USC DEPARTAMENTO DE ESTATÍSTICA E INVESTIGACIÓN OPERATIVA
UNIVERSIDADE DA CORUÑA Departamento de Matemáticas
Universidade Vigo Departamento de Estatística e Investigación Operativa

TRABAJO FIN DE MÁSTER

Comparación de técnicas de estimación de datos genéticos: imputación genética vs interpolación espacial

MÁSTER UNIVERSITARIO EN TÉCNICAS ESTADÍSTICAS

Adrián Lema Casal

Índice general

Capítulo 1

Introducción y objetivos

Desde hace unos años los marcadores genéticos más ampliamente utilizados, especialmente en los estudios de asociación con enfermedades, son los polimorfismos de un solo nucleótido (SNPs, según sus siglas en inglés). La gran mayoría de los SNPs analizados y considerados en este estudio son bialélicos: en esa posición particular del genoma puede haber uno de dos alelos posibles, supongamos A o B (a efectos del estudio y de la explicación de los métodos es irrelevante qué nucleótido concreto representa A o B). Teniendo en cuenta que somos organismos diploides, cada SNP en cada individuo puede presentarse en forma de 3 combinaciones (genotipos) diferentes, AA, AB o BB. Los individuos con genotipo AA o BB (los dos alelos para un SNP son el mismo) son llamados homocigotos y los individuos con genotipo AB se denominan heterocigotos, siempre en referencia al SNP considerado.

En los estudios de asociación genética con enfermedades se evalúa la diferencia en frecuencia relativa de los genotipos en casos (pacientes) y controles (población sana) en gran número de SNPs contenidos en arrays comerciales. Hay grandes diferencias en los SNPs genotipados en cada array y eso puede plantear problemas a la hora de replicar o combinar resultados de estudios individuales, por lo cual se han desarrollado diferentes métodos de imputación genética que estiman el genotipo para un SNP no genotipado en una serie de individuos. La mayoría de estos métodos de imputación llevan consigo la determinación de los haplotipos (secuencia de alelos a lo largo de un gen o sección cromosómica) para lo cual se tiene en cuenta la relación física entre SNPs (SNPs cercanos no son independientes y tienden a transmitirse juntos, véase Apéndice A), la frecuencia relativa de los SNPs en la población analizada y la comparación con secuencias de referencia en las bases de datos HapMap Project [1] y 1000 Genomes Project[2].

Debemos tener presente que, aunque la imputación se realice a nivel de cada individuo para un SNP determinado, a efectos prácticos el resultado final que interesa es la frecuencia relativa (de los genotipos o de los alelos) en la población en la que ese SNP no fue genotipado. Por otra parte, la frecuencia relativa de

[1] HAPMAP Project: http://hapmap.ncbi.nlm.nih.gov/

[2] 1000Genomes Project: http://www.1000genomes.org/

genotipos para un SNP puede variar no sólo en función de su asociación con una determinada patología, sino en función de la población estudiada: la frecuencia relativa de un SNP en población gallega, por ejemplo, puede ser diferente a la frecuencia en población nórdica, como se ha visto en SNPs de algunos genes que muestran gradientes latitudinales a lo largo de Europa. En casos así es dónde se plantea la duda que da pie a este estudio: si queremos estimar la frecuencia de un SNP en una población ¿es mejor imputarlo a partir de los softwares disponibles (computacionalmente costosos) o estimarlo a partir de la frecuencia relativa en poblaciones cercanas geográficamente? En este texto se presenta una comparativa entre las técnicas empleadas en ambas situaciones, la imputación genética y la interpolación espacial respectivamente.

Esta memoria complementa una estancia de 4 meses (de septiembre a diciembre) en el Departamento de Bioinformática del Centro Singular de Investigación en Medicina Molecular y Enfermedades Crónicas (CIMUS) perteneciente a la Universidad de Santiago de Compostela, a lo largo de los cuales se han manipulado los datos empleados para la elaboración de este trabajo y obtenido los resultados relativos a los mismos. Todo ello bajo la supervisión de Raquel Cruz Guerrero y Rosa María Crujeiras Casais.

Este trabajo fin de máster puede dividirse en dos partes bien diferenciadas; la primera parte orientada a los modelos de Markov ocultos [9] (Capítulo 2) con el objetivo de exponer las técnicas de *imputación genética* [4] [8] (Capítulo 3) basadas en estos procesos estocásticos, para una posterior aplicación de los métodos en un conjunto de datos (Capítulo 4).

La segunda parte del trabajo consta de un capítulo dedicado a los procesos estocásticos espaciales (estadística espacial) para posteriormente introducir el método *kriging* para interpolación espacial óptima [10] (Capítulo 5), herramienta con la que acto seguido se hará de nuevo una aplicación en datos (Capítulo 6) susceptible de ser comparada con los resultados obtenidos mediante imputación (frecuencias de un determinado alelo en un SNP a lo largo de un número elevado de individuos), constituyendo esta comparativa una breve conclusión del texto.

Capítulo 2

Modelos de Markov ocultos

En este capítulo introduciremos el concepto de cadena de Markov en tiempo discreto, detallaremos los elementos básicos de las mismas e ilustraremos su funcionamiento mediante algunos ejemplos, para posteriormente generalizar este concepto a otras situaciones mediante los modelos de Markov ocultos (hidden Markov models), siendo estos últimos fundamentales para entender los métodos de imputación abordados en el Capítulo 3.

2.1. Introducción a los modelos de Markov

Consideremos un sistema que pueda ser descrito en cualquier instante mediante un elemento del conjunto de N de estados distintos $\{S_1, \ldots, S_N\}$. Este sistema puede verse sometido a cambios de estado de acuerdo a un conjunto de probabilidades asociadas al estado[1]. Denotamos los instantes de tiempo asociados a cada cambio de estado por $t = 1, 2, \ldots$ y el estado en el momento t del sistema por q_t. En general, se requiere una especificación completa del estado actual (en tiempo t) de igual modo que de los estados previos para una descripción completa del sistema. En el caso particular de una cadena de Markov discreta de primer orden, la citada descripción se reduce al estado actual y su predecesor, i.e.,

$$\mathbb{P}[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \ldots] = \mathbb{P}[q_t = S_j | q_{t-1} = S_i]. \tag{2.1}$$

Además, solo consideraremos aquellos procesos en los que el lado derecho de la ecuación (2.1) es independiente del tiempo, derivando en un conjunto de probabilidades de transición a_{ij} de la forma:

$$a_{ij} = \mathbb{P}[q_t = S_j | q_{t-1} = S_i], \quad 1 \leq i, j \leq N, \quad \forall t,$$

[1] Los conceptos abordados en esta sección han sido estudiados en la materia del máster Procesos Estocásticos.

verificando

$$a_{ij} \geq 0, \quad \sum_{j=1}^{N} a_{ij} = 1.$$

El proceso estocástico que se cita a continuación puede ser llamado un modelo (o cadena) de Markov observable puesto que la salida del proceso es un conjunto de estados en cada instante de tiempo, donde cada estado corresponde a un evento físico observable.

Para afianzar ideas, consideremos un modelo de Markov dado por tres estados que representan la meteorología en un día en particular [9], esta situación constituye un ejemplo clásico en cadenas de Markov en tiempo discreto (*weather chain*). Se supone que, una vez al día, se observa el tiempo pudiendo resultar en:

- Estado 1: Lluvioso
- Estado 2: Nublado
- Estado 3: Soleado

Suponemos que el tiempo en el día t queda determinado por un único estado de los tres descritos, además la matriz A de probabilidades de transición viene dada por

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

Por ejemplo, la probabilidad de pasar de una día lluvioso a un día soleado es 0.3 y la probabilidad de pasar de un día soleado a otro soleado es 0.8. Sabiendo que el tiempo en el día 1 ($t = 1$) es soleado (estado 3), podemos preguntarnos cuál es la probabilidad (de acuerdo al modelo) de que el tiempo en la siguiente semana venga dado por *soleado-soleado-lluvioso-lluvioso-soleado-nublado-soleado*. Formalmente, definimos la secuencia de observaciones O como $O = \{S_3, S_3, S_1, S_1, S_3, S_2, S_3\}$ correspondiente a $t = 1, \ldots, 8$, y deseamos determinar la probabilidad de O, dado el modelo. Esta probabilidad viene dada por

$$\begin{aligned} \mathbb{P}[O|\text{Modelo}] &= \mathbb{P}[S_3, S_3, S_1, S_1, S_3, S_2, S_3|\text{Modelo}] \\ &= \mathbb{P}[S_3]\mathbb{P}[S_3|S_3]\mathbb{P}[S_3|S_3]\mathbb{P}[S_1|S_3]\mathbb{P}[S_1|S_1]\mathbb{P}[S_3|S_1]\mathbb{P}[S_2|S_3]\mathbb{P}[S_3|S_2] \\ &= \pi_3 a_{33} a_{33} a_{31} a_{11} a_{13} a_{32} a_{23} = 1.536 \cdot 10^{-4} \end{aligned}$$

donde hemos empleado la notación

$$\pi_i = \mathbb{P}[q_1 = S_i], \quad i \in \{1, \ldots, N\}$$

para denotar las probabilidades de estado inicial.

También podemos preguntarnos qué probabildad hay de que, dado un estado, el sistema permanezca en él exactamente d días. Esta probabilidad puede ser evaluada como la probabilidad de secuencia de observaciones $O = \{S_i, S_i, \overset{d}{\ldots}, S_j \overset{d+1}{\neq} S_i\}$ dado el modelo, que es

$$\mathbb{P}[O|\text{Modelo}, q_1 = S_i] = a_{ii}^{d-1}(1 - a_{ii}) = p_i(d).$$

Este valor $p_i(d)$ es la función masa de probabilidad de permanecer d días en el estado i. Basándonos en $p_i(d)$ podemos calcular el número esperado de observaciones (duración) en un estado, condicionado a iniciar en uno dado, como

$$\bar{d_i} = \sum_{d=1}^{\infty} dp_i(d) = \sum_{d=1}^{\infty} da_{ii}^{d-1}(1 - a_{ii}) = \frac{1}{1 - a_{ii}}.$$

De este modo, el número esperado de días consecutivos soleados, de acuerdo al modelo, es $\dfrac{1}{0.2} = 5$.

Acabamos de considerar modelos de Markov en los que cada estado se corresponde a un evento observable. Este modelo, en cambio, es demasiado restrictivo para el contexto de interés de este trabajo fin de máster (véase Capítulo 3). Debido a esto extenderemos el concepto de modelo de Markov a situaciones donde la observación es una función probabilística del estado, i.e., el modelo resultante (modelo de Markov oculto) es un proceso estocástico asociado a otro subyacente no observable (oculto), pero que se puede inferir a través de otro conjunto de procesos estocásticos que producen la secuencia observada. Ilustramos esta idea con un ejemplo:

Supongamos que nos encontramos en una habitación con una barrera opaca, y que al otro lado de la citada barrera hay un individuo lanzando repetidas veces una o varias monedas al aire. La otra persona no nos informará exactamente de lo que está haciendo, solamente del resultado de cada lanzamiento. De este modo, una secuencia observada típica sería

$$O = O_1O_2O_3 \ldots O_T = CXXCXCCXXXX \ldots$$

donde C y X representan cara y cruz respectivamente. Dado este escenario, el problema de interés radica en la construcción de un modelo de Markov oculto (hidden Markov model) para explicar la secuencia observada de caras y cruces. El primer problema al que nos enfrentamos consiste en decidir a qué corresponden los estados del modelo. Una posible solución consiste en asumir que tan solo fue lanzada una moneda; en este caso podríamos modelar el sistema como una cadena de 2 estados, donde cada uno de ellos corresponde a los posibles resultados de la moneda. Una segunda forma de modelar el escenario propuesto (explicar la secuencia observada) consiste en considerar que hay dos estados (correspondientes al lanzamiento de dos monedas distintas). Cada estado está caracterizado por la distribución de probabilidad de cara y cruz, y las transiciones entre estados están modelados por la matriz de transición. De igual modo, el mecanismo físico que determina como se realizan las transiciones entre

estados podría ser un conjunto de lanzamientos independientes, u otro evento probabilístico.

2.2. Modelos de Markov ocultos

El anterior ejemplo nos dan una idea de lo que es un modelo de Markov oculto y como pueden ser aplicado a determinados escenarios. Definimos a continuación (formalmente) los elementos de una cadena de Markov oculta y explicaremos cómo se pueden generar las observaciones del modelo. Una cadena de Markov oculta [9] queda caracterizada por:

1. N, el número de estados del modelo. A pesar de que los estados permanezcan ocultos, para muchas aplicaciones prácticas los estados o conjuntos de estados adquieren relevancia. Generalmente estos estados están interconectados de modo que cualquier estado pueda ser alcanzado por los demás estados. Denotamos los estados individuales por $S = \{S_1, S_2, \ldots, S_N\}$, y el estado en un tiempo t por q_t. Esta notación es igual a la empleada en cadenas de Markov.

2. M, el número de símbolos observados por estado, i.e., el tamaño del alfabeto discreto. Los símbolos observados corresponden a la salida física del sistema a modelar. Se denotan los símbolos individuales por $V = \{v_1, v_2, \ldots, v_M\}$.

3. Las probabilidades de transición $\mathcal{A} = \{a_{ij}\}$ donde

$$a_{ij} = \mathbb{P}[q_{t+1} = S_j | q_t = S_i], \quad i, j \in \{1, \ldots, N\}.$$

   En el caso particular de que cualquier estado pueda alcanzar otro en un solo paso, se tiene $a_{ij} > 0$ para cada i, j. Para otro tipo de modelos de Markov ocultos, se puede tener $a_{ij} = 0$ para uno o más pares (i, j).

4. La distribución de probabilidad del símbolo observado en un estado j, $\mathcal{B} = \{b_j(k)\}$, donde

$$b_j(k) = \mathbb{P}[v_k \text{ en tiempo } t | q_t = S_j], \quad j \in \{1, \ldots, N\}, \quad k \in \{1, \ldots, M\}.$$

5. La distribución de estado inicial $\pi = \{\pi_i\}$ donde

$$\pi_i = \mathbb{P}[q_1 = S_i], \quad i \in \{1, \ldots, N\}.$$

Proporcionando valores apropiados de N, M, $\mathcal{A}$, $\mathcal{B}$ y π, la cadena de Markov oculta puede ser empleada como generadora para dar una secuencia de observaciones

$$O = O_1 O_2 \ldots O_T$$

donde cada O_t es uno de los símbolos de V. Nótese que en lo sucesivo, por simplicidad, los índices i y j harán referencia al estado S_i y S_j respectivamente.

Paso 1. Selecciona un estado inicial $q_1 = S_i$ de acuerdo a la distribución del estado inicial π.

Paso 2. Fija $t = 1$.

Paso 3. Selecciona $O_t = v_k$ de acuerdo a la distribución de probabilidad del símbolo observado en el estado S_i, i.e., $b_i(k)$.

Paso 4. Transita a un nuevo estado $q_{t+1} = S_i$ de acuerdo a la distribución de probabilidades de transición en para el estado S_i, i.e., a_{ij}.

Paso 5. Fija $t = t + 1$; vuelve al Paso 3 si $t < T$; en otro caso finaliza el proceso.

El procedimiento previamente descrito puede ser empleado como generador de las observaciones. En resumen, para una completa especificación de una cadena de Markov oculta se requieren los valores (N, M), la especificación de los símbolos observados y la especificación de las tres masas de probabilidad $\mathcal{A}$, $\mathcal{B}$, y π. En general, usamos la notación compacta

$$\lambda = (\mathcal{A}, \mathcal{B}, \pi)$$

para denotar el modelo.

2.3. Problemas en los modelos de Markov ocultos

Una vez establecida la forma de un modelo de Markov oculto en la sección previa, existen tres problemas básicos de interés que deben ser solventados para que el modelo resulte de utilidad en aplicaciones reales [9]. Estos son:

1. Dado el modelo $\lambda = (\mathcal{A}, \mathcal{B}, \pi)$ calcular la probabilidad de una secuencia observada $O = O_1 \dots O_T$ de símbolos.
2. Encontrar la secuencia de estados "óptima" asociada a las secuencias observadas.
3. Dado el modelo y una secuencia de observaciones, determinar un método para ajustar los parámetros del modelo $(\mathcal{A}, \mathcal{B}, \pi)$ que maximizen la probabilidad de la secuencia de observaciones.

En las siguientes secciones se presentan algunas soluciones a estos problemas.

2.3.1. Probabilidad de una secuencia observada

Se desea calcular la probabilidad de una secuencia observada de símbolos, $O = O_1 \dots O_T$, dado el modelo $\lambda = (\mathcal{A}, \mathcal{B}, \pi)$, i.e., $\mathbb{P}(O|\lambda)$. La forma más directa de realizar este proceso consiste en enumerar cada posible secuencia de estados de longitud T. Consideremos una secuencia de estados dada por

$$Q = q_1 q_2 \dots q_T$$

siendo q_1 el estado inicial. Bajo hipótesis de independencia entre las observaciones, la probabilidad de la secuencia observada O es

$$\mathbb{P}(O|Q,\lambda) = \prod_{t=1}^{T} \mathbb{P}(O_t|Q_t,\lambda).$$

Se tiene, por tanto, que:

$$\mathbb{P}(O|Q,\lambda) = b_{q_1}(O_1)b_{q_2}(O_2)\dots b_{q_T}(O_T).$$

La probabilidad de la secuencia de estados Q viene dada por

$$\mathbb{P}(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T},$$

donde a_{ij} ($i \in \{q_1,\dots,q_{T-1}\}$, $j \in \{q_2,\dots,q_T\}$) son las probabilidades de transición y π_{q_1} es la probabilidad del estado inicial. La probabilidad conjunta de O y Q, i.e., la probabilidad de que O y Q ocurran simultáneamente es, en virtud del Teorema de Bayes:

$$\mathbb{P}(O,Q|\lambda) = \mathbb{P}(O|Q,\lambda)\mathbb{P}(Q,\lambda).$$

La probabilidad de la secuencia observada O, dado el modelo, se obtiene sumando esta probabilidad conjunta sobre las distintas secuencias de estados q:

$$\begin{aligned} \mathbb{P}(O|\lambda) &= \sum_{Q} \mathbb{P}(O,Q,\lambda) = \sum_{Q} \mathbb{P}(O|Q,\lambda)\mathbb{P}(Q|\lambda) \\ &= \sum_{q_1\dots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \dots a_{q_{T-1} q_T} b_{q_T}(O_T). \end{aligned} \tag{2.2}$$

La ecuación (2.2) se interpreta como sigue. En el estado inicial ($t = 1$) la cadena está en q_1 con probabilidad π_{q_1} y se genera el símbolo O_1 en este estado con probabilidad $b_{q_1}(O_1)$. El tiempo se translada de t a $t+1$ ($t = 2$), se realiza la transición del estado q_1 al estado q_2 con probabilidad $a_{q_1 q_2}$ y se genera el símbolo O_2 con probabilidad $b_{q_2}(O_2)$. El proceso continua de este modo hasta que se realiza la transición en tiempo T del estado q_{T-1} al q_T con probabilidad $a_{q_{T-1} q_T}$ y se genera el símbolo O_T con probabilidad $b_{q_T}(O_T)$. El cálculo descrito en la ecuación (2.2) involucra $2TN^T$ cálculos, puesto que para cada $t = 1,\dots,T$, existen N posibles estados que pueden ser alcanzados (N^T posibles secuencias de estados), y para cada una de estas secuencias de estados se requieren sobre $2T$ cálculos para cada término en la suma de (2.2). Por todo lo expuesto, este cálculo resulta inabordable, incluso para valores de N y T bajos. Por ejemplo, para $N = 5$ estados y $T = 100$ (observaciones) se necesitan cálculos del orden de $2 \cdot 100 \cdot 5^{100} \approx 10^{72}$. Se requiere de un procedimiento más eficiente para la ejecución de este cálculo; un procedimiento alternativo se conoce como algoritmo *Forward-Backward* [2], que se describe a continuación.

Consideremos la probabilidad *forward* $\alpha_t(i)$ definida por

$$\alpha_t(i) = \mathbb{P}(O_1 O_2 \dots O_t, q_t = S_i|\lambda)$$

i.e., la probabilidad de una secuencia de observaciones, $O_1 \dots O_t$ (hasta tiempo t) y del estado S_i en tiempo t, dado el modelo λ. Podemos calcular $\alpha_T(i)$ como sigue:

Paso 1. Inicialización

$$\alpha_1(i) = \pi_i b_i(O_1), \quad i \in \{1, \dots, N\}.$$

Paso 2. Inducción

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \quad t \in \{1, \dots, T-1\}, \quad j \in \{1, \dots, N\}.$$

Paso 3. Terminación

$$\mathbb{P}(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

El primer paso inicializa las probabilidades forward como la probabilidad conjunta del estado S_i (dada por π_i) y la observación inicial O_1. El estado S_j puede ser alcanzado en tiempo $t+1$ por N posibles estados S_i, $i \in \{1, \dots, N\}$, en tiempo t. Teniendo en cuenta que $\alpha_t(i)$ es la probabilidad de que la secuencia $O_1 \dots O_t$ sea observada, y que el estado en tiempo t es S_j, el producto $\alpha_t(i)a_{ij}$ es, de este modo, la probabilidad del evento conjunto que $O_1 \dots O_t$ sea observado, y que el estado S_j se alcanze en tiempo $t+1$ mediante el estado S_i en tiempo t. Sumando este producto sobre los N posibles estados S_i, $i \in \{1, \dots, N\}$, en tiempo t resulta la probabilidad de S_j en tiempo $t+1$.

Una vez el proceso finaliza y se conoce S_j, es sencillo ver que $\alpha_{t+1}(j)$ se obtiene contabilizando la observación O_{t+1} en el estado j, i.e., multiplicando la cantidad sumada por la probabilidad $b_j(O_{t+1})$.

El cálculo del Paso 2 se ejecuta para todos los posibles estados $j \in \{1, \dots, N\}$, para un t fijo, para posteriormente realizarlo para cada $t \in \{1, 2, \dots, T-1\}$. Finalmente, el Paso 3 produce el valor $\mathbb{P}(O|\lambda)$ como la suma de las variables terminales forward $\alpha_T(i)$.

En cuanto al coste computacional involucrado en el cálculo de $\alpha_t(j)$, $t \in \{1, \dots, T\}$, $j \in \{1, \dots, N\}$, se necesitan N^2T cálculos, en vez de $2TN^T$ (por cálculo directo). En particular, se necesitan $N(N+1)(T-1)+N$ productos y $N(N-1)(T-1)$ sumas.

Las probabilidades forward se calculan teniendo en cuenta que, puesto que solo hay N estados, todas las posibles sucesiones de estados estarán dentro de los N nodos (viendo los estados como nodos de un grafo, y las transiciones entre ellos como aristas del mismo), sin importar la longitud de dicha secuencia. En $t = 1$, necesitamos los valores de $\alpha_1(i)$, $i \in \{1, \dots, N\}$. En $t = 2, 3, \dots, T$, sólo necesitamos los valores de $\alpha_t(j)$, $j \in \{1, \dots, N\}$, donde cada operación involucra únicamente N valores previos de $\alpha_{t-1}(i)$.

De modo similar, consideremos la probabilidad *backward* $\beta_t(i)$ que se define como:

$$\beta_t(i) = \mathbb{P}(O_{t+1} O_{t+2} \dots O_T | q_t = S_i, \lambda)$$

i.e., la probabilidad de la observación parcial de O desde $t+1$ hasta T, dado el estado S_i en tiempo t y el modelo λ. De nuevo calculamos $\beta_t(i)$ mediante un proceso inductivo:

Paso 1. Inicialización

$$\beta_T(i) = 1, \quad i \in \{1, \ldots, N\}.$$

Paso 2. Inducción

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t \in \{T-1, T-2, \ldots, 1\}, i \in \{1, \ldots N\}.$$

La inicialización del algoritmo define arbitrariamente $\beta_T(i)$ como 1 para todo i. El Paso 2 muestra que, para estar en el estado S_i en tiempo t y tomar en consideración la secuencia de observaciones desde $t+1$ en adelante, deben ser considerados todos los posibles estados S_j en $t+1$, teniendo en cuenta la transición de S_i a S_j mediante a_{ij}, de igual modo que la observación O_{t+1} en el estado j ($b_j(O_{t+1})$) y, por último, la restante observación parcial en el estado j ($\beta_{t+1}(j)$). De nuevo, el cálculo de $\beta_t(i)$, $t \in \{1, \ldots, T\}$, $i \in \{1, \ldots, N\}$, requiere N^2T operaciones.

2.3.2. Secuencia óptima

El segundo problema básico en cadenas de Markov ocultas consiste en encontrar la secuencia de estados óptima asociada con las secuencias observadas. La dificultad descansa en la definición de optimalidad de una secuencia de estados ya que existen numerosos criterios para definirla.

Un posible criterio consistiría en seleccionar los estados ($\{S_1, \ldots, S_N\}$) q_t que son individualmente más probables. Para esto se define

$$\gamma_t(i) = \mathbb{P}(q_t = S_i | O, \lambda),$$

i.e., la probabilidad de encontrarse en el estado S_i en tiempo t, dada la secuencia de observaciones O, y el modelo λ.

La ecuación previa puede ser expresada en términos de probabilidades forward-backward:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\mathbb{P}(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)},$$

teniendo en cuenta que $\alpha_t(i)$ toma en consideración la secuencia parcial de observaciones $O_1 \ldots O_t$ y el estado S_i en tiempo t, mientras que $\beta_t(i)$ considera la restante parte de la observación $O_{t+1} \ldots O_T$, dado el estado S_i en tiempo t. El factor de normalización $\mathbb{P}(O|\lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$ convierte a $\gamma_t(i)$ en una masa de probabilidad sobre el conjunto de estados, entonces

$$\sum_{i=1}^{N} \gamma_t(i) = 1.$$

Empleando $\gamma_t(i)$, podemos calcular el estado individual más probable $\hat{q}_t$ en un instante t como

$$\hat{q}_t = \arg \max_{1 \leq i \leq N} [\gamma_t(i)], \quad t \in \{1, \ldots, T\}. \tag{2.3}$$

A pesar de que esta última ecuación maximiza el número esperado de estados correctos (seleccionando aquellos más probables para cada t), puede haber problemas con la secuencia resultante. Por ejemplo, cuando el modelo de Markov oculto posee probabilidades de transición entre estados nulas ($a_{ij} = 0$, para algún i, j) la secuencia de estados óptima puede, de hecho, no ser válida. Esto se debe a que la solución dada en la ecuación (2.3) simplemente determina el estado más probable en cada instante de tiempo, sin tomar en consideración las probabilidades de ocurrencia de la secuencia de estados.

Una posible solución para este problema radica en modificar el criterio de optimalidad. Por ejemplo, se puede buscar la secuencia de estados que maximize el número esperado de pares de estados correctos (q_t, q_{t+1}) o triplas (q_t, q_{t+1}, q_{t+2}). A pesar de que este criterio pueda parecer razonable a priori, el más empleado consiste en encontrar la mejor secuencia de estados individuales que existe, llamada algoritmo de Viterbi [7], que se presenta en la siguiente sección y se empleará en el Capítulo 3.

2.3.3. Algoritmo de Viterbi

Para encontrar la mejor secuencia de estados individuales, $Q = q_1 \ldots q_T$ dada la secuencia de observaciones $O = O_1 \ldots O_T$, definimos

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} \mathbb{P}[q_1 \ldots q_t = i, O_1 \ldots O_t | \lambda],$$

i.e., $\delta_t(i)$ es la mayor puntuación (probabilidad más elevada) a lo largo de una trayectoria simple, en tiempo t, que considera las t primeras observaciones y finaliza en el estado S_i. Por tanto:

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] b_j(O_{t+1}). \tag{2.4}$$

Para conseguir la secuencia de estados, necesitamos almacenar el argumento que maximiza (2.4) para cada t y para cada j. Esto se realiza mediante la creación de la matriz $\psi_t(j)$, que proporciona la secuencia de estados que proporciona la mejor trayectoria, en tiempo t, al estado S_j. El procedimiento completo para encontrar la mejor secuencia de estados viene dado por:

1. Inicialización:
$$\delta_1(i) = \pi_i b_i(O_1), \quad i \in \{1, \ldots, N\}$$
$$\psi_1(i) = 0$$

2. Repetición
$$\delta_t(j) = \max_{i \in \{1, \ldots, N\}} [\delta_{t-1}(i) a_{ij}] b_j(O_t), \quad t \in \{2, \ldots, T\}, j \in \{1, \ldots, N\},$$
$$\psi_t(j) = \arg \max_{i \in \{1, \ldots, N\}} [\delta_{t-1}(i) a_{ij}], \quad t \in \{2, \ldots, T\}, j \in \{1, \ldots, N\}.$$

3. Terminación

$$P^* = \max_{i\in\{1,\dots,N\}}[\delta_T(i)],$$
$$q_T^* = \arg\max_{i\in\{1,\dots,N\}}[\delta_T(i)].$$

4. Retroceso en la trayectoria (secuencia de estados)

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t \in \{T-1, T-2, \dots, 1\}.$$

En este algoritmo, los valores $\delta_t(j)$ representan la mayor puntuación obtenida a través de una secuencia de estados $(q_1, \dots, q_t)$ hasta llegar, en el instante t al estado S_j. La clave del algoritmo radica en que si conocemos estas variables auxiliares en el instante t para todos los estados, podemos calcularlas para $t+1$. Una vez se alcanza el instante final $t = T$, el retroceso en la trayectoria proporciona el estado que maximiza cada paso de la recursión. Nótese que este procedimiento es similar a la implementación del algoritmo forward-backward (excepto en en Paso 4). La mayor diferencia estructural radica en el uso de la maximización sobre todos los estados previos del algoritmo de Viterbi en vez de la suma.

2.3.4. Ajuste del modelo

El tercer problema de las cadenas de Markov ocultas, y posiblemente el más complicado de ellos, consiste en determinar un método para ajustar los parámetros $(\mathcal{A}, \mathcal{B}, \pi)$ que maximizen la probabilidad de la secuencia de observaciones dado el modelo. De hecho, dada cualquier secuencia de observación finita, no existe un modo óptimo de estimar los parámetros del modelo. Sin embargo, podemos seleccionar $\lambda = (A, B, \pi)$ de modo que $\mathbb{P}(O|\lambda)$ sea maximizado localmente utilizando un procedimiento iterativo como el método de Baum-Welch. Aquí se expondrá un método iterativo, basado en el trabajo de Baum [2], con el fin de estimar los parámetros del modelo.

Con el objetivo de describir el procedimiento iterativo de estimación de los parámetros de una cadena de Markov oculta, definimos como $\xi_t(i,j)$, la probabilidad de estar en el estado S_i, en el instante t, y en el estado S_j en el instante $t+1$, dado el modelo y la secuencia observada, i.e.,

$$\xi_t(i,j) = \mathbb{P}(q_t = S_i, q_{t+1} = S_j | O, \lambda).$$

Teniendo en cuenta las definiciones de las probabilidades forward y backward se puede escribir $\xi_t(i,j)$ de la forma

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\mathbb{P}(O|\lambda)} = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)},$$

donde el término del numerador es exactamente $\mathbb{P}(q_t = S_i, q_{t+1} = S_j, O|\lambda)$ y el cociente entre $\mathbb{P}(O|\lambda)$ lo convierte en una masa de probabilidad. Se ha definido

previamente $\gamma_t(i)$ como la probabilidad de estar en el estado S_i en el instante t, dada la secuencia observada; de igual modo podemos establecer una relación entre $\gamma_t(i)$ y $\xi_t(i,j)$ sumando sobre el índice j,

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j).$$

Si sumamos sobre el índice temporal t, obtenemos una cantidad que puede ser interpretada como el número esperado de veces que la cadena está en el estado S_i (véase Sección 2.1), equivalentemente, el número esperado de transiciones realizadas desde el estado S_i (excluyendo el instante $t = T$). De modo análogo, la suma de $\xi_t(i,j)$ sobre t (de $t = 1$ a $t = T-1$) puede ser interpretada como el número esperado de transiciones del estado S_i al estado S_j. Esto es

$$\sum_{t=1}^{T-1} \gamma_t(i) \equiv \text{número esperado de transiciones desde } S_i,$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) \equiv \text{número esperado de transiciones de } S_i \text{ a } S_j.$$

Empleando todo lo previamente descrito se puede proponer un método para la estimación de los parámetros de un modelo de Markov oculto. El conjunto de parámetros de este tipo de modelos puede ser estimado por

$$\hat{\pi}_i = \text{frecuencia esperada del estado } S_i \text{ en tiempo } t = 1 (\gamma_1(i)), \tag{2.5}$$

$$\hat{a}_{ij} = \frac{\#\text{de transiciones de } S_i \text{ a } S_j}{\#\text{ de transiciones desde el estado } S_i} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \tag{2.6}$$

$$\hat{b}_j(k) = \frac{\#\text{ esperado de obs. del símbolo } v_k \text{ en el estado } j}{\#\text{ esperado de veces en el estado } j} = \frac{\sum_{t=1, O_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}. \tag{2.7}$$

De este modo, si el modelo viene dado por $\lambda = (\mathcal{A}, \mathcal{B}, \pi)$, se emplean las fórmulas (2.5), (2.6), (2.7); y definimos el modelo estimado como $\hat{\lambda} = (\hat{\mathcal{A}}, \hat{\mathcal{B}}, \hat{\pi})$; entonces, se demuestra que, o bien el modelo inicial λ define un punto crítico de la función de probabilidad (en cuyo caso $\hat{\lambda} = \lambda$), o bien el modelo $\hat{\lambda}$ es más verosímil que λ en el sentido que $\mathbb{P}(O|\hat{\lambda}) > \mathbb{P}(O|\lambda)$, es decir, hemos encontrado un nuevo modelo $\hat{\lambda}$ del cual es más probable que haya salido la secuencia observada.

Capítulo 3

Imputación genética

En este capítulo abordaremos la primera de las alternativas planteadas para estimar la frecuencia alélica de un determinado SNP (véase Capítulo 1), la imputación genética. Existe una amplia gama de métodos de imputación: fastPHASE[1], IMPUTE2 [2], MaCH[3], BEAGLE[4], PHASE[5], etcétera. De todos los citados, en este trabajo consideraremos únicamente los algoritmos conocidos como BEAGLE y MaCH (ambos programas de libre acceso). Detallaremos el funcionamiento de ambos con vistas a emplearlos ulteriormente en situaciones concretas y comparar su funcionamiento tanto entre ellos como con respecto a métodos de interpolación espacial que se expondrán posteriormente (véase Capítulo 5).

3.1. Imputación genética con Beagle

Este método de estimación de haplotipos se basa en el modelo propuesto por Sharon R. Browning [4] empleado para estudios de asociación (véase Apéndice A). El modelo de conglomerados de haplotipos localizados (localized haplotype-cluster model) modela de forma empírica el desequilibrio de ligamiento (linkage desequilibrium, LD) y se adapta a la estructura local del conjunto de datos. En comparación con otros métodos, se comporta particularmente bien para tamaños de muestra grandes, donde, en un cierto sentido, la muestra está capacitada para hablar "por sí misma". El modelo puede ajustar los haplotipos mediante el uso de una algoritmo bastante rápido, que se utiliza de forma iterativa para la inferencia de la fase de los haplotipos, en la cual se hace una estimación inicial de esta, el modelo se ajusta, se mejoran las estimaciones de la fase de los haplotipos, y así sucesivamente.

[1] fastPHASE: http://stephenslab.uchicago.edu/software.html
[2] IMPUTE2: https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
[3] MaCH: http:/www.sph.umich.edu/csg/abecasis/MACH
[4] BEAGLE: http://faculty.washington.edu/browning/beagle/beagle.html
[5] fastPHASE: http://stephenslab.uchicago.edu/software.html

Comenzamos describiendo el modelo de conglomerados localizados de haplotipos, un caso particular de un grafo dirigido acíclico. Mostraremos, de igual modo, como este modelo de clusters define una cadena de Markov oculta, que puede ser utilizada para muestrear pares de haplotipos o para encontrar el par de haplotipos más probable para cada individuo condicionado a los genotipos individuales.

3.1.1. El modelo de conglomerados de haplotipos

La correlación entre los marcadores es un fenómeno localizado, teniendo en cuenta que el desequilibrio de ligamiento decae con la distancia. Si esta localización no es tomada en consideración a la hora de realizar la estimación de la fase de los haplotipos sobre una región amplia se introduce ruido como consecuencia de la variación de muestreo, resultando en aparentes correlaciones entre marcadores distantes, hecho que reduce la precisión del proceso inferencial. La solución que presenta Beagle al problema previo radica en el uso del modelo de conglomerados de haplotipos localizados, que modeliza empíricamente las frecuencias haplotípicas a escala local. El citado modelo, genera clusters de haplotipos en cada marcador con el objetivo de mejorar la predicción de los alelos en los marcadores $t+1$, $t+2$, ..., dados los alelos en los marcadores t, $t-1$, $t-2$, ... en un haplotipo. Esto se consigue definiendo los conglomerados de acuerdo a la propiedad de Markov, i.e., dado un elemento del cluster en la posición t, la secuencia de alelos en los marcadores $t-1$, $t-2$, ... es irrelevante para predecir la secuencia de alelos en los marcadores $t+1$, $t+2$, ... Los clusters están localizados, de este modo los haplotipos en el mismo cluster en la posición t tienden a estar en el mismo cluster que en la posición $t+1$, aunque no necesariamente. Este modelo posee un número importante de ventajas. En primer lugar, el hecho de agrupar los haplotipos en clusters permite una agilidad computacional al modelo que deriva en una buena estimación de las frecuencias haplotípicas. Permitiendo al número de clusters y a las relaciones entre conglomerados en diferentes posiciones ser determinados por el conjunto de datos en vez de por un modelo restrictivo. Este se adapta a la muestra, lo que resulta de especial utilidad en el caso de que el número de individuos en la muestra sea grande. Por último, el modelo puede ser ajustado utilizando un algoritmo computacionalmente eficiente.

Supongamos que se tiene una muestra de haplotipos procedentes de M marcadores y, además, que los haplotipos no poseen alelos faltantes. Un modelo de conglomerados de haplotipos localizados para esta muestra es un grafo dirigido acíclico verificando:

1. El grafo posee un nodo inicial sin entradas y un nodo terminal sin salidas. El nodo inicial, nodo raíz, representa todos los haplotipos antes de que cualquier marcador sea procesado, mientras el nodo terminal representa todos los haplotipos una vez todos los marcadores han sido procesados.

2. El grafo posee $M+1$ niveles. Cada nodo tiene asociado un nivel m. Todas las aristas entrantes en ese nodo se originan en un nodo con nivel $m-1$ y

todas las salientes tienen como destino un nodo en el nivel $m+1$. El nodo raíz tiene nivel 0 y el terminal nivel M.

3. Para cada $m \in \{1, \ldots, M\}$, cada arista con nodo inicial en el nivel m está asociada con un alelo por el m-ésimo marcador. Dos aristas originadas por el mismo nodo no pueden ser asociadas con el mismo alelo.

4. Para cada haplotipo en la muestra, existe un camino desde el nodo raíz hasta el terminal tal que el m-ésimo alelo del haplotipo está asociado a la m-ésima arista del camino. Cada arista del grafo tiene al menos un haplotipo en la muestra cuyo camino atraviesa la arista.

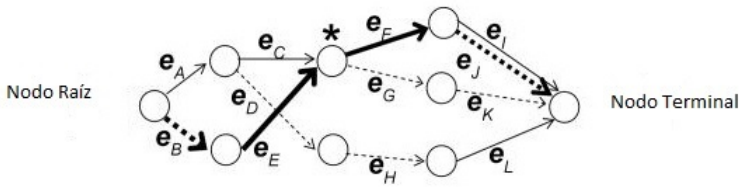

Figura 3.1: Ejemplo de un grafo acíclico dirigido representando el modelo de conglomerados de haplotipos para cuatro marcadores. Para cada marcador, el alelo 1 se representa por una línea continua y el alelo 2 por una discontinua. El nodo marcado con asterisco (*) representa el nodo parental para la arista e_F. Fuente [4]

Cada arista, e, del grafo representa un conglomerado de haplotipos consistente en todos aquellos haplotipos cuyo camino del nodo incial al terminal del grafo atraviesa e. Los haplotipos están definidos a lo largo de todo el cromosoma, pero aquellos haplotipos dentro de un cluster que correspondadan a una arista a nivel m tenderán a poseer patrones alélicos similares en los marcadores contiguos al m-ésimo. De este modo, cada arista define un modelo localizado de conglomerados de haplotipos, que viene determinado por patrones locales de LD (desequilibrio de ligamiento).

Para cada arista, e, de un modelo localizado de conglomerados de haplotipos, se define el recuento de aristas, $n(e)$, como el número de haplotipos en la muestra cuyos caminos atraviesan dicha arista, y definimos el recuento de nodos primarios, $n_p(e)$, como el número de haplotipos en la muestra cuyos caminos atraviesan el nodo inicial de la arista.

Estos modelos están especialmente recomendados para modelizar la estructura de desequilibrio de ligamiento. Asimismo, la recombinación entre haplotipos se modeliza mediante la fusión de las aristas.

3.1.2. El modelo de Markov oculto inducido

El modelo localizado de conglomerados de haplotipos determina una cadena de Markov oculta, para la cual los estados son las aristas del modelo, donde cada estado de la cadena es el alelo que identifica las aristas en el modelo previamente descrito.

Con el objetivo de especificar el modelo, debemos fijar las probabilidades de emisión ($\mathcal{B}$), del estado inicial (π) y las de transición ($\mathcal{A}$). Cada estado (arista) emite con probabilidad 1 el alelo al que está asociado. De este modo el estado queda determinado únicamente por el alelo observado, pero este para un marcador no determina (en general) el estado, puesto que aristas con distinto nodo inicial al mismo nivel en el grafo pueden ser asociadas al mismo alelo.

Las probabilidades iniciales y de transición nacen del recuento de aristas y del recuento de nodo primarios y se calculan como se expone a continuación.

Las probabilidades del estado inicial vienen dadas por

$$\pi_e = \mathbb{P}(e) = \begin{cases} \dfrac{n(e)}{n_p(e)} & \text{si el nodo inicial de la arista es el nodo raíz,} \\ \\ 0 & \text{en otro caso.} \end{cases}$$

Las probabilidades de transición son

$$a_{e_2e_1} = \mathbb{P}(e_1|e_2) = \begin{cases} \dfrac{n(e_1)}{n_p(e_1)} & \text{si el nodo inicial de } e_1 \text{ es el nodo final de } e_2, \\ \\ 0 & \text{en otro caso.} \end{cases}$$

Nótese que, si las aristas e_2 y e_3 tienen el mismo nodo final, entonces $\mathbb{P}(e_1|e_3) = \mathbb{P}(e_1|e_2)$.

Hasta el momento hemos descrito una cadena de Markov oculta haploide; sin embargo, en la práctica necesitaremos una cadena de Markov diploide debido a la observación de genotipos. Crearemos un modelo diploide de pares ordenados de aristas en cada nivel del grafo. Teniendo en cuenta que el grafo está subdivido en niveles, los estados de la cadena de Markov oculta haploide (las aristas del grafo) pueden ser particionadas en clases L_m, donde $m = 1, 2, \ldots, M$ es el nivel del nodo inicial asociado a dicha arista en el grafo. Para una cadena de Markov oculta diploide, el espacio de estados es la unión sobre m de $(L_m \times L_m)$ (si denotamos por S el espacio de estados, este viene dado por $S = \cup_m (L_m \times L_m)$), y el símbolo emitido por cada estado es el par de alelos desordenados que están asociados a los pares de aristas ordenados (estados). Si (e_1, e_2) es un estado del modelo diploide, entonces el genotipo desordenado determinado por los dos alelos asociados a las aristas e_1 y e_2 es emitido con probabilidad 1. Asumiremos en lo sucesivo que se verifica el equilibrio de Hardy-Weinberg (véase Apéndice A), con lo que las probabilidades iniciales y de transición son el producto de las correspondientes probabilidades haploides:

$$\mathbb{P}(e_1, e_2) = \mathbb{P}(e_1)\mathbb{P}(e_2), \quad \mathbb{P}[(e_1, e_2)|(e_3, e_4)] = \mathbb{P}(e_1|e_3)\mathbb{P}(e_2|e_4).$$

3.1.3. Muestreando de un modelo de Markov oculto

Este algoritmo muestrea de la cadena de Markov oculta, condicionado a la muestra observada, mediante el uso de algoritmos "forward-backward" (véase Sección 2.3.3).

Para un individuo dado, sea g_m el genotipo observado en el marcador m, y sea el estado $S_m = (e_1, e_2)$ un par ordenado de aristas en $L_m \times L_m$. Para cada S_m en el modelo de Markov oculto diploide con genotipos individuales $\{g_1, g_2, \ldots, g_M\}$, se definen las variables forward como $\alpha_m(S_m) = \mathbb{P}(g_1, g_2, \ldots, g_m, S_m|\lambda)$. El valor $\alpha_m(S_m)$ se calcula por inducción como sigue.

Paso 1. Inicialización:

$$\alpha_1(S_1) = \mathbb{P}(g_1, S_1) = \mathbb{P}(S_1)\mathbb{P}(g_1|S_1).$$

Paso 2. Inducción:

$$\begin{aligned}\alpha_{m+1}(S_{m+1}) &= \mathbb{P}(g_1, g_2, \ldots, g_{m+1}, S_{m+1}) \\ &= \sum_{S_m} \mathbb{P}(g_1, g_2, \ldots, g_m, g_{m+1}, S_m, S_{m+1}) \\ &= \sum_{S_m} \mathbb{P}(g_1, g_2, \ldots, g_m, S_m)\mathbb{P}(g_{m+1}|S_{m+1})\mathbb{P}(S_{m+1}|S_m) \\ &= \mathbb{P}(g_{m+1}|S_{m+1}) \sum_{S_m} \alpha_m(S_m)\mathbb{P}(S_{m+1}|S_m).\end{aligned}$$

Nótese que las probabilidades $\mathbb{P}(g_{m+1}|S_{m+1})$ son siempre 0 ó 1, dependiendo de si el genotipo está asociado con las etiquetas del par ordenado de aristas. Además se tiene:

$$\mathbb{P}(g_m|S_m) = \begin{cases} 1 & \text{si Caso 1 ó Caso 2,} \\ \\ 0 & \text{en otro caso.} \end{cases}$$

Siendo el Caso 1 $\equiv$ "ambos alelos de g_m son faltantes", y el Caso 2 $\equiv$ "un alelo de g_m es faltante y además el alelo no faltante está etiquetado con 1 en las aristas ordenadas de S_m".

El muestreo de los estados ocultos condicionados a los genotipos individuales se realiza por un procedimiento "backward" por inducción como sigue:

Paso 1. Iniacialización: Se selecciona aleatoriamente el estado S_M con probabilidad proporcional a $\alpha_M(S_M)$.

Paso 2. Inducción: Dados los estados $S_{m+1}, S_{m+2}, \ldots, S_M$, se selecciona el estado S_m con probabilidad

$$\begin{aligned}&\mathbb{P}(S_m|S_{m+1}, s_{m+2}, \ldots, s_M, g_1, g_2, \ldots, g_M) \\ &\quad = \mathbb{P}(S_m|S_{m+1}, g_1, g_2, \ldots, g_{m+1}) \\ &\quad = \mathbb{P}(S_m, S_{m+1}, g_1, g_2, \ldots, g_{m+1})/\alpha_{m+1}(S_{m+1}) \\ &\quad = \mathbb{P}(g_{m+1}|S_{m+1})\mathbb{P}(S_{m+1}|S_m)\alpha_m(S_m)/\alpha_{m+1}(S_{m+1}).\end{aligned}$$

La ruta de la muestra de estados ocultos corresponde a un par ordenado de haplotipos que son consistentes con el genotipo del individuo. Se ha descrito un algoritmo de muestreo para la cadena de Markov oculta diploide. También es posible determinar el par de haplotipos más verosímiles en virtud a la muestra de genotipos y el modelo diploide empleando el algoritmo de Viterbi (véase Sección 2.3.3).

3.1.4. El algoritmo Beagle

El algoritmo Beagle es conceptualmente simple: en cada iteración del algoritmo, los datos de la fase se emplean para construir el modelo de conglomerados de haplotipos localizados como se describe previamente. Una vez este modelo ha sido construido, los haplotipos de la fase para cada individuo se muestrean del modelo oculto de Markov diploide inducido condicionado a los genotipos individuales. Los haplotipos muestreados son nuestro "input" para la siguiente iteración. En la última iteracción, en vez de muestrear los haplotipos, se utiliza el algoritmo de Viterbi para seleccionar los haplotipos más verosímiles para cada individuo, condicionados al modelo de Markov oculto diploide y a los genotipos individuales; estos haplotipos más verosímiles serán nuestro "output" en el algoritmo.

Además, este algoritmo tiene implementadas unas mejoras en el proceso que incrementan la precisión del mismo. En primer lugar, en cada iteracción del algoritmo, se invierte el orden de los marcadores, procesando el cromosoma de izquierda a derecha si el número de la iteracción es par y de derecha a izquierda en caso contrario. En segundo lugar, se muestrean mútiples pares de haplotipos por individuo para emplearse en la construcción del modelo en la siguiente iteracción, tomando en consideración la correlación existente entre pares de haplotipos del mismo individuo; todo lo previamente expuesto deriva en una importante mejora de la precisión cuando se trabaja con un número de individuos bajo.

Una vez el modelo de conglomerados de haplotipos localizados ha sido construido, se muestrean los R pares de haplotipos de fase para cada individuo, condicionado a los genotipos del individuo y al modelo de Markov oculto diploide. Los NR pares de haplotipos muestreados se utilizan como entrada en la siguiente iteracción. La salida gradual de pares de haplotipos para cada individuo es el par de haplotipos más probable condicionado a los genotipos individuales y al HMM diploide en la última iteracción del algoritmo.

3.2. Imputación genética con MaCH

Este método de imputación se basa en el trabajo de Abecasis y Li [8] y proporciona (al igual que Beagle), la estimación de los alelos faltantes de una muestra de genotipos. La sección comienza describiendo el modelo de Markov oculto que el algoritmo utiliza, para posteriormente explicar como se generan los haplotipos de una muestra de genotipos. Continuando con la estimación de los

parámetros del modelo (en la cual, a diferencia de Beagle, sí figura explícitamente la recombinación genética, mutación y el error de genotipado) para finalizar con una breve explicación de como se imputan aleatoriamente los alelos.

3.2.1. El modelo de Markov oculto

Este modelo [8] resuelve un conjunto de genotipos, G, en un mosaico de varios haplotipos plantilla (de referencia). Suponemos que tenemos H haplotipos de referencia genotipados en N *loci* y sea $T_j(i)$ el alelo observado en la posición j en el haplotipo de referencia i. Además definimos una serie de valores $S = \{S_1, S_2, \ldots, S_N\}$ que denotan un hipotético (e inobeservable) conjunto de estados ocultos con respecto a los genotipos. En una posición j, existen H^2 posibles estados. Un estado específico, por ejemplo, $S_j = (x_j, y_j)$ indica que el primer cromosoma usa el haplotipo x_j como plantilla o referencia mientras el segundo cromosoma usa el haplotipo y_j como plantilla.

Estamos interesados hacer inferencia acerca de la secuencia de estados S que mejor describe los genotipos observados. Se define la probabilidad conjunta de los genotipos observados y estados ocultos como:

$$\mathbb{P}(G, S) = \mathbb{P}(S_1) \prod_{j=2}^{N} \mathbb{P}(S_j|S_{j-1}) \prod_{j=1}^{N} \mathbb{P}(G_j|S_j).$$

En el modelo dado, $\pi_1 = \mathbb{P}(S_1)$ denota la probabilidad a priori del estado inicial y se toma habitualmente del mismo modo para cualquier configuración empleada, $a_{j-1j} = \mathbb{P}(S_j|S_{j-1})$ denota la probabilidad de transición entre dos estados y reflejan la probabilidad de los eventos de recombinación en el intervalo entre j y $j-1$, $b_j(j) = \mathbb{P}(G_j|S_j)$ denota la probabilidad de los genotipos observados en cada posición condicionado al mosaico de estados ocultos subyacente y refleja los efectos combinados de la conversión genética, mutaciones y error de genotipado.

3.2.2. Procedimiento de haplotipado por Monte-Carlo

Para estimar haplotipos en una muestra de individuos genotipados se asigna, en primer lugar, un par de haplotipos aleatorios a cada individuo, de acuerdo a los genotipos observados. Este procedimiento involucra la ordenación aleatoria de alelos en cada posición heterocigota y el muestreo de alelos en zonas no genotipadas de acuerdo a las frecuencias poblacionales. Entonces, se actualizan los haplotipos para cada individuo sucesivamente mediante el actual conjunto de haplotipos estimados para todos los individuos como plantillas y se muestrea S proporcional a la verosimilitud $L(S|G) \propto \mathbb{P}(G, S)$. Nótese que, teniendo en cuenta que S_j (junto con la Sección 3.2.3) define una cadena de Markov, este muestreo puede ser realizado convenientemente empleando el algoritmo "forward-backward" de Baum (véase Sección 2.3). A continuación, se define un nuevo conjunto de haplotipos para un individuo de acuerdo con la muestra y

editado para asegurarse de que coincide con los genotipos observados. Finalmente se repite el procedimiento de actualización varias veces sobre todos los individuos (un mayor número de actualizaciones deriva en un refinamiento en la estimación de los haplotipos).

3.2.3. Estimación de los parámetros

Los elementos necesarios en el procedimiento que se expone a continuación son las probabilidades de transición $\mathbb{P}(S_j|S_{j-1})$ y las probabilidades de emisión $\mathbb{P}(G_j|S_j)$. Definimos las probabilidades de transición como una función del parámetro de recombinación θ_j:

$$\mathbb{P}(S_j|S_{j-1}) = \begin{cases} \theta_j^2/H^2 & x_j \neq x_{j-1} \text{ y } y_j \neq y_{j-1}, \\ (1-\theta_j)\theta_j/H + \theta_j^2/H^2 & x_j \neq x_{j-1} \text{ o } y_j \neq y_{j-1}, \\ (1-\theta_j^2) + 2(1-\theta_j)\theta_j/H + \theta_j^2/H^2 & x_j = x_{j-1} \text{ y } y_j = y_{j-1}. \end{cases}$$

Los posibles valores de $\mathbb{P}(S_j|S_{j-1})$ reflejan tanto la tasa general de los cambios en el mosaico para el intervalo, dados por θ_j, como el hecho de que cuando un cambio se produce un nuevo estado se selecciona al azar de entre todos los posibles.

Sea $T(S_j) = T(x_j) + T(y_j)$ el genotipo dado por el estado S_j, definimos las probabilidades de emisión como una función del parámetro de error ε_j:

$$\mathbb{P}(G_j|S_j) = \begin{cases} (1-\varepsilon_j)^2 + \varepsilon_j^2 & \text{si } T(s_j) = G_j \text{ con } G_j \text{ heterocigoto}, \\ 2(1-\varepsilon_j)\varepsilon_j & \text{si } T(s_j) \neq G_j \text{ con } G_j \text{ heterocigoto}, \\ (1-\varepsilon_j)^2 & \text{si } T(s_j) = G_j \text{ con } G_j \text{ homocigoto}, \\ (1-\varepsilon_j)\varepsilon_j & \text{si } T(s_j) \text{ heterocigoto y } G_j \text{ homocigoto}, \\ \varepsilon_j^2 & \text{si } T(s_j) \text{ y } G_j \text{ son homocigotos opuestos}. \end{cases}$$

Inicialmente, se selecciona $\theta_j = \theta = 0.01$ y $\varepsilon_j = \varepsilon = 0.01$. A medida que se crea un nuevo mosaico de estados por individuo mediante muestreo, se toma en consideración tanto del número como de la localización de los puntos de cambio en el mosaico de estados como el número de veces que el genotipo implicado por el mosaico de estados casa o no con el genotipo observado. Estas cantidades se utilizan posteriormente con el objetivo de actualizar los valores de los parámetros θ_j y ε_j para la siguiente iteración. Resulta de especial importancia el hecho de no fijar estas constantes a 0, pues impide a nuestra cadena contemplar diferentes configuraciones del conjunto de estados. Para evitar esto, se estima un parámetro combinado para aquellos intervalos con un número reducido de cambios en el conjunto de estados, un procedimiento análogo se utiliza para marcadores con un número reducido de diferencias entre el mosaico de estados y los genotipos observados.

En general, se espera que el θ_j refleje una combinación de tasas de recombinación de población y la relación entre los haplotipos imputados y los verdaderos haplotipos subyacentes. Se considera el uso de la distancia entre los marcadores flanqueantes para actualizar las estimaciones de θ_j (ya que θ_j son generalmente más grandes en intervalos mayores), pero no se detectan notables mejorías [8]. En general, esperamos que ε_j refleje una combinación de error de genotipado, los eventos de conversión génica, mutación recurrente y, cuando se utiliza el ge-

notipo de datos desde múltiples plataformas o laboratorios, las inconsistencias de ensayo entre las distintas plataformas.

3.2.4. Simulación mediante HMM

En la mayoría de los estudios de secuenciación los genotipos no se observan directamente; en este caso asumiremos que la muestra consiste en recuentos A_j y B_j indicando el número de veces que la base A o B fue observada en la localización j. Definimos de este modo nuestro modelo de Markov oculto:

$$\mathbb{P}(A, B, S) = \mathbb{P}(S_1) \prod_{j=2}^{N} \mathbb{P}(S_j|S_{j-1}) \prod_{j=1}^{N} \left\{ \sum_{G_j} \mathbb{P}(S_j|S_{j-1}) \mathbb{P}(A_j, B_j|G_j) \right\}.$$

Entonces, sumamos sobre todos los posibles genotipos en cada posición y calculamos la probabilidad de los rasgos observados para cada posible conjunto de genotipos. Además, definimos la probabilidad de observar un conjunto de rasgos específicos dado el genotipo subyacente como

$$\mathbb{P}(A_j, B_j|G_j) = \begin{cases} Bi(A_j, A_j + B_j, 1-\delta), & G_j = A|A, \\ Bi(A_j, A_j + B_j, \frac{1}{2}), & G_j = A|B, \\ Bi(A_j, A_j + B_j, \delta), & G_j = B|B. \end{cases}$$

El parámetro δ denota la tasa de error de secuenciación por base y puede ser separada de los efectos de la mutación y la conversión genética que sí engloba el término ε.

En lo relativo a la eficiencia del algoritmo, es posible mejorar la eficiencia computacional de nuestro modelo, por ejemplo, teniendo en cuenta que los estados están desordenados se pueden considerar $H(H+1)/2$ estados distintos en cada localización, en vez de H^2 estados.

3.3. Consideraciones generales

Una cadena de Markov oculta posee estados ocultos subyacentes que no son observados directamente (véase Capítulo 2). En la fase de estimación de los haplotipos, estos estados representan de algún modo los genotipos reales subyacentes. Las probabilidades de transición determinan el modo en que los estados ocultos ven modificada su posición a lo largo del cromosoma, y las probabilidades de emisión relacionan los estados ocultos con la muestra observada.

En MaCH, los estados ocultos son plantillas de haplotipos. Estas plantillas son haplotipos estimados de la muestra. Durante cada iteración del proceso de estimación, cada haplotipo individual se estima tomando en consideración plantillas de haplotipos construidas con los restantes individuos. A medida que se suceden las iteraciones la estimación de los haplotipos mejora. MaCH emplea subconjuntos aleatorios de los haplotipos como plantilla. Además, como parte del proceso de estimación este software también estima las probabilidades de transición de los estados ocultos (esencialmente tasas de recombinación) y las de emisión (representando las tasas de mutación).

BEAGLE forma el modelo de Markov oculto mediante clusters locales de haplotipos en cada posición de un marcador a lo largo del cromosoma. Los haplotipos se clusterizan de tal forma que aquellos que constituyan un mismo conglomerado tienden a tener probabilidades similares para los alelos en marcadores intermedios. Los conglomerados de haplotipos conforman en este modelo los estados ocultos de la cadena. En cada iteración del algoritmo, se muestrean nuevos haplotipos estimados del estado actual de la cadena de Markov oculta, condicionado a los genotipos individuales; siendo estos haplotipos estimados empleados en la posterior construcción de un nuevo modelo.

El modelo propuesto en el software BEAGLE es, sin embargo, escaso en algunos sentidos; en primer lugar, el hecho de agrupar los haplotipos en conglomerados mantiene el número de estados ocultos relativamente bajo. En segundo lugar, el modelo considera únicamente un conjunto pequeño de transiciones entre los estados en una posición y la siguiente; mientras que MACH permite más flexibilidad a este respecto considerando todo el abanico de transiciones entre estados de una posición a la siguiente. La aproximación de BEAGLE permite que distintas posiciones posean un número de clusters diferente.

Por último, al contrario que MaCH, BEAGLE está basado en un HMM que no captura explícitamente la recombinación y mutación génica (aunque sí de forma implícita). El modelo de conglomerados de haplotipos de BEAGLE se adapta a la cantidad de información presente de tal modo que el número de clusters aumenta globalmente con el tamaño muestral y localmente con el aumento del desequilibrio de ligamiento. Para más información a este respecto véase [5].

Capítulo 4

Resultados de la imputación genética

4.1. Introducción

Ilustraremos en este momento los métodos expuestos en el Capítulo 3 con el objetivo de calcular las frecuencias alélicas de tres SNPs: *rs2015035*, *rs6006405* y *rs5749090* (cuyas frecuencias del alelo menor, MAF, son del 5 %, 15 % y 30 % respectivamente). Mediante una muestra de 379 individuos pertenecientes a la comunidad autónoma gallega (obtenidos del consorcio EPICOLON [1]) y genotipados con el Array 6.0 de Affymetrix, seleccionaremos muestras de tamaños $N = 100$, 200 y 379 con el objetivo de ejecutar los algoritmos Beagle y MaCH para distintos tamaños de una misma muestra. De igual modo, dentro de cada uno de esos tamaños de muestra (número de individuos) ejecutaremos Beagle y MaCH para distintos tamaños de ventana, i.e., centrándonos en la posición física del SNP, abriremos ventanas de tamaño $50Kb$ $100Kb$ y $500Kb$. Por ejemplo, para la ventana de $50Kb$, y denotando por x la posición física del SNP, la muestra seleccionada sería $x \pm 25000$ pares de bases. Por último, para cada uno de estos tamaños muestrales y para cada tamaño de ventana, crearemos mediante muestreo aleatorio simple muestras de diferente densidad: densidad baja (1 SNP por cada $10Kb$) y densidad alta (1 SNP por cada $3Kb$). Nótese que en el caso que nos ocupa, la densidad alta equivale a seleccionar la muestra completa de SNPs que caigan en esa ventana particular. Cabe destacar que, en las muestras previamente descritas, el SNP que nos interesa imputar en cada escenario fue eliminado de la muestra.

En cada uno de los escenarios descritos, ejecutaremos los softwares Beagle y MaCH con el objetivo de imputar cada uno de los citados SNPs del cromosoma 22 para cada uno de los tamaños de muestra dados, las ventanas y las densidades seleccionadas.

Supongamos un cierto SNP que posee como alelos A y G, denotaremos en lo sucesivo por $f_A \equiv$ frecuencia real del alelo A en el SNP, $f_G \equiv$ frecuencia real

del alelo G y por $\hat{f}_A \equiv$ frecuencia estimada del alelo A, análogamente $\hat{f}_G \equiv$ frecuencia estimada del alelo G. En caso de que no figure un subíndice, se interpretará como la frecuencia del alelo mayor (bien sea real o estimada).

En el caso que nos ocupa, los valores de las frecuencias reales se muestran en el Cuadro 4.1:

f	N=100	N=200	N=379
rs2015035	0.9800	0.9675	0.9433
rs6006405	0.8500	0.8650	0.8588
rs5749090	0.6800	0.6800	0.7005

Cuadro 4.1: Frecuencia real del alelo mayor (G para el SNP rs2015035, A para el SNP rs6006405 y A para el SNP rs5749090) para cada tamaño de muestra ($N = 100, 200, 379$).

Nótese que esta estimación de frecuencias se realiza como sigue; en primer lugar, el algoritmo en cuestión imputa todos los alelos faltantes del SNP, para posteriormente, mediante un recuento, calcular la frecuencia alélica estimada en ese SNP.

4.2. Resultados de imputación con Beagle

Se adjuntan a continuación los resultados realtivos a la estimación realizada con Beagle (Cuadros 4.2, 4.3 y 4.4) de la frecuencias del alelo mayor (G, A y A) y del alelo menor (T, G y G) para los SNPs *rs2015035*, *rs6006405* y *rs5749090* respectivamente.

SNP rs2015035

N=100	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_G$	1.0000	1.0000	1.0000	1.0000	0.9700	0.9850
$\hat{f}_T$	0.0000	0.0000	0.0000	0.0000	0.0300	0.0150
$\lvert f_G - \hat{f}_G \rvert$	0.0200	0.0200	0.0200	0.0200	0.0100	0.0050
N=200	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_G$	1.0000	1.0000	1.0000	1.0000	0.9750	0.9750
$\hat{f}_T$	0.0000	0.0000	0.0000	0.0000	0.0250	0.0250
$\lvert \hat{f}_G - f_G \rvert$	0.0325	0.0325	0.0325	0.0325	0.0075	0.0075
N=379	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_G$	1.0000	1.0000	1.0000	1.0000	0.9578	0.9617
$\hat{f}_T$	0.0000	0.0000	0.0000	0.0000	0.0422	0.0383
$\lvert \hat{f}_G - f_G \rvert$	0.0567	0.0567	0.0567	0.0567	0.0145	0.0185

Cuadro 4.2: Aproximación de las frecuencias del alelo mayor y menor, diferencia en valor absoluto entre la frecuencia real y la estimada, para el SNP *rs2015035* mediante el software Beagle. Tamaños de muestra empleados N=100,200,379; se utilizan ventanas de $50Kb$, $100Kb$ y $500Kb$, para densidades alta y baja (p.e., 100.baja hace referencia a que la muestra imputada utilizaba una ventana de 100Kb. alrededor del citado SNP con una densidad baja).

SNP rs6006405

N=100	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.9450	0.8550	0.8550	0.8650	0.8350	0.8400
$\hat{f}_G$	0.0550	0.1450	0.1450	0.1350	0.1650	0.1600
$\lvert\hat{f}_A - f_A\rvert$	0.0950	0.0050	0.0050	0.0150	0.0150	0.0100
N=200	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.8800	0.8700	0.8250	0.8625	0.8425	0.8750
$\hat{f}_G$	0.1200	0.1300	0.1750	0.1375	0.1575	0.1250
$\lvert\hat{f}_A - f_A\rvert$	0.0150	0.0050	0.0400	0.0025	0.0225	0.0100
N=379	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.8839	0.8760	0.8496	0.8720	0.8707	0.8641
$\hat{f}_G$	0.1161	0.1240	0.1504	0.1280	0.1293	0.1359
$\lvert\hat{f}_A - f_A\rvert$	0.0251	0.0172	0.0092	0.0132	0.0119	0.0053

Cuadro 4.3: Aproximación de las frecuencias del alelo mayor y menor, diferencia en valor absoluto entre la frecuencia real y la estimada, para el SNP *rs6006405* mediante el software Beagle. Tamaños de muestra empleados N=100,200,379; se utilizan ventanas de $50Kb$, $100Kb$ y $500Kb$, para densidades alta y baja.

SNP rs5749090

N=100	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.7200	0.8550	0.6900	0.7000	0.6800	0.6850
$\hat{f}_G$	0.2800	0.1450	0.3100	0.3000	0.3200	0.3150
$\lvert\hat{f}_A - f_A\rvert$	0.0400	0.1750	0.0100	0.0200	0.0000	0.0050
N=200	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.7600	0.8700	0.6900	0.6950	0.6550	0.6850
$\hat{f}_G$	0.2400	0.1300	0.3100	0.3050	0.3450	0.3150
$\lvert\hat{f}_A - f_A\rvert$	0.0800	0.1900	0.0100	0.0150	0.0250	0.0050
N=379	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.7375	0.8760	0.7230	0.7203	0.7137	0.7018
$\hat{f}_G$	0.2625	0.1240	0.2770	0.2797	0.2863	0.2982
$\lvert\hat{f}_A - f_A\rvert$	0.0369	0.1755	0.0224	0.0198	0.0132	0.0013

Cuadro 4.4: Aproximación de las frecuencias del alelo mayor y menor, diferencia en valor absoluto entre la frecuencia real y la estimada, para el SNP *rs5749090* mediante el software Beagle. Tamaños de muestra empleados N=100,200,379; se utilizan ventanas de $50Kb$, $100Kb$ y $500Kb$, para densidades alta y baja.

Cabe destacar, a la vista de los Cuadros 4.2, 4.3 y 4.4 el hecho de que el tamaño de la ventana (el número de SNPs involucrados en el proceso de imputación) influye sobre la precisión del método a la hora de aproximar las frecuencias alélicas, i.e., a mayor ventana, mayor precisión de $\hat{f}$ como estimador de f, este hecho se ve reflejado en que los valores de $|\hat{f}_X - f_X|$ $(X = A, G)$ disminuyen a medida que aumenta el tamaño de la ventana; si bien en algún caso la densidad de muestra que se tome (alta o baja) puede influír en la estimación que Beagle nos proporcione (esto se debe a que, de entre las muestras de densidad alta, las de densidad baja fueron seleccionadas aleatoriamente como un subconjunto de las primeras).

4.3. Resultados de imputación con MaCH

SNP rs2015035

N=100	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_G$	1.0000	1.0000	1.0000	1.0000	0.9850	0.9850
$\hat{f}_T$	0.0000	0.0000	0.0000	0.0000	0.0150	0.0150
$\lvert\hat{f}_G - f_G\rvert$	0.0200	0.0200	0.0200	0.0200	0.0050	0.0050
N=200	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_G$	1.0000	1.0000	1.0000	1.0000	0.9725	0.9750
$\hat{f}_T$	0.0000	0.0000	0.0000	0.0000	0.0275	0.0250
$\lvert\hat{f}_G - f_G\rvert$	0.0325	0.0325	0.0325	0.0325	0.0050	0.0075
N=379	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_G$	1.0000	1.0000	1.0000	1.0000	0.9617	0.9631
$\hat{f}_T$	0.0000	0.0000	0.0000	0.0000	0.0383	0.0369
$\lvert\hat{f}_G - f_G\rvert$	0.0567	0.0567	0.0567	0.0567	0.0185	0.0198

Cuadro 4.5: Aproximación de las frecuencias del alelo mayor y menor, diferencia en valor absoluto entre la frecuencia real y la estimada, para el SNP *rs2015035* mediante el software MaCH. Tamaños de muestra empleados N=100,200,379; se utilizan ventanas de $50Kb$, $100Kb$ y $500Kb$, para densidades alta y baja.

SNP rs6006405

N=100	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.9450	0.8700	0.8600	0.8750	0.8850	0.8550
$\hat{f}_G$	0.0550	0.1300	0.1400	0.1250	0.1150	0.1450
$\lvert\hat{f}_A - f_A\rvert$	0.0950	0.0200	0.0100	0.0250	0.0350	0.0050
N=200	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.9325	0.8750	0.9275	0.8800	0.8650	0.8700
$\hat{f}_G$	0.0675	0.1250	0.0725	0.1200	0.1350	0.1300
$\lvert\hat{f}_A - f_A\rvert$	0.0675	0.0100	0.0625	0.0150	0.0000	0.0050
N=379	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.8799	0.8747	0.9591	0.8826	0.8905	0.8707
$\hat{f}_G$	0.1201	0.1253	0.0409	0.1174	0.1095	0.1293
$\lvert\hat{f}_A - f_A\rvert$	0.0211	0.0158	0.1003	0.0237	0.0317	0.0119

Cuadro 4.6: Aproximación de las frecuencias del alelo mayor y menor, diferencia en valor absoluto entre la frecuencia real y la estimada, para el SNP *rs6006405* mediante el software MaCH. Tamaños de muestra empleados N=100,200,379; se utilizan ventanas de $50Kb$, $100Kb$ y $500Kb$, para densidades alta y baja.

SNP rs5749090

N=100	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.7200	0.7000	0.7050	0.7000	0.6950	0.6750
$\hat{f}_G$	0.2800	0.3000	0.2950	0.3000	0.3050	0.3250
$\lvert\hat{f}_A - f_A\rvert$	0.0400	0.0200	0.0250	0.0200	0.0150	0.0050
N=200	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.7300	0.6975	0.6975	0.6925	0.6850	0.6825
$\hat{f}_G$	0.2700	0.3025	0.3025	0.3075	0.3150	0.3175
$\lvert\hat{f}_A - f_A\rvert$	0.0500	0.0175	0.0175	0.0125	0.0050	0.0025
N=379	50.baja	50.alta	100.baja	100.alta	500.baja	500.alta
$\hat{f}_A$	0.7414	0.7203	0.7216	0.7098	0.7018	0.6979
$\hat{f}_G$	0.2586	0.2797	0.2784	0.2902	0.2982	0.3021
$\lvert\hat{f}_A - f_A\rvert$	0.0409	0.0198	0.0211	0.0092	0.0013	0.0026

Cuadro 4.7: Aproximación de las frecuencias del alelo mayor y menor, diferencia en valor absoluto entre la frecuencia real y la estimada, para el SNP *rs6006405* mediante el software MaCH. Tamaños de muestra empleados N=100,200,379; se utilizan ventanas de $50Kb$, $100Kb$ y $500Kb$, para densidades alta y baja.

El software MaCH (Cuadros 4.5, 4.6 y 4.7) presenta un comportamiento similar a Beagle en las condiciones empleadas puesto que, al igual que en Beagle (véase Sección 4.2) el tamaño de la ventana afina la precisión del método al aumentar y la disminuye al reducirse este. Al igual que sucede con Beagle, para esta muestra, el tamaño muestral (número de individuos) no parece influir demasiado en la precisión de la estimación, si bien sería necesario tener una muestra de mayor tamaño a la dada para poder ser más concluyentes en este sentido.

A modo de comparativa, para el SNP *rs2015035*, los resultados de frecuencias estimadas por imputación na varían prácticamente de un método a otro (Cuadros 4.2 y 4.5), resultando sus diferencias (en valor absoluto) frente a la frecuencias reales similares. A este respecto, las frecuencias alélicas imputadas no se estabilizan alrededor a las frecuencias reales hasta valores de N altos (200 y 379) y valores altos de la ventana seleccionada. Este hecho puede deberse a estar trabajando con un SNP de baja frecuencia.

En cambio para el SNP *rs6006405*, y a pesar de no apreciarse diferencias relevantes en las estimaciones dadas por Beagle y MaCH (Cuadros 4.3 y 4.6 respectivamente) para $N = 100$, a medida que este parámetro aumenta, Beagle afina más a la hora de estimar esta frecuencia alélica que MaCH.

Por último, en lo relativo al SNP *rs5749090*, ambas estimaciones producen resultados satisfactorios para los tamaños de muestra empleados (Cuadros 4.4 y 4.7), si bien Beagle es capaz de precisar algo más sobre todo para un número de SNPs alto (ventana grande).

Capítulo 5

Estadística Espacial

La estadística espacial estudia fenómenos aleatorios $Z = \{Z(s), s \in D\}$ indexados por un conjunto espacial D. La localización $s \in D$ puede ser determinista (geoestadística por ejemplo) o aleatorio (proceso puntual). Clásicamente, $D \subseteq \mathbb{R}^2$, pero se puede dar $D \subseteq \mathbb{R}$ o también $D \subseteq \mathbb{R}^3$. Existen tres tipos de datos espaciales :

1. los datos geoestadísticos : $D \subseteq \mathbb{R}^2$ es un subconjunto continuo, Z posee valores reales observados sobre un conjunto finito de puntos $\mathcal{O} = \{s_1, \ldots, s_n\}$ (bien sea regular este o no). La geoestadística se preocupa de la identificación y estimación del modelo, para posteriormente realizar predicción mediante *kriging* en las localizaciones no observadas;

2. los datos sobre una red (reticulares o discretos) : D es un subconjunto discreto, Z posee valores reales observados sobre un conjunto finito de puntos $\mathcal{O} = \{s_1, \ldots, s_n\}$. Las localizaciones representan unidades geográficas o administrativas, y las observaciones del proceso son datos agregados (por ejemplo, número de enfermos por región);

3. los datos puntuales : en este contexto, no sólo las observaciones del proceso son medidas aleatorias, sino que también lo son las localizaciones. Por ejemplo, si se desea observar alguna característica en árboles de un bosque, su posición no viene determinada por el experimentador, y puede presentar un patrón (aleatorio) de interés.

La geoestadística es un término que se acuñó en los años 50 para denominar a las técnicas estadísticas aplicadas al análisis geográfico. Su desarrollo, en esa década y en la siguiente, se debe a su aplicación en la ingeniería de minas, para predecir las reservas de mineral a partir de observaciones espacialmente distribuidas en una región.

Existen una gran variedad de problemas que pueden resolverse utilizando métodos geoestadísticos. La característica común a todos ellos es que los datos pueden verse como una realización, habitualmente parcial, de un proceso estocástico sobre una región espacial continua.

La clave fundamental en la modelización de la relación espacial en el proceso es el variograma que será objeto de modelización y estimación para describir adecuadamente el fenómeno observado. El objetivo principal en la aplicación de la geoestadística es habitualmente la predicción en un punto o en un conjunto de puntos de la región observada. La técnica de predicción espacial más empleada es el *kriging* [10].

En el caso que nos ocupa, y mediante las técnicas introducidas en este capítulo, el objetivo último es predecir la frecuencia alélica de los tres mismos SNPs que en el Capítulo 4 en Galicia, conociendo valores de la frecuencia de esos mismos SNPs en $N = 33$ puntos distintos de España.

5.1. Introducción a los procesos estocásticos espaciales

La formulación básica de un proceso estocástico se concreta a la situación espacial tomando como conjunto de índices una determinada región continua D del espacio,

$$\{Z(s) : s \in D \subset \mathbb{R}^2\}.$$

La principal característica de interés para el estudio espacial es la función de covarianza, que determina, para cada par de puntos, la covarianza entre las variables aleatorias correspondientes,

$$\mathrm{Cov}(Z(s_1), Z(s_2))$$

Cabe destacar que, la predicción es posible si el proceso tiene, en algún aspecto, un comportamiento estable en toda la región de estudio. Esta condición de estabilidad requerida se denomina estacionariedad.

La *estacionariedad estricta* es una condición muy fuerte y poco habitual, pues establece que las distribuciones de probabilidad conjunta permanezcan invariantes ante una traslación. Una condición menos exigente es la *estacionariedad de segundo orden*, o *estacionariedad débil*, que conlleva que la esperanza sea constante y que la función de covarianza sea invariante por traslación,

$$\mathbb{E}(Z(s)) = \mu, \quad \forall s \in D;$$

$$\mathrm{Cov}(Z(s_1), Z(s_2)) = C(s_1 - s_2), \quad \forall s_1, s_2 \in D.$$

De este modo, la función de covarianza de un proceso estacionario se puede expresar en función del vector de diferencia entre los puntos. A la función $C(\cdot)$ se le denomina covariograma. De igual modo, se define el correlograma, o función de autocorrelación, que para cada vector proporciona la correlación entre las variables de dos puntos separados por ese vector.

Una perspectiva diferente de la estacionariedad se obtiene al estudiar la variabilidad de los incrementos del proceso. La propiedad de *estacionariedad*

intínseca se verifica si la varianza de las diferencias entre las variables en dos puntos depende únicamente del vector que los separa (vector diferencia),

$$\mathrm{Var}(Z(s_1) - Z(s_2)) = 2\gamma(s_1 - s_2), \quad \forall s_1, s_2 \in D.$$

Esta condición es más débil que la estacionariedad de segundo orden y se emplea habitualmente en la modelización geoestadística. De este modo, se define el *variograma* como la función 2γ de dicho vector diferencia $(s_1 - s_2)$. A la función γ se le denomina *semivariograma*.

Por otro lado, un proceso intrínsecamente estacionario es *isotrópico* si el variograma depende del vector a través de su longitud $h = ||s_1 - s_2||$, sin importar la dirección. Se denomina *proceso homogéneo* a un proceso intrínsecamente estacionario e isotrópico.

Para realizar una predicción de un proceso intrínsecamente estacionario es conveniente modelizar su variograma utilizando un variograma válido, en el sentido de que cumpla la propiedad de ser condicionalmente definido negativo[1] (esta propiedad es relevante en el proceso de estimación, ya que debemos garantizar que los estimadores del variograma también la verifican), se suele emplear el semivariograma. Se pueden diferenciar varios elementos en el semivariograma: la pepita, el umbral y el rango.

- Se denomina *efecto pepita* (nugget en inglés), término extraído de la aplicación a la minería, a la situación en que el variograma no tiende a 0 al acercarse al origen. Esto puede ser debido al error de medida o a la variación a muy pequeña escala,

$$\lim_{h \to 0} \gamma(h) = c_0 > 0.$$

- De forma lógica, un semivariograma crece con la distancia, recogiendo el fenómeno de que el proceso es similar en puntos próximos, hasta que se estabiliza en un valor llamado umbral, que expresa la variabilidad entre puntos distantes,

$$\lim_{h \to \infty} \gamma(h) = c_s > 0.$$

- El rango es la distancia h_s a la que se alcanza el umbral[2], $\gamma(h) = c_s, \quad \forall h > h_s$.

Entre los muchos modelos isotrópicos de semivariograma que se han propuesto, los más empleados son el lineal, esférico, exponencial, cuadrático racional, ondulado, potencial y Gaussiano (para más detalles a este respecto véase [10]). Estos constituyen una amplia batería representativa de diferentes comportamientos de los procesos espaciales.

[1] γ es condicionalmente definido negativo si:
$\forall a \in \mathbb{R}^n$ verificando $\sum_{i=1}^{n} a_i = 0$ entonces $\sum_{i=1}^{n} a_i a_j \gamma(s_i - s_j) \leq 0, \quad \forall \{s_1, \ldots, s_n\} \subseteq D.$

[2] En algunos modelos, c_s es un valor que se alcanza asintóticamente, por lo que se define el rango como el punto en el que el variograma alcanza el 95 % del valor de c_s.

5.2. Estimación del variograma

La estimación del variograma debe relizarse en un contexto en el cual nuestro proceso estocástico espacial $\{Z(s) : s \in D \subset \mathbb{R}^2\}$ sea estacionario (intrínsecamente), i.e., la varianza de la diferencia entre dos observaciones sea función del vector diferencia (el variograma, γ). Asimismo el proceso estocástico debe verificar la propiedad de isotropía, i.e., el variograma depende del vector tan sólo a través de su longitud, independientemente de la dirección.

5.2.1. Estimación empírica del variograma

La estimación más sencilla del variograma puede obtenerse teniendo en cuenta que el variograma puede escribirse como la varianza del proceso diferencia, proporcionando para cada vector su estimador mediante la varianza muestral de la diferencia del proceso entre los pares de puntos separados por la longitud de ese vector (método de los momentos);

$$2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{N(h)} (Z(s_i) - Z(s_j))^2,$$

siendo

$$N(h) = \{(s_i, s_j) : ||s_i - s_j|| = h; i, j = 1, \ldots, n\}.$$

En la práctica, la estimación se realiza permitiendo cierta región de tolerancia alrededor del valor h. Las regiones de tolerancia deben ser tan pequeñas como se pueda, pero con el número de pares suficiente para ejecutar una estimación estable.

Una objeción a este estimador del variograma es su inestabilidad ante la presencia de valores extremos, por lo que se han propuesto diferentes estimadores robustos mediante la introducción de un factor corrector del sesgo o el uso de la mediana;

$$2\hat{\gamma}(h) = \frac{\left\{ \frac{1}{N(h)} \sum\limits_{N(h)} |Z(s_i) - Z(s_j)|^{\frac{1}{2}} \right\}^4}{0.457 + \frac{0.494}{N(h)}}, \tag{5.1}$$

el estimador dado en la ecuación (5.1) se conoce como estimador robusto de Cressie-Hawkins. También existen métodos no paramétricos para realizar la estimación de esta función (estimadores suavizados) [10].

5.2.2. Estimación paramétrica de modelos de variograma

Las estimaciones dadas en la Sección 5.2.1 para el variograma no pueden ser usadas directamente para la predicción espacial, pues no son necesariamente

condicionalmente definidas negativas. De este modo debemos buscar un modelo válido para el semivariograma que se aproxime a la dependencia espacial encontrada por el semivariograma empírico, seleccionando, de alguna familia paramétrica válida, aquella que mejor describa el comportamiento observado. La estimación de los parámetros puede realizarse por diferentes métodos como los de máxima verosimilitud, máxima verosimilitud restringida, mínima norma cuadrática, mínimos cuadrados y mínimos cuadrados generalizados (algunas válidas o no dependiendo de la normalidad del proceso en cuestión). Exponemos brevemente el método de mínimos cuadrados por ser el que se utiliza en el Capítulo 6.

Método de mínimos cuadrados.

Sea $\{Z(s_1), \ldots, Z(s_n)\}$ una realización del proceso $\{Z(s) : s \in D \subset \mathbb{R}^2\}$, y sea $\gamma_\theta(u)$ un modelo paramétrico de (semi)variograma, siendo θ una colección de parámetros que captan la dependencia espacial. Sea $\{u_1, \ldots, u_k\}$ un conjunto de distancias. De la muestra $\{Z(s_1), \ldots, Z(s_n)\}$ se obtiene un estimador piloto, $\hat{\gamma}$ (empírico, robusto, suavizado etc.). El estimador de mínimos cuadrados de θ se obtiene como sigue

$$\hat{\theta} = \arg\min_{\theta} \sum_{l=1}^{k} (\hat{\gamma}(u_l) - \gamma_\theta(u_l))^2.$$

De esta expresión surge la de mínimos cuadrados generalizados:

$$\hat{\theta} = \arg\min_{\theta} (\hat{\gamma} - \gamma_\theta) \Sigma_{\hat{\gamma}}^{-1} (\hat{\gamma} - \gamma_\theta),$$

donde $\Sigma_{\hat{\gamma}}$ es la matriz de covarianzas del variograma estimado en las distancias $u_1, \ldots, u_k$. De igual modo se obtiene la estimación de mínimos cuadrados ponderados (weighted least squares), que sustiye en la expresión previa el valor $\Sigma_{\hat{\gamma}}(l,l) \approx \dfrac{\gamma_\theta^2(u_l)}{|N(u_l)|}$ (la justificación puede encontrarse en [10]).

5.2.3. Kriging

Kriging Ordinario

El método de predicción espacial más extendido es el *kriging*, término acuñado en honor al trabajo del ingeniero de minas D.G. Krige, que consiste en la predicción espacial óptima (mejor predicción lineal insesgada) empleando un modelo de semivariograma para recoger la estructura de segundo orden del proceso.

El denominado *kriging ordinario* consiste en la predicción lineal insesgada óptima, considerando que el proceso se puede descomponer en la suma de una valor medio fijo y un proceso intrínsecamente estacionario,

$$Z(s) = \mu + \varepsilon(s)$$

con el semivariograma $\gamma(h)$ conocido.

El predictor lineal del proceso en un punto arbitrario s_0 es $p(Z;s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$, donde exigiremos $\sum_{i=1}^{n} \lambda_i = 1$ para que sea insesgado. Existe una versión de kriging denominado *kriging simple* en la que μ es conocida y los coeficientes λ_i no están restringidos a sumar 1.

El kriging consiste en la determinación del mejor de estos predictores en el sentido de que minimice el error cuadrático medio de predicción,

$$\min_{p} \mathbb{E}[Z(s_0) - p(Z;s_0)].$$

Este predictor se obtiene a través de la resolución del sistema de ecuaciones de predicción resultantes de la minimización del error cuadrático medio. Puede ser expresado como

$$p_k(Z,s_0) = \left(\gamma + \mathbf{1}\frac{(1-\mathbf{1}'\Gamma^{-1}\gamma)}{\mathbf{1}'\Gamma^{-1}\mathbf{1}}\right)' \Gamma Z$$

donde $\gamma = (\gamma(s_1 - s_0), \ldots, \gamma(s_n - s_0))'$ y Γ es la matriz $n \times n$ cuyo elemento (i,j) es $\gamma(s_i - s_j)$.

La varianza de predicción puede expresarse como

$$\sigma_k^2(s_0) = \gamma'\Gamma^{-1}\gamma - (\mathbf{1}'\Gamma^{-1}\gamma - 1)^2/(\mathbf{1}'\Gamma\mathbf{1}).$$

A partir de las expresiones anteriores (y siempre que el proceso estocástico sea Gaussiano), se pueden construir *intervalos de predicción* al $100(1-\alpha)\,\%$ mediante

$$p_k(Z,s_0) \pm z_{1-\alpha/2}\sigma_k(s_0)$$

siendo $z_{1-\alpha/2}$ los cuantiles de la normal estandarizada.

Kriging Universal

El *kriging universal* generaliza el kriging ordinario, permitiendo que el valor medio del proceso no sea constante, sino una combinación lineal de funciones conocidas o covariables ligadas a las mismas localizaciones. De esta forma, el kriging universal incorpora términos de regresión y correlación espacial,

$$Z(s) = \beta_0 + \beta_1 f_1(s) + \ldots + \beta_p f_p(s) + \varepsilon(s),$$

donde las $f_j(\cdot)$ son funciones de localización espacial s o variables explicativas asociadas a los puntos.

El vector de datos Z puede escribirse como

$$Z = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon,$$

donde $\boldsymbol{X}$ es la matriz $n \times (p+1)$ cuyo elemento (i,j) es $f_{j-1}(s_i)$.

El predictor lineal insesgado en un punto arbitrario s_0 es $p(Z;s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$, sujeto a las restricciones $\boldsymbol{\lambda}'\boldsymbol{X} = \boldsymbol{x}'$, con el objetivo de garantizar el sesgo nulo del estimador, siendo $\boldsymbol{x} = (f_0(s_0), f_1(s_0) \ldots, f_p(s_0))'$.

La predicción óptima, que minimiza el error cuadrático medio, se realiza de forma similar al caso anterior añadiendo tantos coeficientes como términos de regresión aparecen en la media. La expresión del estimador resultante es

$$p_k(Z;s_0) = \{\gamma + \boldsymbol{X}(\boldsymbol{X}'\Gamma^{-1})\boldsymbol{X})^{-1}(\boldsymbol{x} - \boldsymbol{X}'\Gamma^{-1}\gamma)\}'\Gamma^{-1}Z.$$

La varianza de predicción puede expresarse como

$$\sigma_k^2(s_0) = \gamma'\Gamma^{-1}\gamma - (\boldsymbol{x} - \boldsymbol{X}'\Gamma^{-1}\gamma)'(\boldsymbol{X}'\Gamma^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x} - \boldsymbol{X}'\Gamma^{-1}\gamma),$$

siendo en este caso el intervalo de predicción al $100(1-\alpha)\,\%$

$$p_k(Z;s_0) \pm z_{1-\alpha/2}\sigma_k(s_0).$$

La estimación de los parámetros de la media se obtiene por mínimos cuadrados generalizados, asumiendo que los datos Z satisfacen un modelo lineal general con $\mathbb{E}(Z) = \boldsymbol{X}\boldsymbol{\beta}$ y $Var(Z) = \boldsymbol{\Sigma}$, siendo Σ la matriz de covarianzas del proceso,

$$\hat{\boldsymbol{\beta}}_{gls} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}Z.$$

Nótese que, puesto que Σ es desconocida, se emplea una estimación de la misma (de hecho, el variograma y el covariograma se relacionan a través de una sencilla expresión, por lo que podemos obtener Σ a través de $\Sigma_{\hat{\gamma}}$).

Capítulo 6

Resultados de la interpolación espacial

En virtud a lo expuesto en el Capítulo 5 realizaremos una predicción de las frecuencias alélicas de los SNPs *rs2015035*, *rs6006405* y *rs5749090* en Galicia, conociendo las frecuencias alélicas (en particular, poseemos las frecuencias del alelo menor en cada SNP) en $N = 33$ clusters distribuidos a lo largo de España (salvo Galicia). Estos resultados se obtienen con vistas a ser comparados con los obtenidos en el Capítulo 4. Todo este proceso se realizará mediante el software estadístico R-project empleando las librerías `geoR` y `sm`. Se adjunta a continuación la distribución de los clusters a lo largo de España, véase Figura 6.1.

Figura 6.1: Distribución espacial de las observaciones de frecuencias de los SNPs a lo largo de España (en color rojo).

A modo de paso previo al análisis espacial de los datos de frecuencias, se

adjunta la estimación de la función de densidad de las frecuencias (del alelo menor) para cada SNP:

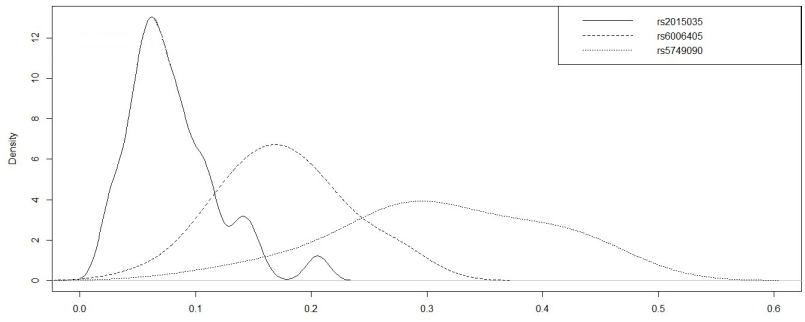

Figura 6.2: Estimación tipo núcleo de la función de densidad para las frecuencias del alelo menor asociadas a los SNPs *rs2015035* (en este caso mediante un estimador tipo núcleo transformado), *rs6006405* y *rs5749090*. Ventana calculada por la regla de validación cruzada, produciendo unos valores de 0.0222, 0.0302 y 0.0501 para las frecuencias alélicas asociadas a los SNPs *rs2015035*, *rs6006405* y *rs5749090* respectivamente.

Se puede observar a la vista de la Figura 6.2 la ligera asimetría de las tres distribuciones de igual modo que el hecho de que no están centradas en puntos similares. A pesar de la Figura 6.2, las frecuencias para los tres SNPs dados verifican normalidad en virtud del test de Shapiro-Wilk

```
Shapiro-Wilk normality test

data: rs2015035
W = 0.9174, p-value = 0.01556


 Shapiro-Wilk normality test

data:  rs6006405
W = 0.9783, p-value = 0.7327


 Shapiro-Wilk normality test

data:  rs5749090
W = 0.9597, p-value = 0.2527
```

Teniendo en cuenta que los p-valores asociados al estadístico de contraste en cada caso son 0.01556, 0.7327 y 0.2527 respectivamente, no existen evidencias para rechazar la hipótesis de normalidad en los datos, si bien cabe destacar que el p-valor asociado a las frecuencias del SNP *rs2015035* únicamente permite no rechazar la hipótesis nula de normalidad al 1 %, este p-valor no se ve modificado demasiado aunque se han probado transformaciones Box-Cox sobre estas frecuencias.

A pesar de los resultados obtenidos mediante el test de Shapiro-Wilk, los datos que poseemos son frecuencias, i.e., valores en el intervalo $[0, 1]$. Podemos estimar la media y la desviación típica mediante el método de máxima verosimilitud para cada una de las muestras de frecuencias asociadas a los SNPs, con esos parámetros estimados calculamos el porcentaje de puntos que caen fuera del citado intervalo. De este modo, si denotamos por $X_1 \sim N(\hat{\mu}_{rs2015035}, \hat{\sigma}_{rs2015035})$, $X_2 \sim N(\hat{\mu}_{rs6006405}, \hat{\sigma}_{rs6006405})$ y $X_3 \sim N(\hat{\mu}_{rs5749090}, \hat{\sigma}_{rs5749090})$ las variables aleatorias de media y desviación típica estimada de la muestra (por máxima verosimilitud) de frecuencias para cada SNP. Pretendemos calcular:

$$1 - \mathbb{P}(0 \leq X_i \leq 1), \quad i \in \{1, 2, 3\},$$

resultando en unos valores de 8.5008 %, 0.0419 % y 0.0156 % para las variables X_1, X_2 y X_3 respectivamente. Cabe destacar como, a pesar de que para los SNPs *rs6006405* y *rs5749090* este porcentaje es reducido, para el SNP *rs2015035* este valor asciende hasta el 8.5 % (un porcentaje elevado), hecho que concuerda con el menor p-valor asociado al estadístico de Shapiro-Wilk. A la vista de los resultados, debemos tomar con cautela las conclusiones de la aplicación de los métodos al SNP *rs2015035*.

Se puede pensar en modelar el comportamiento de las frecuencias alélicas como una regresión lineal que tiene por covariables las latitudes y longitudes de las localizaciones geográficas asociadas a estos puntos y por variable respuesta las frecuencias alélicas,

$$Z(s) = \mu(s) + \varepsilon(s), \quad s \in D \subset \mathbb{R}^2.$$

con $\mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$ y donde s_1 y s_2 son latitud y longitud respectivamente. Ajustando este modelo para los datos asociados a los tres SNPs obtenemos coeficientes estimados asociados a todas las covariables (para los tres modelos) no significativos. Debido a la no significación de los coeficientes estimados, probamos a ajustar un modelo de regresión lineal sin covariables ($\mu(s) = \beta_0, \forall s$), de modo que la predicción encajaría en el kriging ordinario (la versión anterior debería hacerse con kriging universal), para cada una de las frecuencias asociadas a los tres SNPs.

Una vez ajustados estos modelos podemos construír sus residuos espaciales con vistas a determinar su (in)dependencia espacial. Con este objetivo utilizamos el contraste de independencia de Diblasi y Bowman [6] programado en la libería de R `sm`. Este contraste compara el variograma suavizado (mediante splines) con el variograma calculado bajo la hipótesis nula de independencia

(i.e., un variograma plano). Para los residuos de los tres modelos, este contraste devuelve unos p-valores de 1, 0.589 y 0.292, para los modelos asociados a las frecuencias de los SNPs *rs2015035*, *rs6006405* y *rs5749090* respectivamente. Todos los p-valores asociados al estadístico de contraste son mayores que los niveles de confianza usuales, con lo que no existen evidencias para el rechazo de la hipótesis nula de independencia espacial de los residuos. Esto puede ser confirmado mediante la representación del variograma de los residuos de los tres modelos:

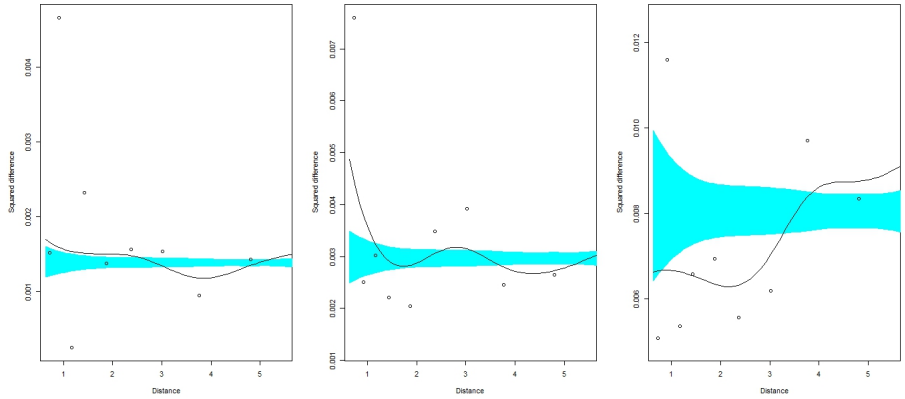

Figura 6.3: Variograma asociado a los residuos del modelo de regresión espacial sin covariables con respuestas las frecuencias de los SNPs *rs2015035* (izquierda), *rs6006405* (centro) y *5749090* (derecha). En azul se representa una banda de variabilidad para la dependencia en cada caso.

Se observa en la Figura 6.3 como los variogramas suavizados caen (más o menos) dentro de la región azul, de igual modo que el hecho de que los dos primeros son planos, lo que nos lleva, junto con el p-valor asociado al estadístico de contraste, a rechazar la dependencia de los residuos.

Con estos resultados, no sería imprescindible utilizar herramientas de predicción espacial, pero por un lado, sabemos que el test de independencia es bastante conservador, y por otro, entra dentro de los objetivos de este trabajo realizar una propuesta de imputación geoestadística general, por lo que continuaremos con el procedimiento, aún sabiendo que los variogramas que obtenemos serán (casi) planos.

Por tanto, vamos a continuar con el proceso de kriging. Para ello, volvemos a las observaciones de partida del proceso. Con vistas a realizar la predicción mediante kriging (Capítulo 5), necesitamos que el proceso sea isotrópico, estacionario (intrínsecamente) e independiente; si bien esta última no es vital a la hora de validar la predicción. De este modo se adjuntan los variogramas empíricos asociados a los tres conjuntos de datos (véase Figura 6.4).

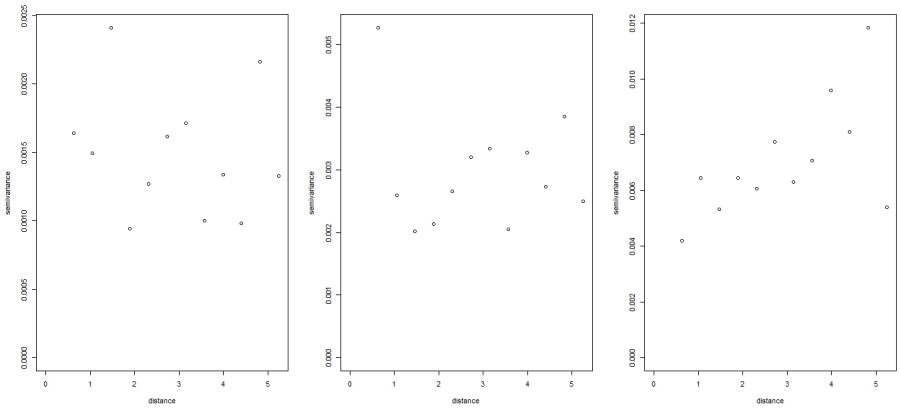

Figura 6.4: Variograma empírico asociado a las frecuencias alélicas de los SNPs *rs2015035* (izquierda), *rs6006405* (centro) y *rs5749090* (derecha).

La Figura 6.4 proporciona los variograma empíricos para los datos de cada uno de los SNPs, mostrando un comportamiento de los mismos casi plano, lo que da una idea de que el proceso pueda ser independiente. Este hecho puede confirmarse mediante la ejecución del test de independencia de la librería `sm`. Este contraste devuelve unos p-valores asociados de 0.62, 0.625 y 0.151 lo que nos indica la independencia del proceso espacial. Con esta misma librería, podemos asesorarnos acerca de la isotropía y la estacionariedad de los procesos mediante los contrastes de isotropía y estacionariedad de la librería `sm` [3]. El primero de ellos, el de isotropía, devuelve unos p-valores asociados de 0.863, 0.944 y 0.898 (para los SNPs *rs2015035*, *rs6006405* y *rs5749090* respectivamente) con lo que no existen evidencias para rechazar la hipótesis nula de isotropía de los tres procesos. Análogamente, ejecutando el test de estacionariedad se obtienen unos p-valores de 0.863, 0.894 y 0.81 indicando la verificación de la propiedad de estacionariedad débil en los tres conjuntos de datos (para los procesos asociados a las frecuencias de los tres SNPs).

Con vistas a obtener nuestra predicción de la frecuencia alélica (mediante kriging), debemos seleccionar un método de estimación, en este caso utilizaremos mínimos cuadrados ponderados. Además, deberemos proporcionar un modelo de variograma válido, puesto que el empírico no verifica estas propiedades necesariamente; para ello emplearemos un modelo paramétrico de variograma, el exponencial:

$$\gamma(u) = \sigma^2(1 - \exp\{-u/\phi\}), \quad u \in \mathbb{R}^+,$$

donde σ^2 es el umbral y ϕ es el rango.

Las estimaciones se realizan mediante la función `variofit` de R que requiere a su vez de unos valores para inicializar el algoritmo de optimización. En el caso que nos ocupa se ha proporcionado una matriz de valores iniciales para el rango

y el umbral. De este modo la función selecciona aquellos valores que minimizan una función de pérdida y fija esos valores como los iniciales para el ajuste.

Una vez ajustado el modelo (sin tomar en consideración la pepita, debido a que no tenemos observaciones demasiado próximas en el espacio), podemos obtener las predicciones en Galicia de las frecuencias para los tres SNPs mediante kriging ordinario (Capítulo 5). Nótese que el proceso de kriging nos proporciona las frecuencias del alelo menor por SNP, sin embargo basta con restarles 1 para obtener la del alelo mayor. Las estimaciones de este método se adjuntan en el Cuadro 6.1:

	rs2015035	rs6006405	rs5749090
$\hat{f}$	0.9469	0.8364	0.7301
f	0.9433	0.8588	0.7005

Cuadro 6.1: Valor estimado de la frecuencia del alelo mayor ($\hat{f}$) para los SNPs *rs2015035* (alelo G), *rs6006405* (alela A) y *rs5749090* (alelo A). Valores de las frecuencias reales del alelo mayor (f) de la población de $N = 379$ gallegos dadas en el Cuadro 4.1.

A la vista del Cuadro 6.1 se observa como los valores de predicción proporcionados por el método de kriging ordinario aproxima con bastante eficacia los valores reales de las frecuencias (para el tamaño muestral $N = 379$). A partir de las predicciones dadas en el Cuadro 6.1 podemos construir intervalos de predicción para nuestras frecuencias alélicas (Capítulo 5), como se ilustra en el Cuadro 6.2:

	LL	UL	$\hat{f}$
rs2015935	0.8583	1.0000	0.9469
rs6006405	0.6808	0.9919	0.8364
rs5749090	0.5407	0.9195	0.7301

Cuadro 6.2: Aproximación bajo normalidad de los límites inferiores (LL), límites superiores (UL) de los intervalos de predicción, centrados en la estimación de la frecuencia del alelo mayor por kriging ordinario ($\hat{f}$) al 95 %.

Además, a modo de complemento, se adjuntan las superficies de predicción y varianzas asociadas a cada modelo ajustado, véanse Figuras 6.5, 6.6 y 6.7:

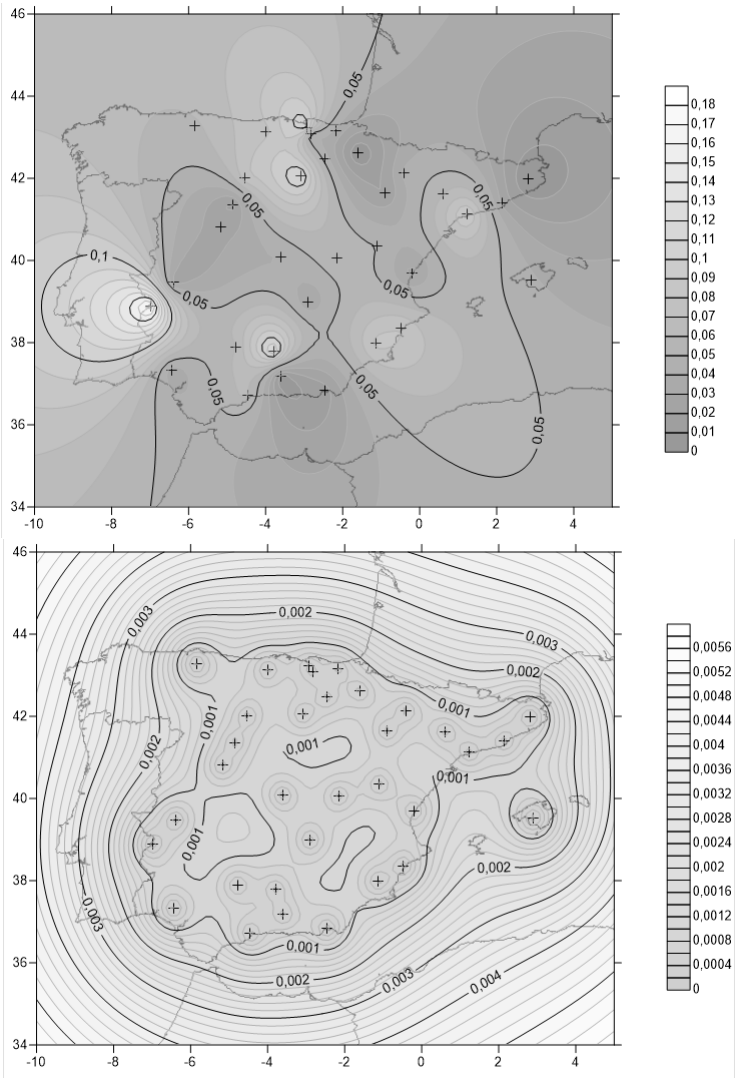

Figura 6.5: Representación de la superficie de predicción (arriba) y superficie de varianzas (abajo), para el kriging ordinario basado en la muestra del SNP *rs2015035*. Se añaden también los puntos de observación.

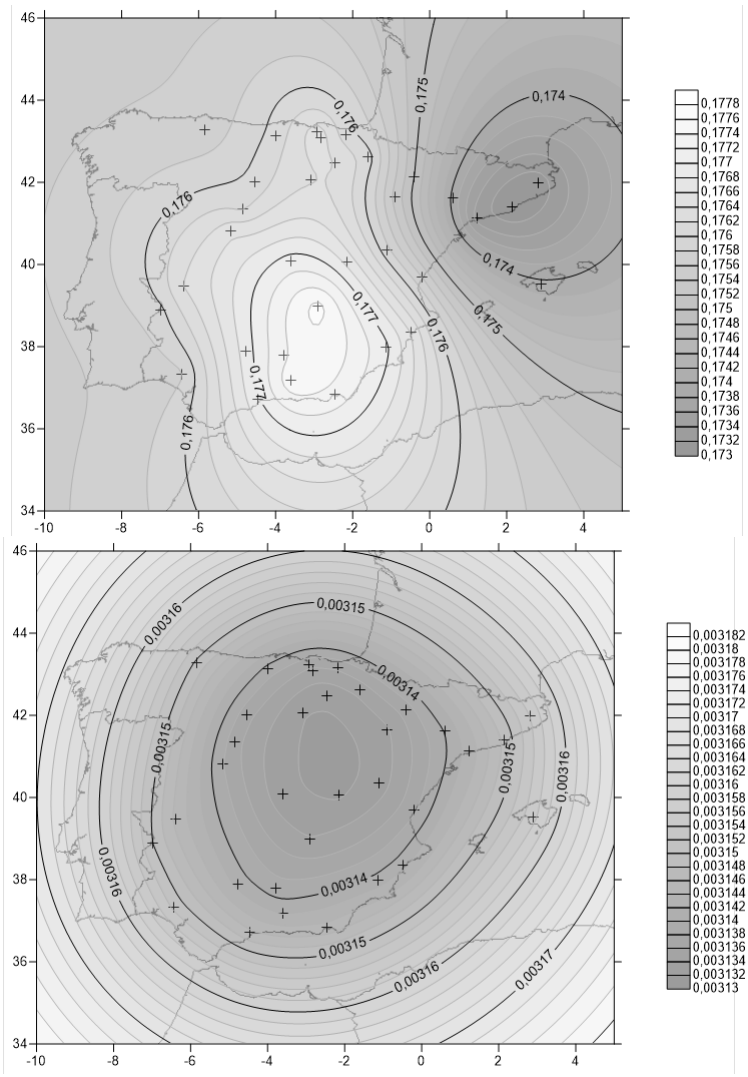

Figura 6.6: Representación de la superficie de predicción (arriba) y superficie de varianzas (abajo), para el kriging ordinario basado en la muestra del SNP *rs6006405*. Se añaden también los puntos de observación.

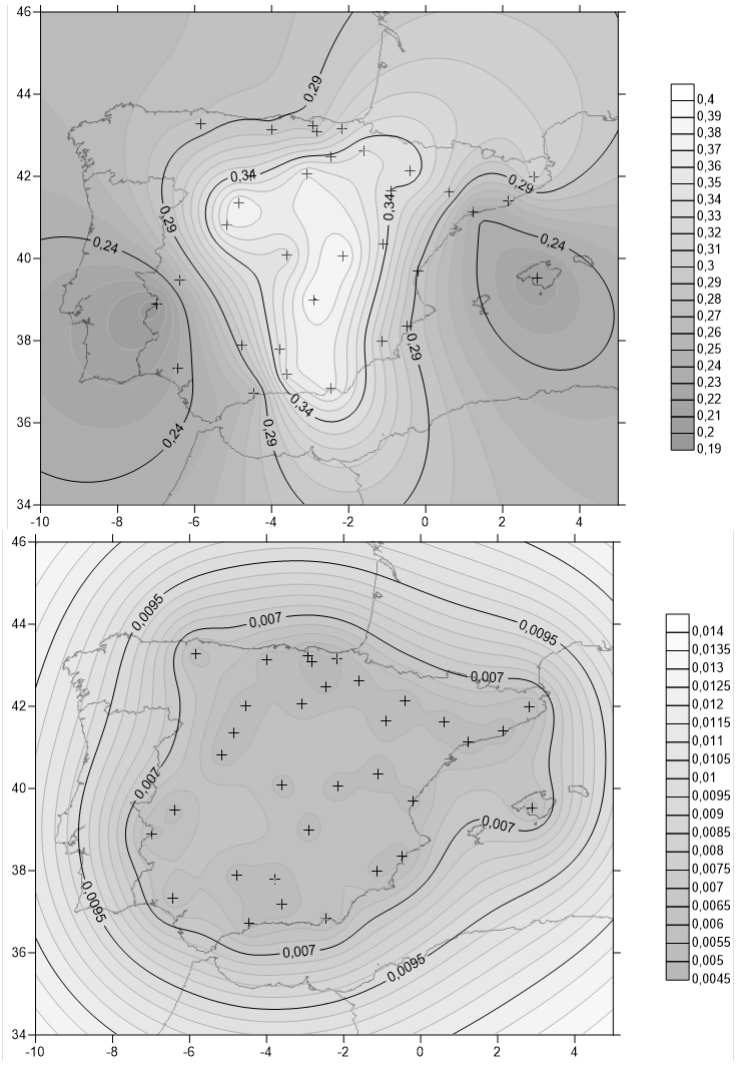

Figura 6.7: Representación de la superficie de predicción (arriba) y superficie de varianzas (abajo), para el kriging ordinario basado en la muestra del SNP *rs5749090*. Se añaden también los puntos de observación.

Capítulo 7

Conclusiones

En aras a realizar una breve comparación de ambas técnicas, nos centraremos en los resultados de frecuencias alélicas obtenidos mediante métodos de imputación (Beagle y MaCH) frente a los dados por el procedimiento de kriging ordinario.

Ante todo, cabe destacar las diferenciales muestrales a la hora de estimar las frecuencias de uno y otro modo, mientras que para imputar se ha seleccionado los genotipos pertenecientes a una sección de SNPs del cromosoma 22 para una población de 379 individuos gallegos; los datos empleados para estimar la frecuencia alélica mediante kriging eran directamente frecuencias alélicas en $N = 33$ puntos distintos de España para cada uno de los tres SNPs, con el objetivo de predecir la asociada a Galicia para cada uno de los tres SNPs citados.

Para el SNP *rs2015035*, los resultados de la imputación (tanto con Beagle como con MaCH) otorgan un valor de la frecuencia alélica (en particular, del alelo mayor) de 1 para $N = 100, 200$ y valores de la ventana $50, 100$ frente a la obtenida mediante kriging (0.9469). A medida que el tamaño de muestra crece ($N = 379$) y, sobre todo, que la ventana aumenta, las estimaciones comienzan a parecerse más (véanse Cuadros 4.2 y 4.5). Además, cabe destacar que, mientras el kriging aproxima con solvencia nuestra frecuencia real en la muestra de $N = 379$ gallegos, los métodos de imputación necesitan un tamaño más elevado de muestra y ventana (i.e., número de SNPs intermedios) para alcanzarla.

A pesar de esto, en los SNP *rs6006405* y *rs5749090*, el kriging produce unas estimaciones más alejadas de los valores de las frecuencias reales de la muestra empleada para imputación de lo que lo están, practicamente todas, las estimaciones dadas por los softwares MaCH y Beagle (véanse Cuadros 4.3, 4.6, 4.4 y 4.7). A modo de curiosidad, los intervalos de predicción normales para la estimación dada por kriging ordinario contienen a todos los valores de las frecuencias alélicas imputados mediante MaCH y Beagle (se debe tomar en consideración la gran amplitud de los intervalos proporcionados en el Cuadro 6.2).

De igual modo, cabe destacar el hecho de que el método de interpolación espacial ha requerido un tiempo de computación menor que los métodos de

imputación, hecho lógico por otra parte debido a que simplemente trabaja con frecuencias y predice directamente una frecuencia alélica. Los métodos de imputación son, en general, más costosos desde el punto de vista computacional, si bien es cierto que trabajan habitualmente con conjuntos de datos de mayores dimensiones y que proporciona una información más completa, a mayores de las frecuencias alélicas imputa alelos faltantes y proporciona una estimación de los genotipos por individuo y SNP.

En lo relativo únicamente a la parte de interpolación espacial, esta podría verse mejorada aumentando el tamaño de la muestra; este tamaño de muestra bajo, $N = 33$, también afecta a los contrastes relativos a las hipótesis sobre el conjunto de datos (en particular, al contraste de independencia) puesto que se trata de un contraste poco potente en general, i.e., se requiere de un tamaño de muestra elevada y que nuestra muestra se encuentre muy alejada de la hipótesis nula de independencia para que pueda ser rechazada. Nótese que esta hipótesis (independencia) no invalida en absoluto la predicción proporcionada mediante kriging, como si habría sucedido en caso de no verificarse las hipótesis de estacionariedad e isotropía. También se ha de tener en cuenta que, por la configuración de las observaciones, la predicción en Galicia supone una "extrapolación" (punto fuera de la nube de localizaciones donde observamos los valores de las frecuencias de los SNPs), es decir, se obtendrán mejores predicciones en otras localizaciones.

Apéndice A

Glosario de términos genéticos

Gen: Secuencia de ADN que constituye la unidad funcional para la transmisión de los caracteres hereditarios.

Alelo: Cada uno de los genes del par que ocupa el mismo lugar en los cromosomas homólogos. Su expresión determina el mismo carácter o rasgo de organización, como el color de los ojos.

Cromosoma: Filamento condensado de ácido desoxirribonucleico (ADN), visible en el núcleo de las células durante la mitosis. Su número es constante para cada especie animal o vegetal. Los humanos, por ejemplo, poseen dos juegos de 23 cromosomas distintos, estos se designan con números del 1 al 22; los últimos (como se distinguen por sexo) se desginan con las letras XY para el varón y XX para la mujer.

Diploides: Relativo a las células, aquellas que poseen un número doble de cromosomas, i.e., tienen dos series de cromosomas.

Locus: En biología, un *locus* es una posición fija en un cromosoma, como la posición de un gen o de un marcador genético.

Genoma: Conjunto de genes contenidos en los cromosomas, i.e., la totalidad de la imformación genética que posee un organismo o una especie en particular.

Genotipo: Información genética que posee un organismo en particular en forma de ADN. Habitualmente el genoma de una especie incluye numerosas variaciones o polimorfismos en muchos de sus genes.

Genotipado: También conocido como genotipificación o caracterización genética, hace referencia al proceso de determinación del genotipo o contenido genómico, en forma de ADN, específico de un organismo biológico.

Polimorfismo de nucleótido simple o SNP: Variación en la secuencia de ADN que afecta a una sola base (adenina, timina, citosina o guanina) de una secuencia del genoma.

Haplotipo: Combinación de alelos de diferentes *loci* de un cromosoma que son transmitidos juntos. En una segunda acepción, un haplotipo es un conjunto

de polimorfismo de un solo nucleótido (SNPs) en un cromosoma particular que están estadísticamente asociados.

Marcador genético: Segmento de ADN con una ubicación física identificable (locus) en un cromosoma y cuya herencia genética se puede rastrear. Un marcador puede ser un gen, o puede ser alguna sección de ADN sin función conocida. Permiten evidenciar variaciones (polimorfismos) en la secuencia de ADN entre dos individuos.

Mutación: Cambio en la información genética (genotipo) de un ser vivo, que produce una variación en las características de este que se presenta de manera espontánea y súbita y que se puede heredar a la descendencia. La unidad genética capaz de mutar es el gen, la unidad de información hereditaria que forma parte del ADN.

Ley de Hardy-Weinberg: Principio que, en genética de poblaciones, establece que la composición genética de una población permanece en equilibrio mientras no actúe la selección natural, mutación, deriva génica, migración y apareamiento no aleatorio. Es decir, la herencia mendeliana, por sí misma, no engendra cambio evolutivo; y consecuentemente la distribución de frecuencias alélicas no varía entre generaciones distintas[1]. Recibe su nombre del matemático inglés G.H. Hardy y del médico alemán W. Weinberg, que establecieron el resultado independientemente en 1908.

Conversión/Recombinación genética: Proceso por el cual una hebra (cadena) de material genético (usualmente ADN) se corta y posteriormente se une a una molécular de material genético diferente.

Progenie: Resultado de la reproducción, i.e., el individuo o individuos producidos mediante la intervención de uno o más parentales.

Segregación: Fenómeno implicado en la transferencia de los distintos *loci* en el ADN a la progenie.

Desequilibrio de ligamiento (LD): Propiedad de algunos genes de las poblaciones genéticas de no segregar de forma independiente, esto es, poseen una frecuencia de recombinación menor del 50 %. Esto suele deberse a que los dos *loci* implicados se encuentran en el mismo cromosoma, lo que imposibilita su transferencia a la progenie de modo aleatorio con la separación de los cromosomas en anafase.

Estudio de asociación del genoma completo: En genética, un estudio de asociación del genoma completo (en inglés, Genome-wide association study o GWAS) es un análisis de una variación genética a lo largo de todo el genoma humano con el objetivo de identificar su asociación a un rasgo observable. Los GWAS suelen centrarse en asociaciones entre los polimorfismos de un sólo nucleótido (SNPs) y rasgos como las principales enfermedades.

MaCH: Software gratuito de imputación genética basado en [8], se obtiene en http:/www.sph.umich.edu/csg/abecasis/MACH.

[1] Una descripción equivalente del equilibrio de Hardy-Weinberg es que los alelos de la siguiente generación para cualquier individuo se eligen aleatoria e independientemente. Consideremos dos alelos A y a con frecuencias p y q respectivamente. De este modo, las tres posibles frecuencias genotípicas finales de la descendencia son: $f_{AA} = p^2$, $f_{aa} = q^2$ y $f_{Aa} = 2pq$.

Beagle: Software gratuito de imputación genética basado en [4], se obtiene en http://faculty.washington.edu/browning/beagle/beagle.html.

HapMap: Proyecto internacional iniciado en el año 2002 y nacido con el objetivo de desarrollar un mapa de haplotipos del genoma humano en el que poder catalogar las regiones de similitudes y diferencias genéticas entre individuos con vistas a comprender mejor la relación entre el genoma y la salud.

Apéndice B

Elementos de un modelo de Markov oculto

A modo de resumen, se introducen los distintos elementos necesarios para caracterizar y explicar un modelo de Markov oculto.

N: Número de estados del modelo de Markov. Se denotan por $S = \{S_1, \ldots, S_N\}$ y el estado en tiempo t por q_t.

M: Número de símbolos observados por estado, estos corresponden a la salida física del sistema. Se denotan los símbolos obervados por $V = \{v_1, \ldots, v_M\}$.

$\mathcal{A}$: Matriz de probabilidades de transición entre estados, $\mathcal{A} = \{a_{ij}\}$, donde

$$a_{ij} = \mathbb{P}[q_{t+1} = S_j | q_t = S_i], \quad i, j \in 1, \ldots, N.$$

$\mathcal{B}$: Distribución de probabilidad del símbolo observado en un estado j, $\mathcal{B} = \{b_j(k)\}$, donde

$$b_j(k) = \mathbb{P}[v_k \text{ en tiempo } t | q_t = S_j], \quad j \in \{1, \ldots, N\}, \quad k \in \{1, \ldots, M\}.$$

π: Distribución del estado inicial, $\pi = \{\pi_i\}$, donde

$$\pi_i = \mathbb{P}[q_1 = S_i], \quad i \in \{1, \ldots, N\}.$$

O: Secuencia de observaciones, $O = \{O_1, \ldots, O_T\}$, donde cada O_t es uno de los símbolos de V.

Q: Secuencia de estados, $Q = \{q_1, q_2, \ldots, q_T\}$, siendo q_1 el estado inicial.

λ: Conjunto de medidas de probabilidad del modelo de Markov oculto, i.e.,

$$\lambda = (\mathcal{A}, \mathcal{B}, \pi).$$

$\mathbb{P}(O|Q, \lambda)$: Probabilidad de la secuencia observada, dada la secuencia de estados Q y el modelo λ, viene dada (bajo independencia) por

$$\mathbb{P}(O|Q, \lambda) = \prod_{t=1}^{T} = b_{q_1}(O_1) b_{q_2}(O_2) \ldots b_{q_T}(O_T).$$

$\mathbb{P}(Q|\lambda)$: Probabilidad de la secuencia de estados, dado el modelo,

$$\mathbb{P}(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T}.$$

$\mathbb{P}(O, Q|\lambda)$: Probabilidad conjunta de O y Q, i.e., de que ambos ocurran simultáneamente,

$$\mathbb{P}(O, Q|\lambda) = \mathbb{P}(O|Q, \lambda)\mathbb{P}(Q, \lambda).$$

$\mathbb{P}(O|\lambda)$: Probabilidad de la secuencia observada, dado el modelo,

$$\begin{aligned}\mathbb{P}(O|\lambda) &= \sum_{Q} \mathbb{P}(O, Q, \lambda) = \sum_{Q} \mathbb{P}(O|Q, \lambda)\mathbb{P}(Q|\lambda) \\ &= \sum_{q_1 \dots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \dots a_{q_{T-1} q_T} b_{q_T}(O_T).\end{aligned}$$

$\alpha_t(i)$: Probabilidad (forward) de una secuencia de observaciones hasta tiempo t, $O_1 \dots O_t$, y del estado S_i en tiempo t, dado el modelo λ,

$$\alpha_t(i) = \mathbb{P}(O_1 O_2 \dots O_t, q_t = S_i|\lambda).$$

$\beta_t(i)$: Probabilidad (backward) de la observación parcial de O desde $t+1$ hasta T dado el estado S_i en tiempo t y el modelo λ,

$$\beta_t(i) = \mathbb{P}(O_{t+1} O_{t+2} \dots O_T|q_t = S_i|\lambda).$$

$\gamma_t(i)$: Probabilidad de encontrarse en el estado S_i en tiempo t, dada la secuencia de observaciones, y el modelo λ,

$$\gamma_t(i) = \mathbb{P}(q_t = S_i|O, \lambda)$$

$\delta_t(i)$: Probabilidad más elevada a lo largo de una trayectoria simple, en tiempo t, que considera las t primeras observaciones y acaba en el estado S_i,

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} \mathbb{P}(q_1 \dots q_t = i, O_1 \dots O_t|\lambda).$$

$\xi_t(i, j)$: Probabilidad de estar en el estado S_i en el instante t, y en el estado S_j en el instante $t+1$, dado el modelo y la secuencia observada,

$$\xi_t(i, j) = \mathbb{P}(q_t = S_i, q_{t+1} = S_j|O, \lambda).$$

Bibliografía

[1] A. Abuli, X. Bessa, J.R. González, C. Ruiz-Ponte et al. Susceptibility genetic variants associated with colorectal cancer risk correlate with cancer phenotype, `Gastroenterology` Vol. 139, 788-796, 2010.

[2] Baum, L.E., Petrie, T., Soules, G. y Weiss, N., A Maximization Technique Ocurring in the Statistical Analysis of Probabilistic Functions of Markov Chains, `The Annals of Mathematical Statistics`, Vol. 41, No. 1, 164-171, 1970.

[3] Bowman, A.W. y Crujeiras, R.M., Inference for variograms, `Computational Statistics and Data Analysis`, Vol. 57, No.1 , 19-31, 2013.

[4] Browning, S.B. y Browning, B.L., Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering, `The American Journal of Human Genetics`, Vol. 81, 2007.

[5] Browning, S.B. y Browning, B.L., Haplotype phasing: Existing methods and new developments, `Nat Rev Genet`; Vol. 12, No. 10, 703-714, 2012.

[6] Diblasi, A. y Bowman, A., On the use of the variogram in checking for independence in spatial data, `Biometrics`, Vol. 57, No. 1, 211-218, 2001.

[7] Forney, G.D., The Viterbi algorithm, `Proceedings of the IEEE`, Vol.61, No. 3, 268-278, 1973.

[8] Li, Y., Willer C.J., Ding J. Scheet P. y Abecasis, G.R., MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes, `Genet Epidemiol`, Vol. 34, No. 8, 816-834, 2010.

[9] Rabiner, L.R., A tutorial on hidden Markov modeles and selected applications in speech recognition, `Proceedings of the IEEE`, Vol. 77, 257-286, 1989.

[10] Schabenberger, O. y Gotway, A.C., *Statistical Methods for Spatial Data Analysis*, Chapman and Hall, 2005.