UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

INGENIERÍA TÉCNICA DE TELECOMUNICACIÓN
SISTEMAS DE TELECOMUNICACIÓN

PROYECTO FINAL DE CARRERA

ANÁLISIS DE IMÁGENES DE RESONANCIA MAGNÉTICA FUNCIONAL DEL CEREBRO HUMANO MEDIANTE PROCESOS GAUSSIANOS

AUTOR: LORENA GÓMEZ DE LA FUENTE
TUTOR: MANEL MARTÍNEZ RAMÓN

8 de Septiembre de 2014

Agradecimientos

A mis padres, Maxi y Lola, por guiarme cuando estoy perdida, por vuestro amor, por no dejar que me rinda y por confiar siempre en mi.

A Vane, por ser mi "hermana"mayor, por tus consejos, por poder compartirlo todo contigo.

A Sara, Belén y Raquel, por haber crecido juntas y porque un trocito de mi es vuestro. Porque sé que siempre puedo y podré contar con vosotras.

A Tati y María, compañeras de carrera pero sobre todo, porque no podéis ser mejores compañeras de vida. Por ser auténticas. Que no me faltéis.

A Ramona y Julia, por ser unas orgullosas abuelas de su nieta. Por todo lo que he aprendido y aún aprendo. Por vuestro cariño incondicional.

A todos ellos, por su ánimo y apoyo siempre.

A Manel, por tu ayuda durante la realización de este proyecto.

Resumen

En este proyecto se estudia el uso de procesos gaussianos como una alternativa a los métodos basados en SPM para la generación de mapas de actividad cerebral a partir de secuencias de resonancia magnética funcional.

Los tests que se han llevado a cabo consisten en la estimación de estímulos en experimentos monosujeto y multisujeto en fMRI. Las gráficas muestran que los valores de los mapas β producidos por el método GP son similares a las estimaciones del método SPM, que consisten en mapas de t de Student.

A la vista de los resultados se puede afirmar que es posible establecer una metodología alternativa a las técnicas actuales de SPM obteniendo resultados competitivos. Las líneas sobre las que se basarían trabajos futuros consistirían en el estudio de modelos alternativos para el ruido, que pueden ser iguales a los usados en SPM. También se podrían establecer métodos de umbralización de los mapas basándose en la estimación de los intervalos de confianza de los resultados.

Índice general

Capítulo 1

MAPAS ESTADÍSTICOS PARAMÉTRICOS

1.1. Introducción

SPM (*Statistical Parametric Mapping*) es una técnica cuya finalidad es la realización de mapas de estadísticos paramétricos para la búsqueda de efectos de interés presentes en imágenes funcionales *PET* (Tomografía por Emisión de Positrones), *SPECT* (Tomografía por Emisión de Fotón Único)o *fMRI* (Resonancia Magnética funcional). *SPM* se utiliza actualmente en departamentos de psiquiatría, psicología, neurología, radiología, medicina nuclear, farmacología, ciencias cognitivas y del comportamiento, bioestadística y física biomédica de todo el mundo para la investigación de enfermedades mentales, cuantificación de efectos farmacológicos, estudios cognitivos, realización de análisis longitudinales, estudios intersujeto, e incluso morfométricos [4] [5].

Este método es univariante, es decir, relaciona cada vóxel independientemente con el estímulo o estímulos aplicados, sin tener en cuenta las relaciones entre las áreas del cerebro, y es lineal, por lo tanto es limitado debido a la posible naturaleza no lineal de los datos.

1.2. Preprocesado

Los mapas paramétricos estadísticos (*SPMs*) son imágenes cuyos vóxeles representan valores que están distribuidos de acuerdo con una función de densidad de probabilidad.

Un estudio de imagen funcional mediante *SPM* requiere de una serie de transformaciones previas de las imágenes para reducir componentes de varianza indeseadas y que sea posible el estudio estadístico propiamente dicho. Este pre-proceso consta de tres etapas: a) realineado, b)

normalización y c) filtrado espacial [4] [5].

Una vez superadas, las imágenes están en disposición de incluirse en el estudio estadístico, que a su vez está dividido en dos etapas más: a) análisis estadístico y b) inferencia estadística.

A continuación se detalla cada una de estas etapas.

1.2.1. Realineado

Este paso de procesado previo tan sólo se aplica en el caso de que se disponga de varias imágenes de un mismo sujeto. Consiste en estimar la diferencia de posición entre las distintas imágenes, debida a la diferente colocación de la cabeza del sujeto dentro del dispositivo de imagen (PET, $SPECT$, $fMRI$). Para corregirla, se aplican las traslaciones y rotaciones adecuadas que compensen esta diferencia, de modo que las imágenes coincidan en el mismo espacio común. Estos movimientos de pacientes podrían estar relacionados con la tarea llevada a cabo en el momento de la adquisición, especialmente en ensayos cognitivos neuropsicológicos, por lo que a veces puede ser interesante incluir las estimaciones del movimiento como variables en el análisis estadístico [16] .

El proceso de realineado corrige las diferencias de posición entre imágenes de un mismo sujeto, pero no es capaz de colocar en un espacio común imágenes de distintos sujetos. Esta es la finalidad de la siguiente etapa, la normalización espacial.

1.2.2. Normalización espacial

Para realizar un análisis vóxel a vóxel, los datos de distintos sujetos deben corresponderse con un espacio anatómico estándar. Establecer esta correspondencia se denomina *normalización espacial*, y permite la comparación entre sujetos y la presentación de los resultados de un modo convencional.

En esta etapa se realiza una deformación elástica de las imágenes de modo que concuerden con un patrón anatómico estandarizado. Para que la transformación espacial sea correcta, las imágenes deben ser razonablemente similares al patrón utilizado, tanto morfológicamente como en contraste. La metodología seguida para construir el patrón en el caso de imágenes PET consiste en un promedio de 12 estudios realizados en sujetos sanos. De forma similar, el patrón de SPECT está formado a partir del promediado de 22 estudios efectuados a mujeres sanas utilizando 99mTc-HMPAO6. En nuestro caso particular de estudio se utilizaron imágenes de resonancia magnética

(fMRI). De este modo, se ponen en correspondencia cada una de las imágenes cerebrales de cada sujeto con una localización homóloga en el espacio estándar. De otro modo, es posible que el algoritmo sea incapaz de encontrar la transformación global óptima.

Esta normalización, además de permitir la comparación vóxel a vóxel de las imágenes, también facilita la localización de las áreas funcionales. El concepto de sistematizar la localización cerebral de las regiones funcionales se debe originalmente a *Talairach* [14], y si bien SPM presenta los resultados finales mediante este método, el sistema de coordenadas empleado para informar acerca de las localizaciones no es el mismo que el que aparece en el atlas de *Talairach*, lo que puede inducir a error.

Es importante destacar que el programa no realiza la verificación automática de la normalización obtenida, por ello la normalización espacial debe validarse mediante comparación visual de las imágenes normalizadas con el patrón utilizado. Las diferencias entre ambas deben encontrarse en los distintos niveles de intensidad, debidos a las características metabólicas individuales del sujeto bajo estudio. También habrá diferencias al ruido presente en la imagen, el cual será reducido en la siguiente etapa filtrado espacial [2] [1].

1.2.3. Filtrado espacial (Suavizado de la imagen)

El filtrado es un proceso por el cual los vóxeles se promedian con sus vecinos, produciendo un suavizado de las imágenes, más o menos pronunciado en función de un parámetro denominado *Amplitud Total a Media Altura o Full Width at Half Maximun* $FWHM$. La $FWHM$ tiene unidades espaciales y mide el grado de suavizado: a mayor $FWHM$, mayor suavizado. Como guía se suele utilizar la regla de que la $FWHM$ sea, al menos, tres veces mayor que el tamaño de vóxel. Debe tenerse en cuenta que el grado de filtrado aplicado afecta a los resultados, siendo necesario establecer un compromiso en función del tamaño esperado de las áreas de activación, el número de pacientes y el ruido de las imágenes.

El suavizado de las imágenes tiene diversos objetivos. En primer lugar, aumenta la relación señal/ruido, ya que elimina fundamentalmente las componentes ruidosas de la imagen. Otro motivo que hace conveniente suavizar las imágenes es que así se garantiza que los cambios entre sujetos se presentarán en escalas suficientemente grandes como para ser anatómicamente significativas, una vez efectuado una normalización en intensidad. Es muy poco probable que se produzcan analogías significativas entre dos sujetos distintos a escalas muy pequeñas. El tercer

motivo para filtrar las imágenes es que así se ajustan mejor a un modelo de campos gaussianos. Esto es importante, ya que la inferencia estadística utilizará la teoría de campos gaussianos para detectar efectos regionales específicos [17].

Una vez filtradas, las imágenes ya están preparadas para ser analizadas estadísticamente. Las etapas de procesado previas al análisis estadístico tan sólo deben efectuarse una vez, después de la cual pueden aplicarse, en principio, en tantos diseños de estudios como se desee.

En la figura 1.1 podemos ver un esquema del sistema SPM.

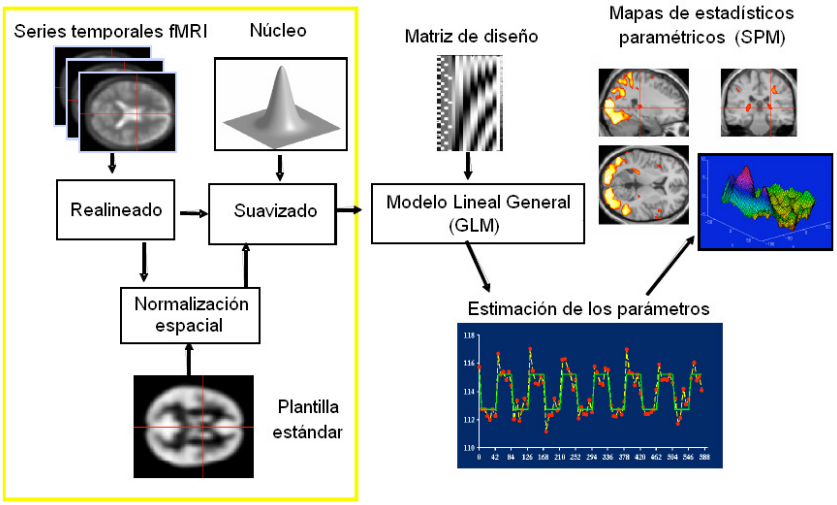

Figura 1.1: Esquema de un sistema SPM [9] [10]

1.2.4. Inferencia estadística

Mediante SPM es posible realizar numerosos test estadísticos, como regresiones, test t *de Student*, test F y análisis de varianza incluyendo covariables y permitiendo el modelado de iteracciones entre ellas.

Para el caso en el que la función de densidad de probabilidad fuera una t de Student o una distribución F (mapas t o F), cuanto mayor sea un valor dado, más improbable es la hipótesis nula de ese vóxel. Para estimar estos valores, se hace un análisis de todos los vóxeles usando un test estadístico y los parámetros estadísticos resultantes son reunidos en una imagen, llamada t-map o F-map.

Todos estos tipos de análisis pueden ser englobados en un modelo general (el *Modelo General Lineal* o GLM), que es utilizado por SPM para efectuar los cálculos matemáticos.

El resultado del análisis estadístico es un valor p para cada vóxel de la imágen (figuras 1.2 y 1.3), el cual representa la probabilidad de ausencia de efectos significativos.

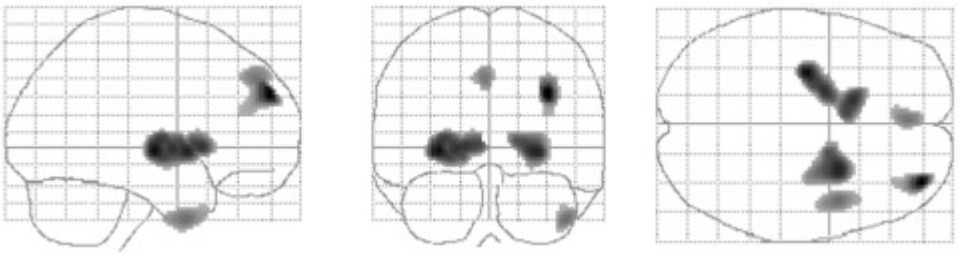

Figura 1.2: Resultado de SPM. La figura muestra zonas de menor actividad metabólica en 12 pacientes de esquizofrenia después de haber sido tratados con un nuevo fármaco neuroléptico

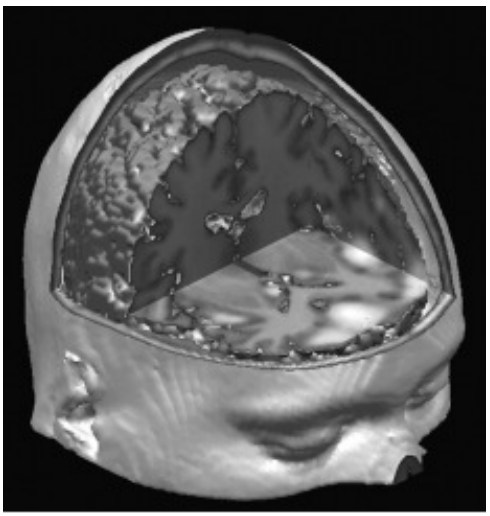

Figura 1.3: Presentación avanzada de resultados mediante SPM. La figura revela zonas de menor actividad metabólica en pacientes de esquizofrenia tratados con un nuevo fármaco neuroléptico con respecto a un grupo de control

Sin embargo, en un estudio PET hay muchos vóxeles, que al ser analizados independientemente, dan lugar a un elevado número de valores p. Al realizar un número tan grande de test estadísticos aparece un cierto numero de valores p que superan un umbral de significación establecido, tan sólo debido al azar. Este es el problema de la comparaciones múltiples. En concreto y dada su propia definición, en caso de que no haya efectos neurofísicos presentes en las imágenes, se espera que aparezca un 1 por ciento de valores p menores de 0,01, un 0,1 por ciento menores que 0,001, etc.

Estos son los denominados falsos positivos o errores tipo I. El problema de las comparaciones múltiples se solventa habitualmente mediante la correción de *Bonferroni*. Con ella, el valor p a partir del cual se acepta la hipótesis de partida, se calcula como α /(número de test). α es la tasa de falsos positivos que se está dispuesto a aceptar para el estudio en su conjunto (habitualmente igual a 0,05, es decir, un falso positivo cada 20 test). Pero la corrección de *Bonferroni* asume que los test son independientes entre sí,lo cual no ocurre en el caso de neuroimágenes, realizando una corrección excesivamente conservadora [3].

Para tratar el problema de la no-independencia entre vóxeles, sobre todo los cercanos, de un modo mas adecuado que la simple corrección de *Bonferroni*, SPM hace uso de la llamada teoría de campos gaussianos. Sus fundamentos son relativamente complejos y basta saber que proporciona un valor de p corregido.

Típicamente, el valor de p corregido por comparaciones múltiples a partir del cual se acepta que un efecto es significativo es de p=0,05. Este umbral se establece a priori y ofrece una protección de un falso positivo cada 20 observaciones, siempre y cuando se cumplan estrictamente todos los supuestos implicados en el proceso. Esto rara vez ocurre en la práctica, por lo que es extremadamente complicado establecer criterios objetivos para determinar un umbral a partir del cual los valores p deban aceptarse como realmente significativos.

1.3. Modelo General Lineal (GLM)

1.3.1. Introducción

A partir de un experimento fMRI, se obtienen N bloques BOLD (señales dependientes del nivel de oxigenación de la sangre), esto es, una serie temporal de N imágenes tridimensionales que indican la variación de oxigenación en el cerebro dentro de cada uno de los vóxeles del tejido cerebral. Estas imágenes se almacenan a efectos de su representación visual en matrices tridimensionales con una ordenación siguiendo los tres ejes del espacio de la imagen. Desde el punto de vista de procesado de la señal, los valores de estas matrices (vóxeles) se redistribuyen de forma arbitraria en vectores $\mathbf{y}\,[n]$ de dimensión M donde $0 \leq n \leq N-1$, es el índice temporal de la serie. La concatenación en columnas de estos vectores componen la *matriz de respuesta hermodinámica* $\mathbf{Y}$, de dimensiones $N \times M$.

El experimento fMRI consiste en la realización de una actividad (visual, motor, cognitiva, auditiva) a intervalos preestablecidos que se señalan mediante algún medio sensorial al sujeto.

Los instantes de inicio y final de los estímulos y actividades en cada bloque se representan mediante unas señales indicadores binarias, denominadas también vectores de referencia o variables explicatorias. Cada una de estas L señales se filtran mediante un filtro gaussiano que aproxima la respuesta hermodinámica del cerebro a la actividad o estímulo en cuestión. El conjunto de vectores de referencia filtrados $\mathbf{x}_l[n]$, $0 \leq l \leq -1$ constituye la llamada *matriz de diseño*, $\mathbf{X}$, de dimensiones $N \times L$.

1.3.2. Modelo General Lineal: Monosujeto

Todos los estadísticos paramétricos están basados en el *Modelo General Lineal* o GLM. Este modelo se diseña de forma que, para cada vóxel, se estima literalmente la respuesta hemodinámica a partir de la matriz de diseño. Concretamente, para cada vóxel m el modelo se puede expresar como:

$$y_m[n] = \mathbf{x}^\top[n]\,\boldsymbol{\beta}_{m_1} + \boldsymbol{\beta}_{m_0} + e_m[n] \tag{1.1}$$

donde $e[n]$ es el error de predicción, $\boldsymbol{\beta}_{m_1}$ es un vector de L componentes correspondientes a otros tantos estímulos y/o actividades y $\boldsymbol{\beta}_{m_0}$ es un factor de sesgo. Usualmente se incluye una componente constante en la matriz de diseño para absorver este sesgo. Alternativamente, se puede eliminar el sesgo de la serie temporal. Este método es univariante, es decir, la estimación de cada vector $\boldsymbol{\beta}_{m_1}$ sólo depende de la respuesta hemodinámica en el vóxel m. Sin embargo, el problema puede resolverse en bloque (sin que por ello el método gane un carácter multivariable), para lo que puede ser escrito en notación matricial de la siguiente forma:

$$\mathbf{Y} = \mathbf{X}^\top\boldsymbol{\beta} + \mathbf{E} \tag{1.2}$$

donde $\mathbf{Y}$ es la respuesta cerebral y $\mathbf{X}$ es la matriz de diseño anteriormente definida. El conjunto de parámetros $\boldsymbol{\beta}$ es una matriz de dimensiones $L{+}1{\times}M$. Las filas de esta matriz pueden ser reordenadas según el orden original de las matrices bloque. Las L primeras filas así reordenadas representan una imagen del cerebro cuyos vóxeles indican el nivel de presencia de cada uno de los estímulos en la correspondiente zona del cerebro. La última fila representa el sesgo de cada uno de los vóxeles en el estimador y, en principio, carace de interés.

La matriz $\mathbf{E}$ contiene el error de estimación por cada vóxel e instante de tiempo. Se supondrá que los errores son independientes e idénticamente distribuidos con media nula y covarianza $\Sigma = \sigma^2\mathbf{I}$.

Para resolver este problema, en el que normalmente el número de parámetros L es menor que el número de observaciones J, ($L < J$), se acude al método de *Least Squares*, o de *Mínimos Cuadrados*, que intenta encontrar la función que mejor se aproxima a los datos (un "mejor ajuste"), de acuerdo con el criterio de mínimo error cuadrático.

1.3.3. Estimación paramétrica

A partir del vector de parámetros estimados $\tilde{\boldsymbol{\beta}} = \left[\tilde{\boldsymbol{\beta}}_1 ... \tilde{\boldsymbol{\beta}}_L\right]^\top$, que cumplen $\tilde{\mathbf{Y}} = \left[\tilde{\mathbf{Y}}_1 ... \tilde{\mathbf{Y}}_N\right]^\top = \mathbf{X}^\top\tilde{\boldsymbol{\beta}}$, y del vector de errores (residuos) $\tilde{e} = [\tilde{e_1} ... \tilde{e_N}]^\top = \mathbf{Y} - \tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$, definimos la suma cuadrática de los errores, $S = \sum_{n=1}^{N} e^2 = e^\top e$. Es decir, esta suma es la diferencia cuadrática de las diferencias entre los valores reales y los estimados. El método de mínimos cuadrados consiste en minimizar esa suma cuadrática, o lo que es lo mismo, minimizar el error de estimación, obtener los valores óptimos de $\boldsymbol{\beta}$ que hacen mínimo el error. En resumen,

$$S = \sum_{n=1}^{N} \left(\mathbf{Y} - \tilde{\mathbf{Y}}\right) = \sum_{n=1}^{N} \left(\mathbf{Y}_n - x_{n_1}\tilde{\boldsymbol{\beta}}_1 - ... - x_{n_l}\tilde{\boldsymbol{\beta}}_L\right)^2 \tag{1.3}$$

La expresión anterior se hace mínima cuando:

$$\frac{\partial S}{\partial \tilde{\boldsymbol{\beta}}} = 2\sum_{n=1}^{N} (-x_{n_l}) \left(\mathbf{Y}_n - x_{n_1}\tilde{\boldsymbol{\beta}}_1 - ... - x_{nl}\tilde{\boldsymbol{\beta}}_L\right) = 0 \tag{1.4}$$

es decir, cuando $\mathbf{XY} = \left(\mathbf{XX}^\top\right)\tilde{\boldsymbol{\beta}}$. Entonces, el estimador de mínimo error cuadrático, descrito como $\hat{\boldsymbol{\beta}}$ satisface la ecuación normal $\mathbf{XY} = \left(\mathbf{XX}^\top\right)\hat{\boldsymbol{\beta}}$.

Para el caso del Modelo General Lineal, los estimadores de mínimo error cuadrático son los estimadores de máxima verosimilitud. Es decir, para aquellos parámetros estimados consistentes en la combinación lineal de datos, en los que se espera que la diferencia entre los valores reales y estimados sea mínima, los estimadores de mínimo error cuadrático son estimadores de mínima varianza.

Entonces, si la inversa de $\mathbf{XX}^\top$ existe, lo que ocurre cuando la matriz de diseño es de rango completo, las estimaciones de mínimos cuadrados son:

$$\boldsymbol{\beta}_{GLM} = \left(\mathbf{X}\mathbf{X}^{\top}\right)^{-1}\mathbf{X}\mathbf{Y} \tag{1.5}$$

1.3.4. Inferencia estadística paramétrica

La función densidad de probabilidad t de Student permite probar la importancia de una combinación lineal de efectos. Es decir, la hipótesis nula de que los efectos contenidos en $\mathbf{X}$ no son significativos se puede probar con el estadístico t usando componentes lineales o contrastes de las estimaciones de los parámetros $\boldsymbol{\beta}$. Un contraste $\mathbf{c}$ es un vector fila de pesos tal que $\mathbf{c}^{\top}$ es de dimensiones $1 \times L$, y cumple:

$$\boldsymbol{\beta_l} = \mathbf{c}^{\top}\boldsymbol{\beta} = [1\ 0..,0]\,\boldsymbol{\beta} \tag{1.6}$$

Si suponemos además un error independiente e idénticamente distribuido, una varianza estimada por mínimos cuadrados como $\hat{\sigma}^2 = \dfrac{e^{\top}e}{N-p}$, donde p es el rango de la matriz de diseño, $p = rank\,(\mathbf{X})$, y unos parámetros $\boldsymbol{\beta}$ normalmente distribuidos $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}\mathbf{X}^{\top}\right)^{-1}\right)$, podemos concluir que:

$$\mathbf{c}^{\top}\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\mathbf{c}^{\top}\boldsymbol{\beta}, \sigma^2\mathbf{c}^{\top}\left(\mathbf{X}\mathbf{X}^{\top}\right)^{-1}\mathbf{c}\right) \tag{1.7}$$

Es más, $\hat{\boldsymbol{\beta}}$ y $\mathbf{c}^{\top}$ son independientes y entonces:

$$\frac{\mathbf{c}^{\top}\hat{\boldsymbol{\beta}} - \mathbf{c}^{\top}\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2\mathbf{c}^{\top}\left(\mathbf{X}\mathbf{X}^{\top}\right)^{-1}\mathbf{c}}} \sim t_{N-p} \tag{1.8}$$

donde t_{N-p} es la distribución t de Student con N-p grados de libertad y $\mathbf{c}^{\top}\boldsymbol{\beta}$ es la hipótesis de la estimación. En SPM se cumple la hipótesis nula, $\mathbf{c}^{\top}\boldsymbol{\beta} = 0$, y por lo tanto:

$$T = \frac{\mathbf{c}^{\top}\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2\mathbf{c}^{\top}\left(\mathbf{X}\mathbf{X}^{\top}\right)^{-1}\mathbf{c}}} \tag{1.9}$$

donde la estimación de la varianza σ^2, es estimada a cada vóxel usando un estimador usual de mínimos cuadrados.

El uso del estadístico t no aporta ningún tipo de información aparte de la contenida en el parámetro $\boldsymbol{\beta}$ y no resulta de interés para la tarea de clasificación.

1.3.5. Estimación matriz de covarianza del error

Hasta ahora habíamos supuesto error independiente e idénticamente distribudo para explicar el Modelo General Lineal. En el caso particular de fMRI los datos están representados por series temporales, lo que implica que cada error en un escaner s, ϵ_s, está correlado con el error del instante siguiente, y por lo tanto no podemos considerar independencia en el rudio. Esto tiene que ser modelado ya que ignorar este hecho implicaría tener estadísticos t no válidos.

La matriz de covarianza del error se ajusta a un *Modelo autorregresivo de orden 1 y ruido blanco*, $AR(1) + w_n$. [1]. Para estimar esta matriz de covarianza del error, es necesario estimar tres hiperparámetros en cada vóxel k.

Esta matriz está formada por una primera componente, la matriz de correlación, y una segunda, la varianza. SPM asume que la matriz de correlación es la misma para todos los vóxeles, por lo que puede ser tratada como una cantidad constante, mientras que la varianza varía para cada uno de ellos.

A continuación vamos a estimar la matriz de covarianza del error, partiendo del modelo lineal

[1] El modelo $AR(1) + w_n$ puede escribirse como:

$$\epsilon(s) = z(s) + \delta_\epsilon(s)$$

$$z(s) = az(s-1) + \delta_z(s)$$

donde $\delta_\epsilon(s) \sim \mathcal{N}(0, \sigma_\epsilon^2)$, $\delta_z(s) \sim \mathcal{N}(0, \sigma_z^2)$, y a es el coeficiente $AR(1)$.

Es decir, se define el error $\epsilon(s)$ en el instante s y en el vóxel k como la suma de una componente autoregresiva $z(s)$ más ruido blanco $\delta_\epsilon(s)$. Existen tres hiperparámetros en cada vóxel k: las varianzas de las componentes de error δ_ϵ y δ_z y el coeficiente autoregresivo a.

La matriz de covarianza del error resultante viene descrita por la expresión siguiente:

$$E(\epsilon\epsilon^\top) = \sigma_z^2 (I_N - A)^{(-1)} (I_N - A)^{(-\top)} + \sigma_\epsilon^2$$

donde $\mathbf{A}$ es una matriz con todos los elementos debajo de la diagonal iguales a a y cero en el resto. $\mathbf{I}_N$ es la matriz identidad de dimensión N.

para el vóxel k:

$$\mathbf{y}^k = \mathbf{X}^\top \boldsymbol{\beta}^k + \epsilon^k \tag{1.10}$$

donde $\mathbf{y}^k$ es el vector de observaciones de tamaño $N \times 1$ en el vóxel k. $\mathbf{X}$ es la matriz de diseño $N \times L$, $\boldsymbol{\beta}^k$ es el vector de parámetros y ϵ^k es el error en cada vóxel k, $\epsilon \sim \mathcal{N}\left(0, \sigma^{k^2}\mathbf{V}\right)$. La principal diferencia con el modelo descrito en la ecuación (1.2) es la distribución del error, donde la matriz identidad $\mathbf{I}$ se sustituye por la matriz de correlación $\mathbf{V}$. Como hemos visto anteriormente, $\mathbf{V}$ no depende del vóxel k, si no que es la misma para todos ellos, $k = 1...K$. Por el contrario, la varianza σ^{k^2} sí es distinta para cada vóxel.

Dado que $\mathbf{V}$ es la misma para cada vóxel podemos aunar los datos para todos los vóxeles y luego estimar $\mathbf{V}$ para ese conjunto de datos. La matriz de covarianza para el total de los vóxeles es $\mathbf{V_y} = 1/K \sum_k \mathbf{y}^k \mathbf{y}^{k^\top}$, que a su vez puede descomponerse en:

$$\mathbf{V_y} = \sum_k \mathbf{X}\boldsymbol{\beta}^k \boldsymbol{\beta}^{k^\top} \mathbf{X}^\top + \epsilon^k \epsilon^{k^\top} \tag{1.11}$$

Para estimar los componentes de la matriz de covarianza del error, $\mathbf{cov}\left(\epsilon^k\right) = \sigma^{k^2}\mathbf{V}$, se utiliza el método de *Máxima Verosilimitud Reducido* (ReML) [7]. El modelo $AR(1) + w_n$ es no lineal, por lo que ReML no se puede aplicar directamente. Usamos una aproximación lineal de la forma:

$$\mathbf{V} = \sum_l \lambda_l \mathbf{Q}_l \tag{1.12}$$

donde $\mathbf{Q}_l$ es una matriz de tamaño $N \times N$ y λ_l, los hiperparámetros. Definimos $\mathbf{Q}_l$ de modo que se ajuste a un modelo adecuado para correlaciones serie en fMRI. El modelo en SPM viene descrito por dos componentes $\mathbf{Q}_1$ y $\mathbf{Q}_2$ (ver la figura 1.4), con $\mathbf{Q}_1$ igual a la matriz identidad de orden N, $\mathbf{Q}_1 = \mathbf{I}_N$ y:

$$\mathbf{Q}_{2_{ij}} = \begin{cases} e^{-|i-j|} : & \text{si } i \neq j \\ 0 : & \text{si } i = j \end{cases} \tag{1.13}$$

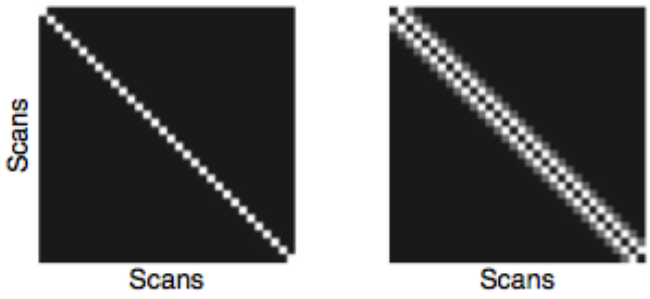

Figura 1.4: Representación de las dos componentes de covarianza. A la izquieda, $\mathbf{Q}_1$ corresponde a la componente de varaianza blanca y estacionaria. A la derecha, $\mathbf{Q}_2$ que implementa $AR(1)$ con coeficiente de autoregresión $1/e$

Para estimar la matriz de covarianza en cada vóxel se utilizan los hiperpárametros (globales) λ_1 y λ_2 y un tercer hiperparámetro (local) (la varianza σ^2), el cual es estimado usando un estimador de mínimos cuadrados univariante, obteniendo:

$$\sigma^{2^k} = \frac{\mathbf{y}^{k^\top}\mathbf{R}\mathbf{y}^k}{trace\,(\mathbf{RV})} \tag{1.14}$$

donde R es la matriz de residuos. Una vez realizada esta estimación sería posible establecer tests estadísticos sobre los parámetros de forma similar al caso con ruido independiente e idénticamente distribuido (ecuación 1.9)

1.3.6. Modelo General Lineal: Multisujeto

Vamos a ver los casos en el que tenemos k sujetos sometidos a experimentos independientes, y dos niveles. Usaremos notaciones diferentes para distinguirlos: (1) para el caso de primer nivel y (2) para el segundo nivel. Partiendo de esta notación, la ecuación (1.2) para el primer nivel queda ahora definida de la forma:

$$\mathbf{Y}^{(1)} = \mathbf{X}^{(1)^\top}\boldsymbol{\beta}^{(1)} + \mathbf{E}^{(1)} \tag{1.15}$$

donde $\mathbf{Y}^{(1)}$ es una matriz con todos los vóxeles y todos los instantes de tiempo para los k sujetos. Por lo tanto, sus dimensiones serán $Nk \times M$ (ver ecuación (1.16)). A su vez, $\mathbf{X}^{(1)}$ es una matriz de diseño compuesta por las matrices $\mathbf{X}$ (ver ecuación (1.17)), $\mathbf{E}^{(1)}$ es el conjunto de errores y $\boldsymbol{\beta}^{(1)}$ es un vector compuesto por las $\boldsymbol{\beta}$ de cada sujeto organizadas en bloques de $(L+1)\,k$ (ver ecuación (1.18)).

$$\mathbf{Y}^{(1)} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_k \end{bmatrix} \tag{1.16}$$

$$\mathbf{X}^{(1)} = \begin{bmatrix} \mathbf{X}_1 & \cdots & \cdots & 0 \\ \vdots & \mathbf{X}_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \mathbf{X}_k \end{bmatrix} \tag{1.17}$$

$$\boldsymbol{\beta}^{(1)} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_k \end{bmatrix} \tag{1.18}$$

Al tener dos niveles, lo que queremos es saber son las ecuaciones de $\boldsymbol{\beta}^{(1)}$ y $\boldsymbol{\beta}^{(2)}$. Existe una relación entre la $\boldsymbol{\beta}$ de primer nivel $\boldsymbol{\beta}^{(1)}$ y la de segundo nivel $\boldsymbol{\beta}^{(2)}$ dada por la ecuación

$$\boldsymbol{\beta}^{(1)} = {\mathbf{X}^{(2)}}^{\top} \boldsymbol{\beta}^{(2)} + \mathbf{E}^{(2)} \tag{1.19}$$

donde $\boldsymbol{\beta}^{(2)}$ es una matriz con un mapa común para todos los sujetos. Si todos los sujetos fueran iguales se cumpliría que $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}$. Por otro lado $\mathbf{X}^{(2)}$ es una matriz compuesta por matrices identidad $\mathbf{I}_1, \mathbf{I}_2, \ldots \mathbf{I}_k$, donde cada matriz identidad tiene unas dimensiones de $(L+1) \times (L+1)$. Esta matriz $\mathbf{X}^{(2)}$ transporta del espacio de $\boldsymbol{\beta}^{(2)}$ a los mapas de todos los sujetos. $\mathbf{E}^{(2)}$ es una

matriz diferencia o un conjunto de parámetros y no una matriz de error como lo es $\mathbf{E}^{(1)}$. A partir de esta relación podemos rescribir la ecuación (1.2) como

$$\mathbf{Y}^{(1)} = \mathbf{X}^{(1)^\top}\left(\mathbf{X}^{(2)^\top}\boldsymbol{\beta}^{(2)} + \mathbf{E}^{(2)}\right) + \mathbf{E}^{(1)} \tag{1.20}$$

$$\mathbf{Y}^{(1)} = \mathbf{X}^{(1)^\top}\mathbf{X}^{(2)^\top}\boldsymbol{\beta}^{(2)} + \mathbf{X}^{(1)^\top}\mathbf{E}^{(2)} + \mathbf{E}^{(1)} \tag{1.21}$$

$$\mathbf{Y}^{(1)} = \begin{bmatrix}\mathbf{X}^{(1)^\top}\mathbf{X}^{(2)^\top} & \mathbf{X}^{(1)^\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}^{(2)} \\ \mathbf{E}^{(2)}\end{bmatrix} + \mathbf{E}^{(1)} \tag{1.22}$$

Una representación de estos pasos, de cómo pasamos de la ecuación (1.19) a la ecuación (1.22) y de las matrices en un modelo de dos niveles la encontramos en la figura 1.5.

Queremos despejar $\boldsymbol{\beta}^{(2)}$ y $\mathbf{E}^{(2)}$ y para ello multiplicamos por la traspuesta de $\begin{bmatrix}\mathbf{X}^{(1)^\top}\mathbf{X}^{(2)^\top} & \mathbf{X}^{(1)^\top}\end{bmatrix}$ con sus bloques también transpuestos a ambos lados de la igualdad.

$$\begin{bmatrix}\mathbf{X}^{(2)}\mathbf{X}^{(1)} \\ \mathbf{X}^{(1)}\end{bmatrix}\mathbf{Y}^{(1)} = \begin{bmatrix}\mathbf{X}^{(2)}\mathbf{X}^{(1)} \\ \mathbf{X}^{(1)}\end{bmatrix}\begin{bmatrix}\mathbf{X}^{(1)^\top}\mathbf{X}^{(2)^\top} & \mathbf{X}^{(1)^\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}^{(2)} \\ \mathbf{E}^{(2)}\end{bmatrix} + \mathbf{E}^{(1)} \tag{1.23}$$

$$\begin{bmatrix}\mathbf{X}^{(2)}\mathbf{X}^{(1)}\mathbf{Y}^{(1)} \\ \mathbf{X}^{(1)}\mathbf{Y}^{(1)}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^{(2)}\mathbf{X}^{(1)}\mathbf{X}^{(1)^\top}\mathbf{X}^{(2)^\top} & \mathbf{X}^{(2)}\mathbf{X}^{(1)}\mathbf{X}^{(1)^\top} \\ \mathbf{X}^{(1)}\mathbf{X}^{(1)^\top}\mathbf{X}^{(2)^\top} & \mathbf{X}^{(1)}\mathbf{X}^{(1)^\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}^{(2)} \\ \mathbf{E}^{(2)}\end{bmatrix} + \mathbf{E}^{(1)} \tag{1.24}$$

Denotaremos $\mathbf{K_{XX}}^{(11)}$ como

$$\mathbf{K_{XX}}^{(11)} = \mathbf{X}^{(1)}\mathbf{X}^{(1)^\top} \tag{1.25}$$

donde $\mathbf{K_{XX}}^{(11)}$ se compone de una matriz de ceros cuya diagonal son las diferentes $\mathbf{K_{XX}}$ de cada sujeto (ver ecuación (1.27)). De la misma forma, denotamos $\mathbf{K_{XY}}^{(11)}$ como

$$\mathbf{K_{XY}}^{(11)} = \mathbf{X}^{(1)}\mathbf{Y}^{(1)} \tag{1.26}$$

donde $\mathbf{K_{XY}}^{(11)}$ sera un vector compuesto por los diferentes $\mathbf{K_{XY}}$ de cada sujeto (ver ecuación (1.28))

$$\mathbf{K_{XX}}^{(11)} = \begin{bmatrix} \mathbf{K_{XX}}(1,1) & \cdots & \cdots & 0 \\ \vdots & \mathbf{K_{XX}}(2,2) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \mathbf{K_{XX}}(k,k) \end{bmatrix} \tag{1.27}$$

$$\mathbf{K_{XY}}^{(11)} = \begin{bmatrix} \mathbf{K_{XY}}(1,1) \\ \mathbf{K_{XY}}(2,2) \\ \vdots \\ \mathbf{K_{XY}}(k,k) \end{bmatrix} \tag{1.28}$$

Si desarrollamos el producto del primer término de la matriz a la derecha de la igualdad en la ecuación (1.24) observamos que será igual a la suma de los productos escalares de cada matriz

$$\mathbf{X}^{(2)}\mathbf{K_{XY}}^{(11)} = \mathbf{K_{XY}}(1,1) + \mathbf{K_{XY}}(2,2) + \cdots + \mathbf{K_{XY}}(k,k) = \sum_{i=1}^{k} \mathbf{K_{XY}}(i,i) \tag{1.29}$$

Si hacemos lo mismo con el primer término de la matriz a la izquierda de la igualdad en la ecuación (1.24) tendremos

$$\mathbf{X}^{(2)}\mathbf{K_{XX}}^{(11)}\mathbf{X}^{(2)^\top} = \mathbf{K_{XX}}(1,1) + \mathbf{K_{XX}}(2,2) + \cdots + \mathbf{K_{XX}}(k,k) = \sum_{i=1}^{k} \mathbf{K_{XX}}(i,i) \tag{1.30}$$

Despejando de la ecuación (1.23) con las nuevas notaciones la solución es igual a

$$\begin{bmatrix} \boldsymbol{\beta}_{\text{GLM}}^{(2)} \\ \mathbf{E}^{(2)} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{k} \mathbf{K_{XX}}(i,i) & \begin{bmatrix} \mathbf{K_{XX}}(1,1) \cdots \mathbf{K_{XX}}(k,k) \end{bmatrix} \\ \begin{bmatrix} \mathbf{K_{XX}}(1,1) \\ \vdots \\ \mathbf{K_{XX}}(k,k) \end{bmatrix} & \begin{bmatrix} \mathbf{K_{XX}}(1,1) \cdots 0 \\ \vdots \quad \ddots \quad \vdots \\ 0 \cdots \mathbf{K_{XX}}(k,k) \end{bmatrix} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{k} \mathbf{K_{XY}}(i,i) \\ \begin{bmatrix} \mathbf{K_{XY}}(1,1) \\ \vdots \\ \mathbf{K_{XY}}(k,k) \end{bmatrix} \end{bmatrix} \tag{1.31}$$

La ecuación anterior puede expresarse finalmente como

$$\begin{bmatrix} \boldsymbol{\beta}^{(2)}_{\mathrm{GLM}} \\ \mathbf{E}^{(2)} \end{bmatrix} = \left(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\top}\right)^{-1} \widetilde{\mathbf{X}}\mathbf{Y} \tag{1.32}$$

donde

$$\widetilde{\mathbf{X}} = \left[\mathbf{X}^{(1)^{\top}}\mathbf{X}^{(2)^{\top}}, \mathbf{X}^{(1)^{\top}}\right] \tag{1.33}$$

Esta expresión tiene la misma forma que la ecuación de primer nivel para un sujeto. Por lo tanto, podemos concluir que el test estadístico se lleva a cabo de forma análoga al caso monosujeto.

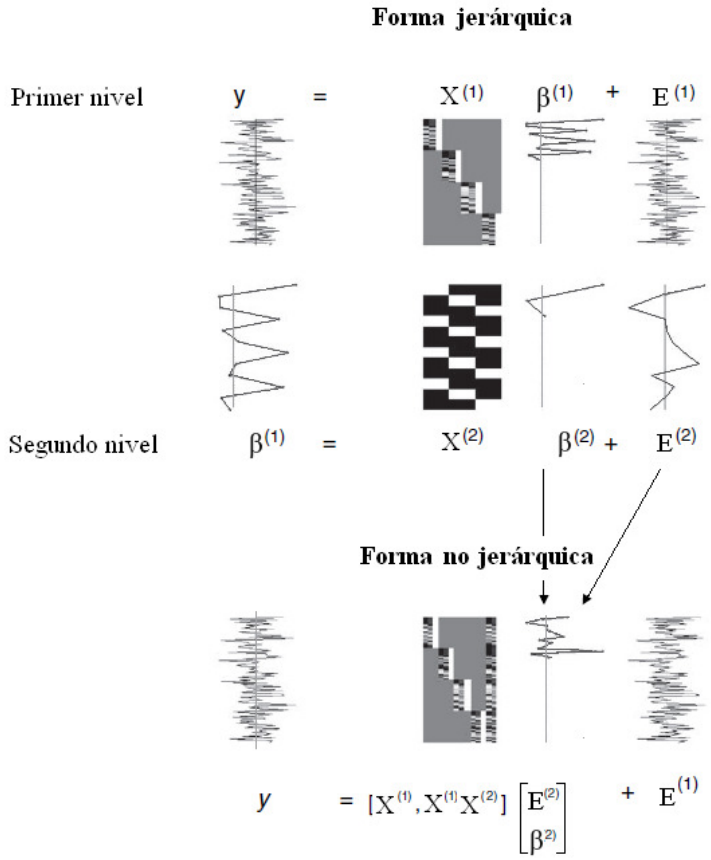

Figura 1.5: Esquema que muestra la forma de las matrices de diseño en un modelo de dos niveles y cómo la forma jerárquica (en la parte superior) se puede reducir a una forma no jerárquica (en la zona inferior). Las matrices de diseño se muestran en formato de imagen con una escala arbitraria de colores. Las variables de respuesta, parámetros y términos de error se representan con celdas. En este ejemplo hay cuatro sujetos o unidades observados en el primer nivel. La respuesta de cada sujeto se modela con los mismos tres efectos, uno de los cuales es un término constante.

Capítulo 2

PROCESOS GAUSSIANOS

2.1. Introducción

En este capítulo vamos a centrarnos en el aprendizaje supervisado que trata de predecir el valor de cualquier entrada a partir de un conjunto de datos observados (entrenamiento). Dependiendo de las características de la salida, el problema se conoce como *regresión* (salida continua) o *clasificación* (salida discreta)

Vamos a definir la entrada como $\mathbf{x}$, y la salida como y. La entrada se suele representar como un vector $\mathbf{x}$ ya que normalmente consideramos múltiples variables. La salida y puede ser continua (regresión) o discreta (clasificación). Las n observaciones vienen dadas por el conjunto de datos $\mathcal{D} = \{(\mathbf{x}_i, y_i) \,|i = 1, ..., n\}$.

A partir de este conjunto de datos de entrenamiento el objetivo es hacer predicciones para nuevas entradas $\mathbf{x}_*$ no consideradas en dicho conjunto. Para ello, tenemos que transformar el conjunto finito $\mathcal{D}$ en una función f que prediga todos los posibles valores de entrada. Esto se consigue haciendo suposiciones de las características de dicha función, o de otro modo cualquier función que sea consistente con el conjunto de entrenamiento sería válida. Existen diferentes métodos con los que tratar el problema del aprendizaje supervisado. Podemos restringir la clase de funciones que vamos a considerar, por ejemplo, utilizando sólo funciones lineales de entrada. El problema lo encontraríamos si la salida no es correctamente modelada por dichas funciones lineales, en cuyo caso las predicciones que estaríamos haciendo serían muy pobres. Podríamos entonces escoger funciones más flexibles, pero correríamos el riesgo de sobreajustar al conjunto de entrenamiento y obtener también malos resultados en las predicciones.

Otro modo de abordar el problema consistiría en dar una probabilidad a priori a cada posible función f, dando mayores probabilidades a aquellas que son más posibles de darse. El problema de éste radica precisamente en cómo tratar un conjunto infinito de posibles funciones. Éste es el objetivo de los *Procesos Gaussianos*.

Un proceso estocástico se define como la generalización de una distribución de probabilidad (la cual describe variables aleatorias finito dimensionales) en funciones. Un proceso gaussiano es un proceso estocástico con distribución de probabilidad normal o gaussiana. Por lo tanto, mientras que una *distribución* de probabilidad describe variables que pueden ser escalares o vectores (distribuciones multivariable), un *proceso* define las propiedades de las funciones. El problema de cómo tratar con un conjunto infinito en el espacio de dimensiones se reduce sencillamente en que si consideramos sólo las propiedades de la función f en un conjunto finito de puntos, la inferencia en un proceso gaussiano nos daría el mismo resultado si ignorásemos el conjunto infinito o si los tuviéramos todos en cuenta. Precisamente, tanto la flexibilidad en su estructura matemática como la facilidad en su implementación, han permitido que los procesos gaussianos hayan ganado gran aceptación en la comunidad de aprendizaje máquina, haciendo que este tipo de modelado probabilístico no paramétrico sea usado principalmente para realizar aprendizaje supervisado como regresión y clasificación.

2.2. Regresión

2.2.1. Modelo lineal estándar

El modelo de regresión lineal donde la salida es una combinación de las entradas ha sido ampliamente analizado y aplicado. Sus principales ventajas son la sencillez de implementación e interpretación.

En primer lugar daremos una perspectiva bayesiana al problema de regresión lineal y por último haremos una generalización en el espacio de características. Aplicaremos entonces el modelo lineal y recurriremos a las funciones *kernel* que reducen el coste computacional cuando la dimensión del espacio de características es grande en comparación con el número de datos.

Partimos de un conjunto $\mathcal{D}$ de n observaciones, tal que $\mathcal{D} = \{(\mathbf{x}_i, y_i) \,|i = 1, ..., n\}$, donde $\mathbf{x}$ es el vector de entradas de dimesión $\mathcal{D}$ e y es la salida (escalar). El vector de entrada es un vector columna para todos los valores n, representada por la *matriz de diseño* de dimensión $\mathcal{D} \times n$, $\mathbf{X}$. Las salidas vienen descritas por el vector $\mathbf{y}$. Es decir, tenemos $\mathcal{D} = (\mathbf{X}, \mathbf{y})$. Se trata entonces de

hacer inferencias sobre la relación entre las entradas y las salidas.

El modelo de regresión lineal estándar con ruido gaussiano tiene la forma:

$$y[n] = \mathbf{x}^\top \boldsymbol{\beta} + e[n] \tag{2.1}$$

donde $\mathbf{x}$ es el vector de entrada, $\boldsymbol{\beta}$ es el vector de pesos (parámetros) del modelo lineal y $e[n]$ es el ruido, que asumimos independiente e idénticamente distribuido siguiendo una distribución gaussiana de media 0 y varianza σ_n^2, es decir,

$$e[n] \sim \mathcal{N}\left(0, \sigma_n^2\right) \tag{2.2}$$

Por otro lado, $y\,[n]$ tiene media $\mathbf{x}^\top \boldsymbol{\beta}$ y varianza σ_n^2.

Asumiendo independencia en las observaciones, podemos definir la *verosimilitud*, o densidad de probabilidad de las observaciones conocidas $\mathbf{x}$ y $\boldsymbol{\beta}$, como:

$$\begin{aligned} p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}\right) &= \prod_{i=1}^{n} p\left(y_i|\mathbf{x}_i, \boldsymbol{\beta}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{2\sigma_n^2}\right) \\ &= \frac{1}{\left(2\pi\sigma_n^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2}|\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta}|^2\right) \sim \mathcal{N}\left(\mathbf{X}^\top \boldsymbol{\beta}, \sigma_n^2 \mathbf{I}\right) \end{aligned} \tag{2.3}$$

Sólo nos hace falta establecer una probabilidad a *priori* sobre los parámetros $\boldsymbol{\beta}$, que siguen una distribución gaussiana de media 0 y matriz de covarianza $\boldsymbol{\Sigma}_p$ desconocida.

$$\boldsymbol{\beta} \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}_p\right) \tag{2.4}$$

Para ello aplicamos el teorema de Bayes,

$$p\left(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}\right) = \frac{p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}\right) p\left(\boldsymbol{\beta}\right)}{p\left(\mathbf{y}|\mathbf{X}\right)} \tag{2.5}$$

donde $p\left(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}\right)$ es la probabilidad a posteriori. El primer término del numerador de la ecuación corresponde a la verosimilitud calculada antes mientras que el segundo término corresponde

precisamente a la probabilidad a priori de los parámetros que estamos buscando. El denominador es la *verosimilitud marginal* que viene descrita por:

$$p\left(\mathbf{y}|\mathbf{X}\right) = \int p\left(\mathbf{y}|\mathbf{X},\boldsymbol{\beta}\right) p\left(\boldsymbol{\beta}\right) d\boldsymbol{\beta} \tag{2.6}$$

Esta probabilidad es constante ya que al desarrollar la integral desaparece la dependencia con $\boldsymbol{\beta}$. Por lo que bastará con conocer la media y la varianza para tener completamente caracterizados los parámetros.

Podemos reescribir la probabilidad a posteriori de la ecuación (2.5) en términos de la verosimilitud y de la probabilidad a priori, obviando la verosimilitud marginal y haciéndola proporcional a los términos dependientes de $\boldsymbol{\beta}$.

$$p\left(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}\right) \propto \exp\left(-\frac{1}{2\sigma_n^2}\left(\mathbf{y}-\mathbf{X}^\top\boldsymbol{\beta}\right)^\top\left(\mathbf{y}-\mathbf{X}^\top\boldsymbol{\beta}\right)\right)\exp\left(-\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\beta}\right) \tag{2.7}$$

La primera exponencial de la ecuación anterior representa la verosimilitud y la segunda, la probabilidad a priori, $p\left(\boldsymbol{\beta}\right)$. Si continuamos simplificando,

$$p\left(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}\right) \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)^\top\left(\frac{1}{\sigma_n^2}\mathbf{X}\mathbf{X}^\top+\boldsymbol{\Sigma}_p^{-1}\right)\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)\right) \tag{2.8}$$

donde

$$\tilde{\boldsymbol{\beta}} = \sigma_n^{-2}\left(\sigma_n^{-2}\mathbf{X}\mathbf{X}^\top+\boldsymbol{\Sigma}_p^{-1}\right)^{-1}\mathbf{X}\mathbf{y}$$

o lo que es lo mismo,

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}\mathbf{X}^\top+\sigma_n^2\boldsymbol{\Sigma}_p^{-1}\right)^{-1}\mathbf{X}\mathbf{y} \tag{2.9}$$

Si comparamos la ecuación (2.9) con la (1.5) podemos concluir que el modelo GLM es un caso particular del modelo de regresión lineal basado en procesos gaussianos. Si suponemos $\sigma_n^2 = 0$ entonces estaríamos en el caso GLM puro, si $\sigma_n^2 0 > 0$ y $\boldsymbol{\Sigma}_p = \mathbf{I}$ entonces tendríamos un GLM regularizado.

La probabilidad a posteriori de la ecuación (2.8) sigue una distribución gaussiana de media $\tilde{\boldsymbol{\beta}}$ y matriz de covarianza $\mathbf{A}^{-1}$

$$p\left(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}\right) \sim \mathcal{N}\left(\tilde{\boldsymbol{\beta}} = \frac{1}{\sigma_n^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{y}, \mathbf{A}^{-1}\right) \tag{2.10}$$

donde $\mathbf{A} = \sigma_n^{-2}\mathbf{X}\mathbf{X}^\top + \boldsymbol{\Sigma}_p^{-1}$.

Para este modelo, la media de la probabilidad a posteriori $p\left(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}\right)$ es también la moda, por lo que se conoce también como estimador *Máximo a Posteriori* (MAP) de los parámetros $\boldsymbol{\beta}$

Para hacer predicciones para un nuevo valor de entrada, $\mathbf{x}_*$, tenemos que promediar todos los posibles valores de los parámetros $\boldsymbol{\beta}$. Así, la distribución predictiva para $f_* \triangleq f\left(\mathbf{x}_*\right)$ en $\mathbf{x}_*$ se obtiene al promediar la salida de todos los posibles modelos lineales con la probabilidad a posteriori gaussiana:

$$\begin{aligned} p\left(f_*|\mathbf{x}_*,\mathbf{X},\mathbf{y}\right) &= \int p\left(f_*|\mathbf{x}_*,\boldsymbol{\beta}\right) p\left(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}\right) d\boldsymbol{\beta} = \int \mathbf{x}_*^\top\boldsymbol{\beta}\; p\left(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}\right) d\boldsymbol{\beta} \\ &\sim \mathcal{N}\left(\frac{1}{\sigma_n^2}\mathbf{x}_*^\top\mathbf{A}^{-1}\mathbf{X}\mathbf{y}, \mathbf{x}_*^\top\mathbf{A}^{-1}\mathbf{x}_*\right) \end{aligned} \tag{2.11}$$

La distribución predictiva es también gaussiana, con una media dada por la media a posteriori de los parámetros de la ecuación (2.10) multiplicada por la entrada $\mathbf{x}_*$. La varianza es la combinación de la entrada $\mathbf{x}_*$ con la matriz de covarianza a posteriori $\mathbf{A}^{-1}$.

Tanto la probabilididad a posteriori como la verosimilitud son medidas que permiten conocer si el modelo es o no lo suficientemente bueno, si podemos hacer buenas predicciones. Para ello, se definirán los intervalos de confianza, que son una medida de la certidumbre (confiabilidad) que expresan la probabilidad de que los límites definidos por el intervalo incluyan el valor real del parámetro. Cuanto mayor sea la varianza más cerca estaremos de la hipótesis nula y de tener gran incertidumbre, y al revés, cuanto más alejado estemos de la hipótesis nula más preciso será nuestro modelo.

2.2.2. Modelo no lineal utilizando núcleos de Mercer

Las aplicaciones reales de regresión requieren hipótesis más complejas que las que se limitan a funciones lineales. Una idea para solucionar este problema consiste en proyectar las entradas

en algún espacio con una mayor dimensión usando un conjunto de funciones básicas y aplicar el modelo lineal en este nuevo espacio. Cuando los datos no son linealmente separables en el espacio de datos original pueden ser linealmente separables en un espacio característico con una dimensión mas alta que el original, llamado *hiperplano* o *espacio de características*(ver figura 2.1)

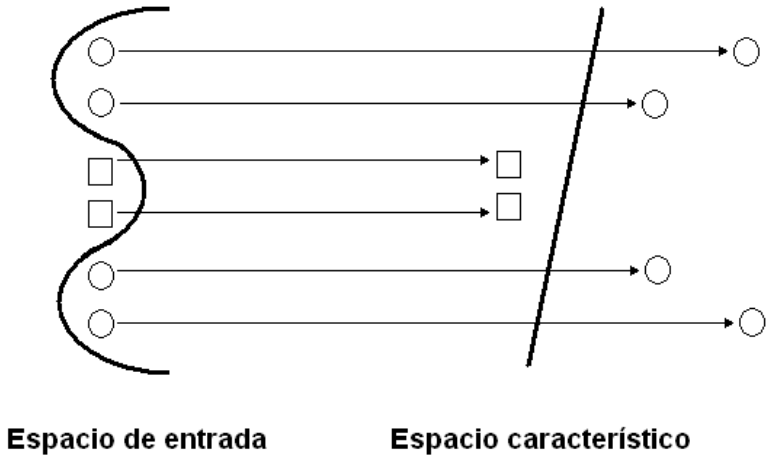

Figura 2.1: Transformación del espacio de entrada (1D) a un hiperplano o espacio característico de mayor dimensión (2D) mediante la transfomación lineal $\phi(\mathbf{x}) = \left(ax^2,\ bx,\ c\right)$

Cuanto mayor sea la dimensión a la que estamos transformando los datos, mayor será la probabilidad de obtener una solución lineal en el hiperplano o espacio característico.

Usaremos la función $\boldsymbol{\phi}(\mathbf{x})$ la cual mapea un vector de entrada de D dimensiones dentro de un espacio característico de N dimensiones llamado *espacio de Hilbert*.

$$\mathbf{x} : \mathbb{R} \ \rightarrow \ \boldsymbol{\phi}(\mathbf{x}) : \mathbb{H} \tag{2.12}$$

Si partimos del estimador de la ecuación (1.2) y hacemos la transformación de $\mathbf{x}$ al espacio de Hilbert podremos escribir el estimador como [8]:

$$\mathbf{Y} = \boldsymbol{\Phi}^{\top}\boldsymbol{\beta} + \mathbf{E} \tag{2.13}$$

Hemos visto que si tenemos un problema en una dimensión en la que la solución no es lineal se puede pasar a un espacio de N dimensiones en el que sea posible una solución lineal gracias a una transformación de $\mathbf{x}$ a $\boldsymbol{\phi}$. Lo malo es que $\boldsymbol{\phi}$ tendrá N dimensiones que serán muy grandes y al hacer su producto escalar tendremos $N \times N$ dimensiones. Lo mas seguro es que sean infinitas, lo que nos supondría un problema. Por ello existe el *truco Kernel* o *truco de los núcleos* que asocia este producto escalar a una función que sólo depende de las entradas que teníamos inicialmente $\mathbf{x}$ y que tienen muchas menos dimensiones (figura 2.2) [13].

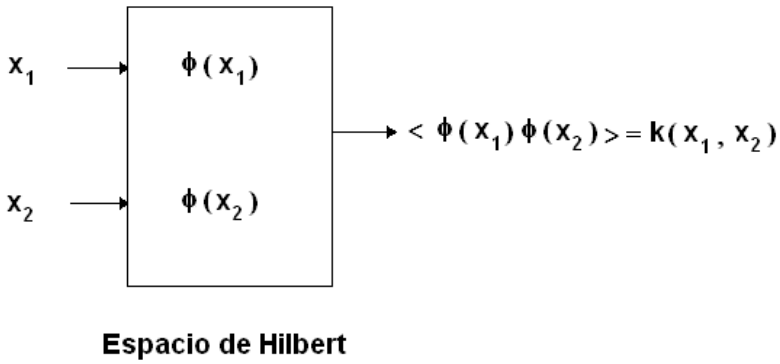

Figura 2.2: Transformación en el espacio de Hilbert

El *kernel* o *núcleo* es toda aquella función $k(\mathbf{x},\ \mathbf{y})$ que verifica el *teorema de Mercer*, el cual muestra que existe una función $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{H}$ y un producto escalar

$$k(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}(\mathbf{x}_i)\, \boldsymbol{\phi}^{\top}(\mathbf{x}_j) \tag{2.14}$$

si y sólo si $k(\cdot,\cdot)$ es un operador integral positivo en un espacio de Hilbert. Para que cualquier función $g(\mathbf{x})$ pueda ser un núcleo tiene que ser una función finita y semi-definida positiva, o lo que es lo mismo que cumpla:

$$\int g(\mathbf{x}) < \infty \tag{2.15}$$

y para la que se cumple que

$$\int k\left(\mathbf{x},\mathbf{y}\right) g\left(\mathbf{x}\right) g\left(\mathbf{y}\right) d\mathbf{x} d\mathbf{y} \;\geq\; 0 \tag{2.16}$$

Si la ecuación (1.2) no tiene una solución lineal, tal y como hemos contado anteriormente, hay que pasarla a un espacio de mayor dimensiones para encontrar una solución mas sencilla en ese nuevo espacio. Si pasamos los datos a un espacio de Hilbert obtenemos una nueva ecuación:

$$\boldsymbol{\phi}\left(\mathbf{Y}\right) = \boldsymbol{\phi}^{\top}\left(\mathbf{X}\right)\boldsymbol{\beta} + \mathbf{E} \tag{2.17}$$

donde $\boldsymbol{\phi}\left(\mathbf{Y}\right)$ es la transformación de $\mathbf{Y}$ en el espacio de Hilbert y $\boldsymbol{\phi}\left(\mathbf{X}\right)$ es la de $\mathbf{X}$. Al ser una ecuación análoga al caso lineal, podemos reescribir la solución lineal (1.5) en términos de núcleos de Mercer sustituyendo los productos escalares lineales implícitos en las matrices $\mathbf{X}\mathbf{X}^{\top}$ y $\mathbf{X}\mathbf{Y}$ por productos escalares en núcleos de Mercer, de la forma:

$$\boldsymbol{\beta}_{\mathrm{GKM}} = \left(\boldsymbol{\phi}\left(\mathbf{X}\right)\boldsymbol{\phi}^{\top}\left(\mathbf{X}\right)\right)^{-1}\boldsymbol{\phi}\left(\mathbf{X}\right)\boldsymbol{\phi}\left(\mathbf{Y}\right) \tag{2.18}$$

Siguiendo la ecuación (2.14) podemos escribir los productos en núcleos de Mercer de forma abreviada

$$\mathbf{K}_{\mathbf{XX}} = \boldsymbol{\phi}\left(\mathbf{X}\right)\boldsymbol{\phi}^{\top}\left(\mathbf{X}\right) \tag{2.19}$$

$$\mathbf{K}_{\mathbf{XY}} = \boldsymbol{\phi}\left(\mathbf{X}\right)\boldsymbol{\phi}\left(\mathbf{Y}\right) \tag{2.20}$$

Así la ecuación (2.18) la podemos abreviar de la siguiente forma

$$\boldsymbol{\beta}_{\mathrm{GKM}} = \mathbf{K}_{\mathbf{XX}}{}^{-1}\mathbf{K}_{\mathbf{XY}} \tag{2.21}$$

El uso del *truco de los núcleos* también se puede aplicar al modelo de regresión lineal estándar, sustituyendo las $\mathbf{X}$ por $\mathbf{\Phi}(\mathbf{X})$ en la (2.11). La distribución predictiva gaussiana quedaría descrita de la forma:

$$f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim \mathcal{N}\left(\frac{1}{\sigma_n^2}\boldsymbol{\phi}^\top(\mathbf{x}_*)\mathbf{A}^{-1}\mathbf{\Phi}\mathbf{y}, \boldsymbol{\phi}^\top(\mathbf{x}_*)\mathbf{A}^{-1}\boldsymbol{\phi}(\mathbf{x}_*)\right) \tag{2.22}$$

con $\mathbf{\Phi} = \mathbf{\Phi}(\mathbf{X})$ y $\mathbf{A} = \sigma_n^{-2}\mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Sigma}_p^{-1}$. Para hacer predicciones utilizando esta ecuación necesitamos invertir la matriz $\mathbf{A}$ de tamaño $N \times N$ lo que puede ser un inconveniente si N, la dimensión del espacio de características, es muy grande. Podemos reescribir la ecuación anterior haciendo uso de los núcleos, de la siguiente manera:

$$f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\phi}_*^\top\mathbf{\Sigma}_p\mathbf{\Phi}\left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1}\mathbf{y}, \boldsymbol{\phi}_*^\top\mathbf{\Sigma}_p\boldsymbol{\phi}_* - \boldsymbol{\phi}_*^\top\mathbf{\Sigma}_p\mathbf{\Phi}\left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1}\mathbf{\Phi}^\top\mathbf{\Sigma}_p\boldsymbol{\phi}_*\right) \tag{2.23}$$

donde $\boldsymbol{\phi}(\mathbf{x}_*) = \boldsymbol{\phi}_*$ y $\mathbf{K} = \mathbf{\Phi}^\top\mathbf{\Sigma}_p\mathbf{\Phi}$ es la matriz de kernel. Para la media, y usando las definiciones de $\mathbf{A}$ y $\mathbf{K}$ tenemos $\sigma_n^{-2}\mathbf{\Phi}\left(\mathbf{K} + \sigma_n^2\mathbf{I}\right) = \sigma_n^{-2}\mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_p\mathbf{\Phi} + \sigma_n^2\mathbf{I}\right) = \mathbf{A}\mathbf{\Sigma}_p\mathbf{\Phi}$. Si multiplicamos ahora cada lado de la igualdad anterior por $\mathbf{A}^{-1}$ a la izquierda y por $\left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1}$ a la derecha tendríamos $\sigma_n^{-2}\mathbf{A}^{-1}\mathbf{\Phi} = \mathbf{\Sigma}_p\mathbf{\Phi}\left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1}$. Para la varianza utilizamos el *lema de inversión de matrices*[1] estableciendo las equivalencias siguientes: $\mathbf{Z}^{-1} = \mathbf{\Sigma}_p^2$, $\mathbf{W}^{-1} = \sigma_n^2\mathbf{I}$ y $\mathbf{V} = \mathbf{U} = \mathbf{\Phi}$. De este modo, en la ecuación (2.23) estaríamos invirtiendo matrices de tamaño $n \times n$ lo que es mucho más manejable que tratar con matrices de tamaño $N \times N$ cuando $n < N$.

2.2.3. Modelo en el espacio de funciones

Un modelo alternativo para obtener los mismos resultados que hemos visto hasta ahora es el de hacer inferencias directamente en el espacio de funciones. Usamos los procesos gaussianos para describir distribuciones sobre funciones. Formalmente:

[1] Este lema, conocido también como la fórmula de Woodbury, Sherman & Morrison [12] establece:

$$\left(\mathbf{Z} + \mathbf{U}\mathbf{W}\mathbf{V}^\top\right)^{-1} = \mathbf{Z}^{-1} - \mathbf{Z}^{-1}\mathbf{U}\left(\mathbf{W}^{-1} + \mathbf{V}^\top\mathbf{Z}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}^\top\mathbf{Z}^{-1}$$

donde la matriz $\mathbf{Z}$ es de tamaño $n \times n$, $\mathbf{W}$ es de tamaño $m \times m$ y $\mathbf{U}$ y $\mathbf{V}$ son ambas de tamaño $m \times n$.

Definición 2.1 Un proceso gaussiano *es un conjunto finito de variables aleatorias que se distribuyen siguiendo una distribución Normal o Gaussiana*

Un proceso gaussiano queda completamente descrito por su función de media y su función de covarianza. Dado un proceso $f(\mathbf{x})$, definimos la media $m(\mathbf{x})$ y la covarianza $k(\mathbf{x}, \mathbf{x}')$ como

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \end{aligned} \tag{2.24}$$

y escribimos el proceso gaussiano como

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k\left(\mathbf{x}, \mathbf{x}'\right)\right) \tag{2.25}$$

Podemos extender la definición de proceso gaussiano al modelo lineal bayesiano $f(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\beta}$ con una probabilidad a priori de $\boldsymbol{\beta}$ tal que $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$. Entonces, la media y la covarianza del proceso quedaría:

$$\begin{aligned} \mathbb{E}[f(\mathbf{x})] &= \boldsymbol{\phi}^\top(\mathbf{x})\,\mathbb{E}[\boldsymbol{\beta}] = 0, \\ \mathbb{E}\left[f(\mathbf{x}) f\left(\mathbf{x}'\right)\right] &= \boldsymbol{\phi}^\top(\mathbf{x})\,\mathbb{E}\left[\boldsymbol{\beta}\boldsymbol{\beta}^\top\right]\boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\phi}^\top(\mathbf{x})\,\Sigma_p\boldsymbol{\phi}(\mathbf{x}') \end{aligned} \tag{2.26}$$

donde $f(\mathbf{x})$ y $f(\mathbf{x}')$ son gaussianas de media cero y covarianza dada por $\boldsymbol{\phi}^\top(\mathbf{x})\,\Sigma_p\boldsymbol{\phi}(\mathbf{x}')$. La covarianza es un núcleo de Mercer.

$$\mathrm{cov}\left(\mathrm{f}\left(\mathbf{x}_\mathrm{p}\right), \mathrm{f}\left(\mathbf{x}_\mathrm{q}\right)\right) = \mathrm{k}\left(\mathbf{x}_\mathrm{p}, \mathbf{x}_\mathrm{q}\right) \tag{2.27}$$

El objetivo es hacer predicciones de un conjunto de posibles valores de entrada $\mathbf{X}_*$ a partir de un conjunto de datos de entrenamiento, $\mathbf{X}$. Si suponemos un modelo con ruido que sigue una distribución gaussiana independiente e idénticamente distribuida, $y = f(\mathbf{x}) + \epsilon$, y varianza σ_n^2, entonces tendríamos una covarianza a priori:

$$\mathrm{cov}\left(\mathrm{y}_\mathrm{p}, \mathrm{y}_\mathrm{q}\right) = \mathrm{k}\left(\mathbf{x}_\mathrm{p}, \mathbf{x}_\mathrm{q}\right) + \sigma_\mathrm{n}^2\delta_\mathrm{pq}$$

o escrito de otra manera:

$$\operatorname{cov}(\mathbf{y}) = \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_{\mathrm{n}}^2 \mathbf{I} \tag{2.28}$$

La diferencia con el caso sin ruido de la ecuación (2.27) es la adición de una matriz diagonal, representada por la delta de Kronecker, δ_{pq}, de valor 1 si y sólo si $p = q$, y 0 en caso contrario. Podemos escribir la distribución a priori conjunta de las observaciones $\mathbf{y}$ y la función $\mathbf{f}_*$ como:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I} & \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \\ \mathbf{K}(\mathbf{X}_*, \mathbf{X}) & \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right) \tag{2.29}$$

Las ecuaciones predictivas para el modelo de regresión basado en procesos gaussianos resultan ser:

$$\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}\left(\overline{\mathbf{f}}_*, \operatorname{cov}(\mathbf{f}_*)\right) \tag{2.30}$$

donde

$$\overline{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_*] = \mathbf{K}(\mathbf{X}_*, \mathbf{X}) \left[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}\right]^{-1} \mathbf{y} \tag{2.31}$$

$$\operatorname{cov}(\mathbf{f}_*) = \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*, \mathbf{X}) \left[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_{\mathrm{n}}^2 \mathbf{I}\right]^{-1} \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \tag{2.32}$$

siendo la ecuación (2.31) la media y la (2.32) la covarianza de dicho procesos gaussiano. Se puede establecer la equivalecia entre estas ecuaciones y el modelo no lineal visto en la ecuación (2.23).

Por otro lado, podemos simplicar las ecuaciones (2.31) y (2.32) a:

$$\overline{f}_* = \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y}, \tag{2.33}$$

$$\mathbb{V}[f_*] = k(\mathbf{x}_*, \mathbf{x}) - \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{k}_* \tag{2.34}$$

definiendo las siguientes igualdades

$$\mathbf{K} = \mathbf{K}(\mathbf{X}, \mathbf{X})$$

,

$$\mathbf{K}_* = \mathbf{K}(\mathbf{X}, \mathbf{X}_*)$$

y $\mathbf{k}(\mathbf{x}_*) = \mathbf{k}_*$ como el vector de covarianzas entre los n datos de entrenamiento y el de test $\mathbf{x}_*$.

La media de la ecuación (2.33) es una combinación lineal de las observaciones $\mathbf{y}$, que también puede interpretarse como una combinación de n funciones de kernel, una por cada dato de entrenamiento. La distribución predictiva puede escribirse entonces como:

$$\overline{f}(\mathbf{x}_*) = \sum_{n=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}_*) \tag{2.35}$$

donde

$$\boldsymbol{\alpha} = \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y}$$

La ecuación (2.35) cumple el *teorema del representador* [11].

Podemos definir la verosimilitud marginal sobre los valores de la función $\mathbf{f}$ como:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X})\, p(\mathbf{f}|\mathbf{X})\, d\mathbf{f} \tag{2.36}$$

La probabilidad a priori es gaussiana, $\mathbf{f}|\mathbf{X} \sim \mathcal{N}(0, \mathbf{K})$, o:

$$\log p(\mathbf{f}|\mathbf{X}) = -\frac{1}{2}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \tag{2.37}$$

y la verosimilitud se define como una gaussiana tal que $\mathbf{y}|\mathbf{f} \sim \mathcal{N}\left(\mathbf{f}, \sigma_n^2 \mathbf{I}\right)$. Entonces, la log verosimilitud marginal queda descrita por la ecuación:

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y} - \frac{1}{2}\log|\mathbf{K} + \sigma_n^2 \mathbf{I}| - \frac{n}{2}\log 2\pi \tag{2.38}$$

Capítulo 3

SELECCIÓN DE MODELO Y AJUSTE DE HIPERPARÁMETROS

3.1. Introducción a la selección de modelo

La elección de una función de covarianza y la definición de sus *hiperparámetros* en ocasiones no es una tarea arbitraria y para ello tenemos que recurrir a lo que se conoce como *selección de modelo*. En este contexto tanto a la selección de unas como de otros se les conoce como *entrenamiento* de un proceso gaussiano.

Por ejemplo, podemos partir de una función de covarianza exponencial cuadrática y expresarla en función de sus hiperparámetros como sigue:

$$k\left(\mathbf{x}_p, \mathbf{x}_q\right) = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\mathbf{x}_p - \mathbf{x}_q\right)^\top \mathbf{M}\left(\mathbf{x}_p - \mathbf{x}_q\right)\right) + \sigma_n^2 \delta pq \tag{3.1}$$

donde $\boldsymbol{\theta} = \left(\{\mathbf{M}\}, \sigma_f^2, \sigma_n^2\right)^\top$ es un vector que contiene todos los hiperparámetros, y de cuyos valores dependen las características de las funciones de covarianza que tratamos de encontrar.

Definimos $\{\mathbf{M}\}$ como los parámetros de la matriz simétrica $\mathbf{M}$, que a su vez es la inversa de la matriz de covarianza, y que podemos expresar también como:

$$\mathbf{M}_1 = \boldsymbol{\ell}^2\mathbf{I}, \qquad \mathbf{M}_2 = \operatorname{diag}\left(\boldsymbol{\ell}\right)^{-2}, \qquad \mathbf{M}_3 = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \operatorname{diag}\left(\boldsymbol{\ell}\right)^{-2}, \tag{3.2}$$

donde $\mathbf{M}_1$ es una gaussiana definida esférica, $\mathbf{M}_2$ es la diagonal del vector $\boldsymbol{\ell}$ que contiene los valores de los hiperparámetros $l_1, \ldots, l_D$. $\mathbf{M}_3$ es una matriz definida positiva pues $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ lo es también, con $\boldsymbol{\Lambda}$ una matriz de dimensión $D \times k$ y $k < D$

3.2. Selección de modelo bayesiano

El problema de selección de modelo se basa en maximizar la verosimilitud en un modelo con hiperparámetros. Los métodos de parámetros lineales no son aplicables al caso de funciones de covarianza con parámetros libres, por lo que es común recurrir a modelos jerárquicos.

En el primer nivel de estos modelos jerárquicos tenemos los parámetros del modelo lineal, $\boldsymbol{\beta}$. En el segundo nivel se sitúan los parámetros no lineales o hiperparámetros, $\boldsymbol{\theta}$, y en el último nivel tenemos un conjunto discreto de *modelos de estructuras* o *familias de funciones* (hipótesis), que denominamos $\mathcal{H}_i$. Con estas premisas, hacemos inferencias desde el punto de vista bayesiano.

Empezamos buscando la *probabilidad a posteriori* de los parámetros $\boldsymbol{\beta}$ del primer nivel:

$$p\left(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \mathcal{H}_i\right) = \frac{p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \mathcal{H}_i\right) p\left(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathcal{H}_i\right)}{p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathcal{H}_i\right)} \tag{3.3}$$

donde $p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \mathcal{H}_i\right)$ es la *verosimilitud* y $p\left(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathcal{H}_i\right)$ la *probabilidad a priori* de los parámetros antes de ver los datos. El denominador de la ecuación, $p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathcal{H}_i\right)$, es la *verosimiltud marginal* de los datos, que es independiente de los parámetros y está descrita por:

$$p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathcal{H}_i\right) = \int p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \mathcal{H}_i\right) p\left(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathcal{H}_i\right) d\boldsymbol{\beta} \tag{3.4}$$

En el siguiente nivel calculamos la probabilidad a posteriori sobre los hiperparámetros de forma análoga al primer nivel, donde ahora la verosimilitud es la verosimilitud marginal del nivel anterior (ecuación (3.4)). La expresión siguiente es equivalente a la (3.3):

$$p\left(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathcal{H}_i\right) = \frac{p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathcal{H}_i\right) p\left(\boldsymbol{\theta}|\mathcal{H}_i\right)}{p\left(\mathbf{y}|\mathbf{X}, \mathcal{H}_i\right)} \tag{3.5}$$

A $p\left(\boldsymbol{\theta}|\mathcal{H}_i\right)$ se la conoce como *hiper-prior* (probabilidad a priori de los hiperparámetros) y la constante independiente de los parámetros (verosimilitud marginal del segundo nivel) queda

descrita por:

$$p\left(\mathbf{y}|\mathbf{X},\mathcal{H}_i\right)=\int p\left(\mathbf{y}|\mathbf{X},\boldsymbol{\theta},\mathcal{H}_i\right)p\left(\boldsymbol{\theta}|\mathcal{H}_i\right)d\boldsymbol{\theta} \tag{3.6}$$

que es equivalente a la ecuación (3.4).

En el último nivel, definimos la probabilidad a posteriori sobre los posibles modelos (número finito de posibles estructuras), $\mathcal{H}_i$, como:

$$p\left(\mathcal{H}_i|\mathbf{y},\mathbf{X}\right)=\frac{p\left(\mathbf{y}|\mathbf{X},\mathcal{H}_i\right)p\left(\mathcal{H}_i\right)}{p\left(\mathbf{y}|\mathbf{X}\right)} \tag{3.7}$$

Queda establecida la equivalencia con las probabilidades a posteriori de las ecuaciones (3.3) y (3.5) de primer y segundo nivel respectivamente. Además, $p\left(\mathbf{y}|\mathbf{X}\right)=\sum_i p\left(\mathbf{y}|\mathbf{X},\mathcal{H}_i\right)p\left(\mathcal{H}_i\right)$.

La implementación del modelo bayesiano implica la evaluación de integrales que pueden no ser muy tratables, por lo que en la práctica se recurre a aproximaciones analíticas.

3.3. Validación cruzada

Los métodos de validación cruzada (CV) se aplican al modelo de selección. La idea consiste en dividir el conjunto de datos de entrenamiento (TR) en dos subconjuntos, uno para entrenamiento (TR') y otro para validación (VAL). Para ello:

1. Se entrena a partir del subconjunto de datos de entrenamiento $\mathbf{X}_{\mathrm{TR}'}\left[n\right]$ y del conjunto de parámetros $\boldsymbol{\theta}_i$
2. Se valida mediante el error del conjunto de test (o validación) para lo que se calcula el error medio o esperanza del error: $\mathrm{E}\left[\ell\left(\mathrm{e}\left[\mathrm{n}\right],\boldsymbol{\theta}_{\mathrm{i}}\right)\right]=\varepsilon_{\mathrm{i}}$
3. Se repite (1) hasta M veces, que es el número de parámetros $\boldsymbol{\theta}_i$
4. Se construye la gráfica de ε_i contra $\boldsymbol{\theta}_i$ y se escoge el parámetro que minimiza el error

Este proceso se puede repetir para un único subconjunto de validación o para varios subconjuntos, de modo que se va variando este subconjunto dentro del total de datos de entrenamiento (TR) y al final se hace la media para obtener el error mínimo. La validación cruzada conocida

como leave-one-out (LOO-CV) considera que si hay n muestras hay n conjuntos de validación para calcular esos parámetros que minimizan el error.

3.4. Regresión basada en procesos gaussianos

3.4.1. Verosimilitud marginal

Los principios de inferencia bayesiana se pueden aplicar a los procesos gaussianos dando como resultados métodos flexibles y análiticamente tratables para encontrar los hiperparámetros de las funciones de covarianza.

El primer paso es aplicar las ecuaciones (3.3) y (3.4) al primer nivel de inferencia. Se establecen equivalencias con las distribuciones predictivas obtenidas mediante núcleos de Mercer (ecuaciones (2.22) y (2.23)), y con la obtenida en el espacio de funciones con procesos gaussianos (ecuación (2.30)). La verosimilitud marginal sobre los hiperparámetros de la (3.4) se puede expresar como (2.38) simplemente con reescribir la primera en términos de logaritmo (log verosimilitud marginal), resultando:

$$\log p\left(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}\right) = -\frac{1}{2}\mathbf{y}^{\top}\mathbf{K}_y^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}_y^{-1}| - \frac{n}{2}\log 2\pi \tag{3.8}$$

donde $\mathbf{K}_y = \mathbf{K}_f + \sigma_n^2\mathbf{I}$ es la matriz de covarianza del vector de salidas $\mathbf{y}$ mientras que $\mathbf{K}_f$ es la matriz de covarianza de la función latente f.

Para calcular los hiperparámetros óptimos maximizamos dicha log verosimilitud marginal derivando la ecuación (3.8) respecto a los hiperparámetros:

$$\begin{aligned}\frac{\partial}{\partial\theta_j}\log p\left(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}\right) &= \frac{1}{2}\mathbf{y}^{\top}\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\theta_j}\mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}\mathrm{tr}\left(\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\theta_{\mathrm{j}}}\right)\\ &= \frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^{\top} - \mathbf{K}^{-1}\right)\frac{\partial\mathbf{K}}{\partial\theta_{\mathrm{j}}}\right)\end{aligned} \tag{3.9}$$

donde $\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}$ es la solución dual y la función "tr"(o traza) es la suma de los elementos de la diagonal.

El coste computacional de la log verosimilitud marginal de la ecuación (3.8) viene impuesto por la necesidad de invertir la matriz de covarianza $\mathbf{K}$, que al ser simétrica y definida positiva

dicho coste es del orden de $\mathcal{O}\left(n^3\right)$. Una vez conocida $\mathbf{K}_{-1}$, el coste computacional de la ecuación (3.9) se reduce al orden de $\mathcal{O}\left(n^2\right)$.

Sin embargo, no hay garantías de que la log verosimilitud no tenga más de un máximo local. Es decir, en la práctica, igualar la expresión anterior (3.9) a cero no garantiza encontrar el hiperparámetro óptimo que minimice el error, por lo que se suele aplicar la validación cruzada.

3.4.2. Validación cruzada

Como ya vimos, la validación cruzada conocida como leave-one-out (LOO-CV) considera que si hay n muestras hay n conjuntos de validación para calcular los parámetros que minimizan el error. Normalmente el coste computacional de este tipo de validación suele ser elevado, pero para el caso regresión basado en procesos gaussianos se reduce utilizando la probabilidad log predictiva negativa. Cuando se deja fuera del entrenamiento la muestra i, dicha probabilidad puede escribirse como:

$$\log p\left(y_i|\mathbf{X}, \mathbf{y}_{-i}, \boldsymbol{\theta}\right) = -\frac{1}{2}\log\sigma_i^2 - \frac{\left(y_i - \mu_i\right)^2}{2\sigma_i^2} - \frac{1}{2}\log 2\pi \tag{3.10}$$

donde $\mathbf{y}_{-i}$ significa todos los posibles valores menos i, y μ_i y σ_i^2 son la media y la covarianza de la ecuaciones (2.31) y (2.32) respectivamente, donde el conjunto de entrenamiento es $(\mathbf{X}_{-i}, \mathbf{y}_{-i})$.

La probabilidad log predictiva del LOO es la suma de las probabilidades de la ecuación (3.10) para todos los n conjuntos de validación, es decir:

$$L_{\text{LOO}}\left(\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}\right) = \sum_{i=1}^{n} \log p\left(y_i|\mathbf{X}, \mathbf{y}_{-i}, \boldsymbol{\theta}\right) \tag{3.11}$$

A L_{LOO} también se la puede denominar *pseudo-verosimilitud* [6]. En cada conjunto de validación (hasta n), la inferencia del modelo basado en procesos gaussianos consiste en invertir la matriz de covarianza para calcular la media y la varianza de las ecuaciones (2.31) y (2.32). Las expresiones de la media y varianza predictivas del LOOCV quedan descritas por:

$$\mu_i = y_i - \left[\mathbf{K}^{-1}\mathbf{y}\right]_i / \left[\mathbf{K}^{-1}\right]_{ii}, \qquad \text{y} \qquad \sigma_i^2 = 1/\left[\mathbf{K}^{-1}\right]_{ii}, \tag{3.12}$$

El coste computacional de estas media y varianza es del orden de $\mathcal{O}\left(n^3\right)$ mientras que el coste una vez conocida la inversa de la matriz de covarianza se reduce al orden de $\mathcal{O}\left(n^2\right)$ para todo el proceso de LOOCV.

Si sustituimos las expresiones de la ecuación (3.12) en (3.10) y (3.11) obtenemos un estimador que podemos optimizar respecto a los hiperparámetros. También podemos trabajar con el gradiente, simplemente calculando las derivadas parciales de la media y varianza de (3.12) con respecto a los hiperparámetros, esto es:

$$\frac{\partial \mu_i}{\partial \theta_j} = \frac{[Z_j \boldsymbol{\alpha}]_i}{[\mathbf{K}^{-1}]_{ii}} - \frac{\boldsymbol{\alpha}_i \left[Z_j \mathbf{K}^{-1}\right]_{ii}}{[\mathbf{K}^{-1}]^2_{ii}}, \qquad \text{y} \qquad \frac{\partial \sigma_i^2}{\partial \theta_j} = \frac{\left[Z_j \mathbf{K}^{-1}\right]_{ii}}{[\mathbf{K}^{-1}]^2_{ii}}, \tag{3.13}$$

donde $\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}$ y $Z_j = \mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \theta_j}$.

Para obtener la derivada parcial de la ecuación (3.11) se aplica la regla de la cadena junto con la ecuación (3.13), dando como resultado:

$$\begin{aligned} \frac{\partial L_{\text{LOO}}}{\partial \theta_j} &= \sum_{i=1}^{n} \frac{\partial \log p\left(y_i | \mathbf{X}, \mathbf{y}_{-i}, \boldsymbol{\theta}\right)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \theta_j} + \frac{\partial \log p\left(y_i | \mathbf{X}, \mathbf{y}_{-i}, \boldsymbol{\theta}\right)}{\partial \sigma_i^2} \frac{\partial \sigma_i^2}{\partial \theta_j} \\ &= \sum_{i=1}^{n} \left(\alpha_i \left[Z_j \boldsymbol{\alpha}\right]_i - \frac{1}{2}\left(1 + \frac{\alpha_i^2}{[\mathbf{K}^{-1}]_{ii}}\right)\left[Z_j \mathbf{K}^{-1}\right]_{ii} \right) / \left[\mathbf{K}^{-1}\right]_{ii} \end{aligned} \tag{3.14}$$

El coste computacional es del orden de $\mathcal{O}\left(n^3\right)$ determinado por el cálculo de la inversa de la matriz de covarianza $\mathbf{K}$ y del orden de $\mathcal{O}\left(n^3\right)$ por cada hiperparámetro de la derivada de la ecuación (3.14).

Así, podemos concluir que la carga computacional del método de validación cruzada leave one out (LOOCV) es mayor que la del modelo de selección desde el punto de vista bayesiano (ecuación (3.9)). Sin embargo, también hay argumentos que demuestran que los métodos de validación cruzada son más robustos que cualquier otro modelo de selección [15]

Capítulo 4

EXPERIMENTOS

4.1. Sujetos y paradigma

Diez sujetos sanos fueron estudiados en un escáner Siemens Sonata de 1.5T y otros diez en un escáner Brucker MedSpec de 4.0 T. El consentimiento basado en las directrices institucionales fue obtenido antes de la participación en el estudio. Los estímulos fueron presentados a través de gafas LCD compatibles con MR y auriculares (Resonancia Technology Inc., Northridge, CA). El paradigma consiste en cuatro tareas entrelazadas: visuales (simulación de damero de 8 Hz), motoras (golpeo con el dedo indice derecho a 2 Hz), auditivas (discriminación sílaba), y cognitivas (cálculo mental). Estas tareas se organizan en un diseño de bloques al azar (8 s por bloque), con un punto de mira sirviendo como base de referencia para un total de 132 s por escaneo (figura 4.1).

Figura 4.1: Representación visual de los estímulos usados en intervalos en el paradigma

La duración total de cada estado es aproximadamente 27 s. El tarea visual consiste en la inversión de un damero blanco y negro con una frecuencia de 8 Hz. La tarea motora consiste en un dedo golpeando al ritmo de un tono a una frecuencia de 1 kHz. Se les pidió a los sujetos golpear con una extensión máxima del dedo en un botón de respuesta (Cedrus corp., San Pedro,

CA). Durante la tarea auditiva, los sujetos escucharon sílabas grabadas (por ejemplo: .ᴬh", "Ba", "Ha", "Ka", Ra") y presionaron un botón cuando escuchaban la sílaba "Ta"(25 % de la sílabas). La tarea cognitiva consiste en cálculos mentales. Se les pidió que sumar tres números presentados auditívamente y dividir la suma por tres, respondiendo mediante la pulsación de un botón cuando el suma es divisible por tres, sin resto (50 % de los ensayos). Los sujetos fueron instruidos para asistir a cada tarea con un esfuerzo constante a través de las exploraciones e intensidades de campo.

4.2. Adquisición de datos

Los datos de la resonancia magnética fueron adquiridos utilizando un sólo disparo eco-planar con TR: 2 s, TE: 50 ms, ángulo ip: 90 -, tamaño de la matriz: 64x64 o 32x32 píxeles, FOV: 192 mm. Los datos con las matrices de 32x32 fueron adquiridos con diferentes anchos de banda, ya sea con 1200 Hz / píxel (BPN) o con 2400 Hz / píxel (HBW), que cambia el grado de distorsión geométrica y la relación señal-ruido. Los cortes fueron de 6 mm de espesor, con un 25 % de diferencia, 66 volúmenes fueron recogidos para un tiempo total de medición de 132 s. El conjunto de datos disponible se compone de 184 t-maps tomados de 18 sujetos diferentes.

4.3. Análisis de datos

El hardware del ordenador utilizado en todos los experimentos consistió en un equipo de trabajo Intel Radeon de 3,2 GHz con memoria caché de 512 que se ejecuta en Linux, y Matlab 6,5 (The MathWorks, Inc) con código ANSI C para los cálculos combinados intensivos (aprendizaje de los parámetros clasificadores multiclase). Usamos el $SPM2$ (Kiebel y Friston, 2004a, b) para generar los t-maps que representan los cambios en la activación cerebral. Los pasos de preprocesamiento incluyen la corrección de movimiento, la corrección de cortes temporales, la normalización espacial y el suavizado espacial. El análisis estadístico usando una matriz de diseño con cuatro condiciones (motor, visual, auditiva, cognitiva) se realizó con umbral de los 1000 vóxeles más significativos y filtro 132 s paso alto.

El análisis mediante procesos gaussianos establece un modelo a priori de tipo gaussiano blanco para el ruido. La varianza de ruido se estima mediante un ajuste por maximización de la verosimilitud de la salida del estimador con respecto a este parámetro. La estimación se lleva

a cabo individualmente para cada uno de los voxels, y posteriormente se filtra espacialmente mediante un kernel gaussiano de varianza 2. Utilizando esta estimación del ruido, se lleva a cabo una estimación de los valores de β del mapa. Los resultados presentados para el método basado en GP consisten en esos valores

4.4. Experimentos Monosujeto

4.4.1. Estímulo visual

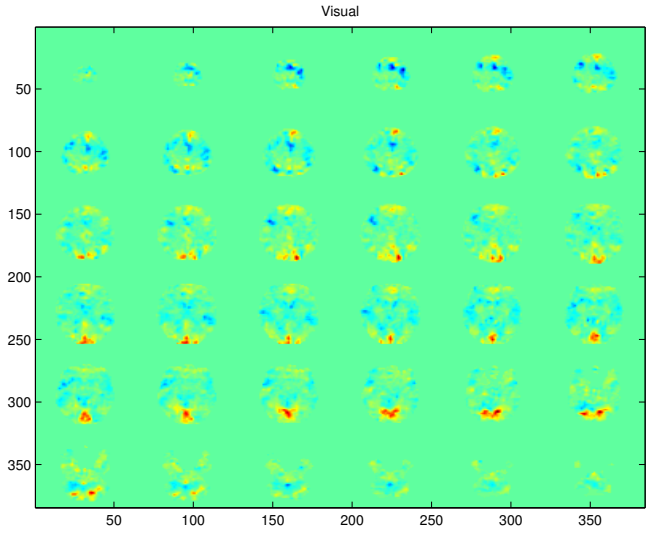

Figura 4.2: Estímulo visual sujeto 1 por GP.

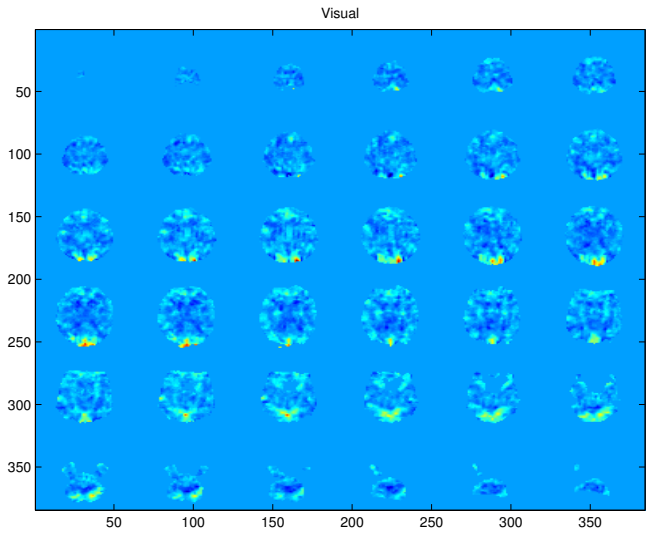

Figura 4.3: Estímulo visual sujeto 1 por SPM.

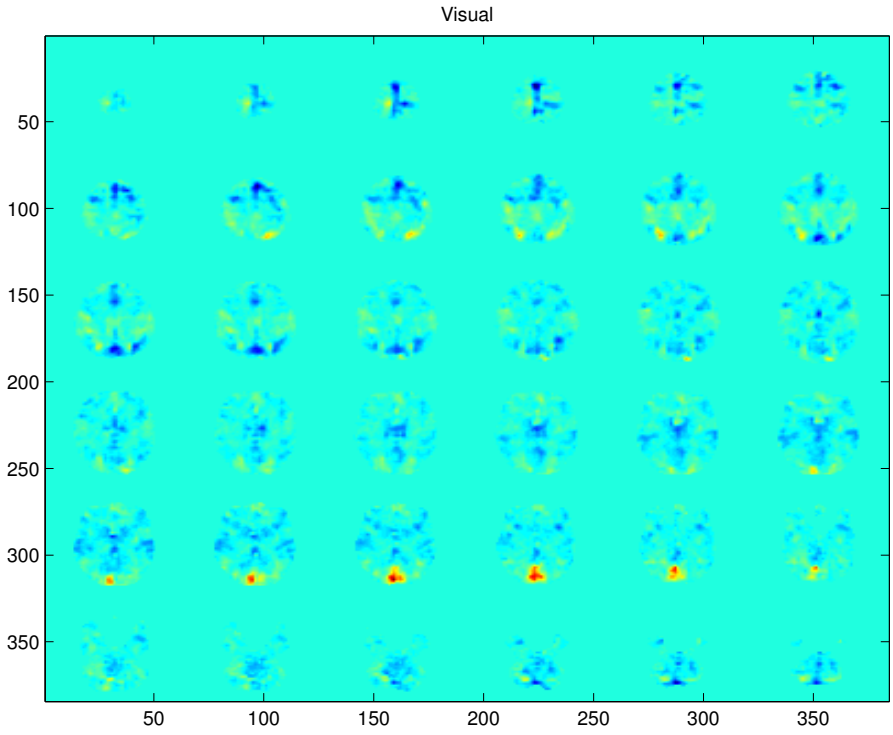

Figura 4.4: Estímulo visual sujeto 2 por GP.

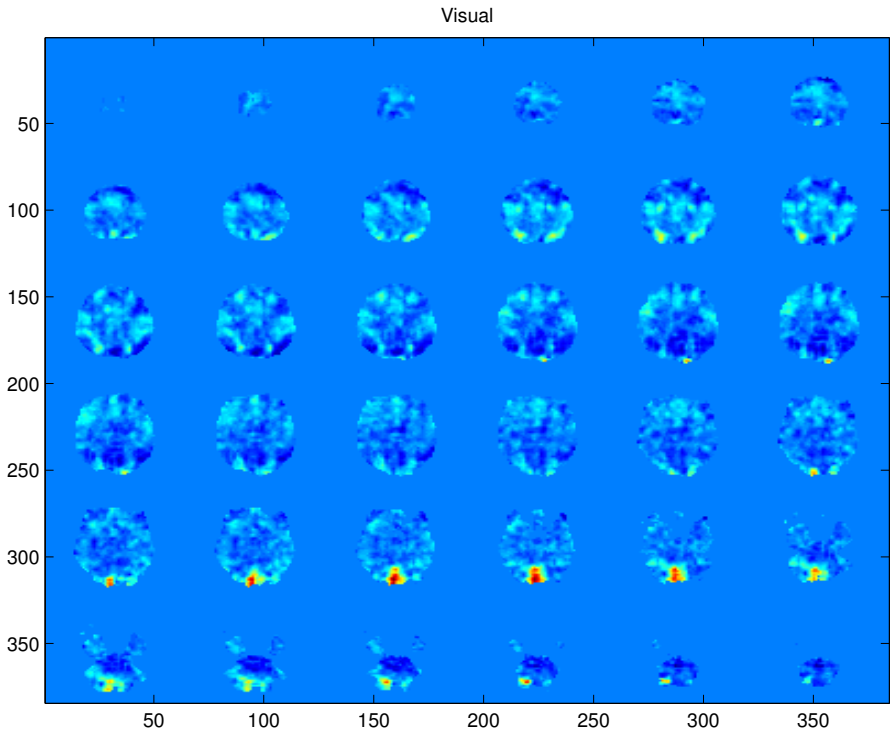

Figura 4.5: Estímulo visual sujeto 2 por SPM.

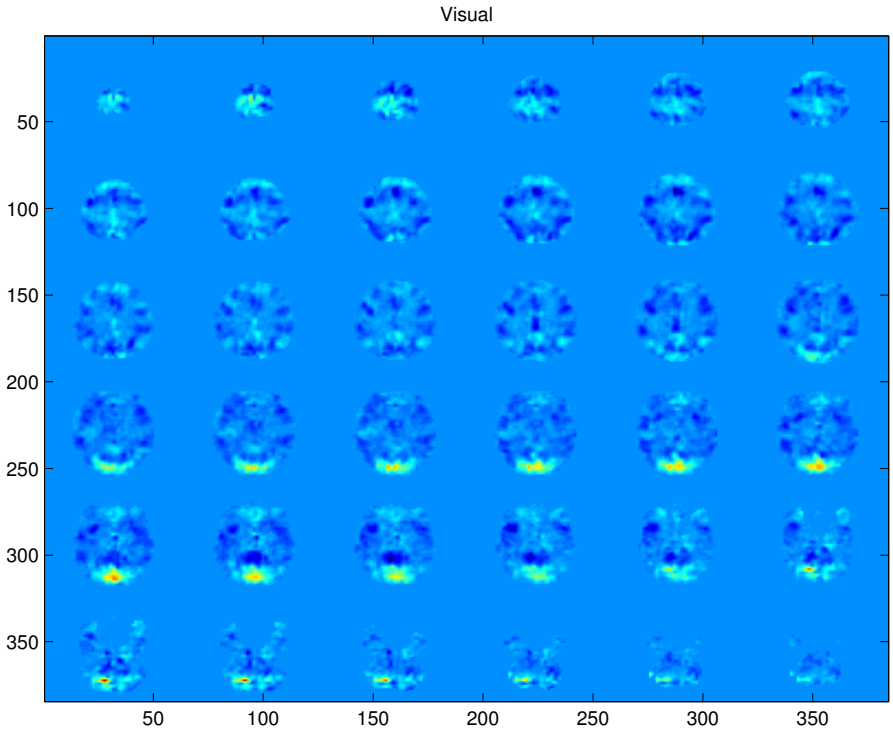

Figura 4.6: Estímulo visual sujeto 3 por GP.

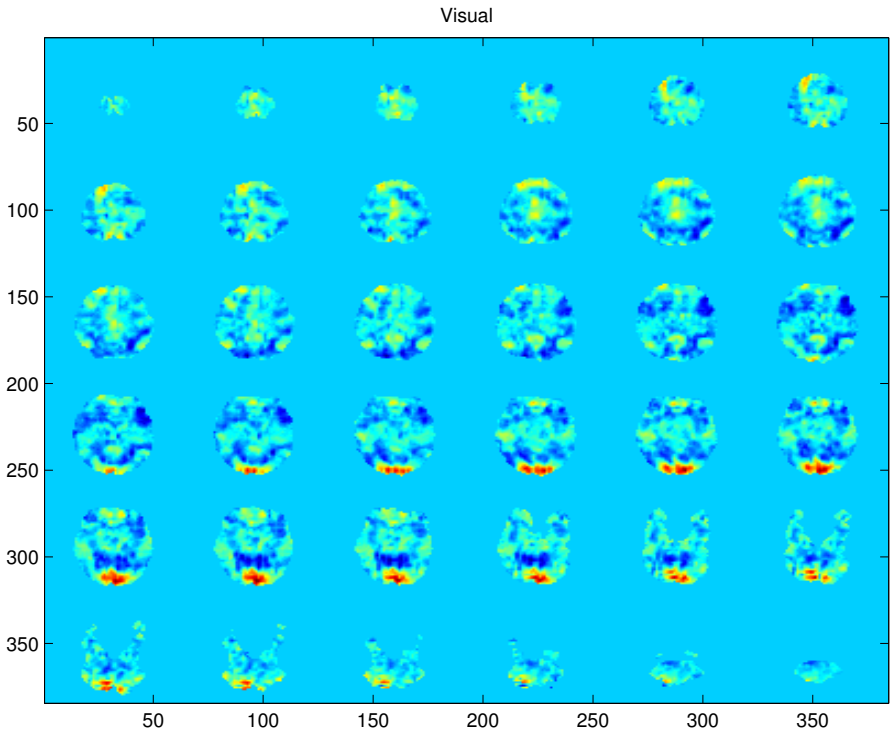

Figura 4.7: Estímulo visual sujeto 3 por SPM.

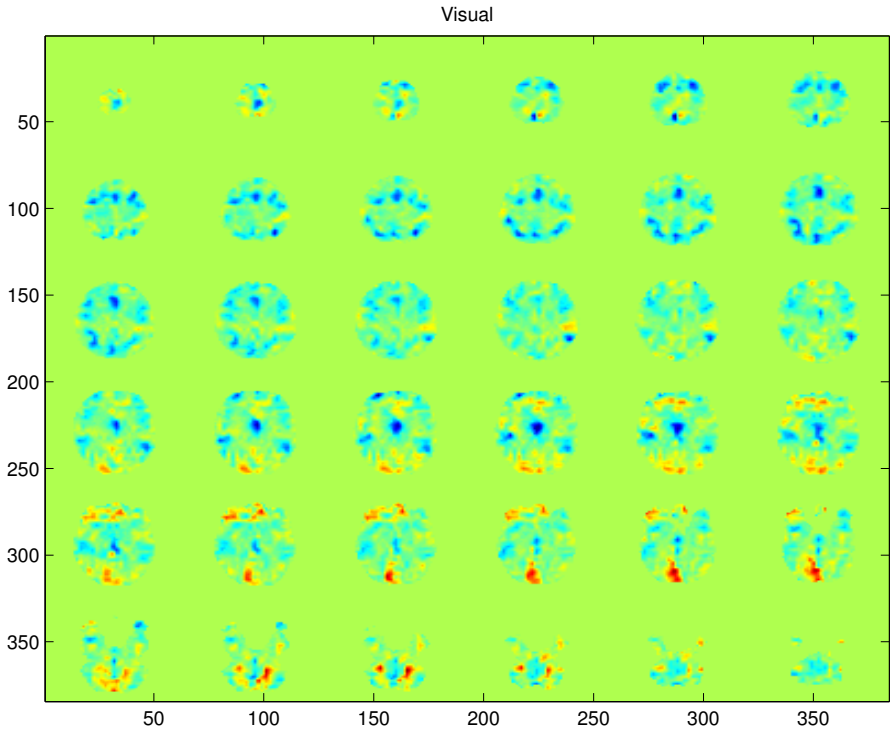

Figura 4.8: Estímulo visual sujeto 4 por GP.

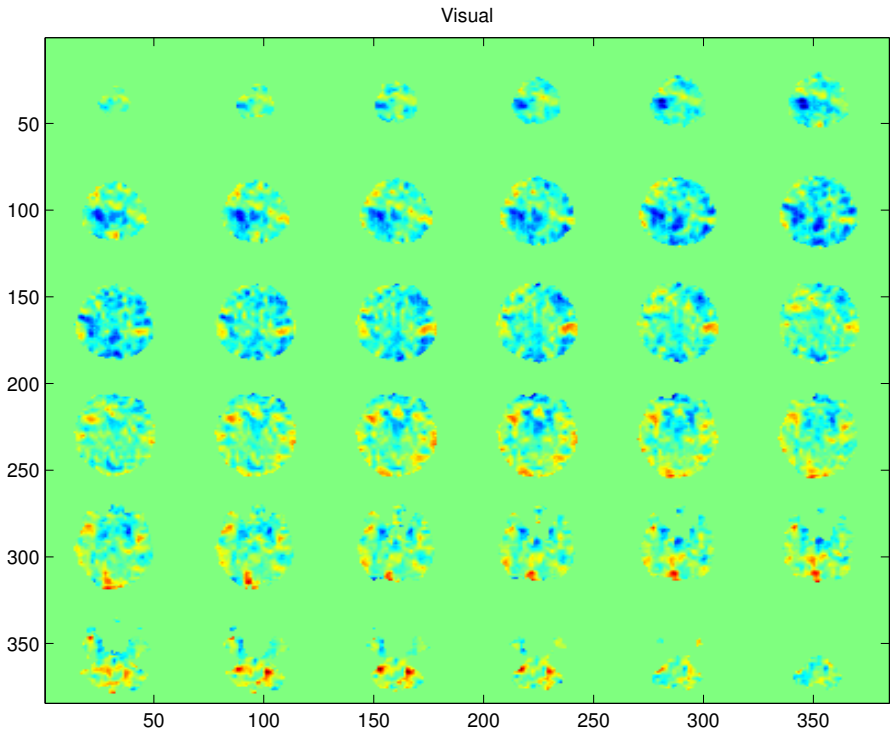

Figura 4.9: Estímulo visual sujeto 4 por SPM.

4.4.2. Estímulo motor

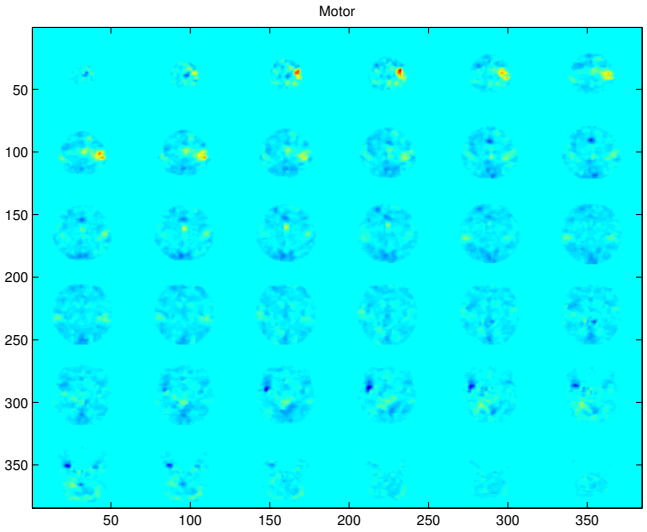

Figura 4.10: Estímulo motor sujeto 1 por GP.

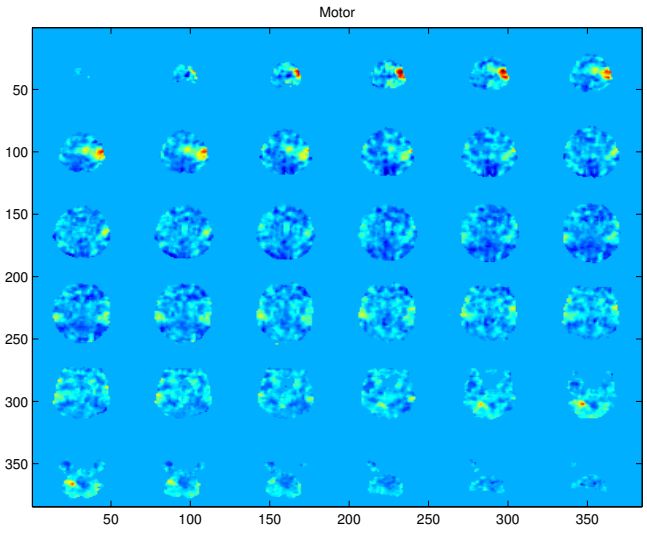

Figura 4.11: Estímulo motor sujeto 1 por SPM.

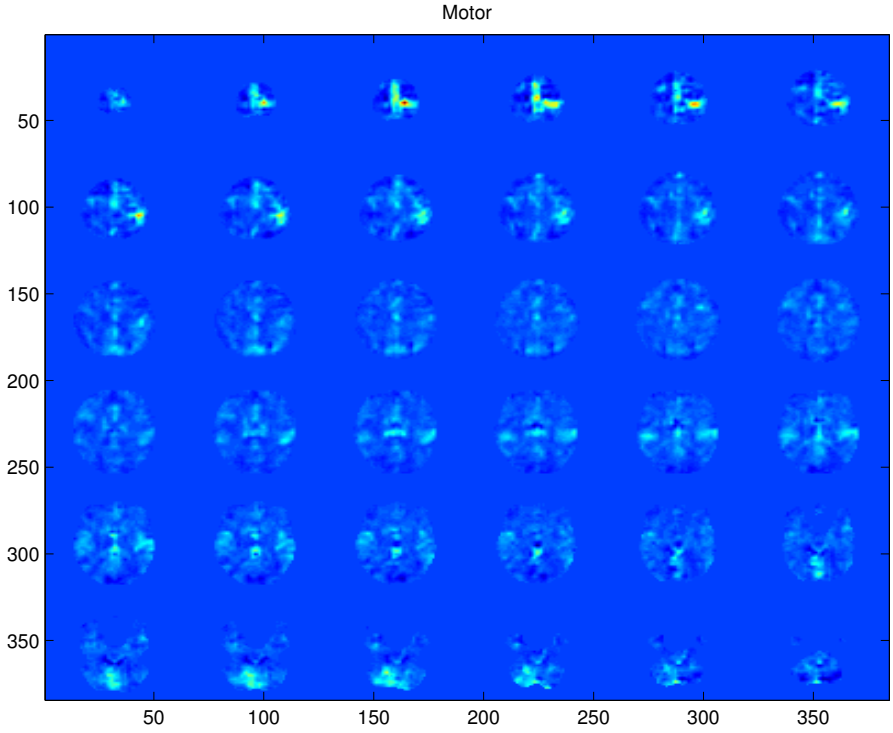

Figura 4.12: Estímulo motor sujeto 2 por GP.

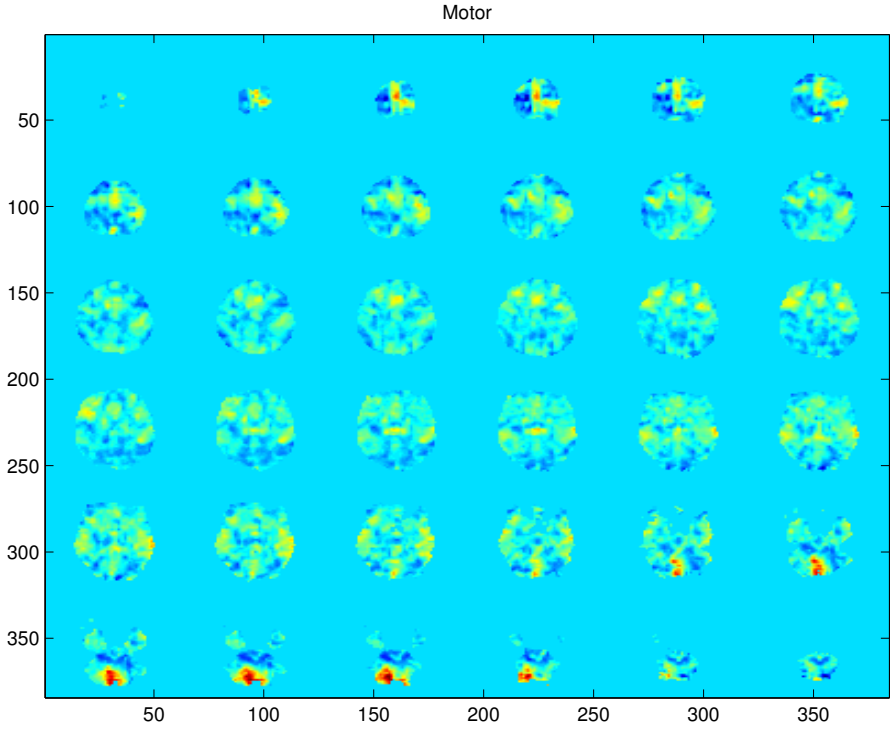

Figura 4.13: Estímulo motor sujeto 2 por SPM.

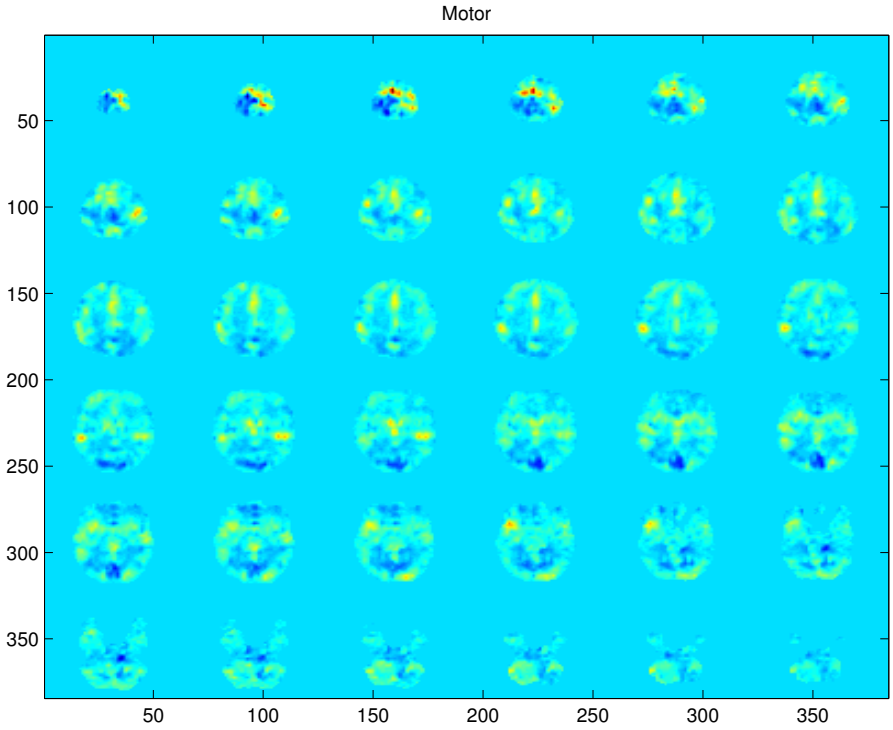

Figura 4.14: Estímulo motor sujeto 3 por GP.

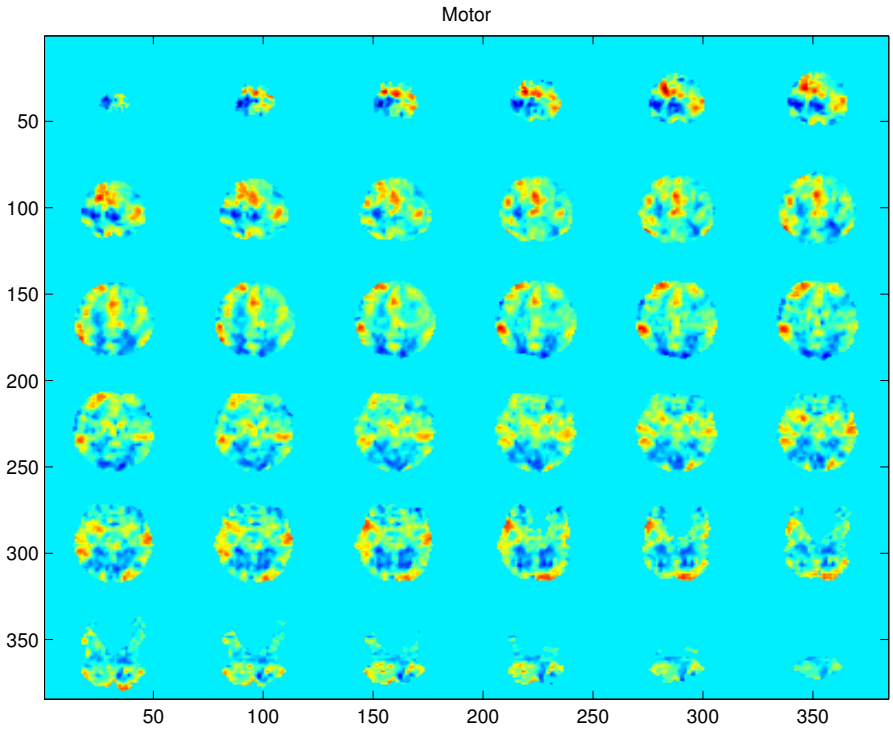

Figura 4.15: Estímulo motor sujeto 3 por SPM.

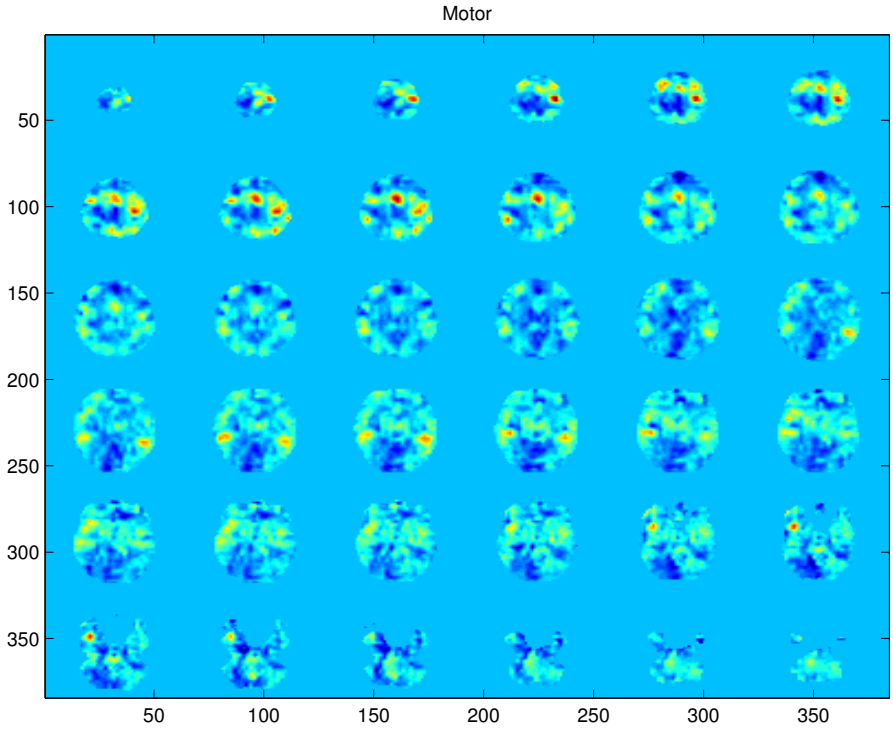

Figura 4.16: Estímulo motor sujeto 4 por GP.

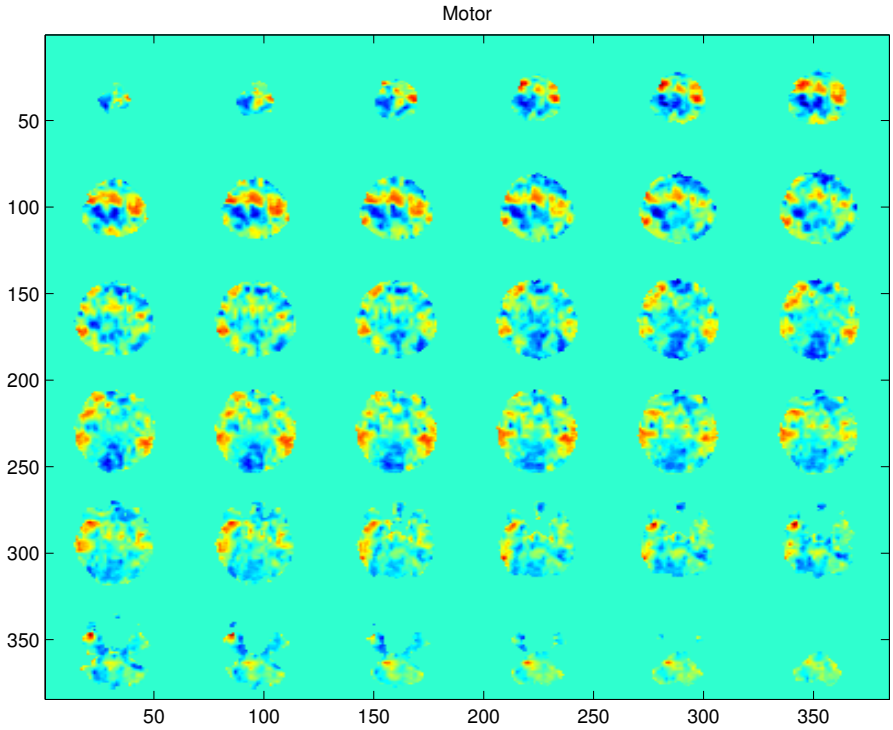

Figura 4.17: Estímulo motor sujeto 4 por SPM.

4.4.3. Estímulo cognitivo

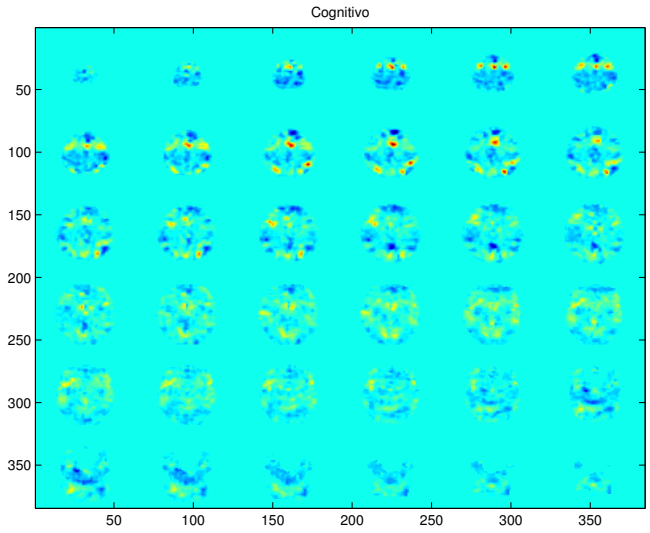

Figura 4.18: Estímulo cognitivo sujeto 1 por GP.

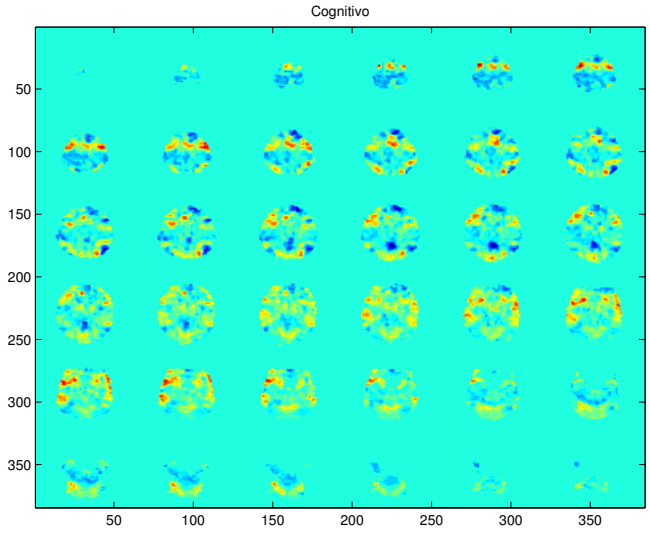

Figura 4.19: Estímulo cognitivo sujeto 1 por SPM.

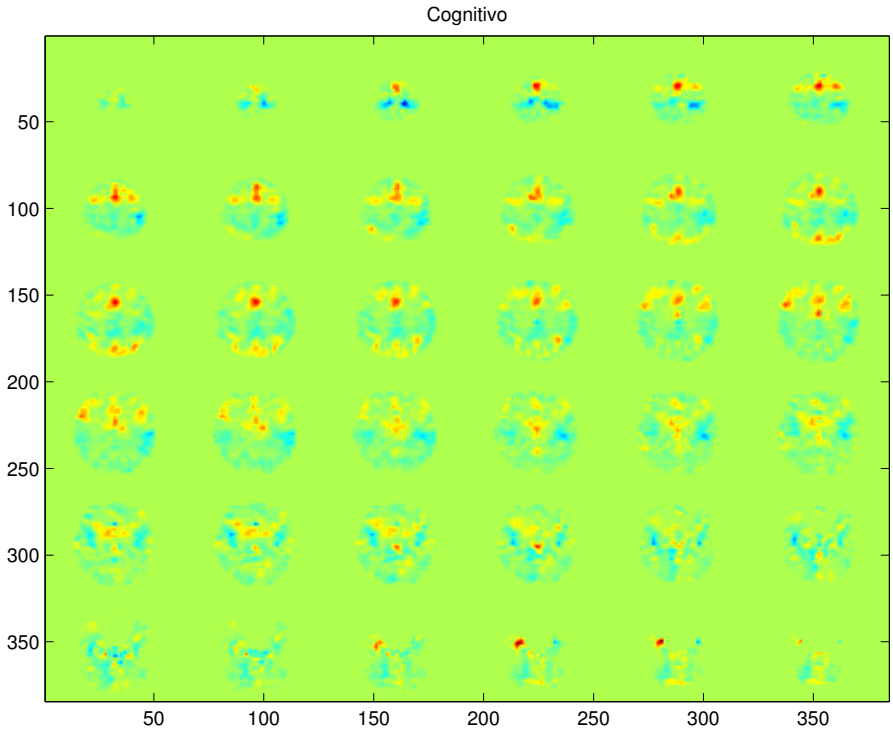

Figura 4.20: Estímulo cognitivo sujeto 2 por GP.

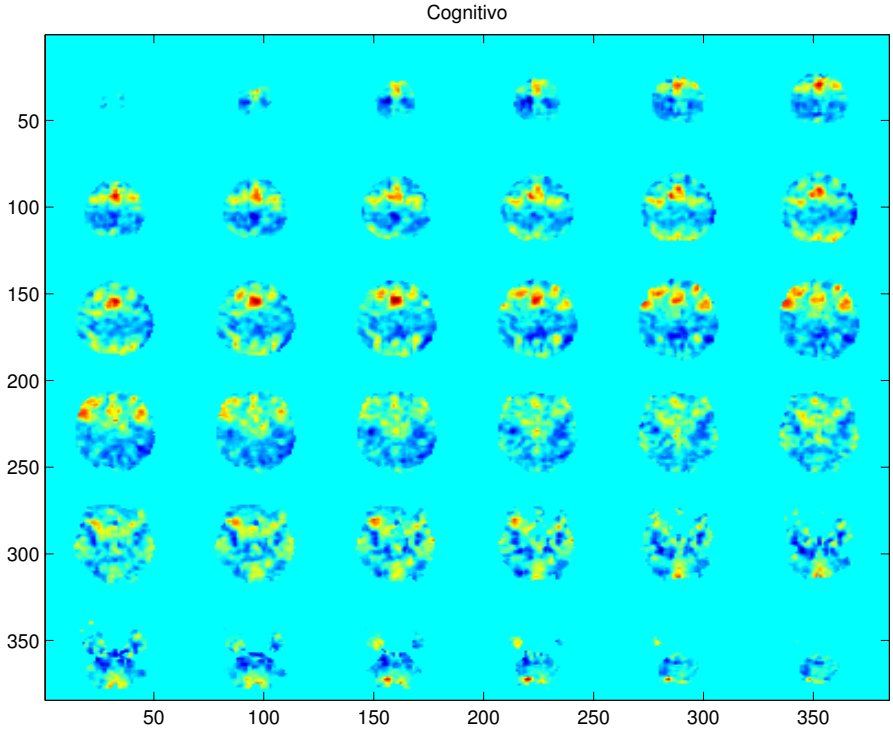

Figura 4.21: Estímulo cognitivo sujeto 2 por SPM.

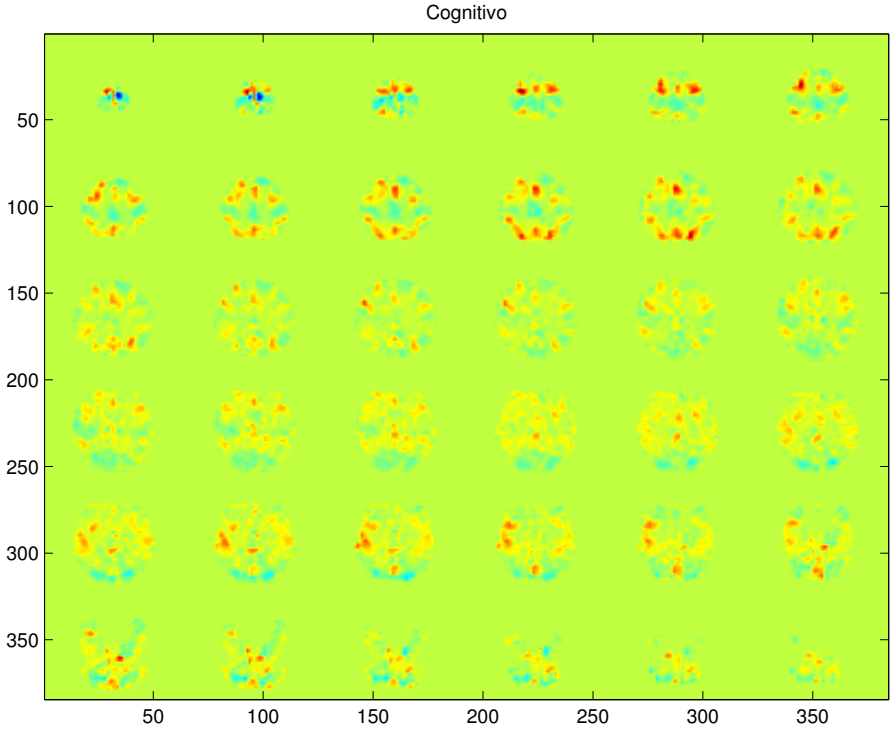

Figura 4.22: Estímulo cognitivo sujeto 3 por GP.

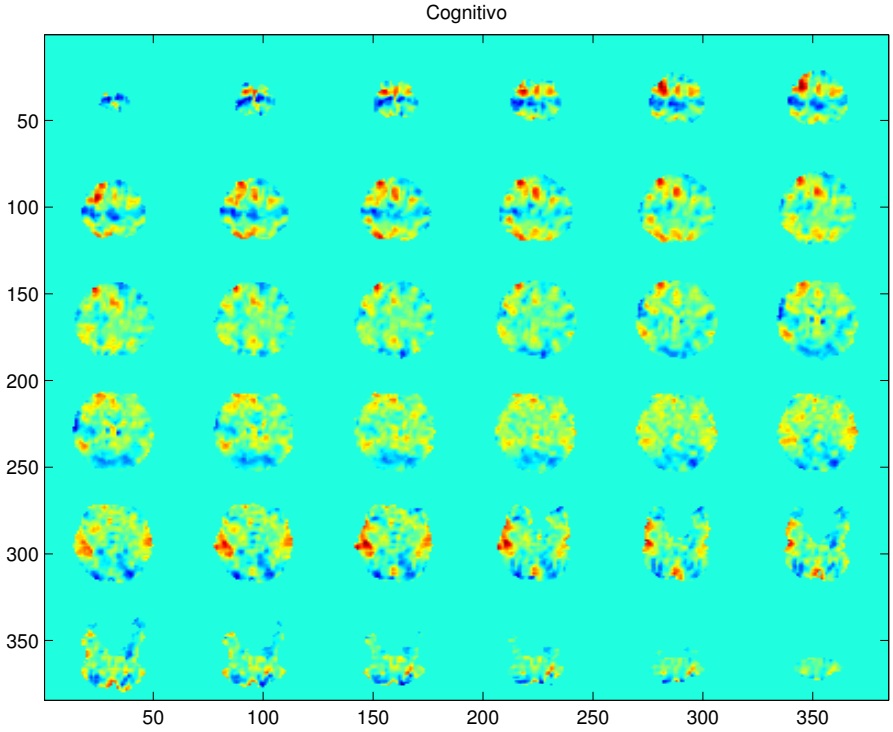

Figura 4.23: Estímulo cognitivo sujeto 3 por SPM.

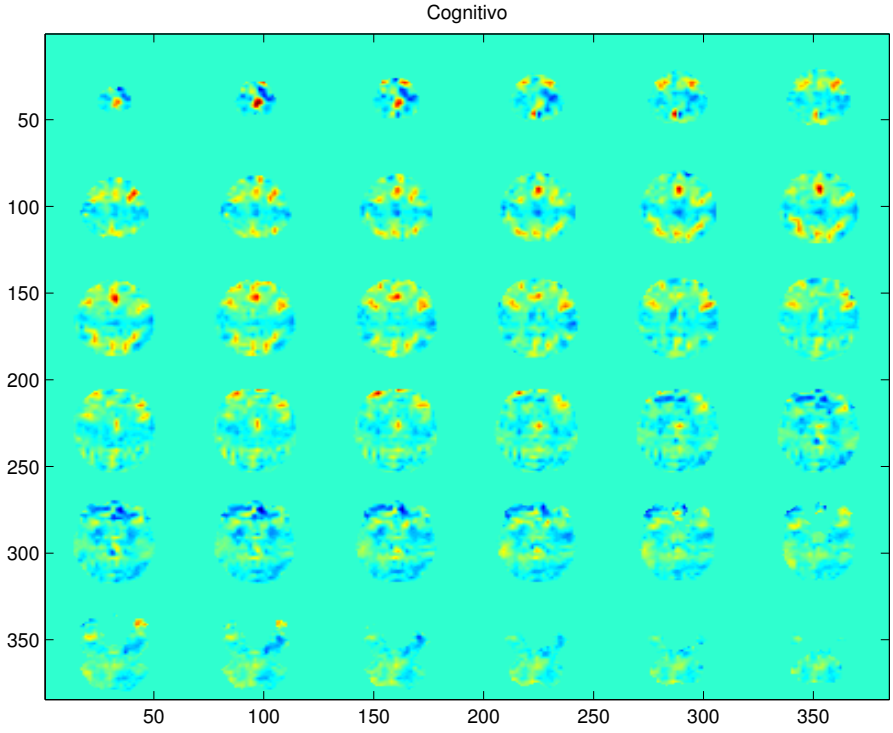

Figura 4.24: Estímulo cognitivo sujeto 4 por GP.

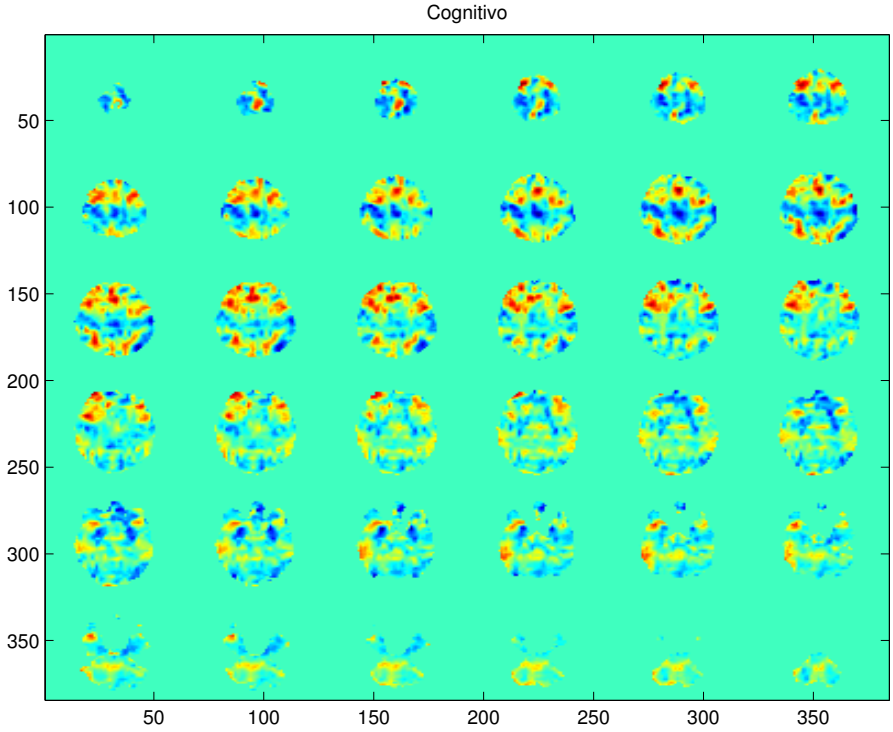

Figura 4.25: Estímulo cognitivo sujeto 4 por SPM.

4.4.4. Estímulo auditivo

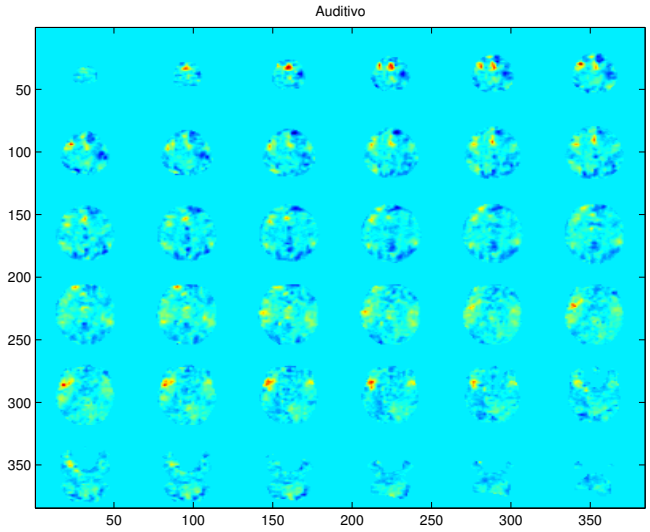

Figura 4.26: Estímulo auditivo sujeto 1 por GP.

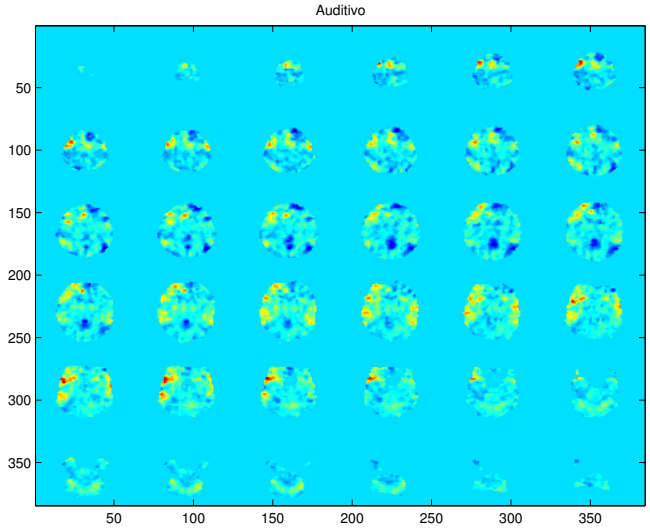

Figura 4.27: Estímulo auditivo sujeto 1 por SPM.

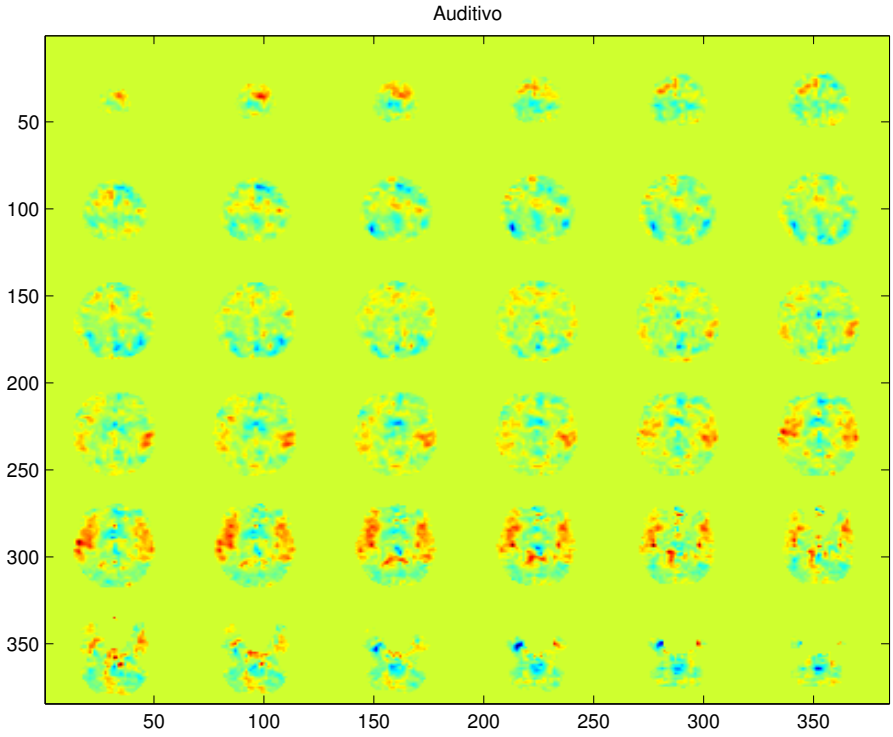

Figura 4.28: Estímulo auditivo sujeto 2 por GP.

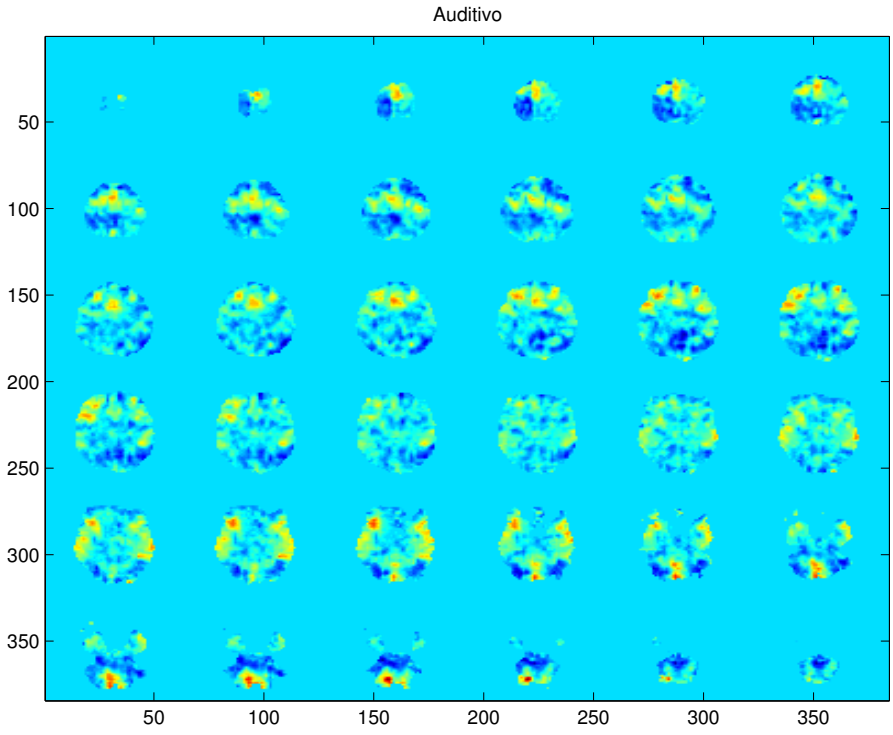

Figura 4.29: Estímulo auditivo sujeto 2 por SPM.

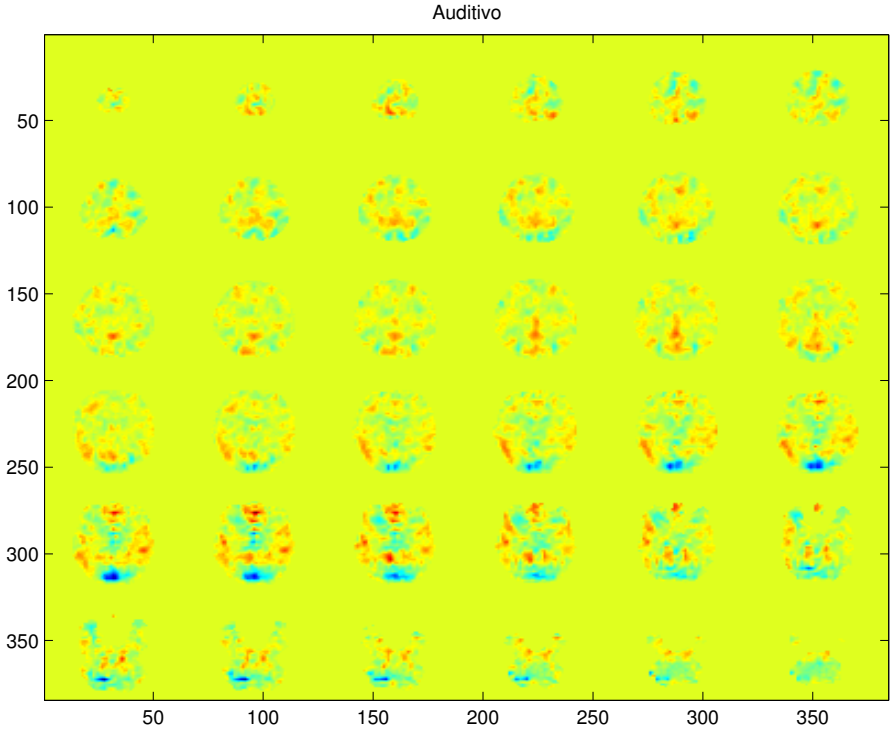

Figura 4.30: Estímulo auditivo sujeto 3 por GP.

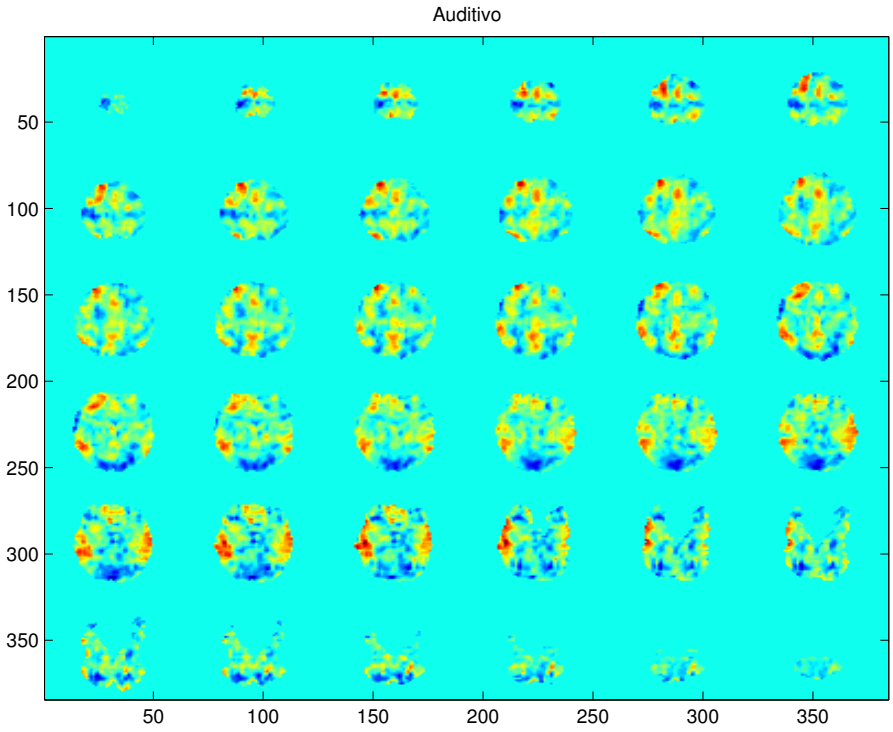

Figura 4.31: Estímulo auditivo sujeto 3 por SPM.

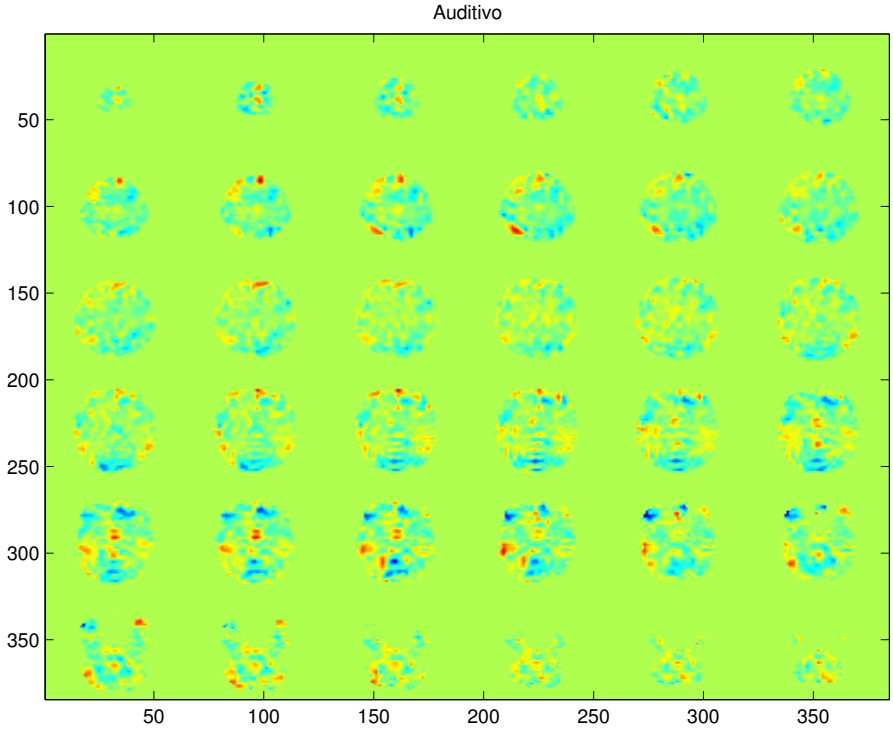

Figura 4.32: Estímulo auditivo sujeto 4 por GP.

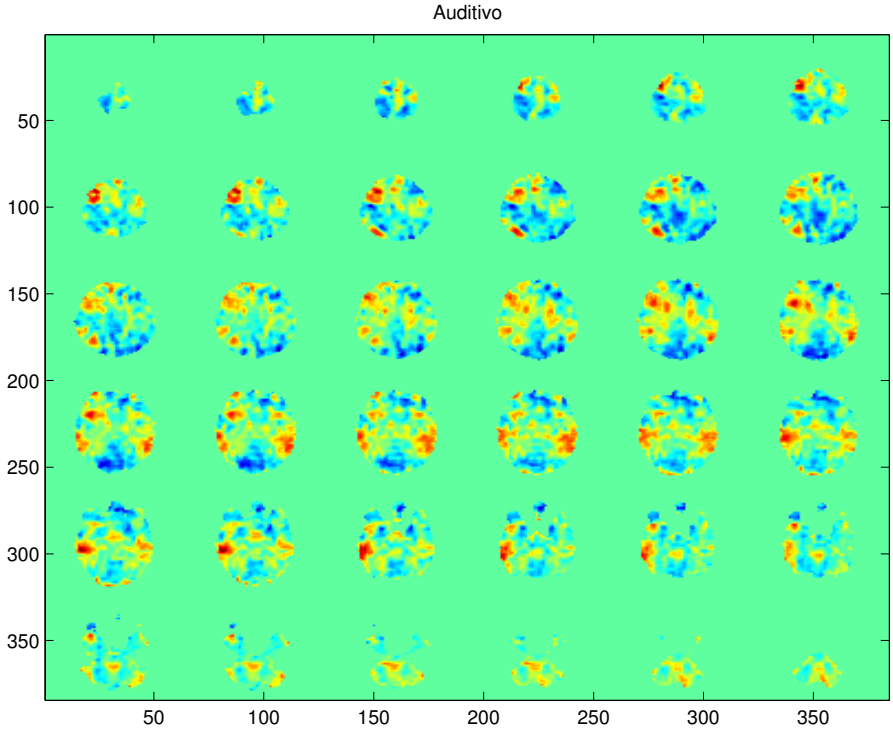

Figura 4.33: Estímulo auditivo sujeto 4 por SPM.

4.5. Experimentos Multisujeto

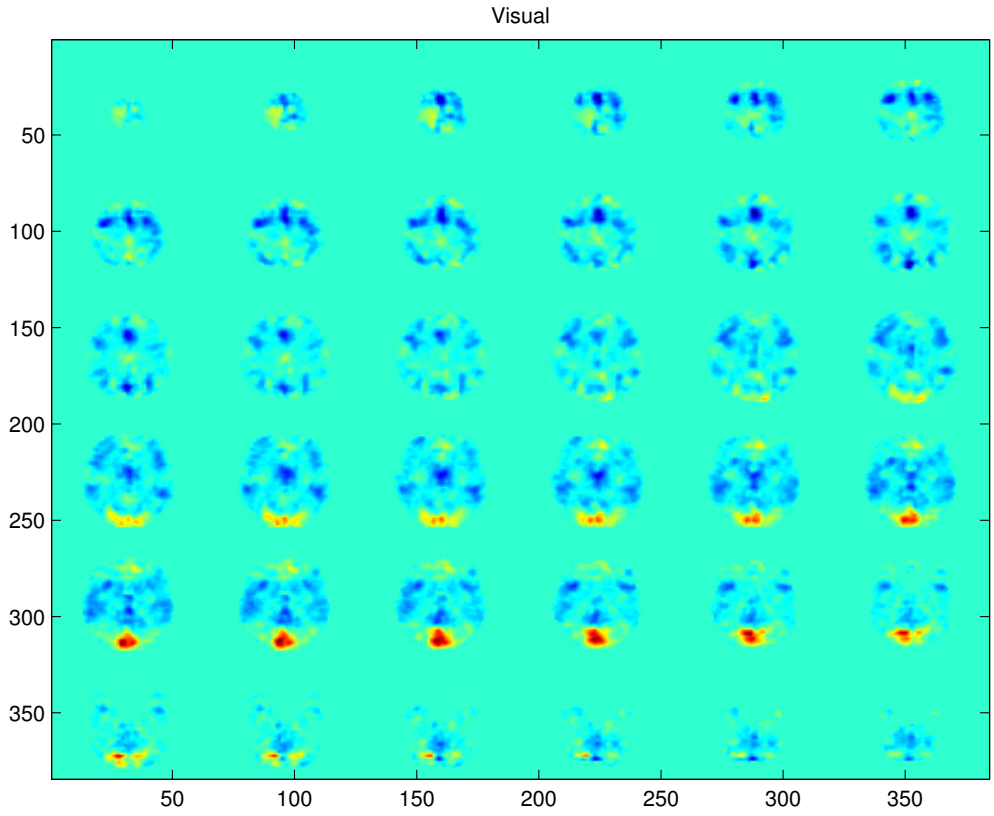

Figura 4.34: Estímulo visual multisujeto por GP.

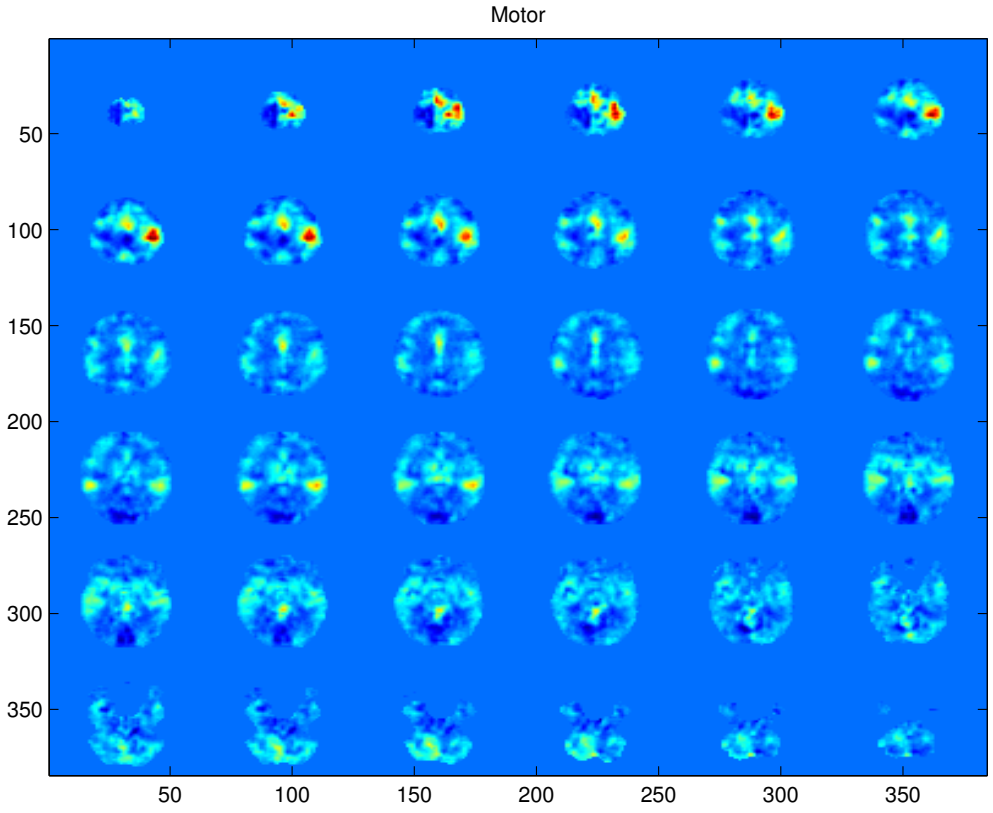

Figura 4.35: Estímulo motor multisujeto por GP.

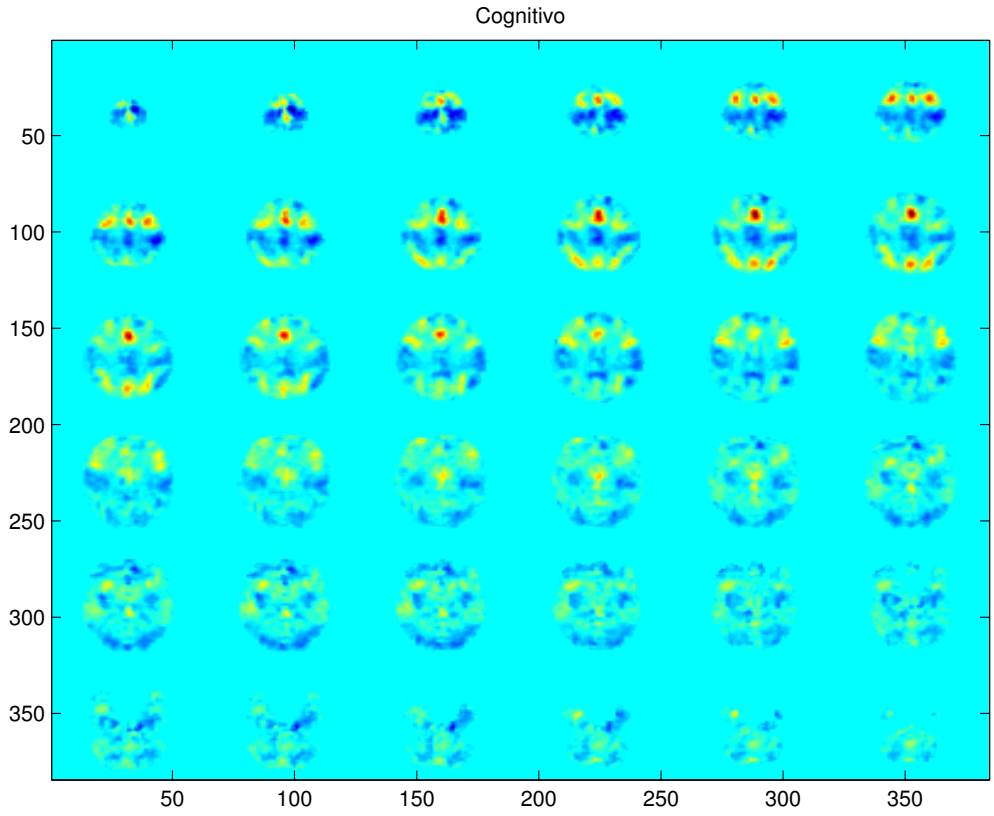

Figura 4.36: Estímulo congnitivo multisujeto por GP.

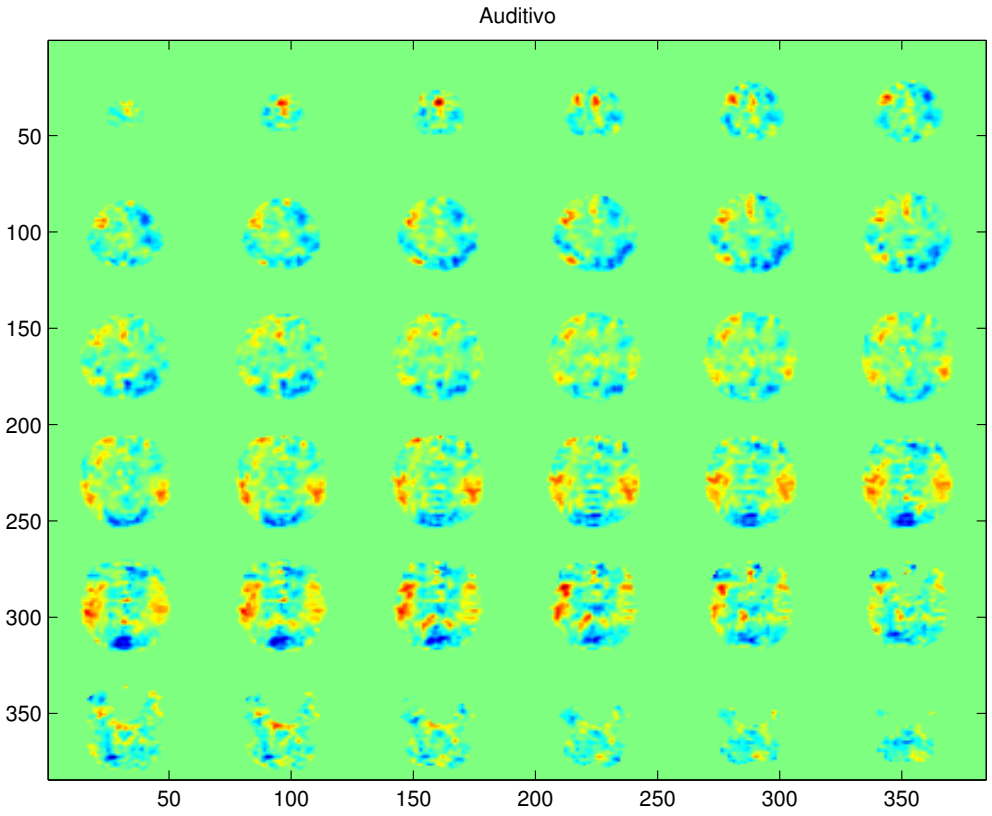

Figura 4.37: Estímulo auditivo multisujeto por GP.

Capítulo 5

CONCLUSIONES Y TRABAJOS FUTUROS

En este proyecto fin de carrera se ha presentado y se ha probado un método de análisis de imágenes por resonancia magnética funcional alternativo al estándar SPM. Este método se fundamenta en la aproximación basada en procesos gaussianos.

La regresión mediante procesos gaussianos establece un modelo estadístico a priori de los parámetros β a estimar consistente en una distribución gaussiana de media nula, para después calcular la salida de la regresión mediante la maximización de la verosimilitud del modelo. Dado que el modelo del error de estimación es también gaussiano, el resultado de la aproximación es idéntico al resultado obtenido mediante el método de minimización del error cuadrático medio, por cuanto es el modelo de máxima verosimilitud para esa distribución.

La diferencia fundamental entre ambos métodos es que el basado en procesos gaussianos permite obtener una estimación de la distribución de la salida, porque se pueden establecer a su vez estimaciones de los intervalos de confianza de la estimación. Además, se pueden ajustar todos los parámetros libres mediante la maximización de la verosimilitud con respecto a esos parámetros. El parámetro libre que se ajusta en los experimentos de este proyecto es el del modelo de ruido, que se supone gaussiano blanco. No obstante, se puede aproximar también un modelo gaussiano blanco combinado con un modelo AR(1) como se hace en la aproximación SPM, lo que se deja como trabajo futuro de este proyecto.

Los tests que se han llevado a cabo consisten en la estimación de estímulos sensorimotores y cognitivos en experimentos monosujeto y multisujeto en fMRI. Las gráficas muestran que los

valores de los mapas β producidos por el método GP son similares a las estimaciones del método SPM, que consisten en mapas de t de Student. En la aproximación presentada no ha sido necesario hacer ningún test estadístico.

Puede observarse en algunas gráficas que la sensibilidad de los mapas basados en GP es igual o incluso superior a la del método SPM, y se registran en general menos artefactos. No obstante, en algunos mapas, sobre todo en los motores, la sensibilidad ha resultado ser algo menor, no observándose, en particular, en algunos casos, activación perimotora, premotora o cerebelar. Los resultados podrían mejorar mediante la sofisticación de los métodos de ajuste de parámetros por maximización de la verosimilitud y, sobre todo, utilizando un moceo para el ruido basado en AR(1), más realista que el utilizado en este proyecto.

Aunque los resultados no son en absoluto concluyentes, lo que queda fuera de los límites establecidos para este proyecto fin de carrera, estos resultados permiten afirmar que se puede desarrollar una metodología basada en procesos gaussianos que produce resultados competitivos con SPM, sin necesidad de llevar a cabo análisis estadísticos del tipo t de Student, pero permitiendo valorar los intervalos de confianza de cada valor de β y, por lo tanto, estableciendo umbralizaciones basadas en un criterio estadístico.

Los trabajos futuros de este proyecto, ya sugeridos arriba, se pueden resumir en dos bloques. El primero de ellos consiste en el estudio de los modelos alternativos para el ruido, que pueden ser iguales a los usados en SPM. Este método utiliza una estrategia llamada ReML, que es una forma de maximización de la verosimilitud, para la estimación de los parámetros del ruido, lo que también se puede utilizar en GP. El segundo consiste en establecer métodos de umbralización de los mapas basándose en la estimación de los intervalos de confianza de los resultados. En otras palabras, el método produce modelos a posteriori de las distribuciones estadísticas de los parámetros, con lo que se pueden umbralizar mediante una estimación estadística razonada.

Bibliografía

[1] J. Ashburner and K. Friston. Multimodal image coregistration and partitioning - a unified framework. *Neuroimage*, 6:209–217, 1997.

[2] J. Ashburner, P. Neelin, D. Collins, A. Evans, and K. Friston. Incorporating prior knowledge into image registration. *Neuroimage*, 6:344–352, 1997.

[3] C. Chatfield and A. Collins. *Introduction to multivariate analysis.* Chapman and Hall, London and New York, 1980.

[4] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny. *Statistical Parametric Mapping: The Analysis of Functional Brain Images.* Academic Press, Department of Cognitive Neurology. University College London, 2007.

[5] K. Friston, A. Holmes, K. Worsley, J. Poline, C. Frith, and R. Frackowiak. Statistical parametric maps in functional imaging in a general linear approach. *Hum Brain Map*, 2:189–210, 1995.

[6] S. Geisser and W. Eddy. *A Predictive Approach to Model Selection*, volume 74(365). 1979.

[7] D. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *American Statistical Association*, pages 320 – 338, 1977.

[8] T. Hill and P. Lewicki. *Statistics: Methods and Applications.* StatSoft, Tulsa, United States of America, 2007.

[9] S. J. Kiebel and K. J. Friston. Statistical Parametric Mapping: I. Generic Considerations. *Neuroimage*, 2:402 – 502, Jun 2004.

[10] S. J. Kiebel and K. J. Friston. Statistical Parametric Mapping: II. A Hierarchical Temporal Model. *Neuroimage*, 2:503 –520, Jun 2004.

[11] G. Kimeldorf and G. Wahba. Some Results on Tchebycheffian Spline Functions. *J. Matematical Analysis and Applications*, 33(1):82 – 95, 1971.

[12] T. S. V. W. Press, W.H. and B. Flannery. *Statistics: Numerical Recipes in C.* Cambridge University Press, 1992.

[13] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis.* Cambridge University Press, Cambridge, United Kingdom, 2004.

[14] J. Tailarach and P. Tournoux. *Co-planar stereotaxic atlas of the human brain: 3-Dimensional proportional.* Thieme Medical Publishers, Stuttgart, 1998.

[15] G. Wahba. Spline Models for Observational Data. *Society for Industrial and Applied Mathematics*, page sec 4.8, 1990.

[16] R. Woods, S. Cherry, and J. Maziota. Rapid automated algorithm for aligning and reslicing pet images. *J Comput Assist Tomogr*, 16:620–633, 1992.

[17] K. Worsley, S. Marrett, P. Neelin, A. Vandal, K. Friston, and A. Evans. A unified statistical approach for determining significant voxels in images of cerebral activation. *Hum brain map*, 4:58–73, 1996.