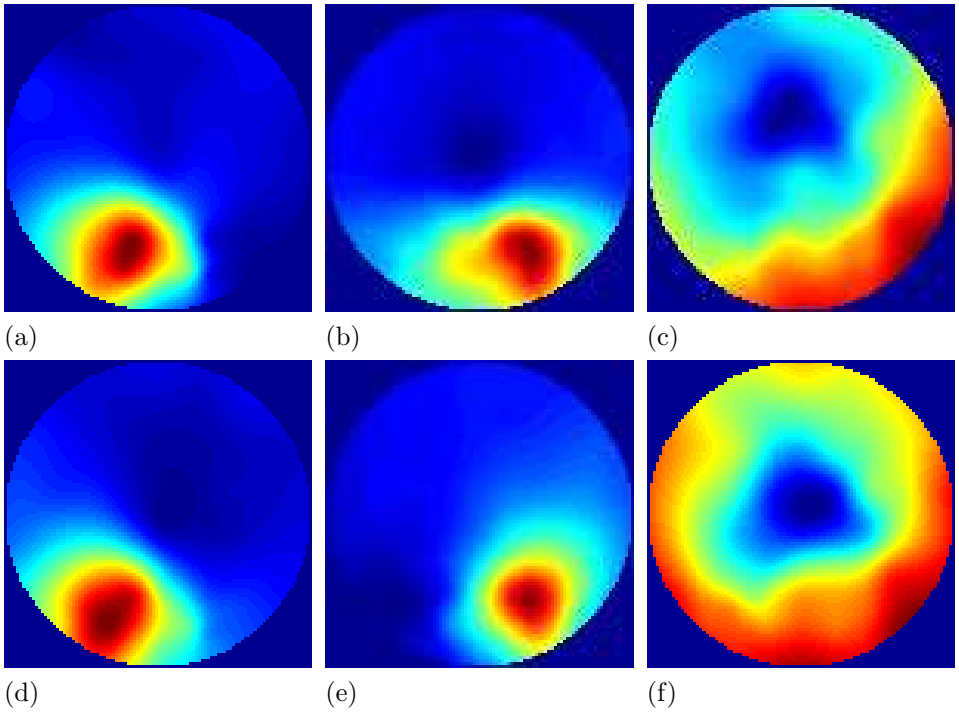


Figure 2.16: Mixing vectors for the three highlighted sources for subject 'LF' and two expressions. Vectors are displayed as topologies over the scalp. Top row: 'Happy' trials'. Bottom row: 'Fear' trials. (a), (d): Left temporal–parietal source. (b), (e): Right temporal–parietal source. (c), (f): Occipito–temporal source

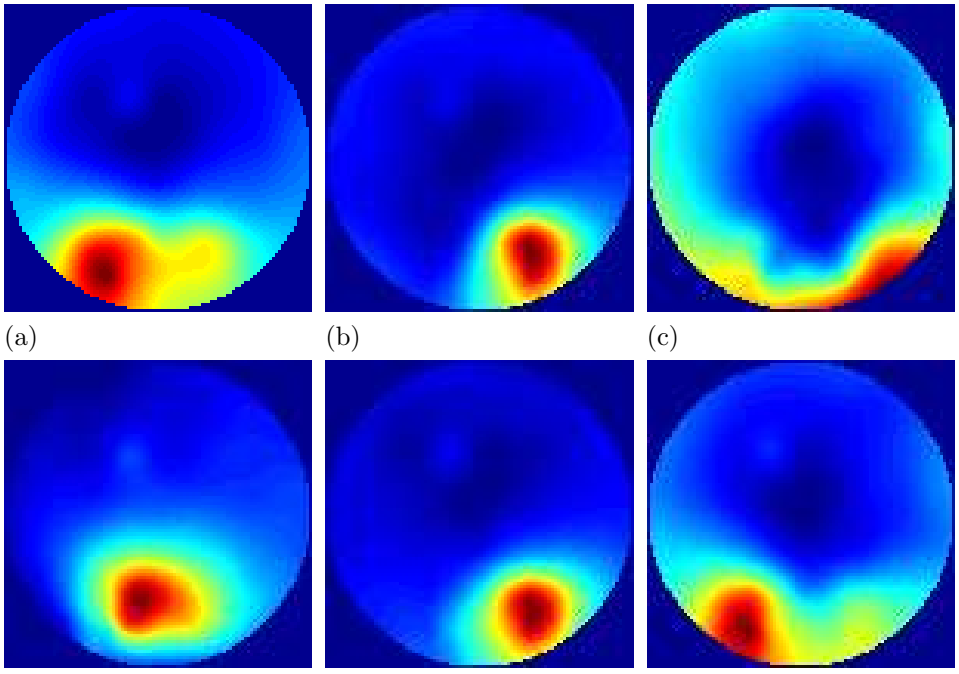


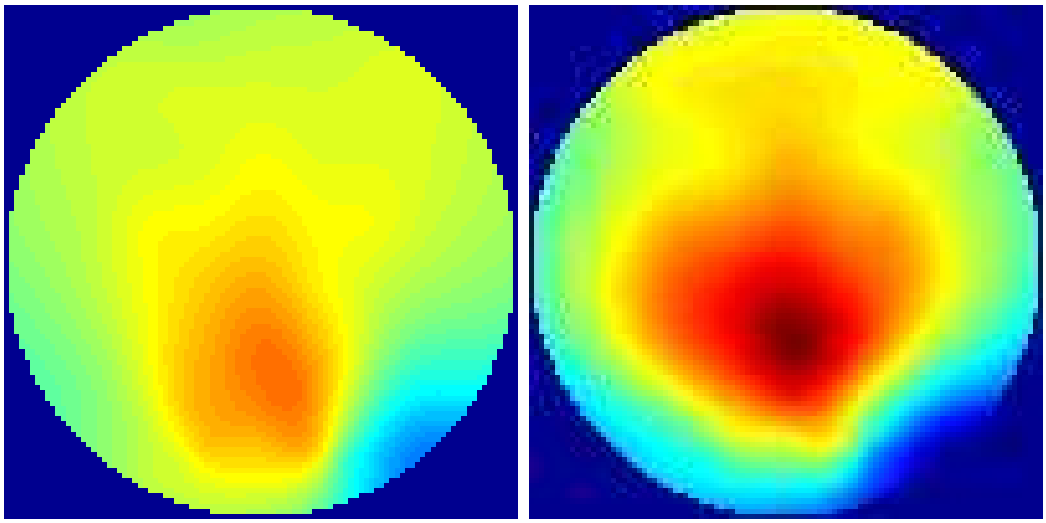
Figure 2.17: Mixing vectors for the three highlighted sources for subject 'UM' and two expressions. Vectors are displayed as topologies over the scalp. Top row: 'Happy' trials'. Bottom row: 'Fear' trials. (a), (d): Left temporal–parietal source. (b), (e): Right temporal–parietal source. (c), (f): Occipito–temporal source

lobe that has been associated with face processing through both fMRI [38] and lesion–based [6] studies. The occipito–temporal sources found here also appear to contain potentials of opposite valency at central/parietal locations, corresponding roughly to electrodes 'Cz/CPz'. This is the site of another famous face–specific ERP effect — the 'VPP' [32][33]. It has been suggested that the VPP and N170 effects are generated by the same underlying process in the brain [36], so we should expect to find them both represented in a single independent source.

Parietal potentials in the EEG tend to be associated with the 'P300' effect [70], which occurs, at $300ms$ following the onset of the stimulus. Although the original study associated the effect with unpredictability in the stimulus, and assumed a single cortical source, it has since been associated with two sources — one (frontal) related to top–down attention mechanisms and the other (temporal–parietal) related to memory, [58]. Given the nature of the task, where incoming stimuli are matched to internal memorial representations, coupled with the relatively low top–down attentional demands of recognizing emotion from the same set of faces thousands of times, we should expect to see high average variance in the temporal–parietal component of the 'P300', but not so much in the frontal component. To allow comparison with the highlighted sources, Figures 2.18a and 2.18b, and Figures 2.18c and 2.18d display the scalp ERP topologies measured from subject 'LP' at $170ms$ and $300ms$ following stimulus onset for expressions 'H' and 'F' respectively. Figures 2.19a – 2.19d and Figures 2.20a – 2.20d show the same results for subject 'LF' and 'UM' respectively.

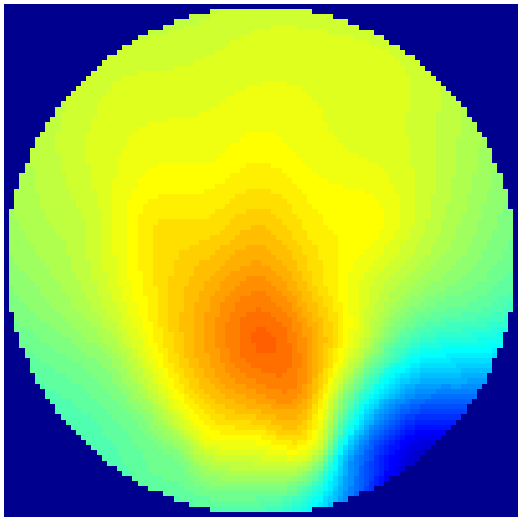
Confirming that the highlighted sources meet with the assumption of Gaussianity in the signal model, Figures 2.21a and 2.21b display the time course of the Jarque–Bera statistic in the three sources for subject 'LP' and expressions 'H' and 'F' respectively. Figures 2.22a and 2.22b and figures 2.23a and 2.23b display the corresponding data for subjects 'LF' and 'UM' respectively. Values of the statistic below the criterion value indicate that the null hypothesis of a Gaussian distribution cannot be ruled out.

Figures 2.24a and 2.24b display the time course of the estimated variance in the three previously highlighted signal sources for subject 'LP' and expressions 'H' and 'F' respectively. Figures 2.25a and 2.25b and figures 2.26a and 2.26b display the corresponding data for subjects 'LF' and 'UM' respectively. All graphs are plotted from stimulus onset ($0ms$) to $400ms$ following stimulus onset. These variance profiles tend to be characterised by at least two peaks in time — the first occurring early, at around $50ms$ following stimulus onset; and the second occurring at around $150 - 180ms$ following stimulus onset. The two temporal–parietal sources appear to act synchronously, while the peaks in the occipito–temporal source variance are perhaps just offset

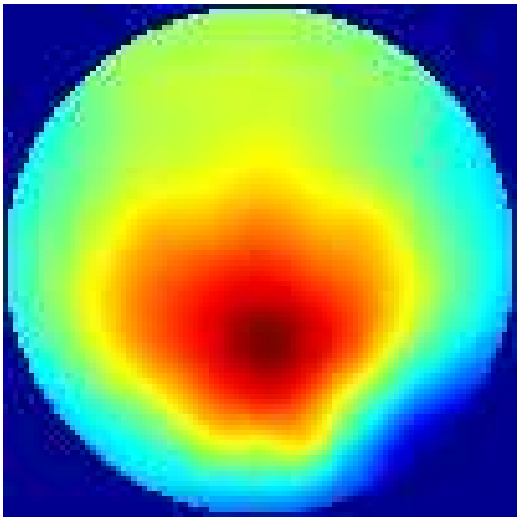


(a)

(b)

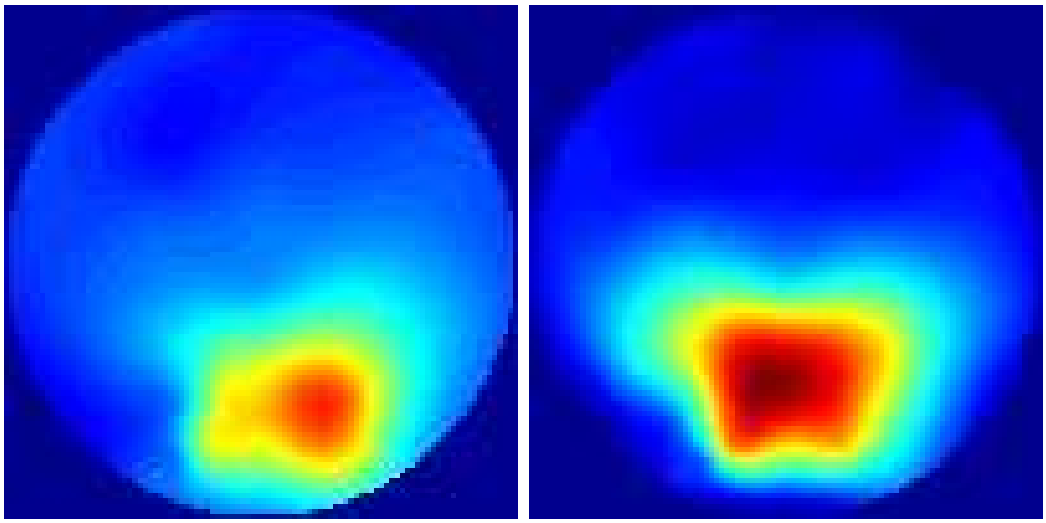


(c)



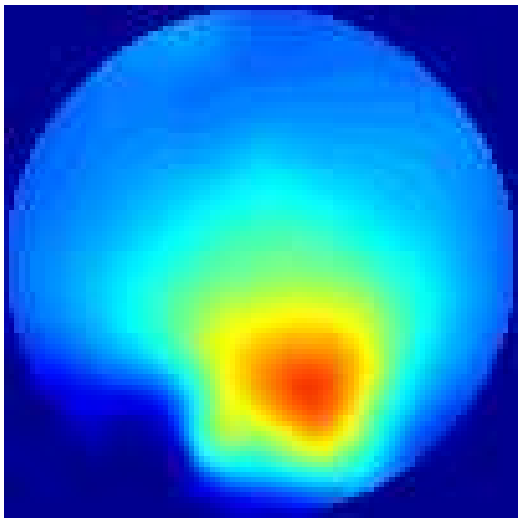
(d)

Figure 2.18: ERP scalp topologies for subject 'LP' during the time windows of the N170 and P300 components. Top row: 'Happy' trials'. Bottom row: 'Fear' trials. (a), (c): N170. (b), (d): P300.

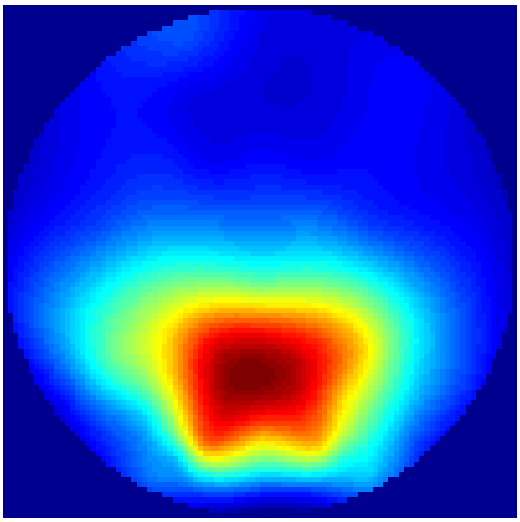


(a)

(b)

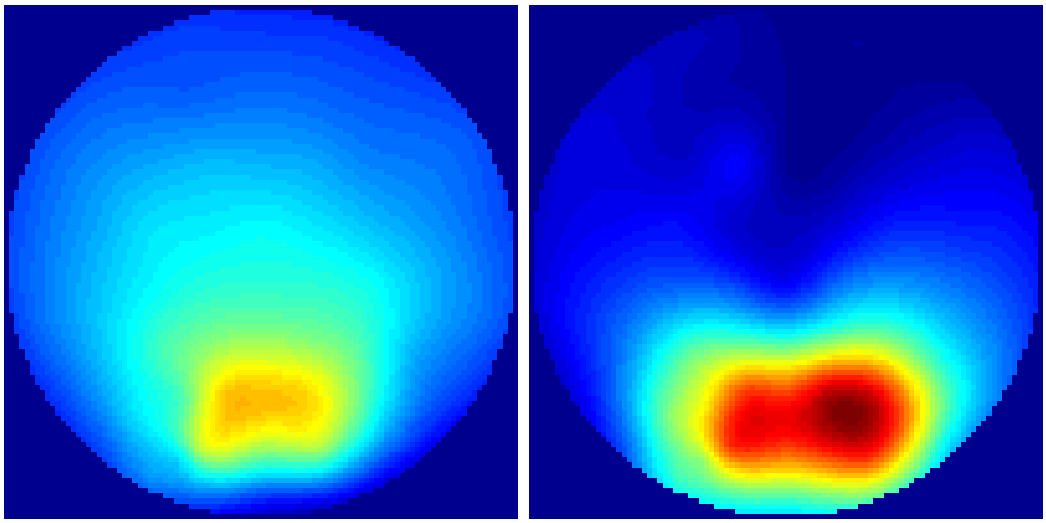


(c)



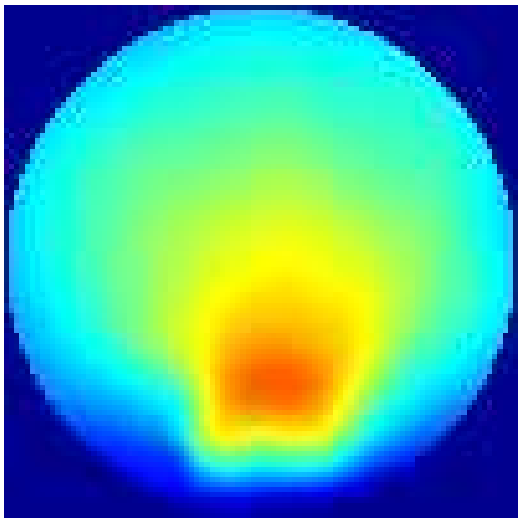
(d)

Figure 2.19: ERP scalp topologies for subject 'LF' during the time windows of the N170 and P300 components. Top row: 'Happy' trials'. Bottom row: 'Fear' trials. (a), (c): N170. (b), (d): P300.

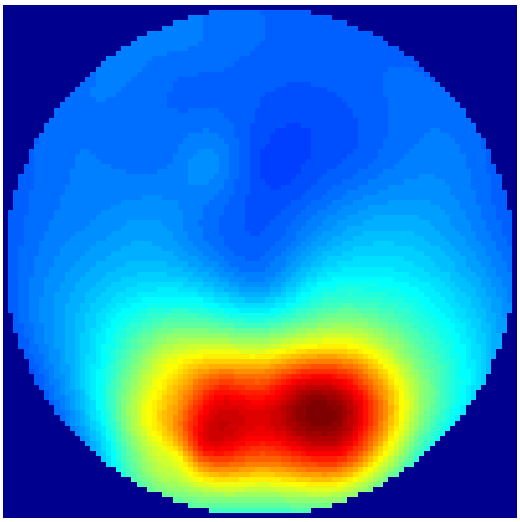


(a)

(b)

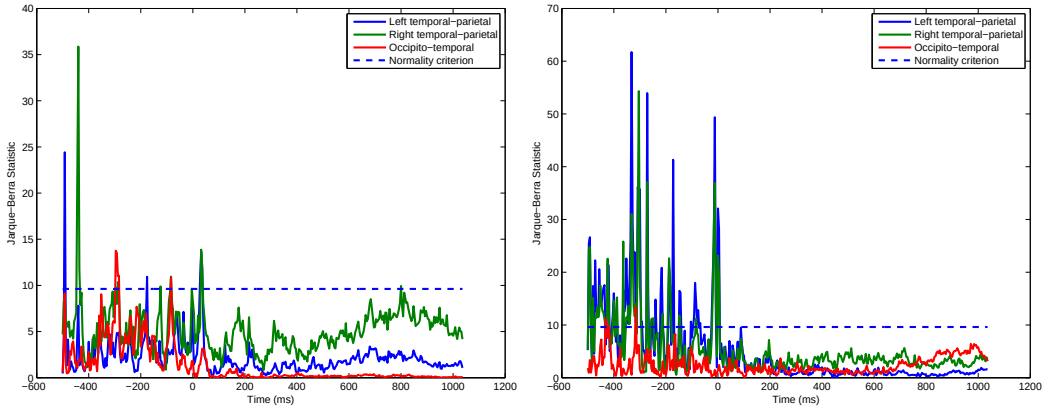


(c)



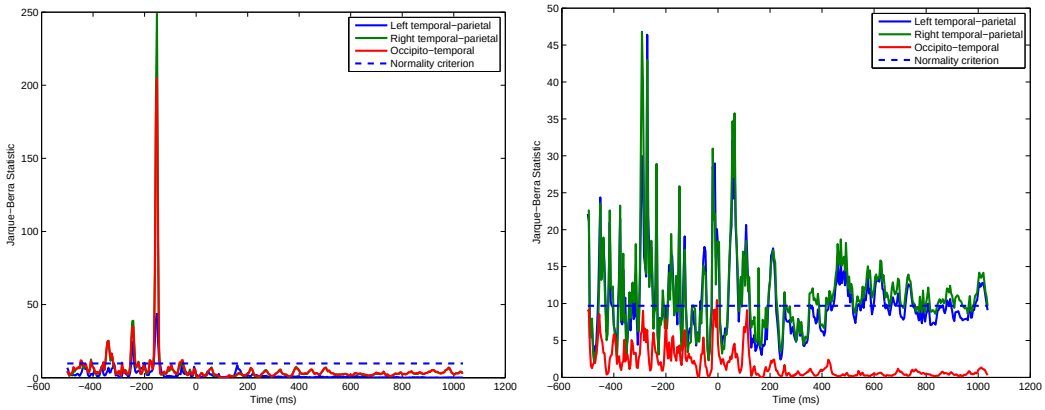
(d)

Figure 2.20: ERP scalp topologies for subject 'UM' during the time windows of the N170 and P300 components. Top row: 'Happy' trials'. Bottom row: 'Fear' trials. (a), (c): N170. (b), (d): P300.



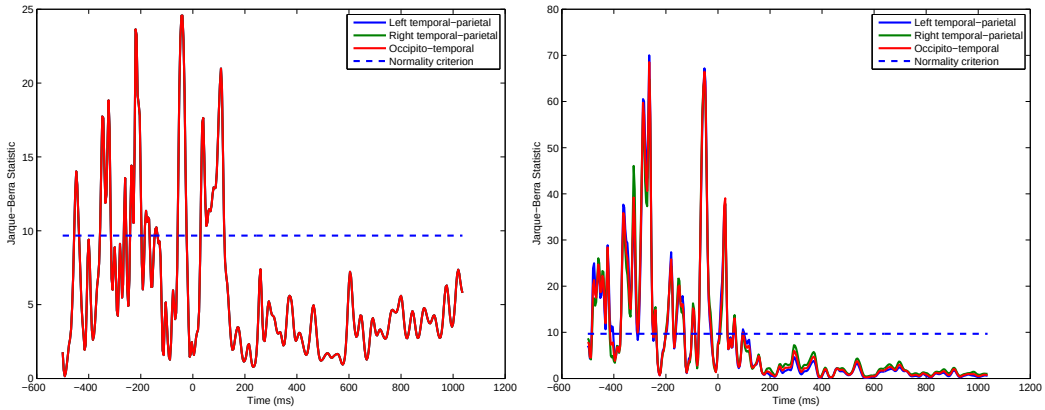
(a) (b)

Figure 2.21: Jarque–Bera test statistic for subject 'LP' and two expressions. Statistic is plotted for the two temporal–parietal sources and the occipito–temporal source. Values below the dotted criterion line indicate normality. (a): Trials where the subject was shown a 'Happy' face. (b): Trials where subject was shown a 'Fear' face.



(a) (b)

Figure 2.22: Jarque–Bera test statistic for subject 'LF' and two expressions. Statistic is plotted for the two temporal–parietal sources and the occipito–temporal source. Values below the dotted criterion line indicate normality. (a): Trials where the subject was shown a 'Happy' face. (b): Trials where subject was shown a 'Fear' face.



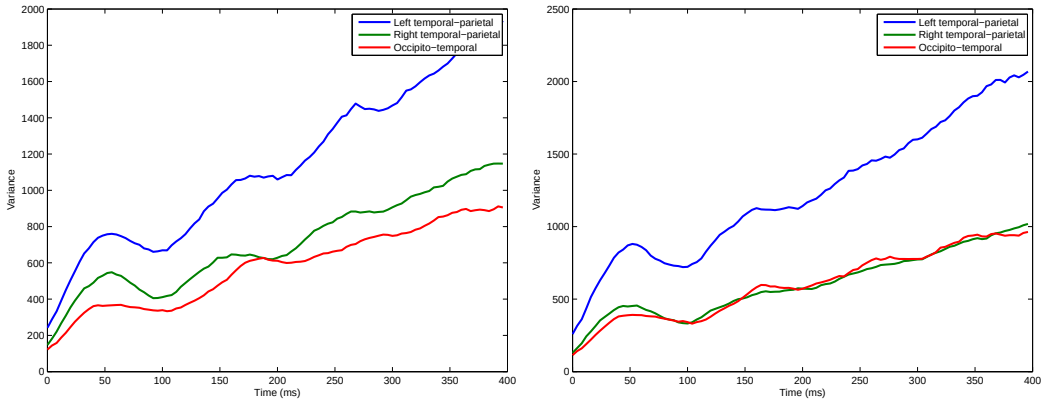
(a) (b)

Figure 2.23: Jarque–Bera test statistic for subject 'UM' and two expressions. Statistic is plotted for the two temporal–parietal sources and the occipito–temporal source. Values below the dotted criterion line indicate normality. (a): Trials where the subject was shown a 'Happy' face. (b): Trials where subject was shown a 'Fear' face.

slightly, and are absent completely in the case of subject 'LF'.

The existence of the second systematic peak in these variance profiles is not unexpected. Given its timing and given the stimuli that are being presented, this fits naturally with the existing literature on the N170 ERP effect. However, the fact that this peak is almost always present in all three of these sources suggests that more than a single process is underlying the N170 effect. The existence of a systematic peak in the variance profiles as early as $50ms$ is more surprising, but has been previously reported as result of low–level stimulus properties, such as luminance [25]. Although the unmasked stimuli were controlled for luminance and contrast, the effect of the random masking would be to alter these properties of the stimulus over the experiment, hence producing a degree of variance in the signal at this time.

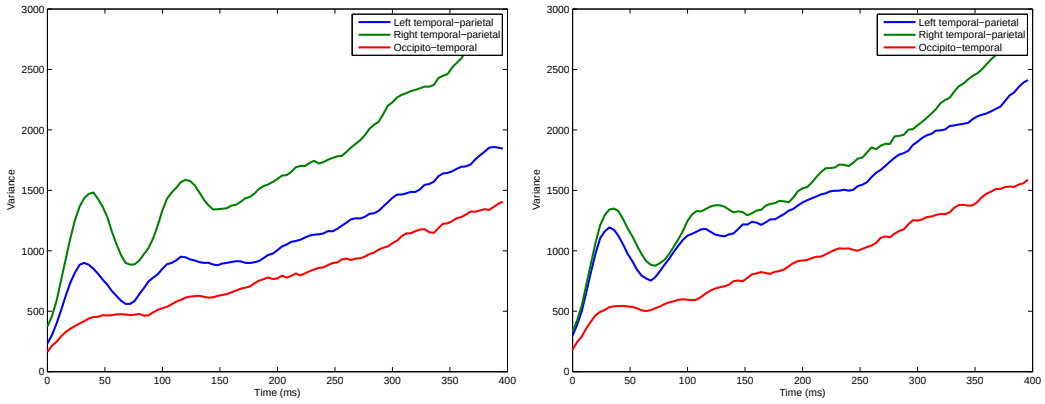
In summary, the estimated signal sources, at least those with the highest average variance, are consistent across multiple subjects and conditions of the experiment. Furthermore, their distributions over the scalp correspond to well studied ERP topologies, their variance profiles over time fit known events in the literature, and the assumption of their normality is justified in 5 out of the 6 cases. The next logical step would be to attempt to localize the cortical generators that might be associated with these sources. The methods that exist to perform this localization (see [46] for a methodological overview) fall under the category of the 'theory–driven' models discussed



(a) (b)

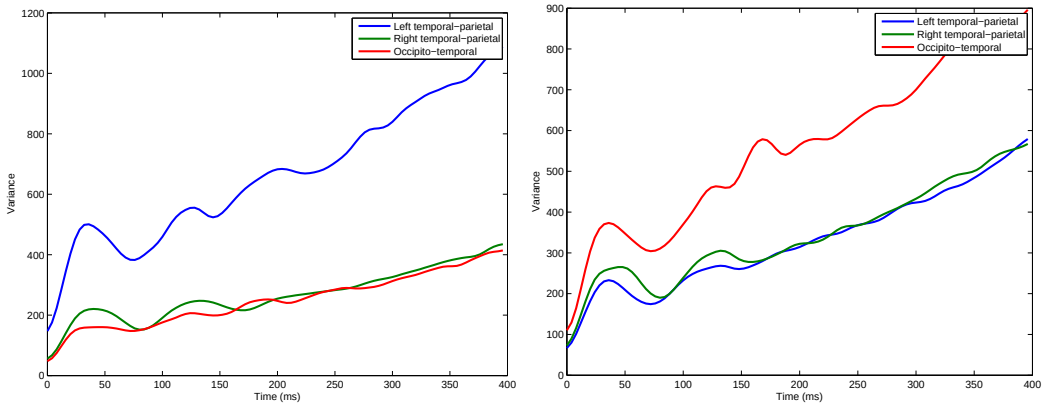
Figure 2.24: Post–stimulus source variance estimate for subject 'LP' and two expressions. Variance is plotted for the two temporal–parietal sources and the occipito–temporal source. (a): Trials where the subject was shown a 'Happy' face. (b): Trials where subject was shown a 'Fear' face.

earlier. An interesting approach might be to combine the maximum independence constraint of the blind source model with other, physiologically motivated, constraints during the estimation process. However, this step was deemed beyond the scope of this thesis, and so the cortical interpretation of these results remains, unfortunately, somewhat speculative.



(a) (b)

Figure 2.25: Post–stimulus source variance estimate for subject 'LF' and two expressions. Variance is plotted for the two temporal–parietal sources and the occipito–temporal source. (a): Trials where the subject was shown a 'Happy' face. (b): Trials where subject was shown a 'Fear' face.



(a) (b)

Figure 2.26: Post–stimulus source variance estimate for subject 'UM' and two expressions. Variance is plotted for the two temporal–parietal sources and the occipito–temporal source. (a): Trials where the subject was shown a 'Happy' face. (b): Trials where subject was shown a 'Fear' face.

Chapter 3

Relating the signal to the information

Introduction

So far we have produced a model of the observed data in terms of some set of maximally independent sources, and we have produced estimates of which of these sources are signal and which are noise. Now that we have this model of the observed responses, our goal is to fit the output of the model to the subject's task performance and to the feature space of the visual input. The estimated model parameters seem to correspond to sensible subspaces of the measured responses, but the true test of the model lies in how well these parameters correspond to real world events that are presumed to be either driving or driven by the brain processes that it represents.

Fitting to the behaviour

The first step is to decide how relevant each of these model parameters are to the output behaviour of the subject during the task. The subject responded by selecting one of the seven possible emotional categories, so the behavioural data are, in theory, of the multi–category classification type described in the introduction. However, incorrect responses were not specifically encoded during the original experiment. Each trial was recorded as either 'incorrect', if the subject responded incorrectly, or as 'N', 'H', 'Su', etc... if the subject responded correctly. Unfortunately, the nature of the incorrect responses must be known if we are to analyse the data in terms of a multi–category classification, so they were instead analysed in terms of a set of independent binary classifications (e.g. 'H' or 'not H'; 'F' or 'not F', etc...). While it

would have been preferable to analyse the full classification problem, this is still OK and corresponds to the behavioural Bubbles analysis in the original paper from which these data were taken.

Logistic regression

Logistic regression was chosen as the method to fit the brain signals to the binary classification data. Under the right conditions, logistic regression is equivalent to a Naive Bayes classifier [39], and should achieve lower error in the case of a large set of data [50]. Furthermore, logistic regression analysis has successfully been applied to classification problems using EEG data in the past ([21]). The basic model for the logistic regression analysis here is that the probability of the subject's response, y_i, being 'correct' on any trial i is a logistic function of the independent signal sources at each time point on this trial, such that:

$$P(y_i = \text{'correct'}|\mathbf{s}_i) = \frac{1}{1 + e^{-f(\mathbf{s}_i(t),t)}}, \tag{3.1}$$

where:

$$f(\mathbf{s}_i(t), t) = \beta_1(t)s_{i1}(t) + \ldots + \beta_{l_B}(t)s_{il_B}(t) + c(t) \tag{3.2}$$

is a continuous linear function of the source activations on trial i and a vector of time–varying parameters, $\boldsymbol{\beta}(t)$. The logistic function can be understood as effectively 'squashing' this continuous function of the source variables into a binary output variable, while the parameters of the continuous function can be understood as the influence of each source variable on the outcome of this process. Large positive values of $\beta_j(t)$ imply that large positive values of source s_{ij} at time t tend to correspond to 'correct' in the output variable, while large negative values of $\beta_j(t)$ imply the opposite, and a value of zero implies that the best guess for the value of the output variable is basically 50/50, regardless of the value of the source.

Model parameters were estimated using the 'Iteratively Re–weighted Least Squares' (IRLS) algorithm of McCullach & Nelder [42]. Once fit, model parameters were evaluated using Cameron & Windmeijer's R^2 formulation for a Bernoulli distribution [15]:

$$R^2 = 1 - \frac{1}{k}\frac{\sum_{i=1}^{k} \hat{y}_i \log(\hat{y}_i) + (1 - \hat{y}_i)\log(1 - \hat{y}_i)}{\bar{y}\log(\bar{y}) + (1 - \bar{y})\log(1 - \bar{y})}, \tag{3.3}$$

where $\bar{y}$ is the expected value of variable y.

Results & Discussion

Results of the behavioural analysis are again discussed for all three subjects and two example expressions ('H' and 'F'). Figures 3.1a to 3.1c display the R^2 for the full model, including all independent signal sources, for subject 'LP' for expressions 'H' and 'F' respectively. Figures 3.2a and 3.2c and figures 3.3a and 3.3c display the corresponding results for subject 'LF' and 'UM' respectively. The overall R^2 values are not particularly impressive, peaking at no higher than 4% explained variance for any subject and any expression, but they do display some systematic variation over time. There is generally a first peak at around the $150 - 180ms$ time window and, for subjects 'LP' and 'UM' at least, occasionally a second peak between 250 and 300 ms following stimulus onset. Again, this first peak fits with the expected time window of the N170 ERP component and although the R^2 values are not particularly high for these time points, the coefficients of the full models are all statistically significant during this window ($p < 0.05$). Note that there is no obvious behavioural relation to the early variance observed at around $50ms$ in the sources.

Figures 3.1b and 3.1d display the R^2 for the three independent signal sources highlighted earlier [left and right temporal–parietal (numbers 1 and 2 on the figure, and bi–lateral occipito–temporal (number 3 on the figure)], for subject 'LP' for expressions 'H' and 'F' respectively. Figures 3.2b and 3.2d and figures 3.3b and 3.3d display the corresponding results for subject 'LF' and 'UM' respectively. Given the timing of the first peak in the R^2 for the full model, we would expect this peak to be localised mainly on source number 3. This is the case in general, although subject 'LF' for expression 'F' is the exception. Simlarly, given the timing of the second peak, we would expect this peak, where it exists, to be localised mainly on the two temporal–parietal sources. Again, in general this is the case, although there is also some sensitivity for the occipato–temporal source during this time window. The significant ($p < 0.05$) peaks in the individual source models include the early peak on source 3 for all cases where it is present (i.e. all cases except for 'LF' and 'F') and the late peak on source 1 wherever it is present ('LP', 'H'; 'LP', 'F'; 'UM', 'H'). However, other peaks, including the early–to–middle peaks for 'LF' and 'F' on source 2 and 'LF' and 'H' on source 2, are significant at this level also.

Overall, what are we to make of these results? The EEG model clearly doesn't provide a great overall fit to the behavioural data of the subject, but some consistent patterns are revealed in the *relative* fit of each of the model parameters when compared against each other. Particularly in the case of subject 'LP', whose results display the highest peak fit between any indi-

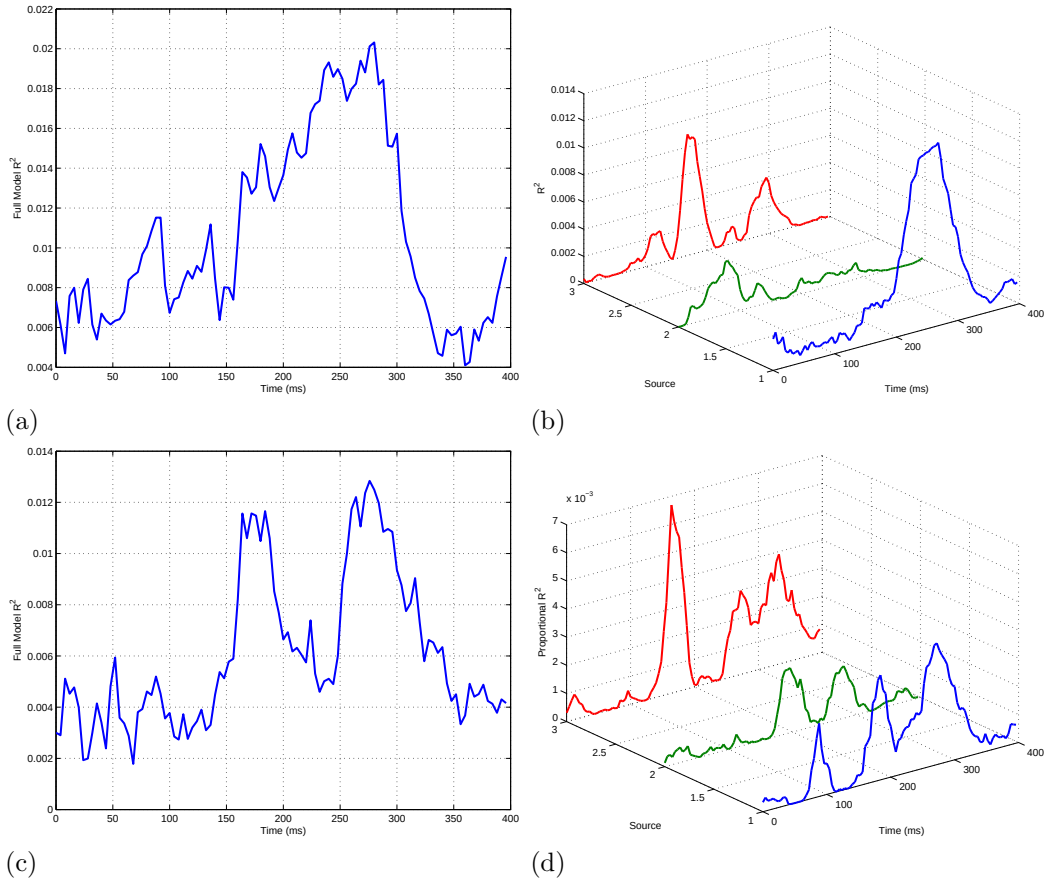


Figure 3.1: R^2 goodness–of–fit for both the full model [(a), (c)] and the three highlighted sources [(b), (d)] over time. Results are shown for subject 'LP' and expressions 'H' [(a), (b)] and 'F' [(c), (d)]. Horizontal axis — time (ms). Vertical axis — R^2.

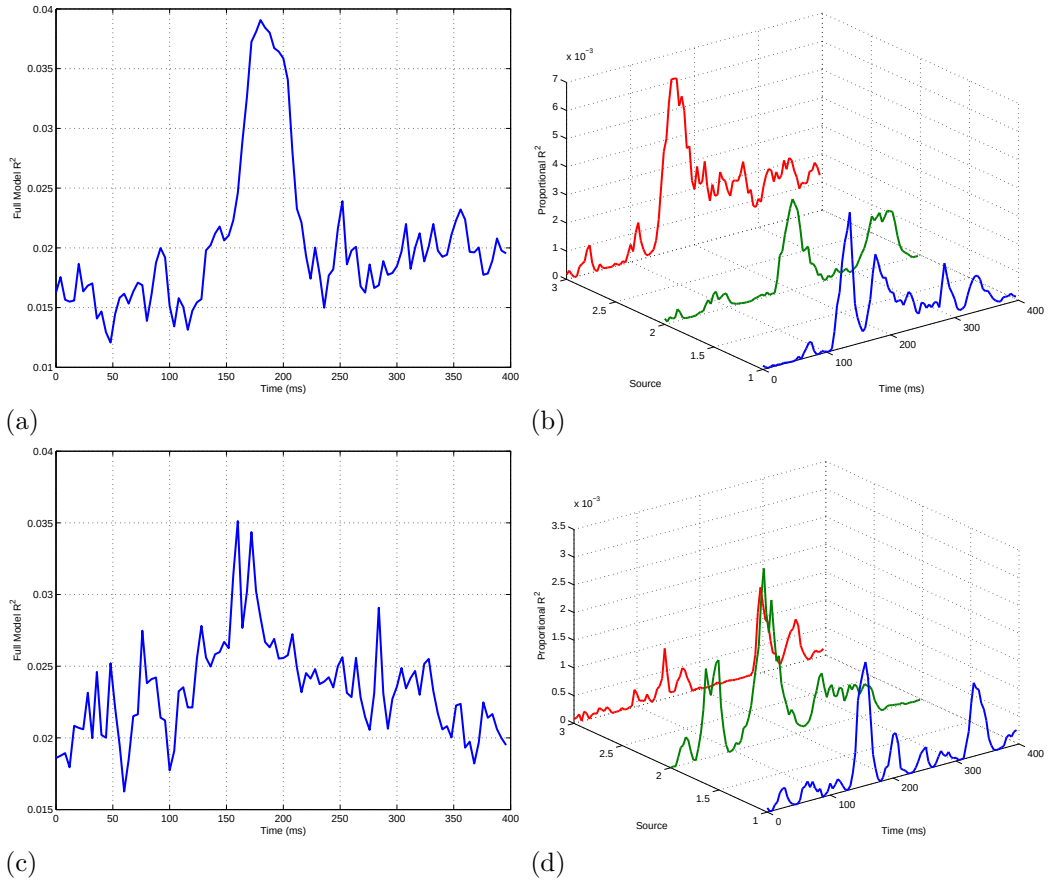


Figure 3.2: R^2 goodness–of–fit for both the full model [(a), (c)] and the three highlighted sources [(b), (d)] over time. Results are shown for subject 'LF' and expressions 'H' [(a), (b)] and 'F' [(c), (d)]. Horizontal axis — time (ms). Vertical axis — R^2.

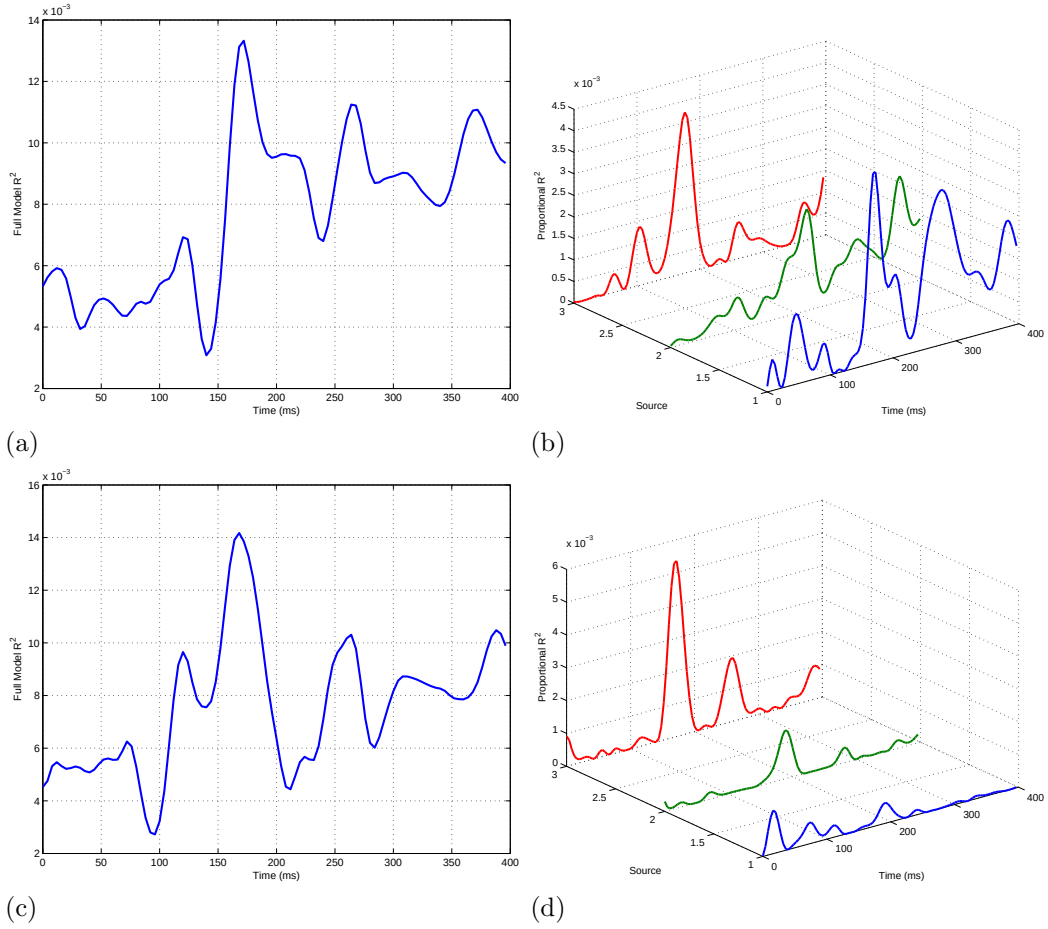


Figure 3.3: R^2 goodness–of–fit for both the full model [(a), (c)] and the three highlighted sources [(b), (d)] over time. Results are shown for subject 'UM' and expressions 'H' [(a), (b)] and 'F' [(c), (d)]. Horizontal axis — time (ms). Vertical axis — R^2.

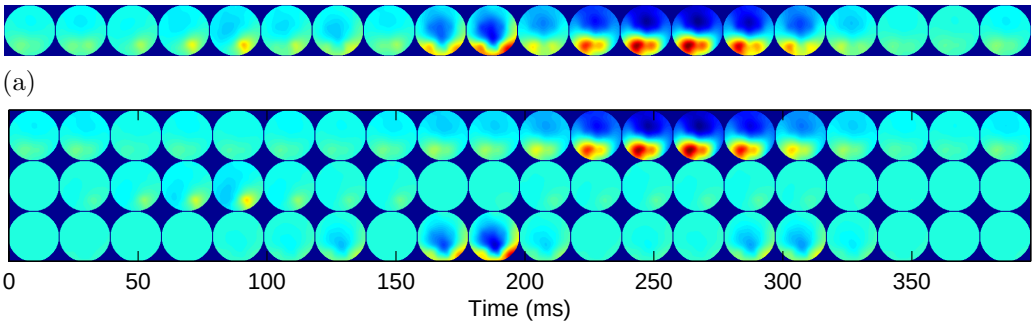


Figure 3.4: Representation of model fit in terms of (a): combined, and (b): individual goodness–of–fit for each of the three highlighted sources. For each time point, the source topologies are weighted by their corresponding R^2 value. Results shown for subject 'LP' and expression 'H'.

vidual source and the behaviour (around 1% of explained variance) the time course of the fit between individual sources and behaviour coincides perfectly with the time courses of both the N170 and P300 ERP components, whose gross topologies match the topologies of the corresponding sources. Figures 3.4a and 3.4b provide a graphical representation of the three highlighted signal sources, each weighted by their fit to the behavioural response over time, for subject 'LP' and expression 'H'. This representation shows quite nicely a shift from early occipital to later parietal processes over the time following stimulus onset.

Perhaps the reason for the inconsistency in these results is that what is being modelled is not the full multi–category classification task that the subject is performing, but rather the set of independent binary classification tasks for each expression. Future experiments designed explicitly around the multi–category framework should be run to see how the model fares. It is not difficult to extend the logistic regression framework to the multi–category case. The probability distribution of the responses should be treated as a multinomial rather than a binomial distribution in this case. Another possible reason for the inconsistency is in the estimation of the logistic regression parameters themselves. One way to improve the estimates is to replace the standard cost function with a regularized version which penalizes over–parametrized solutions (e.g. [39]). However, methods for selecting the appropriate regularization term were deemed beyond the scope of this thesis.

Fitting to the stimulus

EEG Bubbles

The primary inspiration for the analysis presented here is the "EEG Bubbles" method, introduced in [62] and further replicated in [68], [66], and [63], amongst others. Below is an outline of the "EEG Bubbles" method, as presented in [63].

For each subject taking part in the experiment and for each of the 7 expressions presented to the subject:

- All trials on which the subject did not correctly identify the target expression were removed.
- For each EEG measurement channel:
    1. For each time point following stimulus onset:
        (a) The EEG voltage distribution over the selected trials was converted to z–scores.
        (b) The resulting z–score distribution was divided into 13 equally spaced bins and each trial was marked according to its bin.
        (c) The 6 highest and 6 lowest bins were separated from the rest and assigned to two groups — "above mean" and "below mean" respectively.
        (d) Per spatial frequency band of the bubble masks:
            i. For each of the two groups, the bubble masks corresponding to the trials contained within their bins were collected together and summed per pixel.
            ii. The value of each pixel in the "below mean" group was subtracted from the value of the corresponding pixel in "above mean" group. Let us refer to the result of this subtraction as the "difference plane".
    2. The 500ms time window prior to stimulus onset was chosen as a statistical baseline and, for each spatial frequency band of the difference planes:
        (a) The total pixel value mean was computed over the baseline time window.
        (b) The total pixel value standard deviation was computed over the baseline time window.
        (c) For each time point following stimulus onset:

i. Each pixel of the difference plane was normalized to a z–score by removing the baseline mean and dividing by the baseline standard deviation.

ii. A threshold on the absolute values of these z–scores was chosen according to a two–tailed hypothesis ($p < 0.05$), and corrected for multiple comparisons using the "pixel test", [16]. To create a "threshold plane", pixels exceeding threshold were then set to a value of 1 and the rest set to 0.

iii. The threshold planes were then convolved with a Gaussian filter, which was parametrized according to the bubbles in the mask at this spatial frequency.

iv. The final outcome was the result of multiplying, per pixel, each band of the filtered threshold planes with the corresponding band of the filtered stimulus image. The authors referred to this as the "EEG Classification Image".

Before addressing specific points of the analysis in more detail, it makes sense to hold in mind the idea of part 1 as the "meat and potatoes" of the analysis, while part 2 is more like the dessert. All of the numbers that shape the results are derived in part 1; whereas how those numbers are presented is controlled in part 2.

So what is happening in part 1? Well, first of all, the data are being normalized and they are being down sampled - first from thousands of trials to 13 bins, and then from 13 bins to 2 groups. The purpose of normalizing independently at each time point and EEG channel is to ensure that the results are not driven by any global amplitude or power differences along these 2 dimensions. The purpose of down sampling is to create a contrast against which the values of the pixels in the mask can be compared. Subtracting the sum of the masks in the lower bin from the sum of the masks in the higher bin is equivalent to setting each of the EEG amplitudes in the higher bin to 1, each of the the EEG amplitudes in the lower bin to -1, multiplying the bubble mask on each trial with its corresponding EEG amplitude, and adding up the results. So the results contained in the difference plane are similar to the multiplication between the values of each pixel and the amplitudes measured at each EEG channel, normalized by the variance in the EEG data.

The second part is all about determining statistical significance of the values in the difference plane and presenting the final result. Normally claims to statistical significance are reliant upon the distribution of the residual error. However, these data are essentially reduced to two points, per pixel

per time point (above plane and below plane), so there is no distribution from which these statistics can be derived. The authors instead determine the error distribution from the distribution of difference plane values during the pre–stimulus time window, the assumption being that the mean and variance of any correlations during this time period should reflect the mean and variance of any inherent noise in the process. The value of each pixel in the difference planes *post* stimulus are then normalized by these estimates and considered as points on a standard Gaussian cumulative distribution with zero mean and unit standard deviation. They are then deemed to significantly deviate from zero if their absolute value exceeds the reciprocal of some chosen p–value when evaluated by the cdf (e.g. $p < 0.05$).

Linear model of pixel information

As the difference planes from the EEG Bubbles method are equivalent to:

$$x_{\rho,j} = \frac{\mathrm{Cov}(m_j, \rho)}{\mathrm{Var}(m)}, \tag{3.4}$$

calculated for for each pixel, ρ, in the stimulus and each measurement channel, j, the method can basically be seen as a linear model of pixel information in terms of brain activation. To see why, let us denote the set of Bubble mask values for pixel p as the $n \times 1$ vector $\mathbf{b}_\rho$ and the set of brain measurements at electrode j as the $k \times 1$ vector $\mathbf{m}_j^\mathbf{t}$. The linear model is then:

$$\mathbf{b}_\rho = \mathbf{m}_j^\mathbf{t} x_{\rho,j}. \tag{3.5}$$

Multiplying both sides by $\mathbf{m}_j$, we get:

$$\mathbf{m}_j \mathbf{b}_\rho = \mathbf{m}_j \mathbf{m}_j^\mathbf{t} x_{\rho,j}. \tag{3.6}$$

Because the vector product $\mathbf{m}_j \mathbf{m}_j^t$ is a scalar quantity, corresponding to the inner product of brain response vector $\mathbf{m}_j^\mathbf{t}$ with itself, it can be factored out of the right hand side by multiplying both sides by its inverse, leaving us with:

$$(\mathbf{m}_j \mathbf{m}_j^\mathbf{t})^{-1} \mathbf{m}_j \mathbf{b}_\rho = x_{\rho,j}. \tag{3.7}$$

Which, assuming the EEG data for channel j is zero–mean, is simply equation 3.4.

The benefit of re–framing the EEG Bubbles analysis in terms of a linear model is, firstly, that we can more easily compare and transform between the previous linear models estimated up to now for both the signal space and for the relation between the signal and the behavioural space. Secondly, we

can produce goodness–of–fit estimates for each pixel in the input and each electrode and time point of the brain signal, according to:

$$\begin{aligned} R^2 &= 1 - \frac{\mathrm{Var}(\epsilon)}{\mathrm{Var}(\rho)} \\ &= 1 - \frac{\mathrm{Var}(\rho - \hat{\rho})}{\mathrm{Var}(\rho)} \\ &= 1 - \frac{(\mathbf{b}_\rho - \hat{\mathbf{b}}_\rho)^\mathbf{t}(\mathbf{b}_\rho - \hat{\mathbf{b}}_\rho)}{(\mathbf{b}_\rho - \bar{\mathbf{b}}_\rho)^\mathbf{t}(\mathbf{b}_\rho - \bar{\mathbf{b}}_\rho)}, \end{aligned} \tag{3.8}$$

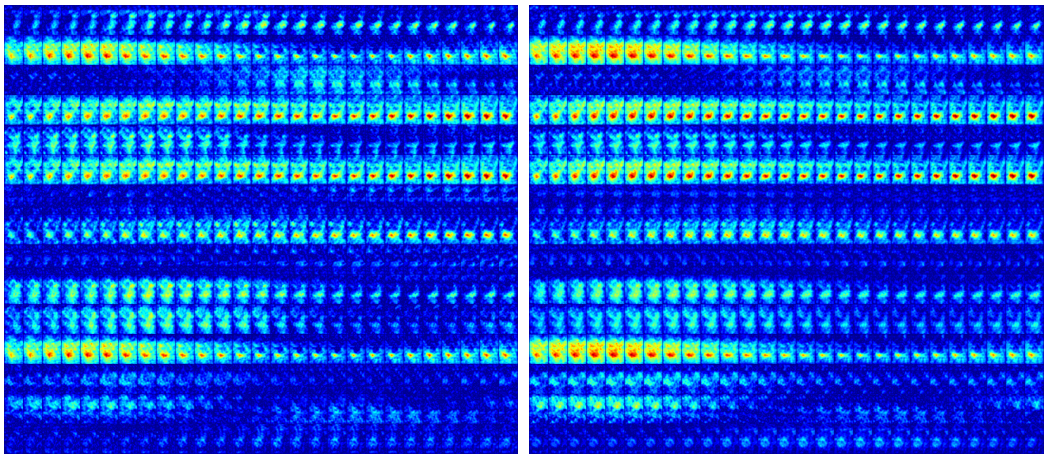
and we can also estimate significance values for each model parameter, allowing us to replace part 2 of the EEG Bubbles method presented above.

Figures 3.5a and 3.5b display the the EEG Bubbles difference planes and linear model coefficients, both applied to the data in [63], for a single subject shown the "happy" expression at 15 random electrodes between 160 and 190 ms. The pixel–wise correlation between the two methods, averaged over electrode, time, expression, and spatial frequency band is 0.832. Figures 3.6a and 3.6b display the time courses of results from the EEG Bubbles difference planes and the linear model coefficients for a single subject and expression, for all electrodes, averaged over the pixels in the mask. Note the extra smoothness in the linear coefficients compared to the difference planes.

Results & Discussion

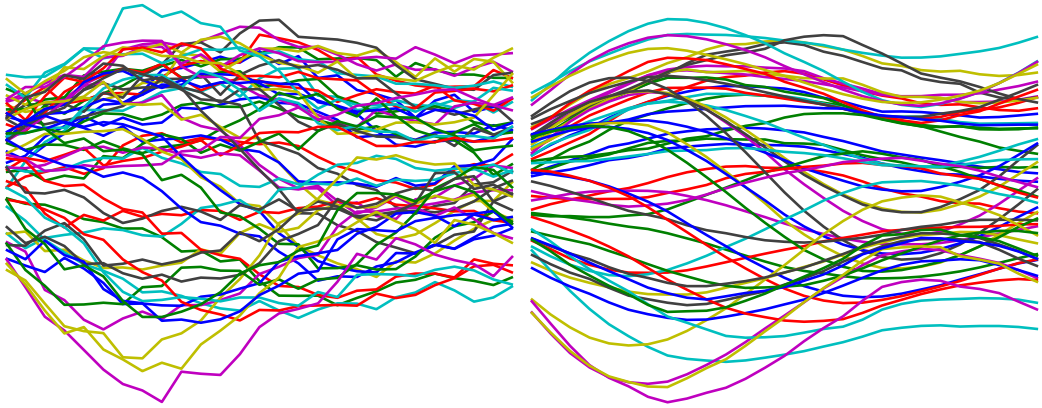
Results of the stimulus–based linear models are presented as per the previous two sections. Figures 3.7a and 3.7c show time courses of the R^2 values for the full linear model containing all signal sources, averaged over all pixels in the stimulus, for subject 'LP' and for expressions 'H' and 'F' respectively. Figures 3.8a and 3.8c and figures 3.9a and 3.9c show the corresponding results for subjects 'LF' and 'UM' respectively. Again, the overall magnitude of the R^2 statistics is far from impressive, peaking at no higher that 1.5% of the explained variance. However, the temporal evolution of the statistic is far more pronounced than it was in the behavioural model. For each subject and expression, with the exception of subject 'LF' and expression 'F', there is a single clear, strong peak between 150 and 180ms.

Figures 3.7b and 3.7d show time courses of the R^2 values for the three individual highlighted signal sources, averaged over all pixels in the stimulus, for subject 'LP' and for expressions 'H' and 'F' respectively. Figures 3.8b and 3.8d and figures 3.9b and 3.9d show the corresponding results for subjects 'LF' and 'UM' respectively. As was the case with the behavioural results, the early peak in the full model R^2 statistic can be primarily attributed to the fit



(a) (b)

Figure 3.5: (a): EEG Bubbles 'difference planes' for subject 'LP' and expression 'Happy'. X–axis is time ($160-190ms$), Y–axis is electrode (15 electrodes chosen at random), colour is absolute value of result. Each axis is further broken down into sub–cells of X–pixels × Y–pixels in the bubble masks. (b): Linear model coefficients for subject 'LP' and expression 'Happy'. X–axis is time ($160-190ms$), Y–axis is electrode (15 electrodes chosen at random), colour is absolute value of coefficient. Each axis is further broken down into sub–cells of X–pixels × Y–pixels in the bubble masks.



(a) (b)

Figure 3.6: Comparison of EEG Bubbles and linear model coefficient time courses [(a) is EEG Bubbles, (b) is linear model coefficients]. X–axis is time, Y–axis is arbitrary because results are adjusted to be in comparable scale. Results are displayed as averages over mask pixels and spatial frequency bands for subject 'LP' and expression 'Happy'.

from source number 3 – the occipito–temporal source. Most of these graphs, excluding subject 'LP' for expression 'H' and subject 'LF' for expression 'F', also display a secondary set of peaks in sources 1 and 2 directly after the initial peak in source 3.

So the majority of the sensitivity to the stimulus in the EEG data appears to be centred on the occipito–temporal source during the time period of the N170 ERP component. But what local information in the stimulus does this correspond to? Figures 3.10a and 3.10b display the relative R^2 statistic for each pixel in the stimulus on source number 3 for subject 'LP' and expressions 'H' and 'F' respectively. These data are displayed between 120 and 180ms following stimulus onset. Figures 3.11a and 3.11b and figures 3.12a and 3.12b show the corresponding results for subjects 'LF' and 'UM' respectively. For subjects 'LP' and 'UM', these data are quite striking. Both cases clearly show a differential local information sensitivity for the happy and the fearful stimuli. Fear is processed early, and with facial information mostly around the eyes, while happy is processed slightly later and utilizing information from the whole face. The behavioural Bubbles results in [63] show that these are the local regions in the face corresponding to correct classification for these two expressions. The sensitivity shown by the occipito–temporal source is therefore to the information relevant for the emotional expression during this time window. Along with the evidence for early sensitivity of this source to the behavioural response of the subject, this goes against any idea that the early occipital EEG reflects only a bottom–up encoding process.

However, for the temporal–parietal sources, the opposite seems to be the case. Where the temporal–parietal sources show the strongest peak in sensitivity to the behavioural outcome ('LP', 'H'), there is little or no sensitivity in these sources to the stimulus information. Why might this be? A speculative answer might be that the parietal sources in fact represent processes of motor planning, rather than stimulus categorization. Given the high number of possible key responses, motor planning was possibly a factor in the response accuracy, and parietal neurons in monkeys have been shown to be sensitive to the locations of targets in a grasping task [54]. All three subjects were right handed, and the left parietal source tends to show the higher sensitivity to the behaviour, but not necessarily to the stimulus information.

Now, how do these results compare with the findings of the original EEG bubbles analysis by Schyns et al.[63]? The main finding presented in [63] are: that category–specific stimulus information is integrated over the time course of the N170 at bi–lateral occipito–temporal electrode sides; and that this integration corresponds to a goal–directed trajectory in the feature space (i.e. moving across the face towards the features that are diagnostic for the emotion). The first of these findings is clearly replicated here, as discussed

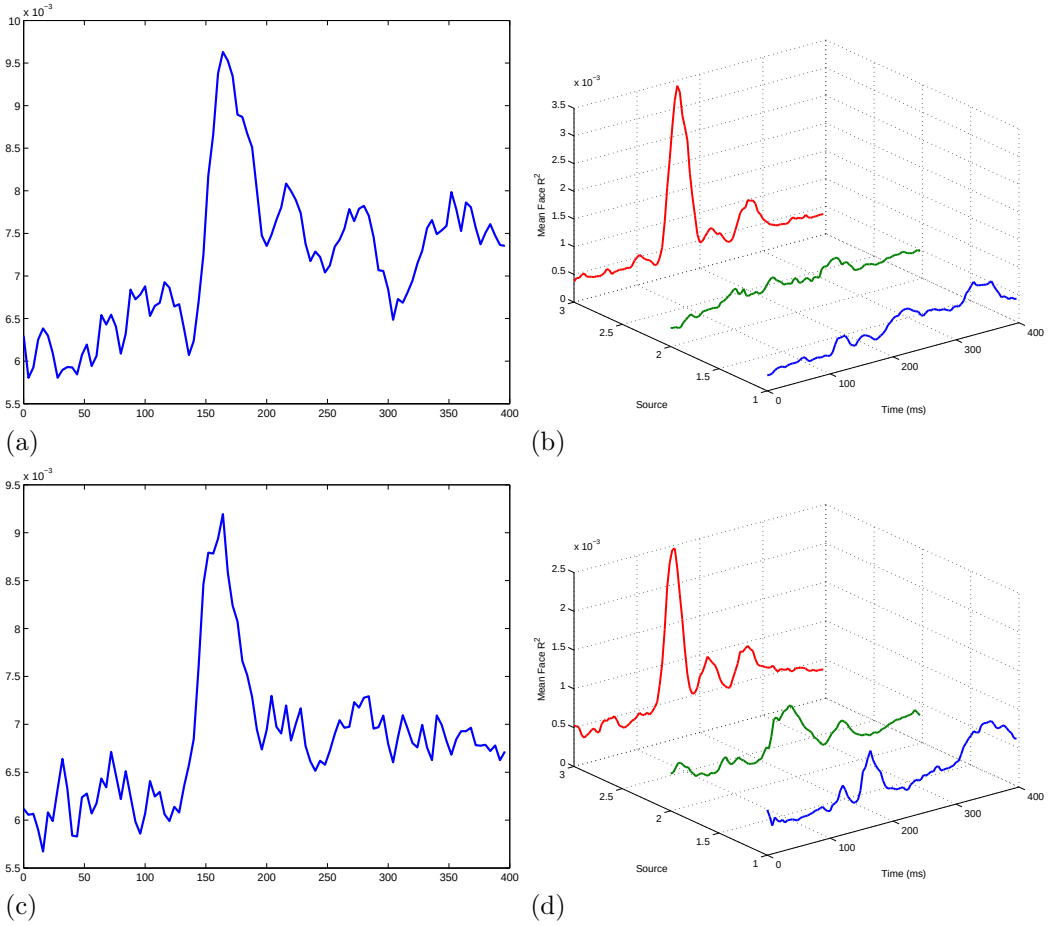


Figure 3.7: R^2 goodness–of–fit, averaged over all pixels in the stimulus, for both the full model [(a), (c)] and the three highlighted sources [(b), (d)] over time. Results are shown for subject 'LP' and expressions 'H' [(a), (b)] and 'F' [(c), (d)]. Horizontal axis — time (ms). Vertical axis — R^2.

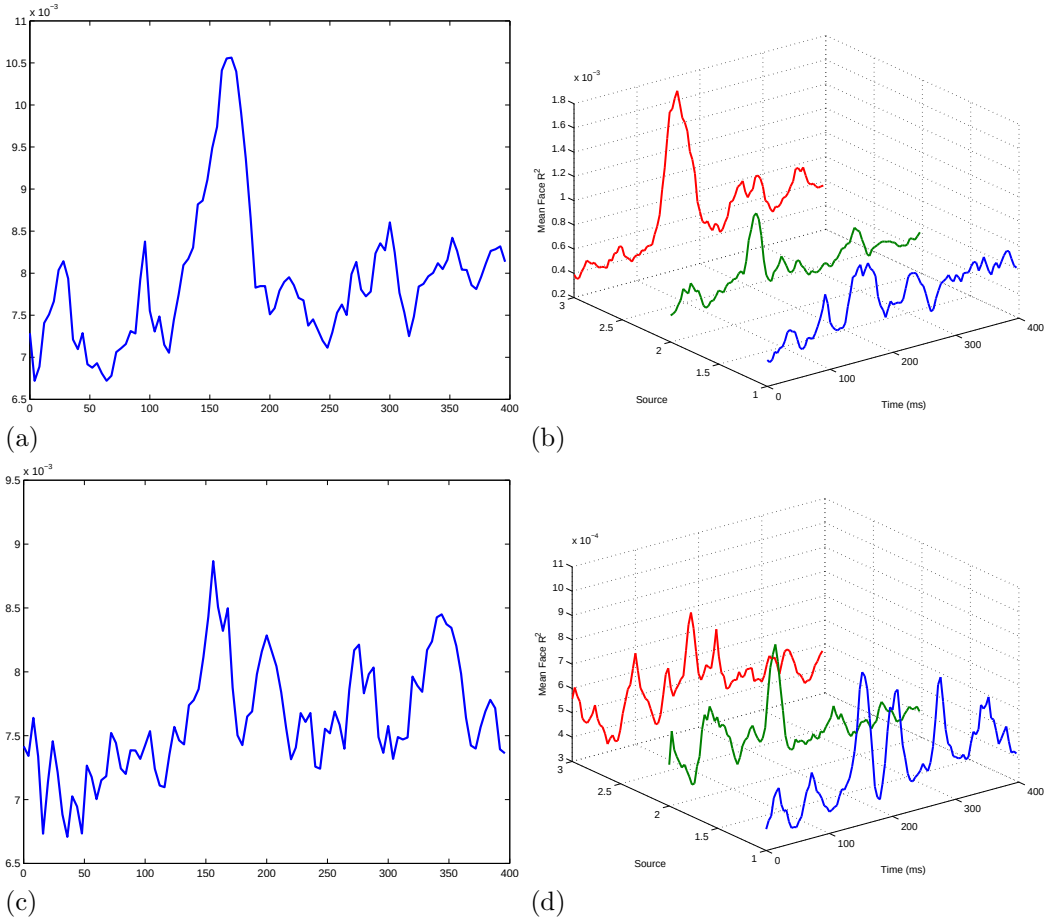


Figure 3.8: R^2 goodness–of–fit, averaged over all pixels in the stimulus, for both the full model [(a), (c)] and the three highlighted sources [(b), (d)] over time. Results are shown for subject 'LF' and expressions 'H' [(a), (b)] and 'F' [(c), (d)]. Horizontal axis — time (ms). Vertical axis — R^2.

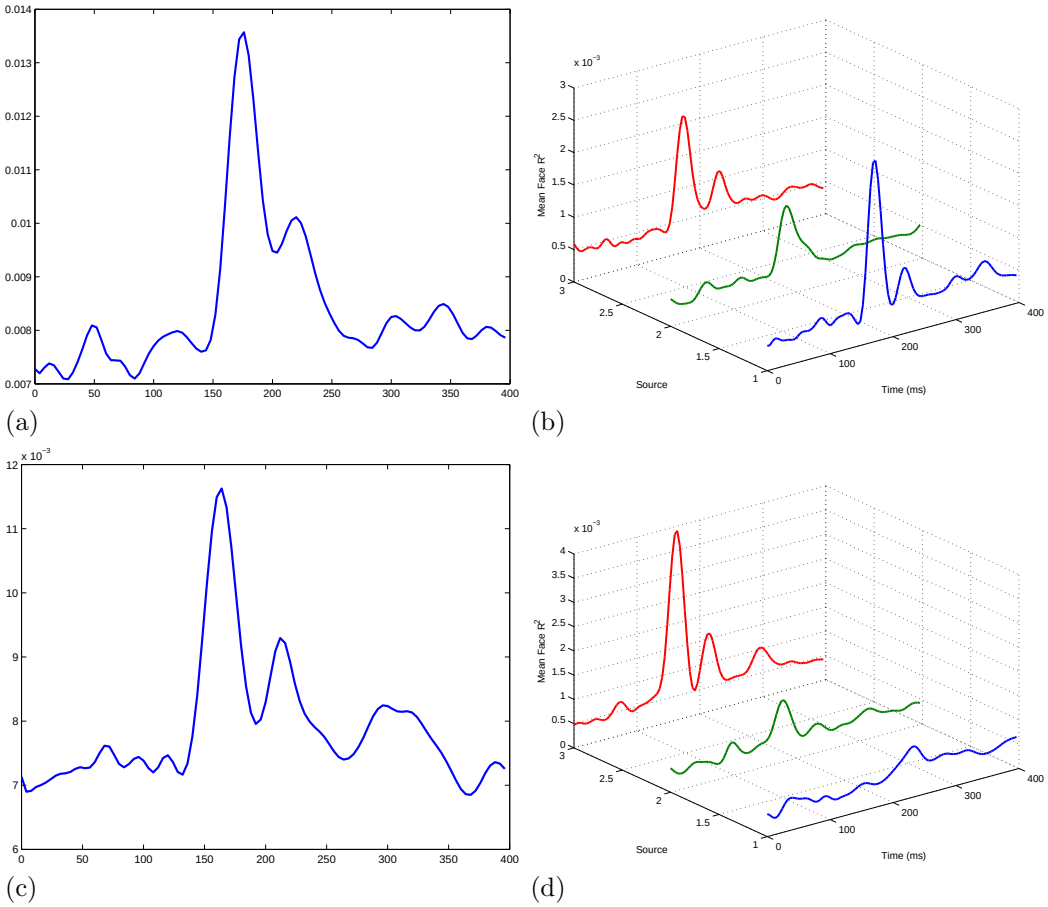


Figure 3.9: R^2 goodness–of–fit, averaged over all pixels in the stimulus, for both the full model [(a), (c)] and the three highlighted sources [(b), (d)] over time. Results are shown for subject 'UM' and expressions 'H' [(a), (b)] and 'F' [(c), (d)]. Horizontal axis — time (ms). Vertical axis — R^2.

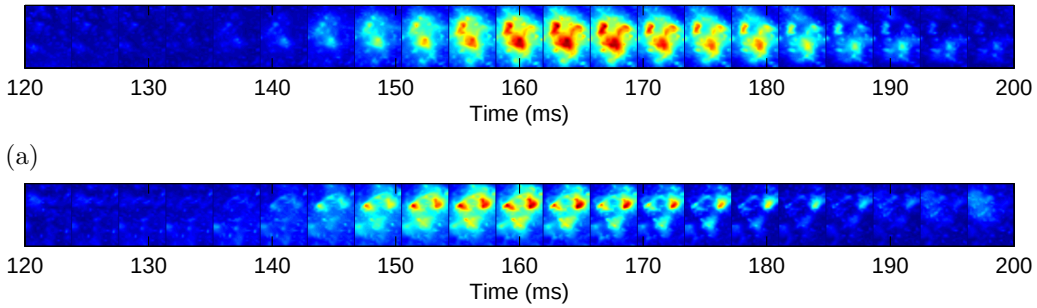


Figure 3.10: Relative maps of individual R^2 statistics for each pixel in the stimulus on occipito–temporal source over time for subject 'LP' for expression (a): 'H' and (b): 'F'. Values shown are averaged over spatial frequency.

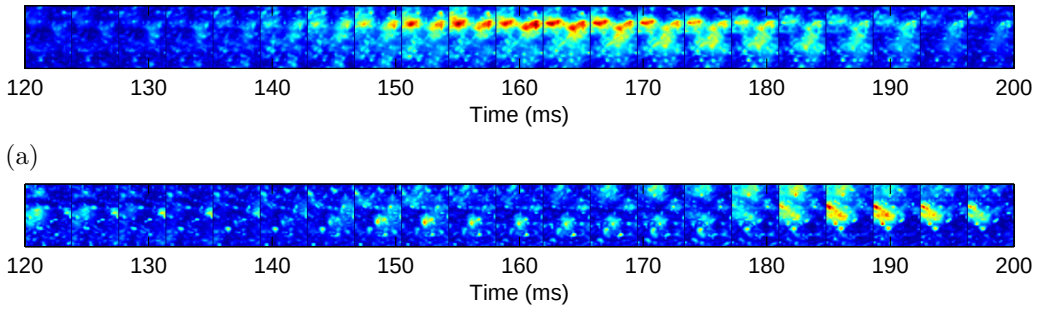


Figure 3.11: Relative maps of individual R^2 statistics for each pixel in the stimulus on occipito–temporal source over time for subject 'LF' for expression (a): 'H' and (b): 'F'. Values shown are averaged over spatial frequency.

above. For the second of these findings, certainly in the case of the happy expression, local stimulus information sensitivity appears to move down from the eyes and towards the mouth, while it stays fixed at the eyes for the fearful expressions (in the two subjects where the eyes are the location of maximum stimulus sensitivity at least). The difference between the results presented here and those presented in [63], however, is that the occipito–temporal source here was selected automatically as one of the top few signal sources, based upon the CV–EV subspace partition and the MEE sorting criterion, while the occipito–temporal electrodes in the original study were selected on the basis of a combination of prior knowledge and trial–and–error.

One possible advantage of the non–model approach to EEG analysis is that the space of the brain can be explored at will, possibly revealing some interesting sensitivity properties that are not revealed at any single source.

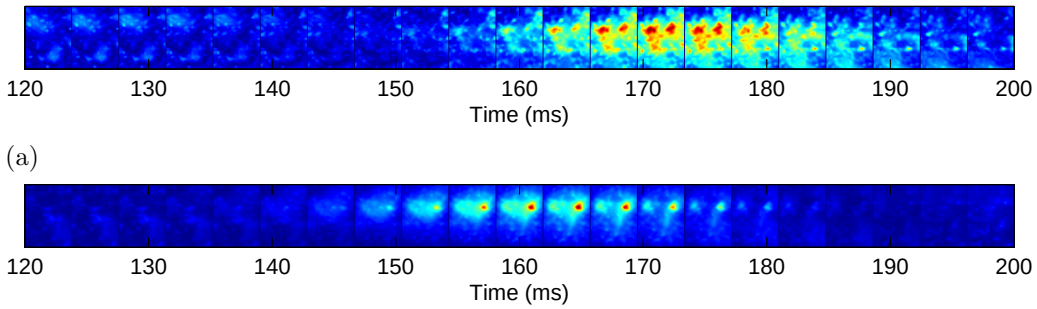
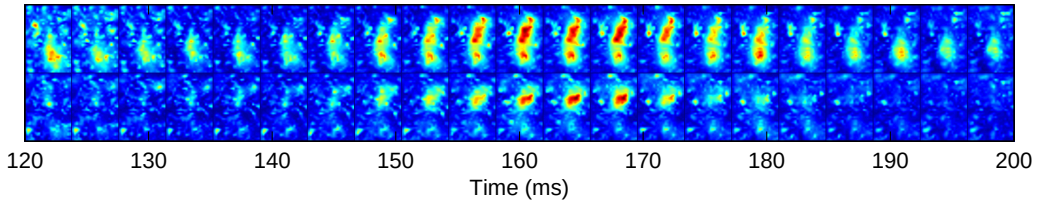
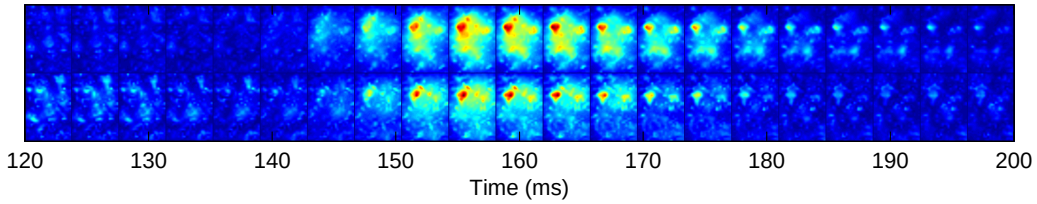


Figure 3.12: Relative maps of individual R^2 statistics for each pixel in the stimulus on occipito–temporal source over time for subject 'UM' for expression (a): 'H' and (b): 'F'. Values shown are averaged over spatial frequency.

For example, our occipito–temporal source is sensitive to visual information bi–laterally, but previous EEG bubbles studies (e.g. [62]), have shown that certain right and left occipito–temporal electrodes are sensitive to contra–lateral information in the stimulus. Well, there is nothing stopping us from inverting the linear transformation from the source space back into the EEG electrode space. As both the independent Gaussian sources model and the linear model of pixel information are simply linear transformations, the coefficients of the pixel model can be transformed into EEG bubbles coefficients by the same unmixing matrix W^t as transforms the sources into EEG electrode measurements. The difference, however, is that the transformed results are truer estimates of the individual channel sensitivities because they are scaled by the appropriate independent variance (i.e. by the variance of the source, rather than the variance of the electrode, which is contaminated by covariance with other electrodes). They also have been estimated in the absence of the noise subspace. Figures 3.13a and 3.13b show a comparison between EEG bubbles data and the linear independent pixel model data, after transformation back into the electrode space, for subject 'LP' and expression 'H' at occipito–temporal electrodes P8 and P7 respectively. The top half of both figures displays the original EEG bubbles data, while the bottom half displays the transformed independent source data.



(a)



(b)

Figure 3.13: Comparison of R^2 statistics from single electrode EEG Bubbles (top half of each figure) and transformation of full source model into electrode space (bottom half of each figure). (a): P8 electrode. (b): P7 electrode.

Chapter 4

General Discussion

In this thesis I have argued that the extension of reverse correlation methods, traditionally applied to study receptive field properties of single cells, to EEG signals measured on the scalp is flawed in the absence of a model relating those scalp measurement to a set of underlying brain processes. Without such a model, interpretation of the results relies upon two assumptions: firstly, that what is measured by any single electrode corresponds to a meaningful brain process in itself; and, secondly, that what are jointly measured on any pair of electrodes correspond to different brain processes. Without prior reason to assume either, the interpretation of reverse correlation results from raw EEG electrode signals is problematic. On the other hand, without detailed structural knowledge of the subject's brain tissue (data taken from a structural MRI scan, for example), we cannot reliable estimate the current density that causes the signals measured at the scalp [34]. Since such data is often unavailable, alternative methods of estimating the functional sources of EEG data are desirable.

Here was considered a particular approach to the 'Blind Source Separation' (BSS) problem, presented in [56], wherein the raw data were modelled as a set of statistically independent signals drawn from an unknown Gaussian distribution with a non–stationary variance profile. Tests on simulated data generated from artificial Gaussian sources demonstrated that the algorithm presented in [56] required only a minimal level of non–stationarity in the source variance for effective separation. The method was further developed to isolate the sources associated with meaningful brain activity from those associated with noise or recording artefact. Tests on simulated data demonstrated that the meaningful sources could be successfully isolated with this method, provided that the variance profile of the noise is close to stationary. When applied to real EEG data, recorded during a facial emotion classification study, the method revealed three sources of the EEG signals

corresponding to consistent scalp topologies across three different subjects — two lateralised temporal–parietal sources and a bilateral (although biased to one side) occipito–temporal source. These topologies fit with previous ERP research in the fields of visual classification and facial emotion processing — the temporal–parietal 'P300' effect ($300ms$ following stimulus) and the occipito–temporal 'N170' effect ($170ms$ following stimulus) — and, more specifically, to each individual subject's ERP topology record during these time windows.

To understand the function of these modelled sources, two forms of regression were performed on their single–trial data — a logistic regression against the subjects' response behaviour, and a linear regression against the random information available through the stimulus masks (equivalent to the 'EEG Bubbles' method developed in [62]). Previous visual classification research [25][26][43] would predict that the occipito–temporal source activity is significantly correlated with general stimulus information (i.e. information suggesting that it is a face), but not with specific information differentiating the particular facial emotion from others, while the two temporal–parietal sources are significantly correlated with both the response behaviour of the subject and the stimulus information specifically differentiating the facial emotion. The results of these two analyses displayed a slightly different pattern, however. In the majority of cases, both the occipito–temporal and the left temporal–parietal sources were significantly correlated with response behaviour, while neither temporal–parietal source tended to show any significant correlation with stimulus information and the occipito–temporal source was correlated specifically with the stimulus information differentiating the particular emotion. These results support the hypothesis that the N170 does, in fact, reflect meaningful classification processes, at least in the case of facial emotion perception [63][64]. They also question the extent to which the P300 reflects any kind of perceptual process at all. It may rather be that what the P300 effect is measuring is a form of motor planning — a hypothesis supported by previous monkey research [54].

Overall, the work presented in this thesis has demonstrated the utility of fitting a principled model to raw measurements of EEG data, even when structural knowledge of the subjects' brain tissue is non–existent and only a few simple constraints are imposed on the solution. To conclude this work, future extensions of the method shall now be discussed, considering both the signal model and the experimental design of the reverse correlation.

Source statistics

Although the simplicity of the non–stationary independent Gaussian source model is appealing, and comes with theoretical justification [29], it is not the only way to go. As mentioned earlier, blind models of non–Gaussian sources are far more common than those of Gaussian sources, and the data generated from such a model should be compared to that presented here. Non–Gaussian BSS tends to be bracketed under the term 'Independent Components Analysis' (ICA). In principle, ICA models can be estimated by a similar joint diagonalization process to that performed for the non–stationary Gaussian sources ([17],[52]). The problem is instead to find a mixing matrix that jointly diagonalizes both the covariance matrix for a given set of data and a second matrix of higher–order cross–statistics, as both must be diagonal for non–Gaussian sources to be independent. However, practically this approach may not be the most effective, due to increasing estimation errors as the order of the statistics increases. Bell & Sejnowski [9] propose an iterative algorithm to maximize the mutual independence between estimated sources, where each source p.d.f. is considered as a particular non–linear transformation of a Gaussian variable, but this method does not consider the joint estimation of sources over a time window.

If the assumption of non–Gaussianity is used to *replace* the assumption of non–stationarity in the sources, then we lose one of the appealing features of this model, which is that it provides a single estimate of the mixing matrix over all time of EEG measurement, and we also lose the ability to compare the results. Ideally, future models of the EEG that consider non–Gaussian sources should also account for the time dimension as the non–stationary Gaussian model of Pham & Cardoso does. However, it should be noted that, in a related work [29], where the non–stationary Gaussian model used a kernel estimator to separate a single time series (as in [56]) and hence the two methods could be compared, the Gaussian BSS model outperformed a variety of non–Gaussian ICA models in isolating meaningful sources over recording artefact.

Principled choice of time window

As was discussed in the text, parameters of the signal model were estimated jointly for the full time course of measurement (i.e. from $-500 - 1024$ms following stimulus presentation.) The full time window was chosen to maximize the number of data points, and hence increase the chance of separating the non–stationary sources from the possibly non–stationary noise. However, if the underlying model did not hold for this entire window, then this could

have been a source of error. The unknowns are: how sensitive is the solution to violations of the model assumptions, and on what criteria should window selection be based? If the JD–BGL algorithm can be shown to be robust under violations of the model assumptions, then the largest available time window should be selected, provided that the number of active sources over this time window does not exceed the number of measurement channels. If it is not so robust, however, then some method of finding, a priori, the windows in which the assumptions are not violated should be developed. Future simulations should be performed to test the robustness of the JD–BGL algorithm when the Gaussian assumption is violated for a subset of the sources. The relationship between the marginal normality of the sources and the marginal normality of the measurement channels, via the Jarque–Bera test, should also be explored as a means for selecting the appropriate window.

Integration of the behavioural correlation with model specification

As the signal model is currently specified, the only constraints on the solution are in terms of the source statistics. However, as we are primarily interested in finding sources that maximally explain both the visual input and the behavioural output of the classification task, it makes sense to introduce those as additional constraints in the joint diagonalization algorithm. In other words, rather than attempting to minimize only the mutual information between sources, attempt also to maximize the mutual information between each source and the task–related variables. This approach makes more sense in the context of the behavioural variable only, as the number of dimensions in the stimulus space is so large.

Extension into the spectral domain

While it is not uncommon for EEG/MEG data to be analysed as a time–varying amplitude signal, there is mounting evidence that many of its interesting behavioural correlates are to be found in its spectral decomposition (e.g. [7][8][35]). Particularly interesting in the context of the research presented here is the spectral relationship between occipito–temporal and temporal–parietal brain regions as a function of both stimulus information modulation and behavioural response.

To extend the BSS model to the spectral domain, it is necessary to replace the trial–averaged, time–varying data covariance matrix with a trial–averaged, time–varying estimate of the cross–spectral density (CSD) matrix at a given frequency. Various methods exist for the estimation of the time–varying CSD matrix, including the short–time Fourier transform [71] and the

wavelet transform [45]. The JD–BGL algorithm must also be extended to perform joint diagonalization on complex matrices.

It should be noted that a recent method has been developed to perform non–Gaussian ICA on time–varying spectral data [29]. However this method, rather than attempting to jointly diagonalize the time–varying CSD, diagonalizes a time– and frequency–averaged CSD matrix jointly with a non–Gaussian constraint. As a result, the method is limited to separating narrow–band sources, but has less parameters to resolve. A comparison of the two approaches would be interesting.

EEG pre–processing

An obvious area for improvement in this analysis is with the early EEG analysis pipeline — in particular, artefact rejection. As shown in the simulations, the most destructive influence on parameter estimation was non–stationary noise in the observed data. A likely source of such noise is eye–blink or eye–movement artefacts. Although EOG potentials were recorded for the purpose of removing such artefacts from the data, there is no guarantee that they always succeeded. Pre–filtering of the EEG data with temporal non–Gaussian ICA components derived from the ERP may be an effective pre–processing stage.

Naturalistic stimulus space

When facial expressions are generated by complex and subtle interactions between muscles in the face, it is an interesting question as to how much we might expect first–order configurations of static pixels in an image to really capture how they are processed. The strength of reverse–correlation methods in psychophysics is that the input space is well specified and the perceptual processes that are being tested are sufficiently low in the processing stream as to be expected to be sensitive to the dimensions of the input space. Can the same be said of emotional face stimuli and processes measured by EEG?

Considering facial emotion as a form of social communication, there is a strong argument that its production and perception should share a common neural basis (see e.g. [65] for an evolutionary argument). In this case we should expect the space of facial muscle activation to modulate the brain response of the observer more strongly than the space of stimulus pixel intensity, which is only indirectly affected by the process of expression generation. This aspect of the current research is more critical than any other to our future understanding of when and how the brain processes facial emotion, as it is the one that directly addresses the question of what it is that the brain

is meant to be processing.

This is an appropriate note to conclude on; a logical extension of the theme of this thesis — for all its potential power to map out multiple associations between features of the brain and the external world with which it interacts, the application of reverse correlation to EEG, MEG, and related brain imaging modalities is useless without first getting the right features from the raw images. While the model presented here may not get those features right, it is at least attempting to find them.

References

[1] R. Adolphs, D. Tranel, H. Damasio, and A. Damasio. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature*, 372(6507):669–672, Dec 1994.

[2] R. Adolphs, D. Tranel, H. Damasio, and A. R. Damasio. Fear and the human amygdala. *J Neurosci*, 15(9):5879–5891, Sep 1995.

[3] A. J. Ahumada and J. Lovell. Stimulus features in signal detection. *The Journal of the Acoustical Society of America*, 49(6B):1751–1756, 1971.

[4] F. Babiloni, C. Babiloni, L. Fattorini, F. Carducci, P. Onorati, and A. Urbano. Performances of surface laplacian estimators: a study of simulated and real scalp potential distributions. *Brain Topogr*, 8(1):35–45, 1995.

[5] S. Baillet and L. Garnero. A bayesian approach to introducing anatomo-functional priors in the eeg/meg inverse problem. *IEEE Trans Biomed Eng*, 44(5):374–385, May 1997.

[6] Jason J S Barton, Daniel Z Press, Julian P Keenan, and Margaret O'Connor. Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. *Neurology*, 58(1):71–78, Jan 2002.

[7] Erol Basar, Canan Basar-Eroglu, Sirel Karakas, and Martin Schumann. Are cognitive processes manifested in event-related gamma, alpha, theta and delta oscillations in the eeg? *Neuroscience Letters*, 259(3):165–168, 1999.

[8] Erol Basar, Canan Basar-Eroglu, Sirel Karakas, and Martin Schumann. Gamma, alpha, delta, and theta oscillations govern cognitive processes. *International Journal of Psychophysiology*, 39(2-3):241–248, 2001.

[9] Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[10] Shlomo Bentin, Truett Allison, Aina Puce, Erik Perez, and Gregory McCarthy. Electrophysiological studies of face perception in humans. *J. Cognitive Neuroscience*, 8(6):551–565, 1996.

[11] R. J. Blair, J. S. Morris, C. D. Frith, D. I. Perrett, and R. J. Dolan. Dissociable neural responses to facial expressions of sadness and anger. *Brain*, 122 (Pt 5):883–893, May 1999.

[12] R. Bro, K. Kjeldahl, A. K. Smilde, and H. A L Kiers. Cross-validation of component models: a critical look at current methods. *Anal Bioanal Chem*, 390(5):1241–1251, Mar 2008.

[13] A. E. Burgess, R. F. Wagner, R. J. Jennings, and H. B. Barlow. Efficiency of human visual signal discrimination. *Science*, 214(4516):93–94, Oct 1981.

[14] A. J. Calder, J. Keane, F. Manes, N. Antoun, and A. W. Young. Impaired recognition and experience of disgust following brain injury. *Nat Neurosci*, 3(11):1077–1078, Nov 2000.

[15] Colin Cameron and Frank Windmeijer. An r-squared measure of goodness of fit for some common non-linear regression models. *Journal of Econometrics*, Jan 1997.

[16] Alan Chauvin, Keith J Worsley, Philippe G Schyns, Martin Arguin, and Frdric Gosselin. Accurate statistical tests for smooth classification images. *J Vis*, 5(9):659–667, 2005.

[17] P. Comon. Independent component analysis: A new concept? *Signal Processing*, 36:287–314, 1994.

[18] Douglas W. Cunningham, Mario Kleiner, Heirich H. Bülthoff, and Christian Wallraven. The components of conversational facial expressions. In *APGV '04: Proceedings of the 1st Symposium on Applied perception in graphics and visualization*, pages 143–150, New York, NY, USA, 2004. ACM.

[19] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *J Pers Soc Psychol*, 17(2):124–129, Feb 1971.

[20] H. Hamdi Eryilmaz, Adil Deniz Duru, Burak Parlak, Ahmet Ademoglu, and Tamer Demiralp. Neuroimaging of event related brain potentials (erp) using fmri and dipole source reconstruction. *Conf Proc IEEE Eng Med Biol Soc*, 2007:3384–3387, 2007.

[21] Adam D Gerson, Lucas C Parra, and Paul Sajda. Cortical origins of response time variability during rapid discrimination of visual objects. *Neuroimage*, 28(2):342–353, Nov 2005.

[22] F. Gosselin and P. G. Schyns. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Res*, 41(17):2261–2271, Aug 2001.

[23] Haxby, Hoffman, and Gobbini. The distributed human neural system for face perception. *Trends Cogn Sci*, 4(6):223–233, Jun 2000.

[24] R. N. Henson, Y. Goshen-Gottstein, T. Ganel, L. J. Otten, A. Quayle, and M. D. Rugg. Electrophysiological and haemodynamic correlates of face perception, recognition and priming. *Cereb Cortex*, 13(7):793–805, Jul 2003.

[25] Christoph S. Herrmann, Axel Macklinger, and Erdmut Pfeifer. Gamma responses and erps in a visual classification task. *Clinical Neurophysiology*, 110:636–642, 1999.

[26] Christoph S. Herrmann and Axel Mecklinger. Gamma activity in human eeg is related to high-speed memory comparisons during object selective attention. *Visual Cognition*, 8(3/4/5):593–608, 2001.

[27] Silvina G Horovitz, Bruno Rossion, Pawel Skudlarski, and John C Gore. Parametric design and correlational analyses help integrating fmri and electrophysiological data during face processing. *Neuroimage*, 22(4):1587–1595, Aug 2004.

[28] Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*, 10:626–634, 1999.

[29] Aapo Hyvarinen, Pavan Ramkumar, Lauri Parkkonen, and Riita Hari. Independent component analysis of short-time fourier transforms for spontaneous eeg/meg analysis. *NeuroImage*, 49(1):257–271, 2010.

[30] Tetsuya Iidaka, Atsushi Matsumoto, Kaoruko Haneda, Tomohisa Okada, and Norihiro Sadato. Hemodynamic and electrophysiological relationship involved in human face processing: evidence from a combined fmri-erp study. *Brain Cogn*, 60(2):176–186, Mar 2006.

[31] Carlos M. Jarque and Anil K. Bera. A test of normality of observations and regression residuals. *International Statstical Review*, 55(2):163–172, 1987.

[32] D. A. Jeffreys. A face-responsive potential recorded from the human scalp. *Exp Brain Res*, 78(1):193–202, 1989.

[33] D. A. Jeffreys. Visual evoked potential evidence for parallel processing of depth- and form-related information in human visual cortex. *Exp Brain Res*, 111(1):79–99, Sep 1996.

[34] Richard M. Leahy John C. Mosher and Paul S. Lewis. Eeg and meg: Forward solutions for inverse methods. *IEEE Transactions on Biomedical Engineering*, 46(3):245–259, 1999.

[35] Daniel Jokisch and Ole Jensen. Modulation of gamma and alpha activity during a working memory task engaging the dorsal or ventral stream. *The Journal of Neuroscience Letters*, 27(12):3244–3251, 2007.

[36] Carrie Joyce and Bruno Rossion. The face-sensitive n170 and vpp components manifest the same brain processes: the effect of reference electrode site. *Clin Neurophysiol*, 116(11):2613–2631, Nov 2005.

[37] Christian Jutten and Jeanny Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[38] N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*, 17(11):4302–4311, June 1997.

[39] Mitchell T. M. Generative and discriminative classifiers: Naive bayes and logistic regression. http://www.cs.cmu.edu/ tom/mlbook/NBayesLogReg.pdf.

[40] P. Z. Marmarelis and G. D. McCann. Development and application of white-noise modeling techniques for studies of insect visual nervous system. *Kybernetik*, 12(2):74–89, Feb 1973.

[41] T Martens, H Naes. *Multivariate calibration*. Wiley, 1989.

[42] P. Mccullagh and J. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC, August 1989.

[43] Axel Mecklinger and Peter Ullsperger. P3 varies with stimulus categorization rather than probability. *Electroenceph clin Neurophysiol*, 86:395–407, 1993.

[44] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72(23):3634–3637, Jun 1994.

[45] J. Morlet, G. Arens, E. Foureau, and D. Glard. Wave propagation and sampling theory. *Geophysics*, 47(2):203–236, 1982.

[46] J. C. Mosher, P. S. Lewis, and R. M. Leahy. Multiple dipole modeling and localization from spatio-temporal meg data. *IEEE Trans Biomed Eng*, 39(6):541–557, Jun 1992.

[47] Richard F Murray, Patrick J Bennett, and Allison B Sekuler. Optimal methods for calculating classification images: weighted sums. *J Vis*, 2(1):79–104, 2002.

[48] Richard F Murray and Jason M Gold. Troubles with bubbles. *Vision Res*, 44(5):461–470, Mar 2004.

[49] J. Narumoto, T. Okada, N. Sadato, K. Fukui, and Y. Yonekura. Attention to emotion modulates fmri activity in human right superior temporal sulcus. *Brain Res Cogn Brain Res*, 12(2):225–231, Oct 2001.

[50] Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems*, 14, 2002.

[51] Bertrand O., Perrin F ., and Pernier J. A theoretical justification of the average reference in topographic evoked potential studies. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 62(6):462–464, 1985.

[52] Lucas Parra and Paul Sajda. Blind source separation via generalized eigenvalue decomposition. *J. Mach. Learn. Res.*, 4:1261–1269, 2003.

[53] Hal Pashler. *Stevens' handbook of experimental psychology*. Wiley, 2001.

[54] Bijan Pesaran, Matthew J Nelson, and Richard A Andersen. Dorsal premotor neurons encode the relative position of the hand, eye, and goal during reach planning. *Neuron*, 51(1):125–134, Jul 2006.

[55] Dinh Tuan Pham. Joint approximate diagonalization of positive definite hermitian matrices. *SIAM J. Matrix Anal. Appl.*, 22(4):1136–1152, 2000.

[56] Dinh-Tuan Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of nonstationary sources. $IEEE_{JS}P, 49(9) : 1837 - -1848, September 2001$.

[57] M. L. Phillips, E. T. Bullmore, R. Howard, P. W. Woodruff, I. C. Wright, S. C. Williams, A. Simmons, C. Andrew, M. Brammer, and A. S. David. Investigation of facial recognition memory and happy and sad facial expression perception: an fmri study. *Psychiatry Res*, 83(3):127–138, Sep 1998.

[58] John Polich. Updating p300: an integrative theory of p3a and p3b. *Clin Neurophysiol*, 118(10):2128–2148, Oct 2007.

[59] K. R. Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1934.

[60] Guillaume A Rousselet, Jesse S Husk, Patrick J Bennett, and Allison B Sekuler. Single-trial eeg dynamics of object and face visual processing. *Neuroimage*, 36(3):843–862, Jul 2007.

[61] Boaz Sadeh, Andrey Zhdanov, Ilana Podlipsky, Talma Hendler, and Galit Yovel. The validity of the face-selective erp n170 component during simultaneous recording with functional mri. *Neuroimage*, 42(2):778–786, Aug 2008.

[62] Philippe G Schyns, Ines Jentzsch, Mark Johnson, Stefan R Schweinberger, and Frdric Gosselin. A principled method for determining the functionality of brain responses. *Neuroreport*, 14(13):1665–1669, Sep 2003.

[63] Philippe G Schyns, Lucy S Petro, and Marie L Smith. Dynamics of visual information integration in the brain for categorizing facial expressions. *Curr Biol*, 17(18):1580–1585, Sep 2007.

[64] Philippe G Schyns, Lucy S Petro, and Marie L Smith. Transmission of facial expressions of emotion co-evolved with their efficient decoding in the brain: behavioral and brain evidence. *PLoS One*, 4(5):e5625, 2009.

[65] Robert M. Seyfarth and Dorothy L. Cheney. Signalers and receivers in animal communication. *Annual Review of Psychology*, 54:145–173, 2003.

[66] Marie L Smith, Garrison W Cottrell, Frdric Gosselin, and Philippe G Schyns. Transmitting and decoding facial expressions. *Psychol Sci*, 16(3):184–189, Mar 2005.

[67] Marie L Smith, P. Fries, F. Gosselin, R. Goebel, and P. G. Schyns. Inverse mapping the neuronal substrates of face categorizations. *Cereb Cortex*, 19(10):2428–2438, Oct 2009.

[68] Marie L Smith, Frdric Gosselin, and Philippe G Schyns. Receptive fields for flexible face categorizations. *Psychol Sci*, 15(11):753–761, Nov 2004.

[69] M. Streit, A. A. Ioannides, L. Liu, W. Wlwer, J. Dammers, J. Gross, W. Gaebel, and H. W. Mller-Grtner. Neurophysiological correlates of the recognition of facial expressions of emotion as revealed by magnetoencephalography. *Brain Res Cogn Brain Res*, 7(4):481–491, Mar 1999.

[70] S. Sutton, M. Braren, J. Zubin, and E. R. John. Evoked-potential correlates of stimulus uncertainty. *Science*, 150(700):1187–1188, Nov 1965.

[71] Peter D. Welch. The use of fast fourier tansform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15:70–73, 1967.