Centro de Investigación en Matemáticas, A.C.

CIMAT

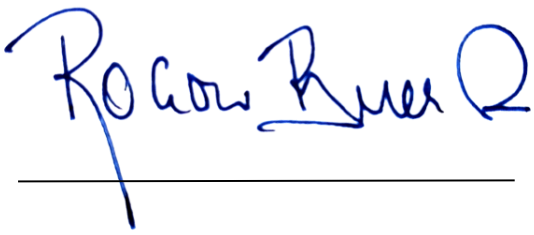
METODOLOGÍA KNOCKOFFS BAJO COVARIABLES CON ESTRUCTURA MARKOVIANA OCULTA

T E S I S

Que para obtener el grado de
Maestro en Ciencias
con especialidad en
Probabilidad y Estadística

P r e s e n t a:
Santiago Correa Pérez

Director de tesis:
Dr. Rogelio Ramos Quiroga



Guanajuato, Gto. 30, noviembre de 2020

Acta de Examen de Grado

.

Acta No.: 167
Libro No.: 002
Foja No.: 167

En la Ciudad de Guanajuato, Gto., siendo las 16:30 horas del día 14 de octubre del año 2020, se reunieron los miembros del jurado integrado por los señores:

DR. JOSÉ ULISES MÁQUEZ URBINA (CIMAT-CONACYT)
DRA. CAROLINA DE JESÚS EUÁN CAMPOS (CIMAT)
DR. ROGELIO RAMOS QUIROGA (CIMAT)

Bajo la presidencia del primero y con carácter de secretario el segundo, para proceder a efectuar el examen que para obtener el grado de

MAESTRO EN CIENCIAS
CON ESPECIALIDAD EN PROBABILIDAD Y ESTADÍSTICA

Sustenta

SANTIAGO CORREA PEREZ

En cumplimiento con lo establecido en los reglamentos y lineamientos de estudios de posgrado del Centro de Investigación en Matemáticas, A.C., mediante la presentación de la tesis

"METODOLOGÍA KNOCKOFFS BAJO COVARIABLES CON ESTRUCTURA MARKOVIANA OCULTA"

Los miembros del jurado examinaron alternadamente al (la) sustentante y después de deliberar entre sí resolvieron declararlo (a)

Aprobado

DR. JOSÉ ULISES MÁRQUEZ URBINA
Presidente

DRA. CAROLINA DE JESÚS EUÁN CAMPOS
Secretario

DR. ROGELIO RAMOS QUIROGA
Vocal

CIMAT
DIRECCIÓN GENERAL

Dr. Victor Manuel Rivero Mercado
Director General

Dedico esta tesis a mi padre que en paz descanse, a mi familia y amigos.

AGRADECIMIENTOS

Expreso mis más sinceros agradecimientos a todas y cada una de las personas quienes con su apoyo y su ayuda hicieron posible mis estudios de maestría.

A mi asesor Rogelio Ramos, por su apoyo, el tiempo compartido y la confianza durante este tiempo. A mi familia y amigos por el apoyo durante estos dos años.

Agradezco el apoyo parcial otorgado por el Proyecto de Ciencia Básica del CONACyT CB-252996 para la realización del presente trabajo de tesis.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT), por la beca para completar mis estudios de maestría.

Al Centro de Investigación en matemáticas, A.C. CIMAT, y a todo el personal, por toda la ayuda recibida y hacer mi estancia placentera.

Al grupo de sinodales, Dr. Ulises Márquez Urbina, Dra. Carolina Euan Campos, por el tiempo dedicado a la revisión de este trabajo y los comentarios brindados.

RESUMEN

Aplicaciones en la cual tenemos un gran conjunto de posibles covariables a una respuesta de forma lineal o no, aparecen con mucha frecuencia en aplicaciones modernas. Este problema de modelamiento se ha estudiado de manera basta, sin embargo, no está claro cómo controlar la fracción de falsos descubrimientos de una manera efectiva. Para abordar este problema el modelo *Knockoffs* proporciona un algoritmo para una inferencia válida en situaciones donde la distribución condicional de la respuesta es arbitraria y desconocida, además sin importar el número de covariables. Esta inferencia es basada en un enfoque probabilístico y no geométrico.

Aunque no es sencillo encontrar un caso concreto en el cual podamos aplicar este algoritmo, una aplicación directa además de modelos lineales, es cuando las covariables se modelan como un *modelo oculto de Markov* (HMM). Teniendo esto, es posible adaptar este tipo de estudios a GWAS (asociación del genoma completo).

En este trabajo veremos cómo se construye el algoritmo y su aplicación directa a GWAS: en el capítulo 1, estudiaremos los conceptos básicos sobre genética, en el capítulo 2 hablaremos sobre los HMM y algoritmos usados para la hallar la fase de haplotipos, en especial el algoritmo *fastPHASE*. En el capítulo 3 estudiaremos la metodología Knockoffs y su construcción, y su aplicación a HMM en el capítulo 4. Por último, en el capítulo 5 verificaremos con un conjunto de datos sintéticos lo visto anteriormente.

Palabras Clave

Knockoffs, GWAS, HMM, Haplotipos, fastPHASE, Tasa de Falsos Descubrimentos.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

CONCEPTOS BÁSICOS DE GENÉTICA

1.1. Nociones Básicas de Genética

Iniciaremos nuestro estudio con el concepto de genoma e iremos implementado cierta terminología que nos será útil en el resto del trabajo. El *genoma* está constituido por una o más moléculas de *ácido desoxirribonucleico* (ADN), que son un tipo de ácido nucleico, el cual es un polímero de compuestos químicos denominados *nucleótidos*, que a su vez, están constituidos por una base nitrogenada, un azúcar pentosa, y de uno a tres grupos de fosfato. En la figura (1.1) se observa que cada nucleótido posee en su estructura un nucleósido formado por una pentosa (amarillo) unida a una base nitrogenada (azul) mediante un enlace glucosídico (verde), más uno a tres grupos fosfato (rojo). Existen dos tipos de nucleótidos de acuerdo con la base nitrogenada (azul) que se incorpora a la molécula:

1. Purinas: adenina (A) y guanina (G);
2. Pirimidinas: citosina (C), timina (T), y uracilo (U)

El uracilo, sólo está presente en otro ácido nucleico, estructural y funcionalmente distinto al ADN, conocido como ácido ribonucleico (ARN).

En la estructura del ADN, los nucleótidos se unen covalentemente por enlaces fosfodiéster formando dos cadenas independientes, las cuales interaccionan entre sí, a través de puentes de hidrógeno entre pares de bases nitrogenadas, siempre dos entre A y T, y tres entre C y G. Lo anterior origina que la secuencia de cada cadena sea inversa y complementaria a la otra,

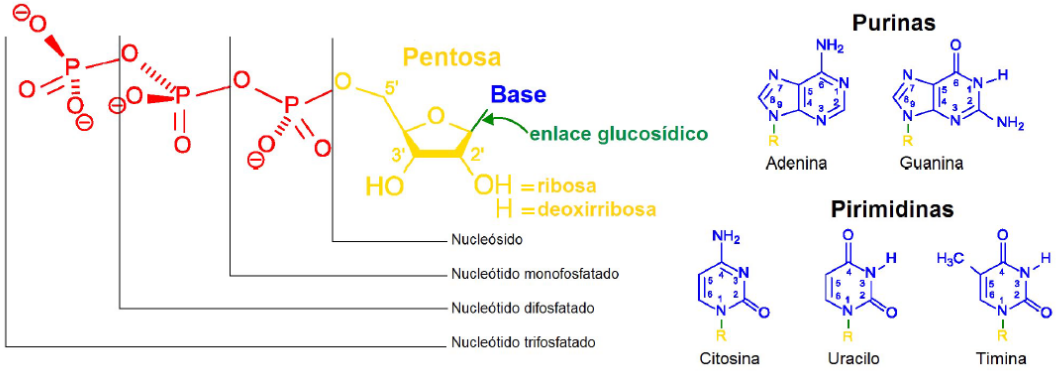


Figura 1.1: Nucleótidos

y da lugar a la estructura de "doble hélice", característica del ADN. En la figura (1.2) vemos que el ADN está compuesto de diferentes subunidades. El esqueleto de la molécula está hecho de dos polímeros de nucleótidos unidos por las pentosas (desoxirribosas) a través de los grupos fosfato. La secuencia de ambas cadenas es complementaria y antiparalela, y entre ellas interaccionan a través de puentes de hidrógeno entre las bases nitrogenadas presentes en su estructura, siempre adenina con timina (dos puentes de hidrógenos), y guanina con citosina (tres puentes de hidrógeno).

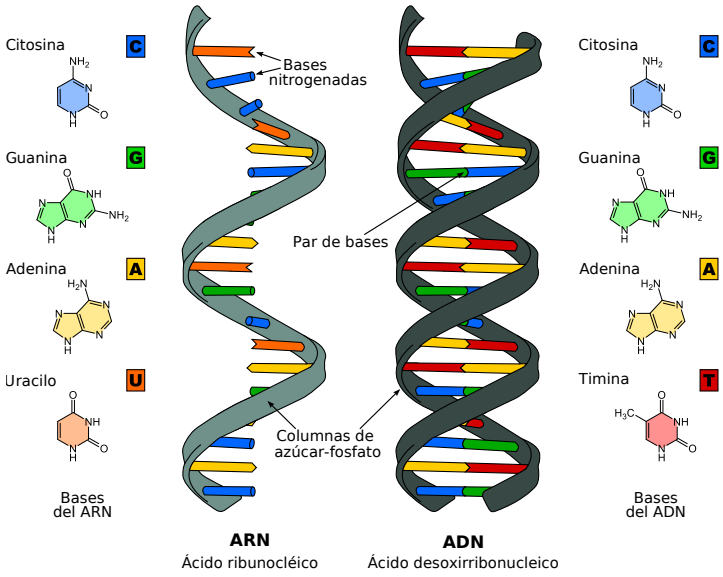


Figura 1.2: Estructura del ADN [1]

[0]Imagen tomada de https://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-ES.svg

El genoma contiene secuencias discretas con la información necesaria para generar moléculas con función biológica a las que denominamos genes, y en el caso del humano, el genoma está organizado en el núcleo en un conjunto de 23 pares de cromosomas. Originalmente la definición de "gen" era "la unidad de herencia", unos años más tarde con el establecimiento del dogma central de la biología molecular, se definió como " un segmento de ADN que posee información necesaria para generar moléculas con función biológica (ARN o proteínas)". Se sabe hoy en día que el ser humano posee alrededor de 22.280 genes que codifican proteínas.

Los seres humanos heredamos $50\,\%$ de esta información a nuestra descendencia y nacemos, crecemos y morimos casi que con la misma información genética en nuestras células. Debido a su naturaleza estable y heredable, la información genética representa una fuente potencial de biomarcadores de amplia aplicación clínica, por lo que existe enorme interés en estudiar el papel de los factores genéticos en la enfermedad humana. Una mutación o variante genética, es una base o secuencia en el ADN que puede diferir entre individuos de la misma especie. Las variantes genéticas pueden clasificarse de manera sencilla en tres grupos (figura (1.3)):

(a) Cambios de una sola base o nucleótido.

(b) Eventos de inserción y deleción.

(c) Rearreglos estructurales.

Se le denomina *alelo* a cada variante o posibilidad de base o secuencia en la que se presenta una posición en el genoma; a su vez, cada variante de acuerdo con el número de formas en las que se presentan en el genoma puede ser bialélica o multialélica. En la figura (1.4) observamos que las variaciones bialélicas se presentan generalmente en sólo dos formas como los polimorfismos de un solo nucleótido (SNPs) (en este caso T$\rightarrow$A) y multialélicas cuando se presentan en tres o más alternativas como los repetidos corto en tándem (STRs).

Un polimorfismo es considerado como tal cuando la frecuencia de uno de sus alelos en la población es superior al $1\,\%$. Los polimorfismos genéticos son variantes del genoma que aparecen por mutaciones en algunos individuos, se transmiten a la descendencia y adquieren cierta frecuencia en la población tras múltiples generaciones. El conjunto de variaciones genéticas comunes y raras presentes en el genoma de un individuo constituyen su *genotipo*, el cual, en estrecha relación con los factores ambientales, da lugar a características individuales (normales o patológicas), es decir a su fenotipo. Las diferencias genéticas entre pares de humanos se han estimado de $0.5\,\%$ a $1\,\%$.

Cuando el objetivo de un estudio es identificar un polimorfismo o variante en un gen que

5'	A	C	T	C	C	T	G	A	G	G	A	G	**3'**
3'	A	C	T	C	C	T	G	T	G	G	A	G	**5'**

5'	A	C	T	C	C	T	G	T	G	G	A	G	**3'**
3'	A	C	T	C	C	T	G	A	G	G	A	G	**5'**

(a) Cambio de un sólo nucleótido: A→T.

5'	A	T	C	A	T	C	T	T	T	G	G	T	**3'**
3'	T	A	G	T	A	G	A	A	A	C	C	A	**5'**

5'	A	T	C	A	T	T	G	G	T	G	T	T	**3'**
3'	T	A	G	T	A	A	C	C	A	C	A	A	**5'**

(b) Deleciones

5'	C	T	C	A	T		G	C	C	C	T	G	**3'**
3'	G	A	G	T	A		C	G	G	G	A	C	**5'**

5'	C	T	C	A	T	T	G	C	C	C	T	G	**3'**
3'	G	A	G	T	A	A	C	G	G	G	A	C	**5'**

(c) Inserción de las bases T→A.

Figura 1.3: Variantes genéticas.

					SNP						STR						
Cromosoma 1	G	T	A	C	T	A	G	A	C	T	A	C	T	A	C	T	G
Cromosoma 2	G	T	A	C	A	A	G	A	C	T	A	C	T	G	C	T	G
Cromosoma 3	G	T	A	C	A	A	G	A	C	T	A	C	T	A	C	T	A

Figura 1.4: Variantes genéticas bialélicas y multialélicas

esté relacionado con una enfermedad se pueden emplear diferentes estrategias. En primer lugar, es importante obtener evidencia de que al menos una fracción de la enfermedad está determinada genéticamente. En segundo lugar, hay que identificar dónde están los genes de

interés para la enfermedad. En esta fase se realizan estudios denominados de *ligamiento (linkage)*, que emplean como *marcadores genéticos* una serie de polimorfismos repartidos por todo el genoma. En estos estudios se suelen emplear familias grandes con varios miembros afectados y sus análisis permiten identificar zonas del genoma de interés, pero tienen poca resolución. En esas zonas identificadas puede haber centenares de genes interesantes y miles de polimorfismos candidatos. Para identificar con mayor precisión los genes de interés y, dentro de esos genes, el o los polimorfismos responsables, se emplean estudios de asociación, en los que se compara la frecuencia relativa de las diferentes variantes de una serie de polimorfismos entre los individuos afectados y un grupo control adecuado. Estos estudios suelen seleccionar "genes candidatos" (aquellos cuya función puede estar relacionada con la enfermedad de interés), y dentro de esos genes se busca como marcadores genéticos a determinados polimorfismos, normalmente de tipo SNP, repartidos a lo largo del gen. En cuanto a la metodología de estudio, se suelen emplear diseños epidemiológicos clásicos basados en individuos no relacionados, como estudios de casos y controles o de cohortes. También se pueden emplear diseños basados en familias, en los que los individuos de control son parientes de los casos, como los diseños de casos y hermanos sanos o tríos (caso y padres).

Los SNPs son las variaciones genéticas más comunes; en las poblaciones humanas se han identificado al menos 10 millones de SNPs con frecuencias mayores que el $1\,\%$ al menos en algún grupo humano.[2] Los *polimorfismos de nucleótido único* (SNPs) son un tipo de polimorfismo que producen una variación en un sólo par de bases. El nombre significa de cierta manera "cortar", y se refiere a los lugares específicos en el genoma donde las personas son diferentes. Estos corresponden a ubicaciones físicas de bases particulares en la cadena de ADN de los cromosomas. Son de doble naturaleza: la primera naturaleza corresponde a la ubicación de la base a lo largo de la cadena de ADN. Las ubicaciones son fijas para todos los seres humanos, por tanto, las posiciones exactas a lo largo del cromosoma son uniformes para todos los humanos. Como su naturaleza es tan estática, en la investigación sirven y se les conoce como *marcadores*. Un marcador es un segmento de ADN con una ubicación física conocida en un cromosoma. Estos pueden ayudar a vincular una enfermedad hereditaria con el gen responsable. Los segmentos de ADN que se encuentran cerca en un cromosoma tienden a heredarse juntos.

La segunda naturaleza corresponde al valor de la base particular en la cadena de ADN,

[2] http://www.ncbi.nlm.nih.gov/snp/.

en la posición del marcador. Los valores de los datos del marcador, para una medición estadística, se organizan en un conjunto de datos tabulares. Un análisis de este tipo de datos es la de comparar las secuencias de los marcadores entre dos cohortes. Esto es posible gracias a la organización tabular mencionada, donde todos los marcadores que pertenecen a un paciente específico se encuentran en una sola fila de la tabla. Las columnas representan los valores presentes en la ubicación del marcador particular, es decir, cada columna es una ubicación de marcador específica dentro del genoma, como consecuencia debido a que el 99 % del genoma tendrá los mismos valores en las posiciones de los marcadores, especialmente en individuos sanos, entonces secuencias de marcadores específicas y diferentes que pertenecen a un grupo de pacientes enfermos se destacan y se localizan fácilmente. Cuando se localizan las diferentes secuencias de SNPs, se registran las ubicaciones físicas de estas diferencias. Esta ubicación es donde probablemente se localiza el gen sospechoso que causa la enfermedad. Los datos de mala calidad, con valores faltantes de SNPs, tienen un impacto negativo en la calidad de los resultados finales del análisis de datos. Por lo cual es de vital importancia eliminar los valores faltantes mediante una sustitución precisa.

Un *haplotipo* en genética se puede referir a una combinación de alelos o a un conjunto de SNPs que se encuentran en el mismo cromosoma. La meiosis es un proceso de división celular, entonces la recombinación en las células sexuales rara vez puede separar los fragmentos alélicos heredables. Un *haplogrupo* es un grupo de haplotipos similares que comparten un ancestro común que tiene la misma mutación SNPs en ambos haplotipos. Debido a que los haplogrupos contienen haplotipos similares, es posible predecir un haplogrupo a partir de haplotipos. Una prueba SNP confirma un haplogrupo con un $100\,\%$ de precisión, pero el haplogrupo puede estimarse estadísticamente con cierto promedio. Los GWAS se utilizan principalmente para identificar cuáles son los factores genéticos que influyen en la salud y enfermedad. Generalmente se centran en la asociación entre los SNPs y características especiales, generalmente aquellas expresadas por enfermedades importantes, aquí se comparan los SNPs en cadenas de ADN homólogas de dos grupos de participantes o *cohortes*.

Una cohorte se refiere a un grupo de personas con cierta enfermedad y personas similares, pero sin esta, es decir, poseen características similares, pero sin la enfermedad. Después de obtener la información del genotipo SNP de estos dos grupos, se verifica si un tipo de alelo SNP o grupo de alelos es más frecuente en personas con la enfermedad que en aquellos que no tienen la enfermedad y se dice que el SNP o el haplotipo está “asociado” con la enfermedad. Ahora bien, se consideran los SNPs asociados que abarcan una región o regiones del genoma

humano que influyen en la expresión de la enfermedad. Los estudios de GWA identifican SNPs y otras variantes en el ADN que están asociadas con una enfermedad, pero que por sí mismas no pueden especificar qué genes son causales, sino que especifican las ubicaciones activas en los cromosomas que son expresivos en individuos enfermos.

Cuando comparamos las secuencias genéticas de grandes grupos de individuos sanos y enfermos, la investigación puede crear un mapa aproximado de los genes defectuosos que causan cierta enfermedad. Este procedimiento de minería de datos se conoce como *estudio de asociación del genoma completo* (GWAS). El punto de partida para comprender y curar ciertas enfermedades es identificar en qué cromosoma se encuentra el gen y en qué parte del cromosoma. Sin embargo, a causa de la volatilidad, la relativa sensibilidad química del material genético y la precisión del equipo de lectura de datos genéticos, el proceso de codificación o traducción responsable de digitalizar los datos genéticos muestreados a datos codificados alfanuméricos discretos se ve frecuentemente comprometido, lo que resulta en valores perdidos de lecturas genéticas. Para resolver este problema, se utilizan algoritmos basados en el aprendizaje automático que se basan en la imputación probabilística para estimar los datos faltantes. Así en este tipo de estudios, los valores perdidos pueden conducir a resultados de análisis muy pobres.

El diseño al que se ha acudido mayoritariamente es el de casos y controles, el cual de manera muy general consiste en comparar la frecuencia de los factores estudiados (en este caso variantes genéticas) en un grupo de individuos afectados y un grupo de individuos no afectados, para determinar si existe una diferencia significativa en la frecuencia de una o más de estas variantes entre ambos grupos. La figura (1.5) ilustra cómo funciona el GWAS, el Panel A muestra un "locus" que corresponde a un fragmento pequeño del cromosoma 9. En el Panel B, la fuerza de la asociación entre cada SNP y la enfermedad se calculan en base a la prevalencia de cada SNP en el grupo de casos y en el de controles. En este ejemplo, los SNPs 1 y 2 del cromosoma 9 están significativamente asociados ($P < 10^{-12}$ y 10^{-8}, respectivamente). La gráfica del Panel C muestra los P-valores para todos los SNPs genotipificados que pasaron todos los controles de calidad; cada cromosoma está representado con un color distinto. Los resultados señalan a un "locus" en el cromosoma 9, claramente marcado por dos SNPS adyacentes 1 y 2 (acercamiento en la gráfica a la derecha), y apoyado por otros SNPs en la misma región.

Existen factores críticos en el diseño y análisis de GWAS, entre los que destacan tres:

(a) La selección cuidadosa de casos y controles para minimizar potenciales sesgos.

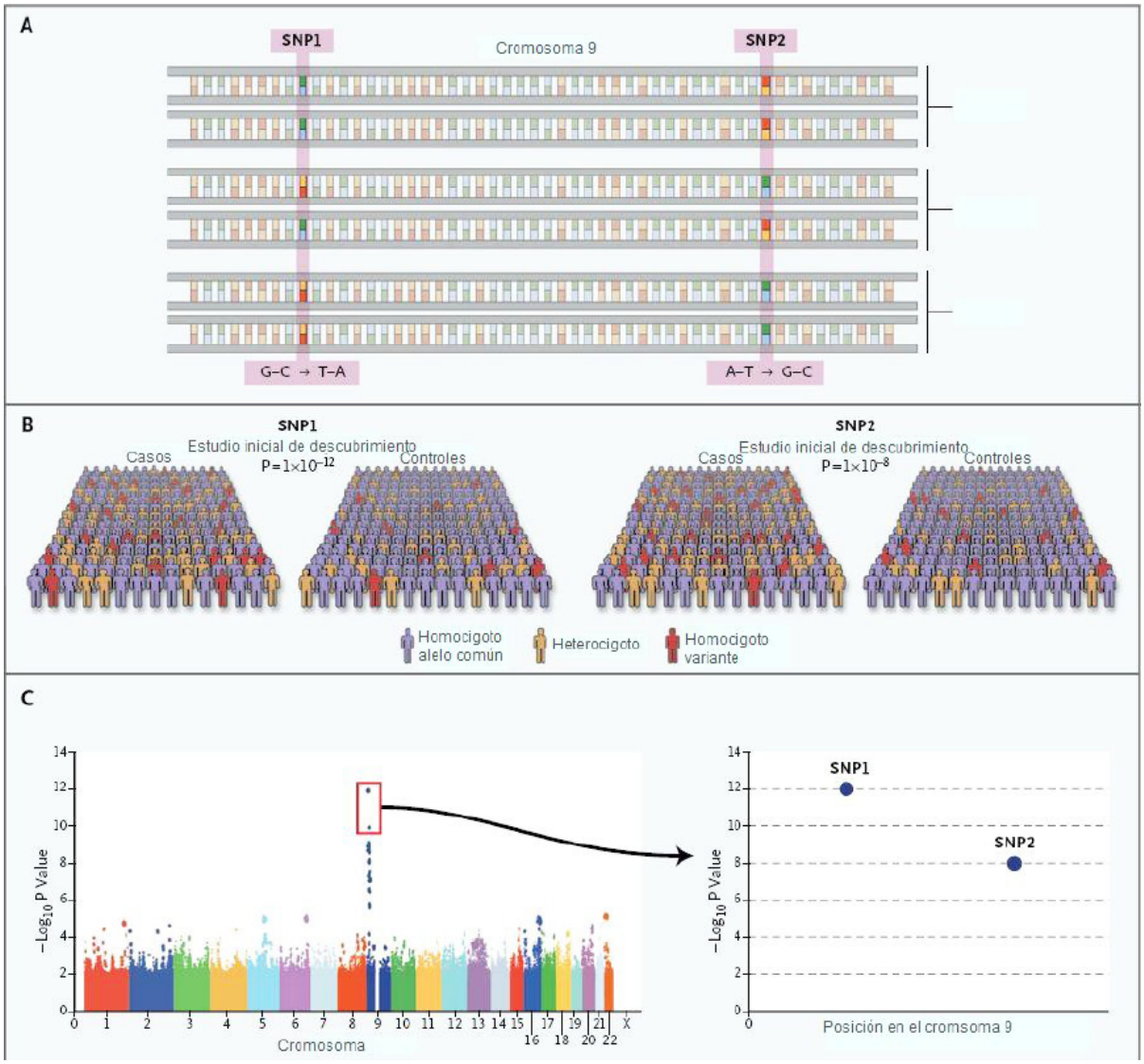


Figura 1.5: GWAS [3]

(b) La selección de los marcadores a caracterizar o genotipificar.

(c) El control de calidad de los datos.

1.2. Análisis Descriptivo de los Haplotipos

Con frecuencia son varios los polimorfismos que se analizan simultáneamente en un gen o región candidata de un gen. El motivo es que el polimorfismo realmente responsable de

[3] Imagen tomada de Silva Zolezzi (2011).

influir o modificar el riesgo de la enfermedad puede ser desconocido; por ello se analizan varios polimorfismos para intentar identificarlo. Entre diferentes polimorfismos localizados en el mismo cromosoma y relativamente próximos entre sí suele observarse cierto grado de correlación o asociación estadística denominada *desequilibrio de ligamiento (linkage disequilibrium)*. Ello es debido a que en el proceso de meiosis que genera los gametos, los cromosomas que se transmitirán serán copias exactas de los del progenitor, a excepción de los entrecruzamientos que generan recombinación. Es decir, cada cromosoma transmitido a la descendencia estará formado por una composición de fragmentos largos que son una copia exacta de los del progenitor, pero combinando partes del cromosoma paterno y del materno. La probabilidad de que entre 2 loci cercanos se dé una recombinación es baja y, por ello, se observa el desequilibrio de ligamiento, que tiende a disminuir en sucesivas generaciones hasta llegar al equilibrio (independencia estadística).

El desequilibrio de ligamiento es muy útil, pues permite localizar polimorfismos relacionados con la enfermedad. Si aparece una mutación que genera un polimorfismo responsable de la enfermedad, es posible que otros polimorfismos cercanos también estén asociados con ella. De hecho, como lo que se transmite de padres y madres a sus hijos son cromosomas, suele ser interesante identificar el conjunto de alelos que se transmiten conjuntamente en cada cromosoma, de manera que sea más fácil así identificar el polimorfismo causal. Un individuo, para un conjunto de loci cercanos, posee 2 haplotipos, cada uno en un cromosoma. Identificar los haplotipos a partir de los genotipos de cada locus suele ser fácil, aunque hay algunos casos excepcionales.

Por ejemplo, si los genotipos para un individuo son TC y AA, sus haplotipos serán T-A y C-A, es decir, el individuo será portador de una pareja de cromosomas, con cada una de estas combinaciones de alelos. En el caso de que los 2 genotipos sean heterocigotos: TC y AG, la combinación de alelos en cromosomas puede ser T-A y C-G o bien T-G y C-A. No se puede saber qué combinación es la correcta para un individuo si no se conocen los genotipos de los progenitores o se emplean técnicas de laboratorio muy sofisticadas que permitan identificar haplotipos. En la práctica, para realizar análisis estadísticos de asociación se recurre a métodos de estimación que tienen en cuenta variables con incertidumbre para resolver este problema.

Lo anterior ilustra que para la estimación de las frecuencias para cada haplotipo sería sencilla si no se dieran casos de incertidumbre, es decir, casos en que no es posible determinar la pareja de haplotipos que lleva el individuo debido a que éste tenga 2 o más loci heteroci-

gotos. Si, además, hay valores perdidos en la determinación de alguno de los genotipos, la incertidumbre aumenta. Los métodos estadísticos existentes permiten identificar o estimar, para cada individuo, la parjea de haplotipos que posee en función de los genotipos. Estos haplotipois puede analizarse entonces en relación con la enfermedad.

1.3. Modelo Matemático

En la sección anterior se habló un poco sobre los valores faltantes en conjuntos de datos GWAS. Veremos cómo estos valores son imputados o inferidos probabilísticamente. Nos referimos a que son imputados a una sustitución algorítmica de espacios en blanco o valores perdidos desconocidos dentro del conjunto de datos, este proceso se conoce como "imputación de unidades". Tras la imputación se puede realizar el análisis de GWAS, sin embargo, en este proceso existe un grado de incertidumbre debido a los valores imputados. Por tanto, el objetivo es encontrar métodos de imputación para eliminar esta incertidumbre lo mejor posible. Los algoritmos de imputación sofisticados y de alto rendimiento que se basan en modelos probabilísticos como los modelos gráficos probabilísticos de conjuntos de datos. El objetivo es entonces estudiar *modelos ocultos de Markov* (HMM) y ver cómo estos son útiles a la hora de modelar este tipo de problemas.

CAPÍTULO 2

MODELOS OCULTOS DE MARKOV (HMM)

2.1. Introducción

Recordemos que un haplotipo es una secuencia de nucleótidos a lo largo de un solo cromosoma. Como humanos, tenemos 23 pares de cromosomas. Sin embargo, con la tecnología actual, es difícil separar los dos cromosomas de un par y, a menudo, obtenemos información combinada de haplotipos o genotipos. El objetivo de la fase del haplotipo es resolver los haplotipos dada la información de los genotipos. Conocer los haplotipos no solo nos da una imagen completa del genoma de un individuo, sino que también tiene otras motivaciones biológicas significativas.

Durante las últimas dos décadas, ha habido un interés significativo en comprender la composición genética de los humanos. Con esfuerzos internacionales como el *Proyecto Genoma Humano* (Consortium et al., 2004) y el *Proyecto Internacional HapMap* (Consortium et al., 2005), la tecnología para leer el genoma humano se ha desarrollado rápidamente. Sin embargo, estas tecnologías siguen siendo limitadas, y se deja a los métodos computacionales para detectar, corregir errores y reunir información parcial de la tecnología.

Debido a que los humanos son diploides, tenemos dos copias de cada tipo de cromosoma: una de nuestra madre y otra de nuestro padre, para un total de 46 cromosomas. Las dos copias

son altamente homólogas entre sí y solo difieren en una pequeña fracción ($0.1\,\%$) de los sitios variantes. Para un cromosoma con k variantes, podemos representar su haplotipo como una cadena del conjunto $\{A, C, G, T\}^k$. De hecho, suponemos que las variantes son bialélicas, es decir, cada variante toma uno de los dos valores alélicos posibles. Por lo tanto, sin pérdida de generalidad, podemos representar haplotipos como una cadena del conjunto $\{0, 1\}^k$, donde 0 y 1 representan los dos posibles valores alélicos en cada ubicación variante. Con la tecnología actual, es difícil separar un par de cromosomas y, a menudo, mezclamos los dos haplotipos. Un genotipo es la información combinada del haplotipo para un par de cromosomas. Podemos representarlo como una lista ordenada de pares de longitud k donde cada par pertenece al conjunto $\{(0, 0), (1, 1), (0, 1)\}$, esta codificación se hace cuando tenemos haplotipos de referencia. Entonces asignamos 0, cuando el valor del genotipo es igual al de la referencia y asignamos el valor de 1, cuando es igual al alternativo. La lista está ordenada de acuerdo con la posición cromosómica de cada par, pero los pares mismos no están ordenados. En la figura (2.1) vemos que se muestran los sitios variantes y los sitios no variantes están representados por "—". En este ejemplo, hay cinco variantes. Los haplotipos son "ACATT" ("00000") y "ATACG" ("01011"). El genotipo se representa como una lista de pares desordenados y es {(A, A), (C, T), (A, A), (T, C), (T, G)} ($\{(0, 0), (0, 1), (0, 0), (0, 1), (0, 1)\}$) .

—	—	A	—	—	C	—	—	A	—	—	—	T	T	—
—	—	A	—	—	T	—	—	A	—	—	—	C	G	—

Figura 2.1: Haplotipos y Genotipos

En una región genómica con k sitios, hay $2k - 1$ posibles haplotipos. El objetivo del problema de la fase del haplotipo es recuperar los dos haplotipos (de los $2k - 1$ posibles haplotipos) de un individuo. La siguiente formulación biológica de la fase del haplotipo, teniendo en cuenta que la entrada cambiará con diferentes algoritmos, pero la salida y el objetivo biológico seguirán siendo los mismos, se da a continuación:

Algoritmo 1 Problema de Fases del Haplotipo

Input: Genotipo de un individuo, $G = (g_1, g_2, \ldots, g_k)$, donde $g_i \in \{(0, 0), (1, 1), (0, 1)\}$ para $1 \leq i \leq k$.

Output: Par de haplotipos, $H = \{h_1, h_2\}$ para el individuo, donde $h_1, h_2 \in \{0, 1\}^k$ y H son consistentes con G.

El proceso es entonces el siguiente: se recibe primero una copia de cada tipo de cromosoma, una de la madre y una del padre. Sin embargo, la evolución de los haplotipos se complica por mutaciones y recombinaciones. Una mutación cambiará el valor alélico en un sitio del cromosoma de padres a hijos. La tasa de mutación, se usa para medir la probabilidad de que un sitio en particular mute. En humanos, la tasa de mutación es de alrededor de 2.5×10^{-8}. La baja tasa de mutación es la razón por la cual dos cromosomas del mismo tipo solo difieren en una pequeña fracción de sitios variantes ($0.1\,\%$). Hay dos tipos de sitios de variantes: polimorfismos de nucleótido único (SNPs) y variantes de nucleótido único (SNV). La diferencia entre los dos se define de manera algo arbitraria. Los SNP son sitios variantes comunes en una población. Los SNV son sitios variantes que son exclusivos de un individuo.

La diversidad genética también es causada por la recombinación, que es cuando los dos cromosomas de un par intercambian regiones de su genoma. La recombinación ocurre durante la meiosis, que es el proceso donde se forman las células reproductivas en los padres. Como ejemplo, una recombinación en el sitio r hará que la región $[1, r]$ de un cromosoma se combine con la región $[r + 1, l]$ del otro cromosoma, donde ambos cromosomas tienen longitud l (ver figura 2.2).

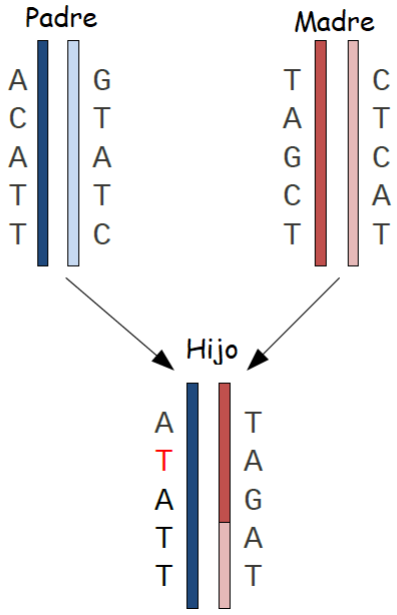


Figura 2.2: Un niño que hereda un cromosoma del padre y uno de la madre. Aquí hay una mutación en el segundo sitio del cromosoma paterno (C→T). También hay una recombinación en el cromosoma materno entre el tercer y cuarto sitio.

Cada posición a lo largo del cromosoma está asociada con una probabilidad de recombinación; esto se denomina tasa de recombinación. Ciertas áreas en el genoma son más propensas a la recombinación que otras. Las regiones de sitios con una alta probabilidad de recombinación se conocen como puntos críticos de recombinación.

Los haplotipos nos dan una descripción completa del genoma humano, y son mucho más informativos que los genotipos. Los haplotipos nos permiten encontrar asociaciones entre un gen particular y una enfermedad. Las asociaciones que podemos detectar con haplotipos no siempre se pueden detectar solo con genotipos. Por ejemplo en la figura (2.1) muestra que sin conocer los haplotipos, no hay diferencia entre un individuo con la enfermedad y su genotipo. Pero si conocemos los haplotipos, entonces podemos detectar una asociación con el haplotipo "GG" ("00") y la enfermedad.

Muestra	Genotipo	Haplotipo
Enfermedad	{(G,C),(G,A)} $\{(0,1),(0,1)\}$	'GG' y 'CA' '00' y '11'
Enfermedad	{(G,C),(G,G)} $\{(0,1),(0,0)\}$	'GG' y 'CG' '00' y '10'
No Enfermedad	{(G,C),(G,A)} $\{(0,1),(0,1)\}$	'GG' y 'CG' '01' y '10'

Tabla 2.1: Asociación entre un individuo con la enfermedad y su genotipo.

Una segunda aplicación de la información del haplotipo es detectar la selección positiva. Si una variante particular está bajo selección neutral, pasará mucho tiempo antes de que muchos individuos en una población tengan la variante. Durante este tiempo, es probable que ocurra una recombinación alrededor de la variante, interrumpiendo el haplotipo. Sin embargo, bajo una selección positiva, la frecuencia de la variante en una población aumentará más rápidamente y hay menos tiempo para que ocurra la recombinación alrededor de la variante. Por lo tanto, podemos detectar una selección positiva buscando haplotipos largos que sean comunes en la población.

Tercero, también ayudan a estimar la tasa de recombinación. Si conocemos la genealogía y los haplotipos de una población, entonces podemos detectar dónde ocurre la recombinación. Por ejemplo, en la figura (2.2) sabemos dónde se produjo la recombinación al observar los haplotipos del niño y de la madre. La información sobre dónde ocurren las recombinaciones

puede usarse para estimar la tasa de recombinación. También podemos detectar puntos críticos de recombinación buscando regiones del genoma donde haya muchas recombinaciones en la población.

Una cuarta aplicación de los haplotipos es ayudar a comprender la función de un gen. La función de un gen se puede determinar por la forma en que ocurren las mutaciones en los dos cromosomas. Si las mutaciones ocurren en el mismo cromosoma (en cis), solo se altera un gen, pero si las mutaciones están en trans, ambos genes se alteran. Ciertos eventos solo ocurren si las mutaciones están en trans y las proteínas que codifican los dos genes alterados no se producen debido a las mutaciones; este evento se conoce como heterocigosidad compuesta. Por otro lado, ciertos eventos conocidos como eventos reguladores cis solo ocurren si el gen está en cis. La información del genotipo no es suficiente para diferenciar si un gen está en cis o trans; Necesitamos conocer los haplotipos.

Finalmente, estudiar información de haplotipos de regiones genómicas que están relacionadas funcionalmente tiene muchas aplicaciones útiles. Ciertas regiones del genoma contienen grupos de genes que están relacionados funcionalmente. Estas regiones están asociadas con enfermedades autoinmunes e infecciosas. Conocer los haplotipos de estas regiones puede ayudar a unir donantes de órganos con receptores.

2.2. Modelos Ocultos de Markov (HMM)

Un modelo oculto de Markov (HMM) es un proceso estocástico generado por dos mecanismos probabilísticos interrelacionados. El primer proceso es un conjunto finito de estados, cada uno de ellos generalmente asociado a una distribución de probabilidad multidimensional.El segundo es aquel en que cualquier estado puede ser observado, es decir, analizaremos lo observado sin ver en qué estado está ocurriendo.

Cuando los tiempos son discretos, se supone que el proceso está en algún estado y la función aleatoria correspondiente al estado actual genera una observación. La cadena de Markov luego cambia su estado de acuerdo a su matriz de transición y el observador sólo ve la salida de las funciones aleatorias asociadas con cada estado, es decir, no puede observar directamente los estados de la cadena de Markov. Es de nuestro interés entonces trabajar con el tratamiento a los procesos de salida con un alfabeto finito discreto.

2.2.1. Definiciones

Un HMM puede verse como una familia de modelos para una secuencia de símbolos de un alfabeto $\mathcal{O} = \{o_1, o_2, \ldots, o_k\}$. El modelo se basa en la idea de una secuencia oculta de transiciones de estados que siguen una cadena de Markov. Más formalmente, un HMM se caracteriza por:

(I) **Cadena de Markov oculta:** una cadena de Markov $\{X_n\}_{n\geq 0}$ que toma valores en un espacio de estados finito $S = \{1, 2, \ldots, J\}$ con J estados. Las probabilidades de transición

$$\mathcal{Q}_{ij} = P\left(X_n = j \mid X_{n-1} = i\right), \quad n \geq 1, i, j \in S$$

son homogéneas. La matriz de transición es denotada por

$$\mathcal{Q} = (\mathcal{Q}_{ij})_{ij} \quad i, j = 1, \ldots, J$$

es tal que $\mathcal{Q}_{ij} \geq 0$ y $\sum_{j=1}^{J} \mathcal{Q}_{ij} = 1$ para todo $i \in S$. Al tiempo $n = 0$ el estado X_0 es especificado por

$$\pi_j(0) = P\left(X_0 = j\right) \quad \text{con} \quad \pi(0) = (\pi_1(0), \ldots, \pi_J(0)).$$

(II) **Un proceso aleatorio observable:** un proceso estocástico $\{Y_n\}_{n\geq 0}$ con espacio de estados finito $\mathcal{O} = \{o_1, o_2, \ldots, o_K\}$, donde K no es necesariamente igual a J. Los procesos $\{Y_n\}_{n\geq 0}$ y $\{X_n\}_{n\geq 0}$ están relacionados para cualquier n fijo por las distribuciones de probabilidad condicional

$$\mathcal{B}_j(k) = P\left(Y_n = o_k \mid X_n = j\right)$$

Sea

$$\mathcal{B} = (\mathcal{B}_j(k))_{j,k} \quad j = 1, \ldots, J; k = 1, \ldots, K$$

esta matriz es llamada la matriz de *probabilidad de emisión*. Esta es una matriz estocástica pues $\mathcal{B}_j(k) \geq 0$ y $\sum_{k=1}^{K} \mathcal{B}_j(k) = 1$.

(III) **Independencia condicional:** para cualquier secuencia de estados $j_0 j_1 \cdots j_n$, la probabilidad de la secuencia $o_0 o_1 \cdots o_n$ es

$$P\left(Y_0 = o_0, \ldots, Y_n = o_n \mid X_0 = j_0, \ldots, X_n = j_n, \mathcal{B}\right) = \prod_{l=0}^{n} \mathcal{B}_{j_l}(l)$$

La expresión anterior significa que los símbolos emitidos son condicionalmente independientes dada la secuencia de estados.

Una consecuencia de la condición (iii) es que el proceso $\{Y_n\}_{n\geq 0}$ siempre se puede representar como función de una cadena de Markov y, en general, no es una cadena de Markov. De esto también se sigue que podemos escribir la distribución de probabilidad conjunta de $o_0 \cdots o_n$ y $j_0 \cdots j_n$ como

$$\begin{aligned} P(\mathbf{Y}, \mathbf{X}; \mathcal{Q}, \mathcal{B}, \pi(0)) &= P(Y_0 = o_0, \ldots, Y_n = o_n, X_0 = j_0, \ldots, X_n = j_n; \mathcal{Q}, \mathcal{B}, \pi(0)) \\ &= P(Y_0 = o_0, \ldots, Y_n = o_n \mid X_0 = j_0, \ldots, X_n = j_n, \mathcal{B}) \\ &\qquad \times P(X_0 = j_0, \ldots, X_n = j_n, \mathcal{Q}, \pi(0)) \\ &= \pi_{j_0}(0) \prod_{i=0}^{n} \mathcal{B}_{j_i}(i) \prod_{i=1}^{n} \mathcal{Q}_{j_{i-1} j_i} \\ &= \pi_{j_0}(0) \mathcal{B}_{j_0}(0) \prod_{i=1}^{n} \mathcal{Q}_{j_{i-1} j_i} \mathcal{B}_{j_i}(i) \end{aligned}$$

Por tanto la probabilidad conjunta de $o_0 \cdots o_n$ estemos que

$$P(Y_0, \ldots, Y_n; \mathcal{Q}, \mathcal{B}, \pi(0)) = \sum_{j_0=1}^{J} \cdots \sum_{j_n=1}^{J} \pi_{j_0}(0) \mathcal{B}_{j_0}(0) \prod_{i=1}^{n} \mathcal{Q}_{j_{i-1} j_i} \mathcal{B}_{j_i}(i)$$

En consecuencia las distribuciones finito dimensionales de $\{Y_n\}_{n\geq 0}$ están completamente especificadas por nuestra elección de las matrices estocásticas $\mathcal{Q}, \mathcal{B}$ y la distribución inicial $\pi(0)$. Por tanto, podemos usar la notación compacta para el modelo

$$\lambda = (\mathcal{Q}, \mathcal{B}, \pi(0))$$

De todo esto tenemos que, para la familia de modelos condicionado a $\lambda = (\mathcal{Q}, \mathcal{B}, \pi(0))$, la cadena $\mathbf{o} = o_0 \cdots o_n$ tiene probabilidad

$$\begin{aligned} P(\mathbf{o}) &= P(Y_0 = o_0, \ldots, Y_n = o_n; \lambda) \\ &= \sum_{j_0=1}^{J} \cdots \sum_{j_n=1}^{J} P(Y_0 = o_0, \ldots, Y_n = o_n, X_0 = j_0, \ldots, X_n = j_n; \lambda) \\ &= \sum_{j_0=1}^{J} \cdots \sum_{j_n=1}^{J} \pi_{j_0}(0) \mathcal{B}_{j_0}(0) \prod_{i=1}^{n} \mathcal{Q}_{j_{i-1} j_i} \mathcal{B}_{j_i}(i) \end{aligned}$$

En la figura (2.3) cada nodo representa una variable aleatoria que describe el estado X_n o la observación Y_n en algún momento n y las flechas representan influencias directas.

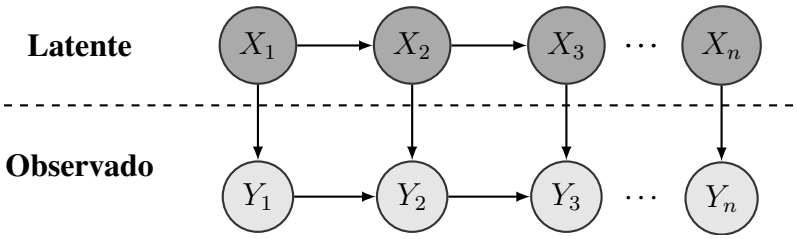


Figura 2.3: Descripción gráfica de un HMM.

2.3. Algoritmo Backward-Forward

Veremos en esta sección el algoritmo cuyo procedimiento computacional es conocido como el *algoritmo Backward-Forward*. El objetivo es por tanto calcular eficientemente $P(\mathbf{o} \mid \lambda)$, es decir,

$$P\left(Y_0, \ldots, Y_N; \lambda\right) = \sum_{j_0=1}^{J} \cdots \sum_{j_N=1}^{J} \pi_{j_0}(0)\mathcal{B}_{j_0}(0) \prod_{i=1}^{N} \mathcal{Q}_{j_{i-1}j_i}\mathcal{B}_{j_i}(i) \tag{2.1}$$

El algoritmo permite calcular la probabilidad descrita en la ecuación (2.1). En adelante obviaremos el término λ en la expresión. Los siguientes lemas cuya prueba se encuentra en (Koski, 2001), nos ayudarán para ciertos cálculos.

Lema 2.3.1. *Para todo $n = 0, 1, \ldots, N$, se cumple*

$$P\left(Y_0, \ldots, Y_N \mid X_n\right) = P\left(Y_0, \ldots, Y_n \mid X_n\right) P\left(Y_{n+1}, \ldots, Y_N \mid X_n\right)$$

Lema 2.3.2. *Para todo $n = 0, 1, \ldots, N$, se cumple*

$$P\left(Y_n, Y_{n+1}, \ldots, Y_N \mid X_n\right) = P\left(Y_n \mid X_n\right) P\left(Y_{n+1}, \ldots, Y_N \mid X_n\right)$$

Lema 2.3.3. *Para todo $n = 0, 1, \ldots, N-1$, se cumple*

$$P\left(Y_0, \ldots, Y_N \mid X_n, X_{n+1}\right) = P\left(Y_0, \ldots, Y_n \mid X_n\right) P\left(Y_{n+1}, \ldots, Y_N \mid X_{n+1}\right)$$

Lema 2.3.4. *Para enteros n y m tales que $0 \leq n \leq m \leq N$, se cumple*

$$P\left(Y_m, \ldots, Y_N \mid X_n, \ldots, X_m\right) = P\left(Y_m, \ldots, Y_N \mid X_m\right)$$

Veamos que

$$P\left(Y_0 = o_0, \ldots, Y_N = o_N\right) = \sum_{j=1}^{N} \alpha_n(j)\beta_n(j) \tag{2.2}$$

donde la variable *forward* $\alpha_n(j)$ se define como la probabilidad simultánea de que la secuencia emitida hasta un tiempo $n \leq N$ y la cadena oculta de Markov este en el estado j en el tiempo n, es decir

$$\alpha_n(j) = P\left(Y_0 = o_0, \ldots, Y_n = o_n, X_n = j\right)$$

y la variable *backward* $\beta_n(j)$ se define como la probabilidad de que la subsecuencia desde el tiempo $n+1$ hasta el final N condicionado a que la cadena oculta de Markov este en el estado j al tiempo n, esto es

$$\beta_n(j) = P\left(Y_{n+1} = o_{n+1}, \ldots, Y_N = o_N \mid X_n = j\right)$$

Tomaremos $\beta_N(j) = 1$ para todo j. Para ver la igualdad en (2.2), usaremos el lema (2.3.1), notemos que

$$\begin{aligned} P\left(Y_0 = o_0, \ldots, Y_N = o_N, X_n = j\right) &= P\left(X_n = j\right) P\left(Y_0 = o_0, \ldots, Y_N = o_N \mid X_n = j\right) \\ &= P\left(X_n = j\right) P\left(Y_0 = o_0, \ldots, Y_n = o_n \mid X_n = j\right) \\ &\quad \times P\left(Y_{n+1} = o_{n+1}, \ldots, Y_N = o_N \mid X_n = j\right) \\ &= P\left(Y_0 = o_0, \ldots, Y_n = o_n, X_n = j\right) \\ &\quad \times P\left(Y_{n+1} = o_{n+1}, \ldots, Y_N = o_N \mid X_n = j\right) \end{aligned}$$

Como $P\left(Y_0 = o_0, \ldots, Y_N = o_N\right) = \sum_{j=1}^{N} P\left(Y_0 = o_0, \ldots, Y_N = o_N, X_n = j\right)$, se sigue

$$= \sum_{j=1}^{N} \alpha_n(j)\beta_n(j)$$

Veamos ahora los algoritmos recursivos para calcular $\alpha_n(j)$ y $\beta_n(j)$.

2.3.1. Recursión Forward

El objetivo de esto es expresar $\alpha_{n+1}(j)$ en términos de $\alpha_n(i)$., en efecto,

$$\begin{aligned}\alpha_{n+1}(j) &= P\left(Y_0 = o_0, \ldots, Y_{n+1} = o_{n+1}, X_{n+1} = j\right) \\ &= \sum_{i=1}^{J} P\left(Y_0 = o_0, \ldots, Y_{n+1} = o_{n+1}, X_n = i, X_{n+1} = j\right) \\ &= \sum_{i=1}^{J} P\left(X_n = i, X_{n+1} = j\right) P\left(Y_0 = o_0, \ldots, Y_{n+1} = o_{n+1} \mid X_n = i, X_{n+1} = j\right)\end{aligned}$$

del lema (2.3.3), tenemos

$$\begin{aligned}&= \sum_{i=1}^{J} P\left(X_n = i, X_{n+1} = j\right) P\left(Y_0 = o_0, \ldots, Y_n = o_n \mid X_n = i\right) \\ &\qquad \times P\left(Y_{n+1} = o_{n+1} \mid X_{n+1} = j\right) \\ &= \sum_{j=1}^{J} P\left(Y_0 = o_0, \ldots, Y_n = o_n, X_n = i\right) \mathcal{Q}_{ij} \mathcal{B}_j(o_{n+1})\end{aligned}$$

ya que $P\left(X_n = i, X_{n+1} = j\right) P\left(Y_{n+1} = o_{n+1} \mid X_{n+1} = j\right) = \mathcal{Q}_{ij}\mathcal{B}_j(o_{n+1}) P\left(X_n = i\right)$

$$\begin{aligned}&= \sum_{i=1}^{J} \alpha_n(i) \mathcal{Q}_{ij} \mathcal{B}_j(o_{n+1}) \\ &= \left(\sum_{i=1}^{J} \alpha_n(i) \mathcal{Q}_{ij} \right) \mathcal{B}_j(o_{n+1})\end{aligned}$$

De esta manera tenemos el algoritmo Forward:

Algoritmo 2 Algoritmo Forward

1: Considerar la variable forward $\alpha_n(j) = P\left(Y_0 = o_0, \ldots, Y_n = o_n, X_n = j \mid \lambda\right)$.
2: **Iniciar:** $\alpha_0(j) = \mathcal{B}_j(o_0)\pi_j(0)$ para $j = 1, \ldots, J$.
3: **Recursión:** $\alpha_{n+1}(j) = \left(\sum_{i=1}^{J} \alpha_n(i)\mathcal{Q}_{ij}\right) \mathcal{B}_j(o_{n+1})$, para $j = 1, \ldots, J, \quad 1 \leq n \leq N-1$

2.3.2. Recursión Backward

Por definición tenemos

$$\begin{aligned}\beta_n(j) &= P\left(Y_{n+1} = o_{n+1}, \ldots, Y_N = o_N \mid X_n = j\right) \\ &= \frac{P\left(Y_{n+1} = o_{n+1}, \ldots, Y_N = o_N, X_n = j\right)}{P\left(X_n = j\right)} \\ &= \frac{\sum_{i=1}^{J} P\left(Y_{n+1} = o_{n+1}, \ldots, Y_N = o_N, X_n = j, X_{n+1} = i\right)}{P\left(X_n = j\right)} \\ &= \frac{\sum_{i=1}^{J} P\left(Y_{n+1} = o_{n+1}, \ldots, Y_N = o_N \mid X_n = j, X_{n+1} = i\right) P\left(X_n = j, X_{n+1} = i\right)}{P\left(X_n = j\right)}\end{aligned}$$

Del lema (2.3.4) tenemos que

$$P\left(Y_{n+1} = o_{n+1}, \ldots, Y_N = o_N\right) = P\left(Y_{n+1} = o_{n+1}, \ldots, Y_N = o_N \mid X_{n+1} = i\right)$$

y del lema (2.3.2) se sigue que

$$\begin{aligned}&= P\left(Y_{n+1} = o_{n+1} \mid X_{n+1} = i\right) \\ &\quad \times P\left(Y_{n+2} = o_{n+2}, \ldots, Y_N = o_N \mid X_{n+1} = i\right)\end{aligned}$$

por tanto

$$\begin{aligned}\beta_n(j) &= \sum_{i=1}^{J} P\left(Y_{n+1} = o_{n+1} \mid X_{n+1} = i\right) P\left(Y_{n+2} = o_{n+2}, \ldots, Y_N = o_N \mid X_{n+1} = i\right) \mathcal{Q}_{ji} \\ &= \sum_{i=1}^{J} \mathcal{B}_i(o_{n+1}) \beta_{n+1}(i) \mathcal{Q}_{ji}\end{aligned}$$

De esto obtenemos el siguiente algoritmo:

Algoritmo 3 Algoritmo Backward

1: Considerar la variable backward $\beta_n(j) = P\left(Y_{n+1} = o_{n+1}, \ldots, Y_N = o_N \mid X_n = j, \lambda\right)$.
2: **Iniciar:** $\beta_N(j) = 1$, para $j = 1, \ldots, J$.
3: **Recursión:** $\beta_n(j) = \sum_{i=1}^{J} \mathcal{B}_i(o_{n+1}) \beta_{n+1}(i) \mathcal{Q}_{ji}$, para $j = 1, \ldots, J, \quad n = N-1, N-2, \ldots, 0$

2.4. Algoritmo fastPHASE

La fase del haplotipo es el problema de inferir información sobre el haplotipo de un individuo. Para resolver este problema, hay muchos métodos. El algoritmo *fastPHASE* es un algoritmo probabilístico que sirve para encontrar la fase del haplotipo, lo que significa determinar haplotipos a partir de genotipos. El supuesto básico del algoritmo es que si se conocen los genotipos de los padres, entonces la fase del haplotipo de la descendencia generalmente se puede determinar. La tarea que se le asigna a dos haplotipos parentales determina el posible haplotipo infantil, también conocido como la fase. Determinar la fase de la descendencia puede contribuir a imputar con alta probabilidad el valor que podría faltar en los genotipos de la descendencia. Algo que se asume en este tipo de problemas es el *equilibrio de Hardy-Weinberg* (HWE). Este afirma que las proporciones de marcadores SNP, permanecen constantes de generación en generación. En conjunto de datos GWAS se asume este equilibrio pues el conjunto de datos es fijo e inmutable de las secuencias de genotipo SNP registradas en varios individuos. Uno de los benerificios de suponer HWE es que en presencia de proporciones alélicas constantes, también están presentes proporciones constantes de haplotipo, lo que facilita la previsibilidad de un haplotipo que ocurre dentro de un conjunto de datos y, por lo tanto, la aparición de genotipos SNP específicos dentro de ellos.

Para entender un poco mejor cómo funciona el algoritmo, supongamos que se nos da una matriz de datos con secuencias SNP de muchas personas. Agregar nuevos individuos o modificar SNPs no vacíos no está permitido, es decir, el conjunto de datos no cambia, esto también supone que la generación futura mantendrá las mismas proporciones que las generaciones anteriores, entonces debido a que el conjunto de datos no cambiará, podemos suponer HWE.

Para el algoritmo fastPHASE, la utilidad de HWE es grande, pues con esta podemos comprender las tasas de herencia de los marcadores SNP. Como conocemos las proporciones alélicas, también conoceremos las proporciones del haplotipo, por tanto, no surgirán nuevos haplotipos de la combinación de ninguno de los padres en la población. Como conocemos las proporciones del haplotipo, conocemos sus frecuencias y las estimaciones de probabilidad de que ocurran dentro de una población. Dadas las probabilidades de los haplotipos individuales, podemos inferir mejor un haplotipo completo o parcialmente perdido al encontrar la probabilidad de que la secuencia conocida dentro del haplotipo parcialmente perdido, ocurra en cualquiera de los haplotipos conocidos.

El modelo estadístico se basa en la idea de que, en regiones cortas de ADN, los haplotipos en una población tienden a agruparse en grupos de haplotipos similares, entonces entre más cortos son los haplotipos, más comunes pueden ser en una población. En (Scheet and Stephens, 2006) observaron que haplotipos cortos observables similares parecen agruparse. Como resultado de la recombinación, los haplotipos que están estrechamente relacionados entre sí y similares variarán a medida que uno se mueve a lo largo del cromosoma, es decir, esta agrupación tiende a ser de naturaleza local. Se puede pensar que cada grupo representa (localmente) un haplotipo común, o una combinación de alelos, y la suposición HMM para la pertenencia al grupo da como resultado que cada haplotipo observado se modele como un mosaico de un número limitado de haplotipos comunes. La figura (2.4) muestra esto, cada columna representa un marcador SNP, con los dos alelos indicados por cuadrados con cruz y sin cruz. Pares sucesivos de filas representan el par estimado de haplotipos para individuos sucesivos.

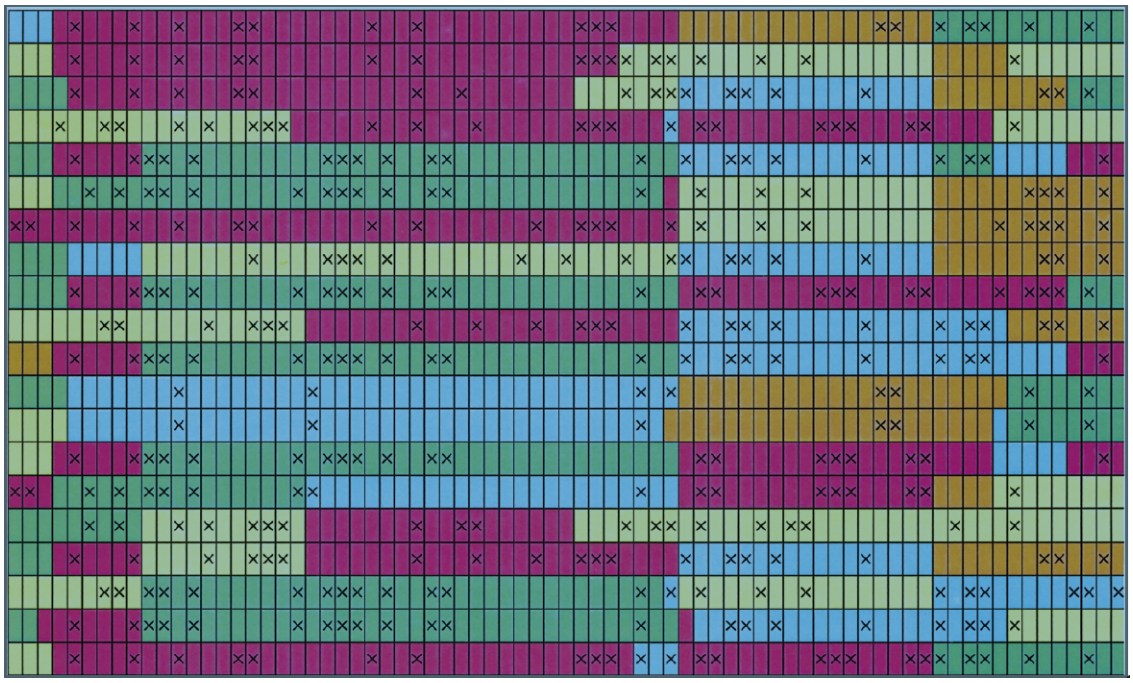


Figura 2.4: Ilustración del algoritmo fastPHASE[1]

Los colores representan la membresía estimada del grupo de cada alelo, que cambia a medida que uno se mueve a lo largo de cada haplotipo. Localmente, se puede pensar que

[1]Imagen tomada de (Scheet and Stephens, 2006).

cada grupo representa una combinación (común) de alelos en SNP estrechamente vinculados, y la figura ilustra cómo cada haplotipo se modela como un mosaico de estas combinaciones comunes.

El propósito de modelos como fastPHASE es capturar patrones complejos de correlación entre marcadores densos en muestras. Modelos de este tipo que emplean correlación de marcadores se prueban en los datos del genotipo para ver si el patrón de variación se ha capturado con precisión. Se prueba contra datos de genotipo de matriz con valores de genotipo faltantes. El modelo primero estima los genotipos faltantes e infiere la fase del haplotipo a partir de los datos del genotipo sin fase. Como se trata de un método en fases, analiza los alelos presentes en los dos haplotipos de un par cromosómico homólogo, es decir, modela el genotipo de cada par de alelos individuales. Hay muchos haplotipos en un cromosoma. Esta secuencia de haplotipos trasciende a una secuencia de grupos que pueden ser modelados por un HMM. Cada grupo representa un haplotipo resumido para el grupo de haplotipos similares encontrados dentro del conjunto de datos completo. La figura (2.5) muestra haplotipos similares basados en alelos comunes (sombreados en gris) y un resumen de haplotipos similares, dentro de la estructura de un grupo. El grupo definido por las secuencias más largas observadas de genotipos coincidentes, para este caso GGGGGAA.

Cada grupo se ve como un haplotipo modelo, es decir, un resumen de los haplotipos similares que contiene. Cuando se visualizan haplotipos a lo largo de un cromosoma, cada haplotipo es miembro de algún grupo de haplotipos definido. La suposición de HMM para las secuencias de marcadores de haplotipos en un grupo de haplotipos observados da como resultado que cada haplotipo observado se origine a partir de un número finito de otras secuencias de haplotipos observadas similares. Este modelo se modifica para permitir que la membresía del grupo cambie a lo largo de cada haplotipo y así capturar el hecho de que, aunque los haplotipos muestreados exhiben patrones similares a los del grupo, tienden a ser de naturaleza local.

Las secuencias completas de SNPs no forman haplotipos. Dado el grupo de origen de cada haplotipo, se supone que los alelos en cada marcador del haplotipo observado son muestras independientes basados en las frecuencias alélicas del marcador del grupo de origen correspondiente, representadas por θ como veremos. El número de haplotipos únicos es considerablemente menor que el número de haplotipos observados en todas las filas de los datos. Por lo tanto, los haplotipos observables pueden tener alguna probabilidad de ocurrencia dentro de los datos. Esta probabilidad está asociada con la probabilidad de su respectivo grupo de

G	T	G	G	G	A	A
G	G	G	G	G	A	A
G	G	G	G	G	A	A
G	G	G	G	G	A	T
C	G	G	G	G	A	T
C	G	A	G	G	A	T
C	T	A	G	G	A	A
C	C	G	G	G	G	T
G	G	G	G	G	A	T
C	T	A	G	G	A	A
C	G	A	G	G	A	A
G	G	G	G	G	A	A
C	T	G	G	G	A	T
G	G	A	G	G	A	A
C	T	G	A	G	A	A
C	G	G	G	G	A	A
G	T	G	G	G	A	A
C	G	G	G	G	A	A

(a) Haplotipos

G	G	G	G	G	A	A
G	G	G	G	G	A	A
G	G	G	G	G		
G	G	G	G	G	A	
G	G	G	G	G		
G	G	G	G	G		A
	G	G	G	G	A	A
	G	G	G	G	A	A

(b) Agrupamiento

Figura 2.5: Agrupamiento de haplotipos similares.

origen.

Vamos ahora a formalizar un poco el modelo. Sea h el conjunto de todos los haplotipos observados dentro de la matriz de datos, $h = (h_1, \ldots, h_n)$, con n el número de haplotipos observados presentes. Si un haplotipo observado aparece más de una vez dentro del conjunto de datos, se dice que la secuencia del alelo SNP del haplotipo específico es un haplotipo único y cualquier aparición de esta secuencia dentro del conjunto de datos se dice que es un haplotipo observado o una instancia. Cada haplotipo es una secuencia de valores SNP en las ubicaciones de los M marcadores, donde cada h_{im} denota el valor del alelo en el i-ésimo haplotipo observado en el marcador m, entonces $h_i = (h_{i1}, \ldots, h_{im})$ denota un haplotipo que consiste en m marcadores. Los grupos de haplotipos se definen por un haplotipo único. Cada haplotipo único es la etiqueta de un grupo individual, por ejemplo, la primera fila del

agrupamiento en la (2.5).

Cada grupo contiene las etiquetas que corresponden al haplotipo único. De esta manera, podemos extraer una medida de probabilidad de pertenecer a un haplotipo específico observable dentro de un grupo, es decir, una frecuencia de haplotipo. Los marcadores que se encuentran en sus ubicaciones h_{im} son *bialélicos*, lo que significa que pueden tomar uno de los dos valores 0 o 1. Cada haplotipo observable muestreado se origina en uno de los K grupos posibles. Al comienzo de la imputación, no hay información disponible sobre grupos o haplotipos. Los grupos se definen sobre la marcha utilizando el *algorítmo EM* para la estimación de los parámetros.

Durante la ejecución de fastPHASE, los haplotipos sin fase a los que no se les podía asignar membresía de grupo determinan sus grupos de origen mediante un HMM. El algoritmo usa un conjunto de datos de alelos en lugar de un conjunto de datos de genotipo, donde cada dos filas consecutivas del conjunto de datos se asemejan a las secuencias de alelos homólogas de un solo individuo, es decir, cada par de alelos alineados verticalmente, uno de cada fila, se compone de un genotipo SNP como se observa en la figura (2.4).

Cada alelo se trata de forma independiente, es decir, se permite que dos alelos de un genotipo pertenezcan a dos haplotipos diferentes, por lo que pueden tener diferentes grupos de origen. La variable z_i es una variable que denota el grupo de origen para el haplotipo h_i. El vector $\alpha = (\alpha_1, \ldots, \alpha_K)$ es el vector de frecuencias de los grupos:

$$\alpha_k = \frac{|k|}{\sum_{i \in K} |i|}$$

Esta frecuencia relativa es el número de haplotipos observados en el grupo divido entre todos los haplotipos observados en todo el conjunto de datos. Ahora bien, la probabilidad de que un grupo de origen sea k, corresponde a la frecuencia del grupo, esto es, $P(z_i = k \mid \alpha) = \alpha_k$, donde $\alpha = (\alpha_1, \ldots, \alpha_K)$.

Teniendo en cuenta que los SNP de los haplotipos en fastPHASE son de naturaleza bialélica, se implementa un esquema de codificación, donde 0 se usa para el genotipo (C-G) y 1 para (A-T) o viceversa. Denotaremos el el vector θ como el que registra las frecuencias del alelo 1 en todas las posiciones de cada marcador y para todos los grupos. Entonces, la frecuencia del alelo 1 en el grupo k en la posición del mercador m se denota por θ_{km}. Para cualquier grupo de haplotipos dado, estas frecuencias están muy cercanas a 0.01 o 0.99. Esto nos permite resumir o ver el grupo como un haplotipo, por tanto, se dice que un haplotipo único como se

mencionó anteriormente etiqueta el grupo, por tanto, la etiqueta es un haplotipo, que denota el grupo y denota todos sus miembros.

Ahora, dado el origen de cada haplotipo, los alelos observados en cada marcador son muestras independientes de frecuencias alélicas específicas del grupo, denotado por θ, es decir,

$$P(h_i \mid z_i = k, \theta) = \prod_{m=1}^{M} \theta_{km}^{h_{im}} (1 - \theta_{km})^{1-h_{im}} \tag{2.3}$$

La ecuación (2.3) es la probabilidad de seleccionar h_i, dado que tenemos el grupo k como nuestro grupo de origen y θ_k de θ para nuestras frecuencias alélicas. Esta distribución puede mostrar las asociaciones entre alelos en las posiciones de marcador vecinas, donde $h_{im} \in \{0, 1\}$ y $\theta_{km}^{h_{im}} \in \{1, \theta_{km}\}$. Si $h_{im} = 1$ entonces $P(h_i \mid z_i = k, \theta) = \theta_{km}(1)$ y si $h_{im} = 0$ entonces $P(h_i \mid z_i = k, \theta) = 1 - \theta_{km}$. Esto significa que cuando h_{im} es 0, la frecuencia del valor 0 en posición del marcador m para el grupo de origen $z_i = k$ es igual a $1 - \theta_{km}$, donde θ_{km} representa la frecuencia del valor alélico 1 en la posición del marcador m para el grupo k.

Notemos que en la ecuación (2.3) estamos suponiendo que conocemos el grupo de origen z_i para h_i. Por lo tanto, debe determinarse. La única forma en que podemos determinar si z_i es realmente el grupo de origen de h_i es probar todos los grupos posibles. Como el grupo de origen del haplotipo no se conoce, entonces tenemos la siguiente ecuación

$$P(h_i \mid \alpha, \theta) = \sum_{j=1}^{K} P(z_i = j \mid \alpha) P(h_i \mid z_i = j, \theta) = \sum_{j=1}^{K} \alpha_j \prod_{m=1}^{M} \theta_{jm}^{h_{im}} (1 - \theta_{jm})^{1-h_{im}} \tag{2.4}$$

Sabemos entonces que el grupo de origen z_i para algún haplotipo h_i es k cuando la regla del producto entrega un resultado final de uno, dado el grupo k.

2.4.1. HMM para fastPHASE

El algoritmo fastPHASE utiliza HMM para modelar el hecho de que los alelos en los marcadores cercanos probablemente surgen del mismo grupo o no. La cadena de Markov del modelo HMM, es modelada por los estados ocultos z_i que corresponden a los grupos de origen, mientras que el valor del marcador es el estado observado del HMM. Cada alelo en cada marcador puede tener su propio grupo de origen, por tanto, se supone que $z_i = (z_{i1}, \ldots, z_{iM})$ forma una cadena de Markov en $\{1, \ldots, K\}$ grupos.

z_{i1}	$*$	$*$	z_{im}	$*$	$*$	z_{iM}
1	1	1	1	1	1	1
$*$	$*$	$*$	$*$	$*$	$*$	$*$
$*$	$*$	$*$	$*$	$*$	$*$	$*$
K	K	K	K	K	K	K

Tabla 2.2: Tabla de grupos de origen que se asemejan a la cadena de Markov.

La tabla (2.2) muestra que para cada marcador, hay un grupo de origen que corresponden a uno de los K posibles grupos. En ella, la primera fila z_{im} para $1 \leq m \leq M$, representa el grupo de origen del alelo m del haplotipo i, es decir, z_{im} representa el grupo de origen del marcador h_{im}. Para cada z_{im} hay K posibles grupos para elegir. Encontrar el grupo de origen para M marcadores y K grupos tiene un costo de computacional del orden de $O(nMK)$, donde n es el número de haplotipos observados dentro del conjunto de datos, M es el número de marcadores por haplotipo y K es el número total de grupos.

Ahora bien, supongamos que z_{im} denota el grupo de origen para el marcador h_{im}, y que $z_i = (z_{i1}, \ldots, z_{iM})$ forma una cadena de Markov en $\{1, \ldots, K\}$ con probabilidades iniciales

$$P(z_{i1} = k) = \alpha_{k1}$$

que corresponde a probabilidad de comenzar con el alelo h_{i1} en el grupo k. Las probabilidades de transición son $P_m(k \to k')$ dada por

$$\begin{aligned} P_m(k \to k') &= P\left(z_{im} = k' \mid z_{i(m-1)} = k, \alpha, r\right) \\ &= \begin{cases} e^{-r_m d_m} + (1 - e^{-r_m d_m})\alpha_{k'm} & si \quad k' = k \\ (1 - e^{-r_m d_m})\alpha_{k'm} & si \quad k' \neq k \end{cases} \end{aligned}$$

Expliquemos un poco lo anterior: la probabilidad de estado inicial corresponde a la frecuencia del grupo. Los grupos de origen de cada alelo siguiente dependen del grupo de origen del presente alelo. Si el grupo de origen del alelo siguiente no es el mismo que el del alelo actual, entonces se toma la medida de probabilidad inferior $(1 - e^{-r_m d_m})\alpha_{k'm}$. Si el grupo de origen del alelo siguiente es el mismo que el del alelo actual, entonces se toma la medida de probabilidad $e^{-r_m d_m} + (1 - e^{-r_m d_m})\alpha_{k'm}$.

La variable d_m que especifica la distancia física entre los marcadores $m - 1$ y m en el cromosoma es conocida y $r = (r_1, \ldots, r_M)$ y α_{km} son parámetros a estimar. El valor

r_m es una tasa promedio a la cual m y $m+1$ no están asociados. Este valor, se considera informalmente como la tasa de recombinación entre m y $m+1$. Como se menciona en (Scheet and Stephens, 2006), los valores de r_m se pueden configurar para que sean todos iguales, constantes, y si las distancias entre m y $m+1$ no se conocen, el parámetro d_m se puede eliminar.

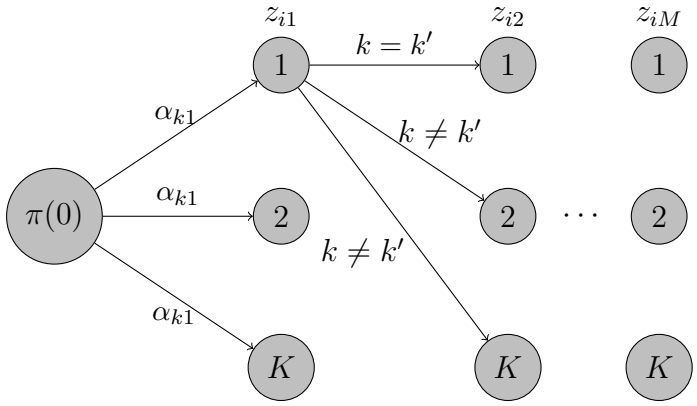


Figura 2.6: HMM para fastPHASE.

La figura (2.6) muestra el flujo probabilístico a medida que cambian los estados de alelo a alelo dentro de un haplotipo. Los cambios de estado pueden conducir a cambios de grupos donde el grupo de origen del estado actual k no es el grupo de origen del alelo en el siguiente estado k'.

2.4.2. HMM para dos alelos: el genotipo

Teniendo en cuenta que estamos tratando con genotipos y los genotipos son tuplas de alelos, se extiende de manera natural el HMM para aceptar dos grupos de origen para un genotipo. Sea $g = (g_1, \ldots, g_n)$ el conjunto de datos de genotipo de n individuos diploides, donde $g_i = (g_{i1}, \ldots, g_{iM})$ y cada $g_{im} \in \{0, 1, 2\}$ representa el genotipo en el marcador m para el individuo i. La codificación de estos valores es que 0 corresponde genotipos homocigotos X_aX_a, 1 para heterocigotos X_aX_b o X_bX_a y 2 para homocigotos X_bX_b. Dado que los genotipos son una tupla de alelos que no necesariamente se originan en el mismo grupo, se deben determinar los grupos de origen para ambos alelos marcadores SNP de un genotipo.

Los dos grupos de origen para un genotipo están representados por una tupla: $\dot{z} = (k_1, k_2)$, este es un par de grupos no necesariamente ordenados, a partir del cual se origina g_{im}. Para

este caso, el tiempo de ejecución para determinar los grupos de origen de los haplotipos para los genotipos SNP es del orden $O(K^2M)$. Para un genotipo $g = (g_1, \ldots, g_n)$, $\dot{z}$ es el conjunto de tuplas de agrupación de origen para cada genotipo en los marcadores 1 a M para el individuo i. Por tanto $\dot{z} = (\dot{z}_{i1}, \ldots, \dot{z}_{iM})$ forma una cadena de Markov con probabilidades de estado iniciales

$$P(\dot{z}_{i1} = \{k_1, k_2\}) = \begin{cases} (\alpha_{k_1 1})^2 & si \quad k_1 = k_2 \\ 2\alpha_{k_1 1}\alpha_{k_2 1} & si \quad k_1 \neq k_2 \end{cases}$$

y probabilidades de transición

$$P_m(\{k_1, k_2\} \to \{k'_1, k'_2\}) \begin{cases} P_m(k_1 \to k'_1)P_m(k_2 \to k'_2) + P_m(k_1 \to k'_2)P_m(k_2 \to k'_1) & \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{si} \quad k_1 \neq k_2 \text{ y} k'_1 \neq k'_2 & \\ P_m(k_1 \to k'_1)P_m(k_2 \to k'_2) \qquad\qquad \text{otro caso} & \end{cases}$$

Dado que cada uno de los dos alelos de un genotipo tiene un grupo de origen independiente, entonces la probabilidad de muestrear los dos es el producto de las probabilidades de muestrear los dos alelos. La probabilidad de extraer un alelo de un grupo de origen, k, corresponde a la frecuencia α_k.

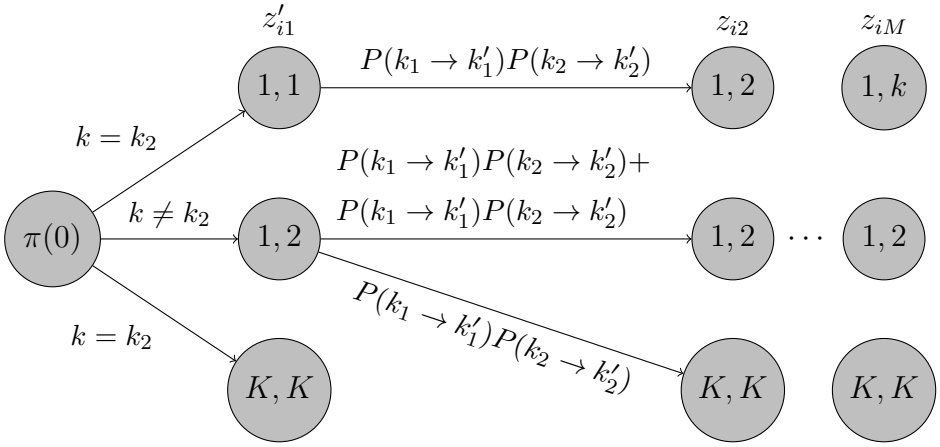


Figura 2.7: HMM para dos alelos.

La Figura (2.7) muestra un HMM y cómo la pertenencia a un grupo de origen de genotipo, atraviesa el modelo probabilístico de Markov a medida que observamos cada tupla de alelo de cada genotipo sucesivo a lo largo del cromosoma. A medida que se atraviesa el árbol y

se dibuja una ruta, cualquier ruta horizontal recta entre dos o más nodos, que representan los estados, significa que los grupos de origen de los genotipos sucesivos son los mismos. La presencia de rutas en zigzag entre dos o más nodos significa que los grupos de origen para los genotipos son diferentes. Estos modelos son de gran ayuda para resolver la fase de haplotipo faltante o para determinar las propiedades hereditarias, como el enlace y la asociación entre los alelos SNP que comprenden los haplotipos individuales y la asociación y relación entre los propios haplotipos homólogos.

El problema del valor del genotipo faltante plantea grandes dificultades para los mecanismos de reconstrucción de haplotipos como estos. A lo largo del recorrido, se pueden encontrar genotipos faltantes. Cuando se encuentran valores perdidos, se debe modificar el modelo probabilístico establecido para la reconstrucción del haplotipo. Esta modificación lleva demasiado tiempo y, teniendo en cuenta la prevalencia de datos faltantes dentro de dichos conjuntos de datos, la reducción de los tamaños de los conjuntos de datos solo disminuye la precisión de fase del modelo.

La imputación del genotipo faltante es la mejor solución. Esta se realiza utilizando un paso de maximización que se muestra a continuación, donde la probabilidad de un genotipo $g_{im} = x$ dado el conjunto de genotipos g y los parámetros $v = (\theta, \alpha, r)$ es igual a la probabilidad condicional (ecuación (2.5)) de g_{im} dado el(los) grupo(s) de origen $\dot{z_{im}}$ del genotipo g_{im}. Para recapitular, aquí θ es la tabla de frecuencias del alelo menor seleccionado "1" para cada haplotipo reconocido de SNP, es decir, el alelo que aparece con menos frecuencia en cada marcador. El parámetro α es el conjunto de frecuencias de grupo para cada grupo, es decir, el número de instancias de haplotipo que cada grupo ha dividido por el número total de instancias de haplotipo para todos los grupos, y r es la velocidad de salto de los alelos marcadores; este parámetro generalmente se silencia en fastPHASE. El conjunto mencionado de parámetros v se estima a partir de los datos dados utilizando el algoritmo EM.

Sin embargo, dado que los grupos de origen de los genotipos faltantes no se pueden determinar explícitamente, se tienen en cuenta todos los grupos posibles. El grupo que produce una probabilidad $p(g_{im} = x) = 1$ es el grupo de origen que luego se utilizará para maximizar el valor de x

$$P(g_{im} = x \mid g, v) = \sum_{k_1=1}^{K} \sum_{k_2=k_1}^{K} P(g_{im=x|\dot{z}_{im}} = \{k_1, k_2\}, v) p(\dot{z}_{im} = \{k_1, k_2\} \mid g_i, v) \quad (2.5)$$

Debemos tener en cuenta que se nos da una secuencia de genotipo g_i. Dado este hecho, entonces por medio del HMM, podemos determinar si para g_i, en la posición $g_{im}, \dot{z}_{im}$, los grupos de origen han cambiado o no de los del genotipo en g_{im-1}, dado que se observa g_{im-1}. Este supuesto es posible gracias al HMM impuesto al conjunto de datos.

Usando los grupos de genotipos de origen conocidos en la secuencia g_i, podemos suponer que los grupos de origen son probablemente los mismos para g_{im-1} y g_{im}. Esta suposición se pone a prueba probando todos los grupos posibles para $\dot{z}_{im}$. Una vez que se ha encontrado el grupo, se determina una estimación del valor del genotipo para g_{im}. Los genotipos como recordamos están codificados por los valores $\{0, 1, 2\}$. La variable x en $p(g_{im} = x \mid g, v)$ puede asumir cualquiera de estos tres valores en la ecuación (2.6). Cada uno de estos valores se introduce individualmente en la función probabilística anterior a la ecuación (2.5). Luego se selecciona el valor x que produce la probabilidad más grande o máxima. Este proceso se repite T veces, donde en (Scheet and Stephens, 2006), T se hizo constante al valor de $T = 20$. En cada T, se proporciona un conjunto diferente de estimaciones de parámetros v. Al promediar $T = 20$ de este procedimiento de maximización, la ecuación (2.5), se logró una mejor estimación para g_{im} el genotipo faltante

$$\hat{g}_{im} = \underset{x \in \{0,1,2\}}{\arg\max} \frac{1}{T} \sum_{t=1}^{T} P(g_{im} = x \mid g, \hat{v}_t) \tag{2.6}$$

es la mejor estimación del genotipo por maximización en T conjuntos de parámetros v posibles. Las corridas de T en el algoritmo EM se usan para estimar los parámetros, mientras que la ecuación (2.6) se utiliza para maximizar el valor esperado del genotipo faltante g_{im}.

CAPÍTULO 3

MODELO KNOCKOFFS-X

3.1. Introducción

Aplicaciones en las cuales tenemos un gran conjunto de posibles covariables o variables a una una respuesta de forma lineal o no, aparecen con mucha frecuencia. Este problema de modelado se ha estudiado de manera basta, sin embargo, no está claro cómo controlar la fracción de falsos descubrimientos de manera efectiva. Para abordar este problema Candes et al. (2018) propusieron el modelo Knockoffs-X el cual proporciona una inferencia válida de muestras finitas en situaciones donde la distribución condicional de la respuesta es arbitraria y desconocida, además sin importar el número de covariables. Esta inferencia es basada en un enfoque probabilístico y no geométrico. El único requisito del método es que las covariables sean aleatorias i.i.d con una distribución conocida, aunque el procedimiento es robusto con distribuciones desconocidas o estimadas como se muestra en Candes et al. (2018).

Supongamos que la variable Y depende de manera arbitraria de las covariables $X_1, \ldots, X_p$. Se impone la restricción que $(X_{i1}, \ldots, X_{ip}, Y)$ sean i.i.d, es decir,

$$(X_{i1}, \ldots, X_{ip}, Y) \sim F_{XY}, \quad i = 1, \ldots, n$$

Tengamos en cuenta que no es necesario tener conocimiento de la distribución condicional de $Y \mid X_1, \ldots, X_p$, pero sí que la distribución conjunta F_X de las covariables es conocida. En el problema de selección de variables, nos interesa un procedimiento que pueda controlar

un error de tipo I, es decir, controlar el error que se comete cuando se rechaza la hipótesis nula siendo esta verdadera, (controlar los falsos positivos). Por tanto, nos gustaría encontrar tantas variables como sea posible sin tener muchos resultados de falsos positivos. Para esto, nos interesa controlar la tasa de falsos descubrimientos (FDR por sus siglas en inglés).

3.2. Modelo Knockoffs X

Introduciremos algunas definiciones y condiciones para la construcción de las variables Knockoffs.

Definicion 3.2.1 (**Manta de Markov).** Al menor subconjunto $\mathcal{S}$ tal que condicionando en $\{X_j\}_{j\in\mathcal{S}}$ a Y es independente a todas las otras variables es conocido como la *manta de Markov*.

Observación. Para una familia de variables aleatorias, la manta de Markov puede no ser única. Supongamos X_1, X_2 vectores aleatorios i.i.d Gausianos y sea $X_3 = X_1 - X_2$. Supongamos que Y depende de X a través de $X_1 + X_2$, por ejemplo, $Y \mid X \sim N(X_1 + X_2, 1)$. Podemos observar aquí que, la verosimilitud de Y depende de X a través de $(X_1, X_2), (X_1, X_3)$ o (X_2, X_3) e igual son subconjuntos igualmente buenos, lo que significa que la manta no es única.

Como buscamos un subconjunto de variables relevantes que sea único, podemos modificar un poco la definición.

Definicion 3.2.2 (**Variable Nula).** Una variable X_j se dice que es *nula* si y sólo si Y es independiente de X_j condicionada a las demás variables $X_{-j} = \{X_1, \ldots, X_p\} \setminus \{X_j\}$, es decir, $Y \perp\!\!\!\perp X_j \mid X_{-j}$. El conjunto de índices de todas las variables nulas, será denotado por $\mathcal{N} \subset \{1, \ldots, p\}$ y diremos que una variable X_j es *relevante* si $j \notin \mathcal{N}$.

Esta definición significa que una variable es nula si no tiene poder predictivo una vez que tenemos en cuenta todas las demás variables, es decir, no influye en la respuesta de ninguna manera.

El *objetivo* es entonces descubrir la mayor cantidad de variables relevantes, pero además mantener el FDR controlado. Matemáticamente si tenemos una regla de selección que elija

un subconjunto $\hat{\mathcal{S}}$ de covariables relevantes, entonces

$$\text{FDR} = E\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{N}|}{|\hat{\mathcal{S}}|}\right]$$

Procederemos ahora a la construcción del modelo. Sea $X = (X_1, \ldots, X_p)$ una familia de variables aleatorias, construiremos una nueva familia $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ que satisface las dos siguientes propiedades:

1. Para cualquier $S \subset \{1, \ldots, p\}$, se cumple

$$(X, \tilde{X})_{sw(S)}{}^{1} \stackrel{d}{=} (X, \tilde{X}) \tag{3.1}$$

2. Si existe una respuesta Y

$$\tilde{X} \perp\!\!\!\perp Y \mid X \tag{3.2}$$

La condición de intercambiabilidad (3.1) dice que la distribución de $(X, \tilde{X})$ es invariante bajo esta transformación. Como discutiremos más adelante, esta condición es esencial y no siempre es fácil producir un vector no trivial, es decir, diferente de X mismo, que lo satisfaga. Para aclarar un poco la notación, esta condición significa que para cada $j \in S$, vamos a intercambiar las columnas X_j y $\tilde{X}_j$, por ejemplo, si $p = 3$ y $S = \{2, 3\}$, entonces

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{sw(\{2,3\})} \stackrel{d}{=} (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3).$$

Una consecuencia inmediata de (3.1) es que las variables Knockoffs nulas $\{\tilde{X}_j\}_{j \in \mathcal{N}}$ se distribuyen igual que las nulas originales $\{X_j\}_{j \in \mathcal{N}}$, pero se preserva una dependencia, a saber, para $j \neq k$ donde $k \in \mathcal{N}$, entonces

$$(X_j, \tilde{X}_k) \stackrel{d}{=} (X_j, X_k)$$

Esto significa que si conocemos F_X entonces el primer paso para construir una distribución para $\tilde{X}$ condicionado a X tal que se cumpla (3.1), es construir $F_{\tilde{X}|X}(\cdot \mid x)$ de modo que la conjunta $(X, \tilde{X})$ sea igual a $F_X(x)F_{\tilde{X}|X}(\tilde{x} \mid x)$ y sea simétrica tal que satisfaga (3.1).

La condición (3.2) se conoce como la condición de nulidad, pues esta implica que todas las variables *knockoff* son nulas en el modelo aumentado que incluye tanto a X como a

[1] $sw(S)$ significa swap (intercambiar), con esto nos referimos a intercambiar las columnas del conjunto S.

$\tilde{X}$. Una propiedad importante de las variables knockoff es que podemos intercambiar las covariables nulas con su respectiva variable knockoff sin cambiar la distribución conjunta de las covariables originales X y sus knockoffs $\tilde{X}$ condicionadas a Y.

Observación. Usaremos la notación para los pares i.i.d $(X_{i1}, \ldots, X_{ip}, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ de las covariables y la respuesta. Las covariables serán reunidas en la matriz de datos $\boldsymbol{X}$ y las respuestas en el vector de datos $\boldsymbol{y}$. Esto significa que la i-ésima fila de $\boldsymbol{X}$ es $(X_{i1}, \ldots, X_{ip})$ y de $\boldsymbol{y}$ es Y_i.

Lema 3.2.1. *Sea $S \subset \mathcal{N}$ un subconjunto de índices nulos, entonces*

$$[\boldsymbol{X}, \tilde{\boldsymbol{X}}] \mid \boldsymbol{y} \overset{d}{=} [\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{sw(S)} \mid \boldsymbol{y}$$

Demostración. El resultado es equivalente a $([\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{y}]) \overset{d}{=} ([\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{y}])$, pues se trata de la misma marginal $\boldsymbol{y}$. Supongamos que el conjunto $S = \{1, 2, \ldots, l\}$, entonces por la independencia de las filas, es suficiente mostrar que la igualdad en distribución se da en las filas, es decir, $((X, \tilde{X}), Y) \overset{d}{=} ((X, \tilde{X})_{sw(S)}, Y)$ donde X es una fila de $\boldsymbol{Y}$ y Y de $\boldsymbol{y}$. Para esto, por la condición (3.1) vemos que de la definición de distribución condicional, basta demostrar que

$$Y \mid (X, \tilde{X}) \overset{d}{=} Y \mid (X, \tilde{X})_{sw(S)}$$

Notemos que

$$\begin{aligned} F_{Y|(X,\tilde{X})_{sw(S)}}(y \mid (x, \tilde{x})) &= F_{Y|(X,\tilde{X})}(y \mid (x, \tilde{x})_{sw(S)}) \\ &= F_{Y|X}(y \mid \hat{x}) \end{aligned}$$

donde $\hat{x}_i = \tilde{x}_i$ si $i \in S$ y $\hat{x}_i = x_i$ en otro caso. Notemos que la segunda igualdad se da por la condición (3.2). Ahora bien, como $1 \in S$, entonces $Y \perp\!\!\!\perp X_1 \mid X_{2:p}$, lo que implica

$$\begin{aligned} F_{Y|X_{1:p}}(y \mid \tilde{x_1}, \hat{x}_{2:p}) &= F_{Y|X_{2:p}}(y \mid \hat{x}_{2:p}) \\ &= F_{Y|X_{1:p}}(y \mid x_1, \hat{x}_{2:p}) \end{aligned}$$

Esto implica que

$$Y \mid (X, \tilde{X})_{sw(S)} \overset{d}{=} Y \mid (X, \tilde{X})_{sw(S\setminus\{1\})}$$

Continuando inductivamente hasta que el conjunto S sea vacío, se prueba el resultado. ■

Ahora bien, en orden de encontrar variables significativas, necesitamos medir la importancia de las variables, para esto definimos la estadística W_j para $j \in \{1, \ldots, p\}$, la cual para un valor positivo y grande es evidencia en contra de la hipótesis de que X_j es nula. Una característica que debe poseer esta estadística es que depende de la respuesta, la covariable original y la variable knockoff, es decir, para alguna función w_j

$$W_j = w_j\left([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y}\right),$$

además debe de cumplir la *propiedad de cambio de signo*, la cual establece que si intercambiamos la variable j-ésima con su knockoff, esto tiene el efecto de cambiar el signo de W_j, es decir,

$$w_j\left([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{sw(S)}, \boldsymbol{y}\right) = \begin{cases} w_j\left([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y}\right) & si \quad j \notin S \\ -w_j\left([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y}\right) & si \quad j \in S \end{cases} \tag{3.3}$$

Denotando a $W = (W_1, \ldots, W_p)$, el siguiente lema establece que las estadísticas nulas W_j son simétricas.

Lema 3.2.2. *Sea $\epsilon \in \{\pm 1\}^p$ una secuencia de signos independientes de W, con $\epsilon_j = +1$ para toda j no nula y $\epsilon_j \overset{i.i.d}{\sim} \{\pm 1\}$ para toda j nula, entonces*

$$W = (W_1, \ldots, W_p) \overset{d}{=} (W_1 \cdot \epsilon_1, \ldots, W_p \cdot \epsilon_p) = W \odot \epsilon$$

Demostración. Sea ϵ y $S = \{j : \epsilon_j = -1\} \subset \mathcal{N}$. Luego

$$W_{sw(S)} = w([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{sw(S)}, \boldsymbol{y}),$$

de la propiedad (3.3) tenemos que $W_{sw(S)} \overset{d}{=} W \odot \epsilon$ y del lema (3.2.1), se sigue que $W_{sw(S)} \overset{d}{=} W$. ■

Una relación importante como consecuencia de este lema es que

$$\#\{j \text{ nula } : W_j \leq -t\} \overset{d}{=} \#\{j \text{ nula } : W_j \geq t\},$$

pues al condicionar en $|W| = (|W_1|, \ldots, |W_p|)$, ambas variables aleatorias siguen la misma distribución binomial, lo que significa que sus distribuciones marginales son las mismas. Por tanto, para cualquier $t > 0$

$$\#\{j : W_j \leq -t\} \geq \#\{j \text{ nula} : W_j \geq -t\} \overset{d}{=} \#\{j \text{ nula} : W_j \geq t\},$$

podemos entonces estimar la proporción de falsos descubrimientos (FDP)

$$\text{FDP}(t) = \frac{\#\{j \text{ nula} : W_j \geq t\}}{\#\{j : W_j \geq t\}} \tag{3.4}$$

como

$$\frac{\#\{j \text{ nula} : W_j \geq t\}}{\#\{j : W_j \geq t\}} \approx \frac{\#\{j \text{ nula} : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} = \widehat{\text{FDP}}(t). \tag{3.5}$$

El siguiente resultado es no asintótico y nos permite controlar el error de Tipo I.

Teorema 3.2.3. *Sean $\mathcal{W} = \{|W_j| : j = 1, 2, \ldots, p\}$ y $\tau > 0$ el umbral definido como*

$$\tau = \text{mín}\left\{t \in \mathcal{W} : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\right\} \quad (\textbf{Knockoffs}) \tag{3.6}$$

donde q es el nivel tope del FDR. Entonces el proceso de seleccionar las variables

$$\hat{S} = \{j : W_j \geq \tau\}$$

controla el FDR modificado definido como

$$\textit{mFDR} = E\left[\frac{|\{j \in \hat{S} \cap \mathcal{N}\}|}{|\hat{S}| \vee 1}\right] \leq q.$$

Un procedimiento un poco más conservador, dado al incrementar el número de negativos en uno,

$$\tau_+ = \text{mín}\left\{t \in \mathcal{W} : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\right\} \quad (\textbf{Knockoffs +}) \tag{3.7}$$

y $\hat{S} = \{j : W_j \geq \tau_+\}$, controla el FDR usual

$$\textit{FDR} = E\left[\frac{|\{j \in \hat{S} \cap \mathcal{N}\}}{|\hat{S}| \vee 1}\right] \leq q$$

Observación. Notemos que en el umbral (3.7) es mayor o igual al umbral definido en (3.6), por esta razón es más conservativo.

Para entender un poco de cómo el *Knockoffs* + controla el FDR. Sea

$$\tau = \text{mín}\{t : F\hat{D}P(t) \leq q\}$$

y

$$S^+(t) = \{j : W_j \geq t\},\ S^-(t) = \{j : W_j \leq -t\}$$

para el conjunto

$$\hat{S} = \{j : W_j \geq \tau\}$$

tenemos

$$\begin{aligned}
\text{FPD}(\tau) &= \frac{\#\{j \text{ nula } j \in S^+(\tau)\}}{\#\{j :\in S^+(\tau)\} \vee 1} \\
&= \frac{\#\{j \text{ nula } j \in S^+(\tau)\}}{\#\{j :\in S^+(\tau)\} \vee 1} \cdot \frac{1 + \#\{j \text{ nula } j \in S^-(\tau)\}}{1 + \#\{j \text{ nula } j \in S^-(\tau)\}} \\
&\leq q \cdot \frac{\#\{j \text{ nula } j \in S^+(\tau)\}}{1 + \#\{j \text{ nula } j \in S^-(\tau)\}} \\
&= q \cdot \frac{V^+(\tau)1}{1 + V^-(\tau)},
\end{aligned}$$

donde

$$\begin{aligned}
V^+(\tau) &= \#\{j \text{ nula} : \ j \in S^+(\tau)\} \\
V^-(\tau) &= \#\{j \text{ nula} : \ j \in S^-(\tau)\}
\end{aligned}$$

Ahora usaremos un argumento de Martingala para mostrar que

$$E\left[\frac{V^+(\tau)}{1 - V^-(\tau)} \leq 1\right]$$

y por tanto

$$\text{FDR} = E[\text{FDP}] \leq q$$

Para filtración $\mathcal{F}_t = \{\sigma(V^{\pm}(u))\}_{u \leq t}$, debemos mostrar que $\frac{V^+(\tau)}{1+V^-(\tau)}$ es una supermartingala (figura 3.1). Condicionando a $V^+(s) + V^-(s), V^+(s)$ es hipergeométrico, y así

$$E\left[\frac{V^+(\tau)}{1 - V^-(\tau)} \,\middle|\, V^{\pm}(t), V^+(t) + V^-(t)\right] \leq \frac{V^+(t)}{1 + V^{-1}(t)}$$

Entonces por el teorema de muestreo opcional de Doob,

$$\text{FDR} \leq qE\left[\frac{V^+(\tau)}{1 - V^-(\tau)}\right] \leq qE\left[\frac{V^+(0)}{1 - V^-(0)}\right] \leq q$$

donde la última desigualdad se da porque $V^+(0) \sim Bin(\#\text{ nulas}, 1/2)$

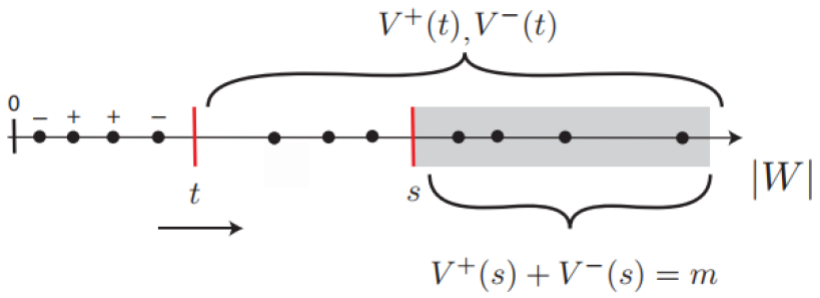


Figura 3.1: Argumento de martingala

3.3. Prueba del Teorema

En esta sección demostraremos el teorema (3.2.3). Para esto necesitamos introducir el concepto de pruebas de hipótesis secuenciales, pues veremos que el procedimiento Knockoffs es un ejemplo de un procedimiento para controlar el FDR en un problema de pruebas de hipótesis secuenciales.

3.3.1. Dos Procedimientos para Pruebas Secuenciales

Supongamos que $p_1, \ldots, p_m$ son los p-valores acerca de las hipótesis $H_1, \ldots, H_m$. Estos p-valores satisfacen que para toda j nula y para todo $u \in [0,1], P(p_j \leq u) \leq u$. Introducimos entonces dos estrategias secuenciales, que controlan el FDR a un nivel fijo q bajo la propiedad de independencia usual.

Definicion 3.3.1 (Primer procedimiento de pruebas secuenciales (FSTP))**.** Para un umbral fijo $c \in (0,1)$ y cualquier subconjunto K y definidos

$$\hat{k}_0 = \max\{k \in K : \frac{\#\{j \leq k : p_j > c\}}{k \vee 1} \leq (1-c)q\}$$

y

$$\hat{k}_1 = \max\{k \in K : \frac{1 + \#\{j \leq k : p_j > c\}}{1+k} \leq (1-c)q\},$$

con la convención de que $\hat{k}_{0/1} = 0$ si el conjunto es vacío: aquí $\hat{k}_{0/1}$ debería leerse como "$\hat{k}_0$" o "$\hat{k}_1$", pues podemos tomar elegir cualquiera de los dos en la definición de arriba. Rechazamos H_j para todo $j \leq \hat{k}_{0/1}$ y entonces obtenemos dos procedimientos llamados *FSTP0* y *FSTP1*.

Para entender un poco esto, consideremos el el caso *FSTP0* y supongamos que todos los p-valores nulos son i.i.d Unif(0,1), entonces

$$\frac{\#\{j \text{ nula} \leq k\}}{k \vee 1} \approx \frac{1}{1-c} \cdot \frac{\#\{j \text{ nula} \leq k : p_j > c\}}{k \vee 1} \leq \frac{1}{1-c} \cdot \frac{\#\{j \leq k : p_j > c\}}{k \vee 1}$$

de nuevo, el procedimiento maximiza el número de rechazos bajo la restricción de que una estimación del FDR se controla a un nivel q. FSP1 corrige FSP0 para garantizar el control FDR.

Definicion 3.3.2 (Segundo procedimiento de pruebas secuenciales (SSTP))**.** Alternativamente, definimos

$$\hat{k}_{0/1} = \max\{k \in K : \frac{0/1 + \#\{j \leq k : p_j > c\}}{\#\{j \leq k : p_j \leq c\} \vee 1} \leq \frac{1-c}{c} q\}$$

y

$$\hat{k}_1 = \max\{k \in K : \frac{1 + \#\{j \leq k : p_j > c\}}{1+k} \leq (1-c)q\},$$

con la convención de que $\hat{k}_{0/1} = 0$ si el conjunto es vacío. Rechazamos H_j para todo $j \leq \hat{k}_{0/1}$ tal que $p_j \leq c$.

Para entender un poco esta definición, para los p-vealores nulos que son i.i.d Unif(0,1), tenemos

$$\frac{\#\{j \text{ nula} \leq k : p_j \leq c\}}{\#\{j \leq k : p_j \leq c\} \vee 1} \approx \frac{c}{1-c} \cdot \frac{\#\{j \text{ nula} \leq k : p_j > c\}}{\#\{j \leq k : p_j \leq c\} \vee 1} \leq \frac{c}{1-c} \cdot \frac{\#\{j \leq k : p_j > c\}}{\#\{j \leq k : p_j \leq c\} \vee 1}$$

De nuevo, el procedimiento maximiza el número de rechazos bajo la restricción de que una estimación del FDR se controla en el nivel q. Tenemos entonces el siguiente resultado teórico.

Teorema 3.3.1. *Supongamos que todos los p-valores nulos son i.i.d. Unif(0,1), y que son independientes de los no nulos. Para cada procedimiento considerado, sea V el número de falsos descubrimientos y R el número total de descubrimientos. Entonces:*

- *Ambos FSTP1 y SSTP1 controlan el FDR, es decir,*

$$E\left[\frac{V}{R \vee 1}\right] \leq q.$$

- *SSTP0 controla una modificación del FDR, a saber,*

$$E\left[\frac{V}{R + \frac{c}{1-c} q^{-1}}\right] \leq q.$$

  cuando $c = 1/2$, el controla $E[V/(R + q^{-1})]$.

- *FSTP0 también controla un FDR modificado,*

$$E\left[\frac{V}{R+\frac{1}{1-c}q^{-1}}\right] \leq q.$$

Observación (Conexión con Knockoffs)**.** El método Knockoffs se puede considerar como un caso especial del SSTP, y las propiedades de control del FDR son solo una consecuencia del teorema (3.3.1). Sea $m = \#\{j : W_j \neq 0\}$, como el método Knockoffs nunca selecciona la variable j cuando $W_j = 0$, podemos ignorar dichas variables. Sin pérdida de generalidad, supongamos que $|W_1| \geq |W_2| \geq \cdots \geq |W_m| > 0$, y hagamos

$$p_j = \begin{cases} 1/2 & si \quad W_j > 0 \\ 1 & si \quad W_j < 0 \end{cases}$$

Luego del lema (3.2.2) que los p-valores nulos son i.i.d. con $P(p_j = 1/2) = 1/2 = P(p_j = 1)$ y son independientes de los demás, obedeciendo así los supuestos del teorema (3.3.1). Sea K los índices de las desigualdades estrictas,

$$K = \{k \in [m] : |W_k| > |W_{k+1}|\} \cup \{m\}$$

podemos ver que este método es equivalente al segundo procedimiento de prueba secuencial en estos p-valores. Para ver por qué esto es cierto, sea $c = 1/2$ y notemos que para cualquier $k \in K$,

$$\frac{0/1 + \#\{j \leq k : p_j > 1/2\}}{\#\{j \leq k : p_j \leq 1/2\} \vee 1} = \frac{0/1 + \#\{j \leq k : W_j < 0\}}{\#\{j \leq k : W_j > 0\} \vee 1} = \frac{0/1 + \#\{j \leq k : W_j \geq -|W_k|\}}{\#\{j \leq k : W_j \geq |W_k|\} \vee 1}$$

La primera igualdad se da por la definición de p_j. La segunda igualdad es válida porque los valores absolutos de W están en orden decreciente. Por definición de K, la desigualdad solo es posible si se cumple que $j \leq k$. Por lo tanto, $W_j \geq |W_k|$ (respectivamente, $W_j < -|W_k|$) es verdadero si y solo si $j \leq k$ y $W_j > 0$ (respectivamente, $j \leq k$ y $W_j < 0$). De esto tenemos que, encontrar la mayor k tal que la relación en el lado izquierdo esté por debajo de q es lo mismo que encontrar la $|W_k|$ más pequeña de tal manera que el lado derecho esté por debajo de q. Esto es equivalente a encontrar el mínimo $t \in \mathcal{W}$ tal que

$$\frac{0/1\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q$$

que son los umbrales Knockoffs y Knockoffs+. Finalmente, rechazamos los p-valores tales que $p_j \leq 1/2$ es lo mismo que rechazar los W_j positivos. El control FDR sigue aplicando el teorema (3.3.1).

Prueba del Teorema (3.3.1)

En esta sección veremos la prueba del teorema (3.3.1) y con esto quedará justificado el teorema (3.2.3). Trabajaremos con $K = \{1, \dots, m\}$, para un $K \subsetneq \{1, \dots, m\}$ es idéntico. Iniciamos nuestro proceso con el siguiente lema:

Lema 3.3.2 (Proceso de Martingala)**.** *Para $k = m, m-1, \dots, 1, 0$, sean*

$$V^+(k) = \#\{j \text{ nula} : 1 \leq j \leq k, p_j \leq c\},\ V^-(k) = \#\{j \text{ nula} : 1 \leq j \leq k, p_j > c\}$$

con la convención de que $V^{\pm}(0) = 0$. Sea $\mathcal{F}_k$ la filtración definida al conocer todos los p-valores no nulos, así como $V^{\pm}(k)$ para todo $k' \geq k$. Entonces el proceso

$$M(k) = \frac{V^+(k)}{1 + V^-(k)}$$

es una supermartingala que retrocede en el tiempo con respecto a $\mathcal{F}_k$. Además para cualquier q fijo, se tiene que $\hat{k}$ definido como en cualquier procedimiento de pruebas secuenciales es un tiempo de parada (stopping time), y como consecuencia

$$E\left[\frac{\#\{j \text{ nulo} \leq \hat{k} : p_j \leq c\}}{1 + \#\{j \text{ nulo} \leq \hat{k} : p_j > c\}}\right] \leq \frac{c}{1-c}$$

Demostración. Notemos que la filtración $\mathcal{F}_k$ nos dice si k es nulo o no, pues el proceso no nulo es conocido exactamente. Por otro lado, si k es no nulo, entonces $M(k-1) = M(k)$, ahora bien, si k es nulo, entonces

$$\begin{aligned} M(k-1) &= \frac{V^+(k-1)}{1 + V^-(k-1)} \\ &= \frac{V^+(k) - I}{1 + V^-(k) - (1 - I)} \\ &= \frac{V^+(k) - I}{(V^-(k) + I) \vee 1} \end{aligned}$$

donde $I = \mathbb{1}_{p_k \leq c}$. El evento $\mathcal{F}_k$ no nos da más información acerca de I, y se sigue de la propiedad de intercambiabilidad de los nulos (pues ellos son i.i.d y por tanto intercambiables) que

$$P(I = 1) = \frac{V^+(k)}{V^+(k) + V^-(k)}$$

Entonces para el caso en que k es nulo, se tiene

$$E\left[M(k-1) \mid \mathcal{F}_k\right] = \frac{1}{V^+(k)+V^-(k)}\left[V^+(k)\frac{V^+(k)-1}{V^-(k)+1} + V^-(k)\frac{V^+(k)}{V^-(k)\vee 1}\right]$$
$$= \begin{cases} \frac{V^+(k)}{1+V^-(k)} & si \quad V^-(k)>0, \\ V^+(k)-1 & si \quad V^-(k)=0 \end{cases}$$

Esto significa que

$$E[M(k-1) \mid \mathcal{F}_k] = \begin{cases} M(k) & si \quad k \text{ es no nulo}, \\ M(k) & si \quad k \text{ es nulo y } V^-(k)>0, \\ M(k)-1 & si \quad k \text{ es nulo y } V^-(k)=0 \end{cases}$$

Esto muestra que $E[M(k-1) \mid \mathcal{F}_k] \leq M(k)$. Por tanto, se cumple la propiedad de ser supermartingala. Por otro lado, como $\{\hat{k} \geq k\} \in \mathcal{F}_k$, entonces $\hat{k}$ es un tiempo de parada con respecto a la filtración que retrocede $\{\mathcal{F}_k\}$. La última parte del teorema, se sigue del teorema del tiempo de detención óptimo para supermartales que establece que

$$EM(\hat{k}) \leq EM(m) = E\left[\frac{\#\{j \text{ nula} : p_j \leq c\}}{1+\#\{j \text{ nula} : p_j > c\}}\right]$$

Sea $X = \#\{j \text{ nula} : p_j \leq c\}$. La independencia de los nulos junto con el dominio estocástico de la distribución de los p_j para los nulos, implica que $X \overset{d}{\leq} Y$, donde $Y \sim Bin(N,c)$, con N el número total de nulas. Más aún, como la función $x \mapsto x/(1+N-x)$ es no decreciente, tenemos

$$\begin{aligned} E\left[\frac{X}{1+N-X}\right] &\leq E\left[\frac{Y}{1+N-Y}\right] \\ &= \sum_{i=1}^{N} P(Y=i)\frac{i}{1+N-i} \\ &= \sum_{i=1}^{N} c^i(1-c)^{N-i}\frac{N!}{i!(N-i)!}\frac{i}{1+N-i} \\ &= \frac{c}{1-c}\sum_{i=1}^{N} c^{i-1}(1-c)^{N-i+1}\frac{N!}{(i-1)!(N-i+1)!} \\ &= \frac{c}{1-c}\sum_{i=1}^{N} P(Y=i-1) \\ &\leq \frac{c}{1-c} \end{aligned}$$

Prueba del Teorema para SSTP

Recordando que $V = \#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}$ y $R = \#\{j \leq \hat{k} : p_j \leq c\}$. Para SSTP1, tomamos $\hat{k} = \hat{k}_1$ y tenemos

$$\begin{aligned} E\left[\frac{V}{R \vee 1}\right] &= E\left[\frac{V}{R \vee 1}\mathbb{1}_{\hat{k}>0}\right] \\ &= E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{1 + \#\{j \text{ nula} \leq \hat{k} : p_j > c\}} \cdot \frac{1 + \#\{j \text{ nula} \leq \hat{k} : p_j > c\}}{\#\{j \leq \hat{k} : p_j \leq c\} \vee 1}\mathbb{1}_{\hat{k}>0}\right] \\ &\leq E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{1 + \#\{j \text{ nula} \leq \hat{k} : p_j > c\}} \cdot \frac{1-c}{c} \cdot q\right] \\ &\leq q \end{aligned}$$

Similar sucede para SSTP0, sea $\hat{k} = \hat{k}_0$, tenemos

$$\begin{aligned} E\left[\frac{V}{R \vee \frac{c}{1-c}q^{-1}}\right] &= E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{1 + \#\{j \text{ nula} \leq \hat{k} : p_j > c\}} \cdot \frac{1 + \#\{j \text{ nula} \leq \hat{k} : p_j > c\}}{R + \frac{c}{1-c}q^{-1}}\right] \\ &\leq E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{1 + \#\{j \text{ nula} \leq \hat{k} : p_j > c\}} \cdot \frac{1 + \frac{1-c}{c} \cdot qR}{\frac{c}{1-c} + qR}\right] \cdot q \\ &= E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{1 + \#\{j \text{ nula} \leq \hat{k} : p_j > c\}}\right] \cdot \frac{1-c}{c} \cdot q \\ &\leq q \end{aligned}$$

Prueba del Teorema para FSTP

Sea $\hat{k} = \hat{k}_1$, entonces

$$E\left[\frac{V}{R \vee 1}\right] = E\left[\frac{V}{R \vee 1} \cdot \mathbb{1}_{\hat{k}>0}\right] = E\left[\frac{\#\{j \text{ nula} \leq \hat{k}\}}{\hat{k} \vee 1} \cdot \mathbb{1}_{\hat{k}>0}\right]$$

y

$$\text{FDP}_+(\hat{k}) = \frac{1 + \#\{j \text{ nula} \leq \hat{k}\}}{1 + \hat{k}}$$

luego

$$
\begin{aligned}
E\left[\frac{V}{R \vee 1}\right] &\leq E\left[\frac{1+\#\{j \text{ nula} \leq \hat{k}\}}{\hat{k}+1} \cdot \mathbb{1}_{\hat{k}>0}\right] \\
&= E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{\hat{k}+1} \cdot \mathbb{1}_{\hat{k}>0}\right] + E\left[\frac{1+\#\{j \text{ nula} \leq \hat{k} : p_j > c\}}{\hat{k}+1} \cdot \mathbb{1}_{\hat{k}>0}\right] \\
&= E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{\hat{k}+1} \cdot \mathbb{1}_{\hat{k}>0}\right] + E\left[\frac{1+\#\{j \text{ nula} \leq \hat{k} : p_j > c\}}{\hat{k}+1} \cdot \mathbb{1}_{\hat{k}>0}\right] \\
&\leq E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{1+\#\{j \text{ nula} \leq \hat{k} : p_j > c\}} \cdot FDP_+(\hat{k}) \cdot \mathbb{1}_{\hat{k}>0}\right] + E\left[FDP_+(\hat{k}) \cdot \mathbb{1}_{\hat{k}>0}\right] \\
&\leq E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{1+\#\{j \text{ nula} \leq \hat{k} : p_j > c\}} \cdot\right] \cdot (1-c) \cdot q + (1-c) \cdot q \\
&\leq \frac{c}{1-c}(1-c) \cdot q + (1-c) \cdot q \\
&= q
\end{aligned}
$$

Por último, para el caso FSTP0, sea $\hat{k} = \hat{k}_0$, entonces

$$
\begin{aligned}
E\left[\frac{V}{\frac{1}{1-c}q^{-1} + R}\right] &= E\left[\frac{\#\{j \text{ nula} \leq \hat{k}\}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] \\
&= E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] + E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j > c\}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] \\
&= E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{1+\#\{j \text{ nula} \leq \hat{k} : p_j > c\}} \cdot \frac{1+\#\{j \text{ nula} \leq \hat{k} : p_j > c\}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] + \\
&\qquad E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j > c\}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] \\
&\leq E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{1+\#\{j \text{ nula} \leq \hat{k} : p_j > c\}} \cdot \frac{1+(1-c) \cdot q\hat{k}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] + E\left[\frac{(1-c) \cdot q\hat{k}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] \\
&\leq E\left[\frac{\#\{j \text{ nula} \leq \hat{k} : p_j \leq c\}}{1+\#\{j \text{ nula} \leq \hat{k} : p_j > c\}}\right] \cdot (1-c) \cdot q + (1-c) \cdot q \\
&\leq \frac{c}{1-c} \cdot (1-c) \cdot q + (1-c) \cdot q \\
&= q
\end{aligned}
$$

■

3.4. Construcción del Modelo Knockoffs X y Algoritmo

Como mencionamos anteriormente una característica fundamental para la construcción del modelo es que se cumpla la propiedad (3.1). El siguiente resultado caracteriza el modelo Knockoffs-X.

Lema 3.4.1. *Las variables aleatorias* $(\tilde{X}_1, \ldots, \tilde{X}_p)$ *son el modelo knockoff X para* $(X_1, \ldots, X_p)$ *si y sólo si para cualquier* $j \in \{1, \ldots, p\}$ *el par* $(X_j, \tilde{X}_j)$ *es intercambiable condicionado a todas las otras variables y sus knockoffs, es decir,*

$$(X_j, \tilde{X}_j, X_{-j}, \tilde{X}_{-j}) \stackrel{d}{=} (\tilde{X}_j, X_j, X_{-j}, \tilde{X}_{-j})$$

Este lema nos redirecciona a lo siguiente: para construir el modelo Knockoffs, debemos construir pares de variables que condicionadas sean intercambiables. Si las componentes del vector X son independientes, entonces cualquier copia de variables independientes de X funcionaría, esto significa que cualquier vector $\tilde{X}$ muestreado independientemente de la misma distribución conjunta de X funcionará. Como consecuencia de esto tenemos el siguiente algoritmo:

Algoritmo 4 Algoritmo secuencial de pares independientes condicionales

$j = 1$
while $j \leq p$ **do**
  Muestrear $\tilde{X}_j$ de $\mathcal{L}(X_j \mid X_{-j}, \tilde{X}_{1:j-1})$
  $j = j + 1$
end while

Aquí, $\mathcal{L}(X_j \mid X_{-j}, \tilde{X}_{1:j-1})$ es la distribución condicional de X_j dado $(X_j, \tilde{X}_{1:j-1})$. Antes de justificar el porqué el algorítmo cumple la condición (3.1), veamos un ejemplo para ilustrar un poco. Tomemos $p = 3$, entonces:

Probaremos por inducción en j que el Algoritmo (4) produce variables knockoffs que satisfacen la propiedad de intercambiabilidad (3.1). Probaremos esto para el caso discreto. La hipótesis inductiva en este caso es: después de j pasos, todo par $X_k, \tilde{X}_k$ es intercambiable en la distribución conjunta de $(X_{1:p}, \tilde{X}_{1:j-1})$ para todo $k = 1, \ldots, j$. Supongamos que la hipótesis inductiva se da hasta $j - 1$, notemos que por hipótesis $\mathcal{L}$ es simétrica en $X_j, \tilde{X}_k$

Ejemplo para el caso $p = 3$

1: Muestrear $\tilde{X}_1$ de $\mathcal{L}(X_1 \mid X_{2:3})$ y obtenemos $\mathcal{L}(X_{1:3}, \tilde{X}_1)$, entonces sabemos $\mathcal{L}(X_2 \mid X_1, X_3, \tilde{X}_1)$.
2: Muestrear $\tilde{X}_2$ de $\mathcal{L}(X_2 \mid X_1, X_3, \tilde{X}_1)$ y obtenemos $\mathcal{L}(X_{1:3}, \tilde{X}_{1:2})$, entonces sabemos $\mathcal{L}(X_3 \mid X_{1:2}, \tilde{X}_{1:2})$.
3: Muestrear $\tilde{X}_3$ de $\mathcal{L}(X_3 \mid X_{1:2}, \tilde{X}_{1:2})$.

para $k = 1, \ldots, j-1$. La distribución conjunta de $\tilde{X}_j$ dado $X_{1:p}, \tilde{X}_{1:j-1}$ es dada por

$$\frac{\mathcal{L}(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u \mathcal{L}(X_{-j}, u, \tilde{X}_{1:j-1})}$$

Por tanto, la distribución conjunta de $(X_{1:p}, \tilde{X}_{1:j})$ es dada por

$$\frac{\mathcal{L}(X_{-j}, X_j, \tilde{X}_j, \tilde{X}_{1:j-1})\mathcal{L}(X_{-j}, \tilde{X}_j, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u \mathcal{L}(X_{-j}, u, \tilde{X}_{1:j-1})} \tag{3.8}$$

La intercambiabilidad de $X_j, \tilde{X}_j$ de estos dos valores, se sigue de la simetría de (3.8). Para $k < j$, notemos que (3.8) sólo depende de $X_k, \tilde{X}_k$ a través de la función $\mathcal{L}$ y la función $\mathcal{L}$ es simétrica en $X_k, \tilde{X}_k$. Por tanto, (3.8) es también simétrica en $X_k, \tilde{X}_k$, y así, el par es intercambiable en la distribución conjunta de $(X_{1:p}, \tilde{X}_{1:j})$.

3.5. Prueba de aleatorización condicional

Esta sección presentamos un enfoque alternativo al problema de selección de variables controladas. Antes de describir el método, veamos qué sucede con un ejemplo. Supongamos que estamos en el caso de una regresión y que $\hat{\beta}_j(\lambda)$ es el valor de la estimación Lasso del j-ésimo coeficiente de regresión. Entonces nos gustaría usar la estadística $\hat{\beta}_j(\lambda)$ para probar si Y es condicionalmente independiente de X_j, ya que valores grandes de $|\hat{\beta}_j(\lambda)|$ son evidencia en contra de la hipótesis nula. Sin embargo, para construir una prueba, necesitaríamos conocer la distribución de muestreo de $\hat{\beta}_j(\lambda)$ bajo la hipótesis nula de que Y y X_j son condicionalmente independientes, y no está claro cómo se obtendría dicho conocimiento.

Una manera de muestrear la covariable X_j condicional en todas las demás covariables, pero no la respuesta, donde por "muestrear" nos referimos explícitamente a extraer una nueva muestra de la distribución condicional de $X_j \mid X_{-j}$ usando un generador de números aleatorios. Luego calculamos el estadístico Lasso $\hat{\beta}_j^*(\lambda)$, donde el superíndice $*$ indica que el

estadístico se calcula a partir del valor muestreado artificialmente de la covariable X_j. Ahora, bajo la hipótesis nula de independencia condicional entre Y y X_j, sucede que $\hat{\beta}_j^*(\lambda)$ y $\hat{\beta}_j(\lambda)$ están distribuidos de forma idéntica y, además, esta afirmación es verdadera condicionando en Y y todas las demás covariables. Esto se demuestra en el siguiente lema. Una consecuencia de esto es que al simular una covariable condicional a las otras, podemos muestrear a voluntad de la distribución condicional de cualquier estadística de prueba y calcular los p-valores como se describe en el algoritmo (5).

Lema 3.5.1. *Sea (Z_1, Z_2, Y) una tripleta de variables aleatorias, y sea (Z_1^*, Z_2, Y) tal que*

$$Z_1^* \mid (Z_2, Y) \stackrel{d}{=} Z_1 \mid Z_2$$

Entonces bajo la hipótesis nula de que $Y \perp\!\!\!\perp Z_1 \mid Z_2$, cualquier estadística de prueba de prueba $T = t(Z_1, Z_2, Y)$ satisface

$$T \mid (Z_2, Y) \stackrel{d}{=} T^* \mid (Z_2, Y),$$

donde $T^* = t(Z_1^*, Z_2, Y)$.

Una consecuencia de este lema es que podemos calcular por ejemplo el percentil de $95\,\%$ de la distribución condicional de T^* denotada por $t^*_{0.95}(Z_2, Y)$. Entonces por definición, bajo la hipótesis nula,

$$P(T > t^*_{0.95}(Z_2, Y) \mid (Z_2, Y)) \leq 0.05$$

Basado en esto, tenemos el siguiente algoritmo:

3.6. Construcción de las Estadísticas

Ahora bien, ya sabemos cómo construir las variables Knockoffs y controlar el error de Tipo I, sin embargo queda la otra parte del problema y es cómo construir con exactitud las estadísticas. Las estadísticas Knockoffs $W = (W_1, \ldots, W_p)$ las podemos pensar en dos pasos: primero, considere una estadística T para cada variable original y knockoffs

$$T = (Z, \tilde{Z}) = (Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p) = t([X, \tilde{X}], y)$$

con la idea de que Z_j(resp. $\tilde{Z}_j$) mide la importancia de X_j (resp. $\tilde{X}_j$). Supongamos la propiedad de que cambiar una variable con su knockoffs simplemente cambia los componentes de

Algoritmo 5 Prueba de aleatorización condicional

Input Un conjunto de n muestras independientes $(X_{i1}, \ldots, X_{ip}, Y_i)_{1 \leq i \leq n}$ puestas en una matriz de datos X y un vector de respuesta y, un test estadístico $T_j(X, y)$ para probar si X_j e Y son condicionalmente independientes.

for $k = 1, \ldots, K$ **do**

Crear una nueva mateiz de datos $X^{(k)}$ simulando la j-ésima columna de X de $\mathcal{L}(X_j \mid X_{-j}$ (y dependiendo (y manteniendo las columnas restantes iguales). Es decir, $X_{ij}^{(k)}$ se muestrea a partir de la distribución condicional $X_{ij} \mid \{X_{i1}, \ldots, X_{ip}\} \setminus \{X_{ij}\}$, y es (condicionalmente) independiente de X_{ij}.

end for

Output Un p-valor

$$p_j = \frac{1}{K+1}\left[1 + \sum_{k=1}^{K} \mathbb{1}_{T_j(X^{(k)}, y) \geq T_j(X, y)}\right]$$

Al igual que con las pruebas de permutación, sumar uno en el numerador y el denominador asegura que los p-valores nulos sean estocásticamente más grandes que las variables uniformes.

T de la misma manera, es decir, para cada $S \in \{1, \ldots, p\}$

$$(Z, \tilde{Z})_{sw(S)} = t([X, \tilde{X}]_{sw(S)}, y)$$

Una propiedad importante es que se deriva de todo esto es que si S es un subconjunto de nulos, tenemos que

$$(Z, \tilde{Z})_{sw(S)} \stackrel{d}{=} (Z, \tilde{Z})$$

y esto se da condicionando en Y, sin importar la relación entre Y y X.

Recordemos que para construir un W_j tal que satisfaga la *propiedad de cambio de signo* (3.3), una forma es simplemente hacer

$$W_j = f_j(Z_j, \tilde{Z}_j)$$

sea cualquier función antisimétrica. Bajo este enfoque, consideremos por ejemplo un problema de regresión y con el enfoque Lasso en el diseño original aumentado con knockoffs, es decir,

$$\min_{b \in \mathbb{R}^{2p}} \frac{1}{2}\|y - [X, \tilde{X}]\beta\|_2^2 + \lambda\|\beta\|_1$$

y denotamos la solución como $\hat{\beta}(\lambda)$ (los primeros p componentes son los coeficientes de las variables originales y los últimos p son para los knockoffs). Entonces el estadístico coeficiente de diferencia Lasso (LCD) establece que $Z_j = |\hat{\beta}_j(\lambda)|$, $\tilde{Z}_j = |\hat{\beta}_{j+p}|$ y

$$W_j = Z_j - \tilde{Z}_j = |\hat{\beta}_j(\lambda)| - |\hat{\beta}_{j+p}(\lambda)|$$

Entonces para un valor grande y positivo de W_j, este proporciona alguna evidencia de que la distribución de Y depende de X_j, mientras que bajo la hipótesis nula, W_j tiene una distribución simétrica y, por lo tanto, es igualmente probable que tome valores positivos y negativos. Sin embargo, observe cuidadosamente que el valor de λ no necesita ser fijo por adelantado, y puede calcularse a partir de y y $[X, \tilde{X}]$ de cualquier manera dependiente de los datos, siempre que permutar las columnas de X no cambie su valor; por ejemplo, se puede seleccionar mediante validación cruzada.

También hay muchas opciones disponibles para la función antisimétrica f_j, como

$$|Z_j| - |\tilde{Z}_j|, sign(|Z_j| - |\tilde{Z}_j|), \text{máx}\{|Z_j|, |\tilde{Z}_j|\} \text{ o } \log(|Z_j|) - \log(|\tilde{Z}_j|).$$

3.7. Construcción para Modelos Lineales

En un modelo lineal generalizado, la respuesta Y tiene una distribución de probabilidad tomada de una familia exponencial, que depende de las covariables solo a través de la combinación lineal $\eta = \beta_1 X_1 + \ldots + \beta_p X_p$. La relación entre Y y X especificada a través de una función de enlace g tal que $E(Y \mid X) = g^{-1}(\eta)$. En este tipo de modelos y en condiciones generales, $Y \perp\!\!\!\perp X \mid X_{-j}$ si y solo si $\beta_j = 0$. En este contexto, probar la hipótesis de que X_j es una variable nula es lo mismo que probar $H_j : \beta_j = 0$. Tenemos entonces el siguiente resultado para MLG.

Teorema 3.7.1. *Supongamos que una familia de variables aleatorias $X_1, \ldots, X_p$ es tal que no podemos predecir perfectamente ninguno de ellos a partir del conocimiento de los demás. Si la probabilidad de Y sigue un MLG, entonces $Y \perp\!\!\!\perp X_j \mid X_{-j}$ si y solo si $\beta_j = 0$. Por lo tanto, $\mathcal{N} = \{j : \beta_j = 0\}$.*

Demostración. Probaremos esto para el caso de regresión logística, el caso general es similar. Aquí, la distribución condicional de Y es Bernoulli con

$$E(Y \mid X) = P(Y = 1 \mid X) = \frac{e^{\eta}}{1 + e^{\eta}} = g^{-1}(\eta),$$

notemos que la suposición acerca de las covariables, implica que el modelo es identificable. Supongamos entonces que $\beta_j = 0$, así

$$p_{Y,X_j|X_{-j}}(y, x_j \mid x_{-j}) = p_{Y|X_j,X_{-j}}(y \mid x_j, x_{-j}) p_{X_j|X_{-j}}(x_j \mid x_{-j})$$

y como el primer factor en el lado derecho no depende de X_j, vemos que la función de distribución de probabilidad condicional se factoriza. Esto implica independencia condicional. Por otro lado, supongamos que Y y X_j son condicionalmente independientes. Entonces la función de probabilidad

$$\frac{\exp\left(Y(\beta_1 X_1 + \ldots + \beta_p X_p)\right)}{1 + \exp\left(\beta_1 X_1 + \ldots + \beta_p X_p\right)}$$

debe, condicionalmente en X_j, factorizar en una función de Y multiplicada por una función de X_j. Una consecuencia de esto es que condicionalmente en X_{-j}, el cociente de posibilidades no debe depender de X_j (debe ser constante). Sin embargo, esta relación es igual a $\exp(\beta_j X_j)$ y es constante solo si $\beta_j = 0$ ya que, por suposición, X_j no está determinado por X_{-j}. ■

3.7.1. Construcción para el Modelo Gaussiano

En esta sección, analizaremos el caso particular para modelos Gaussianos. Vamos a suponer que tenemos un modelo de regresión lineal de la forma

$$y = X\beta + z$$

donde $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^p$ y $z \sim N(0, \sigma^2 I)$ y supondremos que $n \geq p$, pues en otro caso el modelo será no identificable. El procedimiento no requiere conocimiento sobre σ y además es independiente del número de variables que el modelo tenga.

Construcción de Knockoffs

Tomamos la matriz de Gram $\Sigma = X^\top X$, suponiendo que esta variable es invertible y tal que $\Sigma_{jj} = \|X_j\|_2^2 = 1$. La idea es que $\tilde{X}$ tenga la misma estructura de covarianza que la matriz de diseño original, pero que la correlación entre variables diferentes y originales con las knockoffs sean las mismas que las existentes entre las originales, pues requerimos hacer una comparación entre la variable original y su imitación (knockoff). En términos matemáticos necesitamos que se cumpla que

$$\tilde{X}^\top \tilde{X} = \Sigma, \quad X^\top \tilde{X} = \Sigma - \operatorname{diag}\{s\}$$

esto significa que

$$X_j^\top \tilde{X}_k = X_j^\top \tilde{X}_k \quad \text{para } j \neq k$$

y

$$X_j^\top \tilde{X}_j = 1 - s_j$$

Más adelante daremos la forma explícita de la variable $\tilde{X}$ y sus consecuencias.

Calcular la Estadística

Ahora bien, para medir cuál variable es mejor, es decir, para comprar la variable original y su knockoff, necesitamos una estadística W_j para cada $\beta_j \in \{1, 2, \ldots, p\}$. Estas $W_j's$ se construyen de manera que grandes valores positivos son evidencia contra la hipótesis nula $\beta_j = 0$. Por ejemplo, en regresión Lasso, sabemos que el estimador para β es de la forma

$$\hat{\beta}(\lambda) = \arg\min_b \left\{ \frac{1}{2}||y - Xb||_2^2 + \lambda||b||_1 \right\}$$

Notemos que al variar el parámetro de penalización λ, obtenemos diferentes modelos en los que algunas variables tiene un coeficiente distinto de 0, entonces la idea es construir una estadística de tal manera que seleccione variables cuyo coeficiente ajustado esté por encima de algún umbral de significancia. Dado que los coeficientes ajustados están correlacionados entre sí, entonces un umbral incorrecto puede producir una proporción muy alta o muy baja de falsos descubrimientos. La verdadera importancia de una variable explicativa X_j se puede deducir comparando su poder predictivo para y con su copia $\tilde{X}_j$. Definamos

$$Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$$

esperamos entonces que Z_j sea grande para la mayoría de señales y pequeño para las variables nulas. En vez de calcular Lasso para la matriz X, calculamos Lasso para la matriz aumentada $[X, \tilde{X}]$, produciendo así un vector $2p$ dimensional $(Z_1, Z_2, \ldots, Z_p, \tilde{Z}_1, \tilde{Z}_2, \ldots, \tilde{Z}_p)$. Ahora bien, para cada $j \in \{1, \ldots, p\}$ definimos la estadística

$$W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1 & si \quad Z_j > \tilde{Z}_j \\ -1 & si \quad Z_j < \tilde{Z}_j \end{cases}$$

Un valor grande y positivo de W_j, indica que la variable X_j entra al modelo Lasso rápido que su knockoff $\tilde{X}_j$ algún valor λ, por tanto, esto nos indica que la variable pertenece al modelo.

Entonces el filtro de Knockoff funciona comparando los $Z_j's$ con los $\tilde{Z}_j's$, de manera que selecciona sólo las variables que son mejores que su copia knockoff. Por construcción, las estadísticas nulas son intercambiables par a par. El intercambio de Z_j y $\tilde{Z}_j$ correspondiente a variables nulas, deja la distribución conjunta de

$$(Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p)$$

sin cambios.

Más formalmente la estadística W_j es una función tal que $W\left([X, \tilde{X}], y\right) \in \mathbb{R}^p$ tal que para grandes valores positivos de W_j dá evidencia que $\beta_j \neq 0$. La estadística W_j debe satisfacer las dos siguientes propiedades:

Propiedad de suficiencia: W depende sólo de la matriz de Gram y de los productos internos de la respuesta.

$$W = f\left([X, \tilde{X}]^\top [X, \tilde{X}], [X \quad \tilde{X}]^\top y\right), f : S^+_{2p} \times \mathbb{R}^{2p} \mapsto \mathbb{R}^p$$

Propiedad antisimétrica: para $S \subset \{1, \ldots, p\}$

$$W_j\left([X, \tilde{X}]_{swap(S)}, y\right) = W_j\left([X, \tilde{X}], y\right) \cdot \begin{cases} +1 & si \quad j \notin S \\ -1 & si \quad j \in S \end{cases}$$

donde $[X, \tilde{X}]_{swap(S)}$ significa que las columnas X_j y $\tilde{X}_j$ han sido intercambiadas de la matrix $[X, \tilde{X}]$.

Otros ejemplos para W son:

1. $W_j = X_j^\top y - \tilde{X}_j^\top y$: Bajo el modelo gaussiano para y, tenemos que

   $$W \sim N(\operatorname{diag}\{s\}\beta, 2\sigma^2 \operatorname{diag}\{s\}),$$

   lo que significa que las p estadística son independientes. Reescalando cosas para que $W = N(\beta, 2\sigma^2 \operatorname{diag}\{s^{-1}\})$ (esta distribución se puede obtener simplemente tomando $W = \Sigma^{-1}X^\top y + N(0, \sigma^2(2\operatorname{diag}\{s^{-1}\} - \Sigma^{-1}))$), donde los términos en la suma son independientes. Sin embargo, para valores grandes de $|\beta_j|$, W_j puede ser positivo o negativo dependiendo del signo de $|\beta_j|$.

2. Haciendo $W_j = |X_j^\top y| - |\tilde{X}_j^\top y|$, resuelve el problema del signo del caso anterior.

3. Tomar $W_j = |\hat{\beta}_j^{LS}| - |\hat{\beta}_{j+p}^{LS}|$ o $W_j = |\hat{\beta}_j^{LS}|^2 - |\hat{\beta}_{j+p}^{LS}|^2$, donde $\hat{\beta}^{LS}$ es la solución por mínimos cuadrados, obtenido de la regresión lineal clásica

$$\hat{\beta}^{LS} = \left([X, \tilde{X}^\top [X, \tilde{X}]\right)^{-1} [X, \tilde{X}]^\top y.$$

Umbral para la Estadística

Por lo mencionado en anteriormente, seleccionaremos las variables tal que W_j sea grande y positivo, esto significa que necesitamos un $t > 0$ tal que $W_j \geq t$. Para un q dado en el FDR, definimos el umbral T como

$$T = \text{mín}\left\{t \in \mathcal{W} : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\} \tag{3.9}$$

para $\mathcal{W} = \{|W_j| : j = 1, \ldots, p\} \backslash \{0\}$. Una observación aquí es que si tomamos $W_j \geq t$ en la fracción de T, esta es una estimación del FDP, por tanto, la fracción de arriba se llama estimación de knockoff de FDP.

Knockoff para el Modelo Gaussiano

Recordemos que buscamos una matriz $\tilde{X}$ que conserve la covarianza de la matriz original de diseño, pero necesitamos que las correlaciones entre variables diferentes originales y de imitación sea las mismas que las existentes entre los originales. Vamos a suponer que $n \geq 2p$ y se puede extender a la restricción $p \leq n < 2p$. Para el primer caso, la matriz $\tilde{X}$ satisface que

$$[X, \tilde{X}]^\top [X, \tilde{X}] = \begin{bmatrix} \Sigma & \Sigma - diag\{s\} \\ \Sigma - diag\{s\} & \Sigma \end{bmatrix} = G$$

para algún $s \in \mathbb{R}^p$. Recordando que el complemento de Schur nos dice que para

$$M = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix},$$

de dimensión $n \times n$, si $A \succ 0$ entonces $M \succ 0$ sii $C - B^\top A^{-1} B \succ 0$. Por lo tanto, una condición necesaria y suficiente para que $\tilde{X}$ exista es que la matriz G sea semidefinida

positiva y esto ocurre si y solo si

$$\begin{aligned} A &= \Sigma - (\Sigma - \operatorname{diag}\{s\})\,\Sigma^{-1}\,(\Sigma - \operatorname{diag}\{s\}) \\ &= \Sigma - (\Sigma - \operatorname{diag}\{s\}) - \operatorname{diag}\{s\}\,\Sigma^{-1}\Sigma + \operatorname{diag}\{s\}\,\Sigma^{-1}\operatorname{diag}\{s\} \\ &= 2\operatorname{diag}\{s\} - \operatorname{diag}\{s\}\,\Sigma^{-1}\operatorname{diag}\{s\} \end{aligned}$$

es semidefinida positiva, lo que que es equivalente a que

$$\begin{bmatrix} \Sigma & diag\{s\} \\ diag\{s\} & 2diag\{s\} \end{bmatrix} \succ 0 \Longleftrightarrow \begin{matrix} diag\{s\} \succ 0 \\ 2\Sigma - diag\{s\} \succ 0 \end{matrix}$$

Aquí tenemos una condición que debe satisfacer s para que $\tilde{X}$ exista. Ahora bien, la existencia de $\tilde{U} \in \mathbb{R}^{n\times p}$, tal que $U^\top X = 0$ se garantiza por la condición de que $n \geq 2p$, en efecto, por el teorema de la dimensión $n \geq 2p$ es equivalente a que $dim(\tilde{U}) \geq n \geq 2p$, así $n - dim(\tilde{U}) \leq p$, por tanto la existencia de $\tilde{U}$ se dá. Ahora bien, dado que A es semidefinida positiva, por el teorema de Cholesky, se puede factorizar como $A = C^\top C$, donde C es $p \times p$. Cálculos directos muestran que si ponemos a $\tilde{X}$ de manera explícita como arriba, esta satisface la estructura que tiene la matriz G. La elección de s además de cumplir la condición anterior, debe satisfacer que $\tilde{X}_j^\top X_j = 1 - s_j \approx 0$, pues necesitamos que la variable original y su imitación sean diferentes (esto nos está dando una ortogonalidad entre los vectores). Entonces como se sugiere en (Barber et al., 2015), se pueden considerar dos tipos particulares:

Knockoffs equi-correlacionados: $s_j = 2\lambda_{min}(\Sigma) \wedge 1$ para todo j.

Knockoffs SDP: resolver el problema convexo

$$minimizar \underset{\substack{s_j \geq 0 \\ 2\Sigma \succ diag\{s\}}}{} \sum_j |1 - s_j| \Longleftrightarrow \text{minimizar} \underset{\substack{0 \leq s_j \leq 1 \\ 2\Sigma \succ diag\{s\}}}{} \sum_j (1 - s_j)$$

CAPÍTULO 4

IDENTIFICACIÓN DE GENES CON HMM KNOCKOFFS

4.1. Introducción

En esta sección explicaremos una aplicación directa de la metodología Knockoffs. Veremos un algoritmo que permite muestrear variables knockoffs cuando las covariables se modelan como un HMM. Este permite realizar inferencia en estudios GWAS controlando la tasa de falsos descubrimientos. El objetivo en GWAS es identificar qué marcadores en una variación genética influyen en el riesgo de una enfermedad o rasgo particular, eligiendo entre millones de SNPs. La idea es pues, tener un algoritmo de selección que sea capaz de detectar tantas variables relevantes como sea posible utilizando sólo un pequeño número de muestras, pero además debe garantizar replicabilidad.

Los GWAS presentan dos grandes desafíos. Primero, muchos fenotipos dependen de las variantes genéticas a través de mecanismos que son en su mayoría desconocidos y pueden involucrar interacciones. Segundo, debido a la presencia de correlaciones entre las variables explicativas, ya que los polimorfismos que ocupan posiciones cercanas en el genoma están estrechamente vinculados. Este tipo de problemas motivan la necesidad de métodos que puedan identificar variables importantes para fenómenos complejos, al tiempo que brindan garantías rigurosas de control de errores Tipo I, bajo supuestos más leves y bien justificados.

Cuando en Knockoffs conocemos la distribución de las covariables F_X, es posible generar un conjunto de variables artificiales, que sirven como un control negativo de las variables originales, y por tanto es posible estimar y controlar la tasa de falsos descubrimientos. En GWAS debido a la naturaleza de las relaciones entre las variantes genéticas y fenotipos, el supuesto de conocer F_X está bien fundamentada, pues a través de varios estudios, estos tienen a su disposición ricos modelos sobre cómo surgen y se propagan las variantes de ADN a través de las poblaciones humanas con el tiempo, es decir, la combinación de conocimiento teórico y datos da una buena comprensión de F_X.

En este capitulo estudiaremos el algoritmo propuesto por Sesia et al. (2019), para muestrear variables knockoffs cuando las variables originales están distribuidas como un modelo oculto de Markov (HMM). Como vimos en capitulos anteriores, los HMM se han adoptado ampliamente para describir los haplotipos, es decir, la secuencia de alelos en una serie de marcadores a lo largo de un cromosoma. El éxito de estos algoritmos en la reconstrucción de genotipos parcialmente observados se puede probar empíricamente, y su precisión realizada es un testimonio del hecho de que los modelos ocultos de Markov ofrecen una buena descripción fenomenológica de la dependencia entre las variables explicativas en los GWAS.

4.2. Knockoffs para Candenas de Markov

Trabajaremos sobre cadenas de Markov a tiempo discreto. Sea $X = (X_1, \ldots, X_p)$ un vector de variables aleatorias, cada una tomando valores en el espacio de estados finito S. Diremos que X es una cadena de Markov a tiempo discreto si su función másica de probabilidad puede escribirse como

$$P\left(X_1 = x_1, \ldots, X_p = x_p\right) = q_1(x_1) \prod_{j=2}^{p} \mathcal{Q}_j(x_j \mid x_{j-1}) \tag{4.1}$$

donde $q_1(x_1)$ denota la distribución marginal del primer elemento de la cadena y las matrices de transición entre variables consecutivas son $\mathcal{Q}_j(x_j \mid x_{j-1}) = P\left(X_j = x_j \mid X_{j-1} = x_{j-1}\right)$. El siguiente resultado proporciona una forma de muestrear copias exactas knockoffs de una cadena discreta de Markov.

Teorema 4.2.1. *Supongamos que X se distribuye como la cadena de Markov en (4.1), con parámetros conocidos $(q_1, \mathcal{Q})$. Entonces, se puede obtener una copia Knockoff $\tilde{X}$ mediante*

muestreo secuencial, con sólo una iteración sobre $j = 1, \ldots, p$, la j-ésima variable knockoff $\tilde{X}_j$ tiene la forma:

$$P(\tilde{X}_j = \tilde{x}_j \mid x_{-j}, \tilde{x}_{1:(j-1)}) = \begin{cases} \frac{q_1(\tilde{x}_1)\mathcal{Q}_2(x_2|\tilde{x}_1)}{\mathcal{N}_1(x_2)} & si \quad j = 1 \\ \frac{\mathcal{Q}_j(\tilde{x}_j|x_{j-1})\mathcal{Q}_j(\tilde{x}_j|\tilde{x}_{j-1})\mathcal{Q}_{j+1}(x_{j+1}|\tilde{x}_j)}{\mathcal{N}_{j-1}(\tilde{x}_j)\mathcal{N}_j(x_{j+1})} & si \quad 1 < j < p \\ \frac{\mathcal{Q}_p(\tilde{x}_p|x_{p-1})\mathcal{Q}_p(\tilde{x}_p|\tilde{x}_{p-1})}{\mathcal{N}_{p-1}(\tilde{x}_p)\mathcal{N}_p(1)} & si \quad j = p \end{cases} \tag{4.2}$$

donde la función de normalización $\mathcal{N}_j : S \longrightarrow \mathbb{R}_+$ es definida recursivamente como

$$\mathcal{N}_j(k) = \begin{cases} \sum_{l \in S} q_1(l)\mathcal{Q}_2(k \mid l) & si \quad j = 1 \\ \sum_{l \in S} \frac{\mathcal{Q}_j(l|x_{j-1})\mathcal{Q}_j(l|\tilde{x}_{j-1})\mathcal{Q}_{j+1}(k|l)}{\mathcal{N}_{j-1}(l)} & si \quad 1 < j < p \\ \sum_{l \in S} \frac{\mathcal{Q}_p(l|x_{p-1})\mathcal{Q}_p(l|\tilde{x}_{p-1})}{\mathcal{N}_{p-1}(l)} & si \quad j = p \end{cases} \tag{4.3}$$

Por tanto, el algoritmo (6) muestra un procedimiento exacto para muestrear copias Knockoffs de una cadena de Markov.

Algoritmo 6 Copias Knockoff de una cadena de Markov discreta.

for $j = 1$ to $j = p$ **do**
  for $k \in S$ **do**
    Calcular $N_j(k)$ de acuerdo a (4.3)
  end for
  Muestrear $\tilde{X}_j$ de acuerdo a (4.2)
end for

En cada paso j del Algoritmo (6), la evaluación de la función de normalización $\mathcal{N}_j(k)$ implica una suma sobre todos los elementos del espacio de estado finito S y depende solo del anterior $\mathcal{N}_{j-1}(\cdot)$. Dado que esta operación debe repetirse para todos los valores de k, el muestreo de la variable j-ésima knockoff requiere $O(|S|^2)$ tiempo. Este procedimiento es secuencial y genera una variable de imitación a la vez. Por lo tanto, el tiempo total de cálculo es $O(p|S|^2)$, mientras que la memoria requerida es $O(|S|)$. Veamos ahora la prueba del teorema.

Demostración. Por el lema (4) (pág 47) basta con mostrar que el Algoritmo (6) muestrea $\tilde{X}_j$ de la distribución condicional de X_j dadas las otras variables originales X_{-j} y todas las copias knockoff $\tilde{X}_{1:(j-1)}$ que ya se han muestreado: $\tilde{X}_j \sim P(X_j \mid X_{-j}, \tilde{X}_{1:(j-1)})$, para $j = 1, \dots, p$, procederemos por inducción para demostrar el resultado. Supongamos que la hipótesis inductiva vale para algún $j \in \{1, \dots, p-1\}$, entonces el algoritmo muestrea todas las copias knockoff $\tilde{X}_i$ para $i \leq j$, de $P(X_i \mid X_{-i}, \tilde{X}_{1:(i-1)})$, respectivamente. Veamos que $\tilde{X}_{j+1}$ es muestreada de $P(X_{j+1} \mid X_{-(j+1)}, \tilde{X}_{1:j})$. Definamos $\mathcal{Q}_p(k \mid l) = 1$, para todo $k, l \in \{1, \dots, K\}$. Por propiedades tenemos

$$
\begin{aligned}
&P(X_{j+1} = \tilde{x}_{j+1} \mid X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:j} = \tilde{x}_{1:j}) \\
&\quad \propto P(X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:j} = \tilde{x}_{1:j}) \\
&\quad \propto P(X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:(j-1)} = \tilde{x}_{1:(j-1)}) \\
&\qquad\qquad \times P(\tilde{X}_j = \tilde{x}_j \mid X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:(j-1)} = \tilde{x}_{1:(j-1)}) \\
&\quad \propto P(X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}) P(\tilde{X}_{1:(j-1)} \mid X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}) \\
&\qquad\qquad \times P(\tilde{X}_j = \tilde{x}_j \mid X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:(j-1)} = \tilde{x}_{1:(j-1)})
\end{aligned}
$$

Como estamos interesados sólo en las dependencias de $\tilde{x}_{j+1}$, entonces el primer término de la expresión anterior se puede simplificar como

$$
P(X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}) \propto \mathcal{Q}_{j+1}(\tilde{x}_{j+1} \mid x_j) \mathcal{Q}_{j+2}(x_{j+2} \mid \tilde{x}_{j+1})
$$

Por otro parte, de la hipótesis de inducción se deduce que el segundo término es constante con respecto a $\tilde{x}_{j+1}$, pues de acuerdo con (4.2), la distribución de $\tilde{X}_i$ sólo depende de X_{i-1}, X_{i+1} y $\tilde{X}_{i-1}$ para todo $i \leq j$. Por tanto, la distribución condicional de $\tilde{X}_{1:(j-1)}$ depende sólo de $X_{1:j}$. Así obtenemos

$$
\begin{aligned}
P(X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}) &= \frac{\mathcal{Q}_j(\tilde{x}_j \mid x_{j-1}) \mathcal{Q}_j(\tilde{x}_j \mid \tilde{x}_{j-1}) \mathcal{Q}_{j+1}(\tilde{x}_{j+1} \mid \tilde{x}_j)}{\mathcal{N}_{j-1}(\tilde{x}_j) \mathcal{N}_j(\tilde{x}_{j+1})} \\
&\propto \frac{\mathcal{Q}_{j+1}(\tilde{x}_{j+1} \mid \tilde{x}_j)}{\mathcal{N}_j(\tilde{x}_{j+1})}.
\end{aligned}
$$

La igualdad anterior se deriva del hecho de que el Algoritmo (6) muestrea $\tilde{X}_j$ condicionalmente independiente de X_j, como se ve en (4.2). Por tanto tenemos que

$$
\begin{aligned}
P(X_{j+1} = \tilde{x}_{j+1} \mid X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:j} = \tilde{x}_{1:j}) \propto \\
\mathcal{Q}_{j+1}(\tilde{x}_{j+1} \mid x_j) \mathcal{Q}_{j+1}(x_{j+2} \mid \tilde{x}_{j+1}) \frac{\mathcal{Q}_{j+1}(\tilde{x}_{j+1} \mid \tilde{x}_j)}{\mathcal{N}_j(\tilde{x}_{j+1})}
\end{aligned}
$$

Con esto hemos probado que se cumple para $j+1$. El paso base $j=1$, se sigue del hecho de que el Algoritmo (6) muestrea $\tilde{X}_1$, independientemente de X_1, de

$$\begin{aligned} P(X_1 = \tilde{x}_1 \mid X_{-1} = x_{-1}) = P(X_1 = \tilde{x}_1 \mid X_2 = x_2) &\propto P(X_1 = \tilde{x}_1, X_2 = x_2) \\ &= q_1(\tilde{x}_1)\mathcal{Q}_2(x_2 \mid \tilde{x}_1) \end{aligned}$$

■

4.3. Knockoffs para Modelos Ocultos de Markov(HMM)

Recordemos que $X = (X_1, \ldots, X_p)$, tomando valores en un espacio de estados finito S, se distribuye como un modelo oculto de Markov con K estados ocultos si existe un vector $Z = (Z_1, \ldots, Z_p)$ tal que

$$\begin{cases} Z \sim \text{MC}(q_1, \mathcal{Q}) & \text{(Cadena de Markov discreta latente)} \\ X_j \mid Z \sim X_j \mid Z_j \sim f_j(X_j \mid Z_j) & \text{(Distribución de emisión)} \end{cases} \tag{4.4}$$

donde $\text{MC}(q_1, \mathcal{Q})$ indica la distribución inicial y probabilidades de transición de la cadena según (4.1), donde cada elemento X_j toma valores en $\{1, \ldots, K\}$. Condicionando a Z cada X_j se muestrea independientemente de la distribución de emisión $f_j(X_j \mid Z_j)$.

Ahora estudiaremos un algoritmo para este tipo de modelos utilizando la metodología Knockoff. Las variables observadas X en el modelo HMM (4.4) no satisfacen la propiedad de Markov, por esta razón como se discutió en capítulos anteriores, vamos a considerar el algoritmo *backward-forward*.

Algoritmo 7 Copias knockoff para un HMM

1: Muestrear $Z = (Z_1, \ldots, Z_p)$ de $P(Z \mid X = x)$ usando el algoritmo (8).
2: Muestrear una copia knockoff $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_p)$ de $Z = (Z_1, \ldots, Z_p)$ usando el algoritmo (6).
3: Muestrear $\tilde{X}$ de $P(X \mid Z = \tilde{z})$, lo cual es fácil por la independencia condicional.

En la figura (4.1) se muestra una representación del algoritmo (7). En la primera etapa, la cadena de Markov latente se imputa mediante muestreo de la distribución condicional de Z dada X. Esto se hace de manera eficiente con el Algoritmo (8). Una vez que se ha muestreado

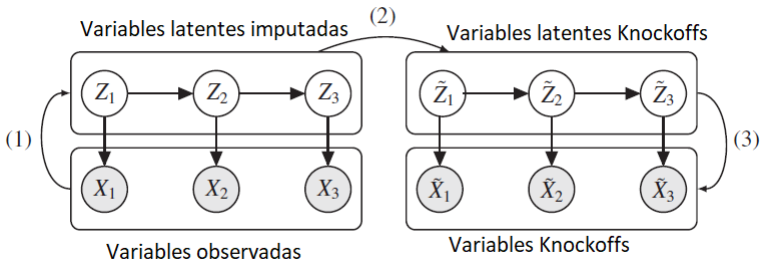


Figura 4.1: Ilustración del algoritmo (7) para el caso en que $p = 3$.

Z, se puede obtener una copia knockoff de $\tilde{Z}$ con el Algoritmo (6). Finalmente, se muestrea $\tilde{X}$ a partir de $P(X \mid Z = \tilde{z})$, lo cual es fácil debido a la independencia condicional entre las distribuciones de emisiones en el modelo oculto de Markov.

Algoritmo 8 Muestreo Forward-Backward para un HMM

Inicializar $\alpha_0(k) = 1$, $\mathcal{Q}_1(k \mid l) = q_1(k)$ y $\mathcal{Q}_{p+1}(k \mid l) = 1$ para todo $1 \leq k, l \leq K$
for $j = 1$ to $j = p$ (paso forward) **do**
  for $k = 1$ to $k = K$ **do**
    $\alpha_j(k) = f_j(x_j \mid k) \sum_{l=1}^{K} \mathcal{Q}_j(k \mid l)\alpha_{j-1}(l)$.
  end for
end for
for $j = p$ to $j = 1$(paso backward) **do**
  Muestrear z_j de acuerdo a $\pi_j(z_j) = \frac{\mathcal{Q}_{j+1}(z_{j+1}|z_j)\alpha_j(z_j)}{\sum_{l=1}^{K} \mathcal{Q}_{j+1}(z_{j+1}|k)\alpha_j(k)}$
end for
Retornar $(z_1, \ldots, z_p)$

El algoritmo (8) es una adaptación del algoritmo Backward-Forward visto en el capitulo 2. Notemos que el teimpo de ejecución requerido por los algoritmos (6) y (8) es de $O(pK^2)$, mientras que la complejidad del estado final es $O(p|S|)$. Por tanto, el algoritmo (7) tarda $O(pK^2|S|)$. La prueba del siguiente teorema se puede encontrar en Sesia et al. (2019).

Teorema 4.3.1. *Supongamos que $X = (X_1, \ldots, X_p)$ es observado a partir de un HMM dado (4.4), con parámetros conocidos $(q_1, \mathcal{Q}, f)$. Entonces, el algoritmo (8) produce una muestra exacta de la distribución condicional de su cadena de Markov latente $Z = (Z_1, \ldots, Z_p)$ dado $X = (X_1, \ldots, X_p)$.*

Teorema 4.3.2. *Supongamos que $X = (X_1, \ldots, X_p)$ se observa a partir del HMM dado en (4.4), con parámetros conocidos $(q_1, \mathcal{Q}, f)$. Entonces $(\tilde{X}, \tilde{Z})$ generado por el algoritmo (7) es una copia knockoff de (X, Z). Es decir, para cualquier $T \subset \{1, \ldots, p\}$,*

$$\left\{(X, \tilde{X})_{sw(T)}, (Z, \tilde{Z})_{sw(T)}\right\} \stackrel{d}{=} \left\{(X, \tilde{X}), (Z, Z)\right\} \tag{4.5}$$

En particular, esto implica que $\tilde{X}$ es una copia knockoff de X.

Demostración. Condicionando sobre los valores de las variables latentes, tenemos

$$\begin{aligned}
&P((X, \tilde{X}) = (x, \tilde{x})_{sw(T)}, (Z, \tilde{Z}) = (z, \tilde{z})_{sw(T)}) \\
&= P((X, \tilde{X}) = (x, \tilde{x})_{sw(T)} \mid (Z, \tilde{Z}) = (z, \tilde{z})_{sw(T)}) P((Z, \tilde{Z}) = (z, \tilde{z})_{sw(T)}) \\
&= P((X, \tilde{X}) = (x, \tilde{x}) \mid (Z, \tilde{Z}) = (z, \tilde{z})) P((Z, \tilde{Z}) = (z, \tilde{z})_{sw(T)}) \\
&= P((X, \tilde{X}) = (x, \tilde{x}) \mid (Z, \tilde{Z}) = (z, \tilde{z})) P((Z, \tilde{Z}) = (z, \tilde{z}))
\end{aligned}$$

La primera igualdad anterior se deriva de la línea 1 del algoritmo (7) y del teorema (4.3.1). La segunda igualdad se deriva de la independencia condicional de las distribuciones de emisiones en un modelo oculto de Markov. Por último, la tercera igualdad se deduce de que $\tilde{Z}$ es una copia knockoff de Z, como se establece en el teorema (4.2.1). ■

4.4. HMM en GWAS

Esta sección está basada en el algoritmo *fastPhase* estudiando en el capítulo anterior. Recordemos que un estudio de asociación del genoma completo (GWAS) la respuesta Y es el estado de una enferemdad o un rasgo cuantitativo de interés, mientras que cada muestra de X consiste en el genotipo de un conjunto de SNPs. Vamos a considerar el caso en que $X \in \{0, 1, 2\}^p$ recoge genotipos sin fase. Inicialmente, nos enfocaremos en un solo cromosoma, pues se extiende de manera natural, ya que se supone que distintos cromosomas son independientes.

El genotipo sin fase de un individuo puede verse como la suma de componente a componente de dos secuencias no observadas, llamadas haplotipos $H = (H_1, \ldots, H_p)$, donde $H_i \in \{0, 1\}$ es una variable binara que representa el alelo en el i-ésimo marcador. Recordemos que el principal supuesto es que los dos haplotipos son independientes y se distribuyen de manera idéntica como HMM. La figura (4.2) muestra este caso para $p = 3$. Esta muestra una secuencia de polimorfismos de genotipo (sombreados) como la suma componente a componente de dos haplotipos de un modelo de Markov oculto (blanco).

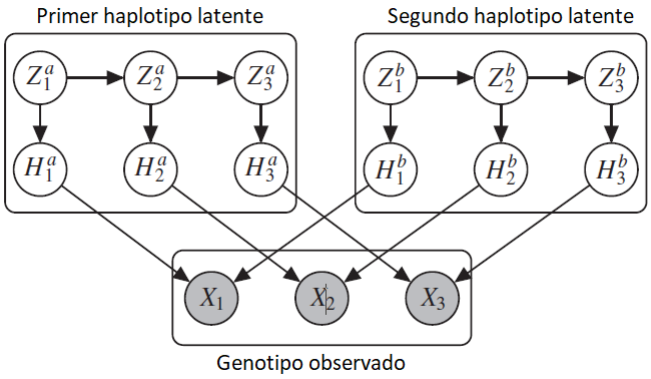


Figura 4.2: Secuencia para $p = 3$.

Iniciamos entonces con el estudio de una sola secuencia H. El HMM es definido como

$$\begin{cases} Z \sim \text{MC}(q_1^h, \mathcal{Q}^h) & \text{(Cadena de Markov discreta latente para un haplotipo,)} \\ H_j \mid Z \sim H_j \mid Z_j \sim f_j^h(H_j \mid Z_j) & \text{(Distribución de emisión del haplotipo)} \end{cases} \tag{4.6}$$

con cadena de Markov latente $Z = (Z_1, \ldots, Z_p)$ cuyos elementos indican la pertenencia a uno de los K grupos de haplotipos estrechamente relacionados. Estos grupos se caracterizan por las frecuencias alélicas específicas en los diversos marcadores, de modo que se puede ver H como un mosaico de segmentos, cada uno de los cuales se origina en uno de los K patrones distintos que se pueden tomar libremente como representantes del genoma de los fundadores de la población. Este modelo proporciona una buena descripción de los patrones locales de correlación, pero es de naturaleza fenomenológica y no debe interpretarse como una representación precisa de la secuencia real de mutaciones y recombinaciones que originan los haplotipos de la población. La distribución marginal del primer elemento de la cadena oculta de Markov Z es

$$q_1^h(k) = \alpha_{1,k}, \quad k \in \{1, \ldots, K\}$$

y las matrices de transición

$$\mathcal{Q}_j^h(k' \mid k) = \begin{cases} e^{-r_j} + (1 - e^{-r_j})\alpha_{j,k'} & si \quad k' = k \\ (1 - e^{-r_j})\alpha_{j,k'} & si \quad k' \neq k \end{cases}$$

Los parámetros $\alpha = (\alpha_{j,k})_{k \in [K], j \in [p]}$ describe la inclinación de patrones diferentes a suceder

mutuamente. La aparición de una transición está regulada por los valores de $r = (r_1, \ldots, r_p)$, que están intuitivamente relacionados con las tasas de recombinación genética. Una vez que se fija una secuencia de segmentos ancestrales, el alelo H_j en la posición j es muestreado de la distribución de emisión

$$f_j^h(h_j; z_j, \theta) = \begin{cases} 1 - \theta_{j,z_j} & si \quad h_j = 0 \\ \theta_{j,z_j} & si \quad h_j = 1 \end{cases}$$

Los parámetros $\theta = (\theta_{j,k})_{k\in[K],j\in[p]}$ representan las probabilidades de que los alelos sean iguales a 1, para cada uno de los p polimorfismos y los K patrones del haplotipo ancestral. Estos pueden estimarse junto con α y r.

Una vez definida la distribución de H, volvemos nuestra atención al vector de genotipo observado. Por definición, el genotipo X de un individuo se obtiene emparejando, marcador por marcador, los alelos en cada haplotipo y descartando información sobre el haplotipo de origen, es decir, la fase. Luego, bajo supuestos estándar como el equilibrio de Hardy-Weinberg (HW), la población de la que se toma al azar el vector genotipo de un sujeto puede describirse como la suma de elemento a elemento de dos distribuciones de haplotipos independientes e idénticas descritas por el modelo anterior.

En consecuencia, su distribución también es un modelo oculto de Markov. La cadena de Markov latente tiene estados bivariados, que corresponden a pares no ordenados de estados latentes de haplotipos. Estos pueden tomar $K(K+1)/2$ valores posibles. Por esta construcción, se deduce que las probabilidades de estado inicial para el modelo de genotipo son

$$q_1^g(\{k_a, k_b\}) = \begin{cases} (\alpha_{1,k_a})^2 & si \quad k_a = k_b \\ 2\alpha_{1,k_a}\alpha_{1,k_b} & si \quad k_a \neq k_b \end{cases}$$

y las matrices de transición son

$$\mathcal{Q}_j^g(\{k'_a, k'_b\} \mid \{k_a, k_b\}) = \begin{cases} \mathcal{Q}_j^h(k'_a \mid k_a)\mathcal{Q}_j^h(k'_b \mid k_b) + \mathcal{Q}_j^h(k'_b \mid k_a)\mathcal{Q}_j^h(k'_a \mid k_b) & si \\ & k_a \neq k_b \\ \mathcal{Q}_j^h(k'_a \mid k_a)\mathcal{Q}_j^h(k'_b \mid k_b) & \text{otro caso} \end{cases}$$

Las probabilidades de emisión para X_j son

$$f_j^h(h_j;\{k_a,k_b\},\theta)=\begin{cases}(1-\theta_{j,k_a})(1-\theta_{j,k_b}) & si \quad x_j=0\\ \theta_{j,k_a}(1-\theta_{j,k_b})+(1-\theta_{j,k_a})\theta_{j,k_b} & si \quad x_j=1\\ \theta_{j,k_a}\theta_{j,k_b} & si \quad x_j=2\end{cases}$$

En este capítulo vimos una aplicación directa de lo estudiado en el capítulo 3. Como se discutió antes, no siempre es fácil construir variables Knockoffs, a pesar de que existe un algoritmo que está bien fundamentado. La construcción de copias de knockoffs requiere conocer la distribución de las covariables, sin embargo, el conocimiento exacto no es realista en aplicaciones prácticas y, en última instancia, es inevitable cierto grado de aproximación. El modelo construido ofrece una descripción sensible y manejable de genotipos reales, tiene sentido estimar los parámetros (r,α,θ) a partir de los datos. Todos los parámetros se pueden estimar eficientemente con en el algoritmo de imputación fastPHASE. Esto se ajusta al modelo descrito, con el propósito original de recuperar las observaciones faltantes, y proporciona convenientemente las estimaciones de $(\hat{r},\hat{\alpha},\hat{\theta})$. Una ventaja importante del modelo oculto de Markov es que el número de parámetros solo crece linealmente en p, lo que reduce en gran medida el riesgo de sobreajuste en comparación con otros métodos por ejemplo con una aproximación gaussiana multivariada. La complejidad de este modelo está controlada por el número de haplotipos K, cuyos valores típicos recomendados en Scheet and Stephens (2006) y pueden ajustarse con validación cruzada. A pesar de que la garantía teórica del control de la tasa de falsos descubrimientos para knockoffs requiere que se conozca F_X, en Sesia et al. (2019), muestran mediante simulaciones que el procedimiento es robusto.

CAPÍTULO 5

ANÁLISIS DE DATOS

5.1. Introducción

En esta sección vamos a ejemplificar todo lo visto anteriormente. Primero veremos cómo es la codificación estándar de los datos. Iniciaremos este capítulo explicando un poco sobre las convenciones comunes que se suele utilizar en este tipo de estudios. Recordemos que la fase o la estimación de haplotipos se refiere al proceso de estimación estadística de haplotipos a partir de datos de genotipos. Las tecnologías de genotipado obtienen información de genotipo de los SNPs que mezcla la información genética de cromosomas. Sin embargo, muchos análisis genéticos requieren información de haplotipo, que es la información genética en cada cromosoma. Consideremos un individuo con un genotipo heterocigoto en cada uno de los tres SNP en una región (ver tabla 5.1). Hay cuatro configuraciones posibles de haplotipos que son consistentes con los datos del genotipo, es decir, posibles patrones de fase (1)-(4). Supongamos que las frecuencias de haplotipos están disponibles de otros individuos en la población en estos sitios, estas frecuencias pueden haberse estimado a partir de datos de la población sin modelos adicionales o de un modelo que explica los procesos biológicos de recombinación y mutación.

El equilibrio de Hardy-Weinberg (HWE) puede verse afectado por una serie de fuerzas, que incluyen mutaciones, selección natural, apareamiento no aleatorio, deriva genética y flujo de genes. Por ejemplo, las mutaciones alteran el equilibrio de las frecuencias alélicas al

introducir nuevos alelos en una población. Del mismo modo, la selección natural y el apareamiento no aleatorio interrumpen el equilibrio de Hardy-Weinberg porque provocan cambios en las frecuencias genéticas. Esto ocurre porque ciertos alelos ayudan o dañan el éxito reproductivo de los organismos que los transportan. Otro factor que puede alterar este equilibrio es la deriva genética, que ocurre cuando las frecuencias de los alelos aumentan o disminuyen por casualidad y generalmente ocurren en poblaciones pequeñas. El flujo de genes, que ocurre cuando la reproducción entre dos poblaciones transfiere nuevos alelos a una población, también puede alterar dicho equilibrio.

Debido a que todas estas fuerzas disruptivas ocurren comúnmente en la naturaleza, el equilibrio de Hardy-Weinberg rara vez se aplica en la realidad. Por lo tanto, el equilibrio de Hardy-Weinberg describe un estado idealizado, y las variaciones genéticas en la naturaleza pueden medirse como cambios de este estado de equilibrio. Sin embargo, como se discutió antes, para estudios de GWAS este principio si es aplicable. Por tanto, la frecuencia de población de un par de haplotipos se obtiene utilizando el principio; el factor de dos en la frecuencia de los pares de haplotipos explica las posibles asignaciones de origen materno y paterno a los dos haplotipos. Las probabilidades posteriores de los datos escalonados se obtienen de las frecuencias de población de los posibles pares de haplotipos. En este ejemplo, la probabilidad posterior de fase (2) ($93\,\%$) es mucho mayor que la de fase (3) ($7\,\%$).

Haplotipos	**Fase (1)**	**Fase (2)**	**Fase (3)**	**Fase (4)**
A C	A C	A C	A C	A C
G T	G T	G T	T G	T G
A T	A T	T A	A T	T A
FHP	$55\,\%$ $0\,\%$	$15\,\%$ $5\,\%$	$2\,\%$ $3\,\%$	$0\,\%$ $20\,\%$
FPPHN	$0\,\%$	$2 \times (15\,\% \times 5\,\%) = 1.5\,\%$	$2 \times (2\,\% \times 3\,\%) = 0.12\,\%$	$0\,\%$
Posterior FPPHN	$0\,\%$	$1.5\,\%/(1.5\,\% + 0.12\,\%) = 93\,\%$	$0.12\,\%/(1.5\,\% + 0.12\,\%) = 7\,\%$	$0\,\%$

Tabla 5.1: Secuencia de haplotipos y posibles fase

FHP: Frecuencia de haplotipos de población

FPPHN: Frecuencia poblacional del par de haplotipos no ordenados

La codificación de este tipo de estudios es como se sigue: supongamos que de dos individuos obtenemos los siguientes haplotipos en fase, tomados del cromosoma 12.

Esta información suele representarse como se muestra en la tabla (5.3), en ella vemos el cromosoma, la posición donde se encuentra, la referencia y alternativo, significan el valor que toman en estudios de *genoma de referencia* (más adelante hablaremos sobre esto) y los

Individuo 1	Individuo 2
C | T	T | T
T | T	T | A
ATC | G – –	ATC | ATC

Tabla 5.2: Ejemplo de haplotipos de dos individuos.

diferentes valores de los genotipos de cada individuo.

Chr	Posición	Ref	Alt	Ind1-H1	Ind1-H2	Ind2-H1	Ind2-H2
12	2147839	C	T	C	T	T	T
12	2147913	T	A	T	T	T	A
12	2152883	G – –	ATC	ATC	G – –	ATC	ATC

Tabla 5.3: Representación de haplotipos en fase

Una forma equivalente de esta representación en forma numérica se muestra en la tabla (5.4). Cuando asginamos 0, significa que el valor del genotipo es igual al de la referencia y cuando asignamos el valor de 1, significa que es igual al alternativo.

Chr	Posición	Ref	Alt	Ind1-H1	Ind1-H2	Ind2-H1	Ind2-H2
12	2147839	C	T	0	1	1	1
12	2147913	T	A	0	0	0	1
12	2152883	G – –	ATC	1	1	1	1

Tabla 5.4: Representación numérica de haplotipos en fase

Sin embargo cuando no se conoce la fase, lo que se hace es representar la información como en la tabla (5.5). Aquí, lo que se hace es contar cuantas instancias del alelo alternativo vemos en cada SNP, por ejemplo, T aparece una vez en el primer SNP C|T del individuo 1, mientras que la T aparece dos veces en el individuo 2 (T|T).

Una forma como se obtienen las referencias alélicas es a través de las frecuencias alélicas, esto consisten en tomar todos los individuos y asignar la menor o mayor frecuencia como la referencia. Por ejemplo, en la tabla (5.6), vemos 5 individuos con 4 posiciones donde difieren las secuencias. Una forma de seleccionar la referencia es tomar la mayor o menor frecuencia.

Chr	Posición	Ref	Alt	Ind1	Ind1
12	2147839	C	T	1	2
12	2147913	T	A	0	1
12	2152883	G – –	ATC	1	2

Tabla 5.5: Matriz de genotipos

Ind 1	C	A	C	G	T	C	A	C	T	T	C	A	C	G	T	A	T	G
	C	T	C	C	T	C	T	C	A	T	C	A	C	–	–	–	T	G
Ind 2	C	T	C	C	T	C	A	C	T	T	C	A	C	G	T	A	T	G
	C	T	C	C	T	C	A	C	T	T	C	A	C	–	–	–	T	G
Ind 3	C	A	C	G	T	C	T	C	A	T	C	A	C	G	T	A	T	G
	C	A	C	G	T	C	T	C	A	T	C	A	C	G	T	A	T	G
Ind 4	C	T	C	T	T	C	A	C	T	T	C	A	C	–	–	–	T	G
	C	T	C	C	T	C	A	C	T	T	C	A	C	–	–	–	T	G
Ind 5	C	T	C	C	T	C	A	C	T	T	C	A	C	–	–	–	T	G
	C	A	C	C	T	C	A	C	T	T	C	A	C	G	T	A	T	G

Tabla 5.6: Frecuencia alélica

Por ejemplo para la primera posición, hay 6 ocurrencias de T y 4 de A, entonces podríamos seleccionar la referencia como la mayor, en este caso T y el alternativo como el menor (A).

Por último, para obtener la frecuencia de población se hace uso de *genoma de referencia* que es una base de datos digital de secuencias de nucleótidos, creada por científicos como un ejemplo representativo del conjunto de genes en un organismo individual idealizado de una especie. A medida que se ensamblan a partir de la secuenciación del ADN de varios donantes individuales, los genomas de referencia no representan con precisión el conjunto de genes de ningún organismo individual. En cambio, una referencia proporciona un mosaico haploide de diferentes secuencias de ADN de cada donante. Existen genomas de referencia para múltiples especies de virus, bacterias, hongos, plantas y animales. Los genomas de referencia se usan como una guía sobre la construcción de nuevos genomas.

El propósito de este análisis estadístico es identificar las variables importantes en la distribución condicional de $Y|X_1, \ldots, X_p$, donde Y es un fenotipo de interés (por ejemplo, el estado de una enfermedad) y X es una familia de SNPs. Esto se realiza probando p hipótesis

multivariadas que consisten en que Y es independiente del j-ésimo SNP X_j condicionado a las otras variables, para $j = 1, \ldots, p$, es decir,

$$Y \perp\!\!\!\perp X_j \mid X_{-j}$$

Cada X_j es un SNP que toma valores en $\{0, 1, 2\}$, suponiendo que la distribución de X puede escribirse como la suma de elementos de dos modelos de Markov ocultos independientes e idénticamente distribuidos, tenemos que para cada $j \in \{1, \ldots, p\}$

$$X_j = H_j^{(a)} + H_j^{(b)}$$

Los vectores $H^{(a)} = (H_1^{(a)}, \ldots, H_p^{(a)})$ y $H^{(b)} = (H_1^{(b)}, \ldots, H_p^{(b)})$, cuyos elementos pueden tomar valores en $\{0, 1, 2\}$, se distribuyen de acuerdo con el modelo oculto de Markov mostrado en el capítulo anterior.

La siguiente análisis se realizó con datos sintéticos creados a partir de un conjunto de datos disponible públicamente. Cabe aclarar, que se hicieron 100 repeteciones del experimento y los resultados fueron muy similares.

5.2. Datos y Análisis

El conjunto de datos consta de 10.000 SNPs de una porción del cromosoma 22 en las posiciones $0 - 26192121$, entre los genotipados por el Proyecto Internacional HapMap (versión $\#22$ - NCBI Build 36), este último siendo nuestro genoma de referencia.

Se analizó un estudio de 200 casos y 350 controles con 10 loci "causales", cuyos efectos multiplicativos para cada copia del alelo de riesgo se distribuyen uniformemente entre 1.25 y 1.5, es decir, el efecto de riesgo para determinar el riesgo de enfermedad de cada individuo. Las posiciones de los 10 loci causales son se resumen en la tabla (5.7).

Una vez que se cargan los datos, estamos listos para eliminar los SNP que no cumplen con los criterios mínimos debido a datos faltantes, baja variabilidad o errores de genotipado. La tabla (5.8) muestra un resumen de los datos a trabajar, donde

- Calls: la cantidad de llamadas válidas
- Call.rate: la proporción de genotipos llamados
- Certain.calls: proporción de SNP llamados con ciertas llamadas

Loci	Posición
1	15002875
2	16599872
3	16890873
4	16907889
5	21731173
6	22152390
7	23531242
8	25186766
9	25640922
10	25738063

Tabla 5.7: Posición de los loci.

- RAF: la frecuencia del alelo de "riesgo" (alelo B)
- MAF: la frecuencia del alelo menor
- P.AA: la frecuencia del genotipo homocigoto 1 (A / A)
- P.AB: La frecuencia del genotipo heterocigoto 2 (A / B)
- P.BB: La frecuencia del genotipo homocigoto 3 (B / B)
- z.HWE: una prueba z para el equilibrio de Hardy-Weinberg

Para un SNP determinado, el `Call.rate` se define como la proporción de individuos en nuestro estudio para los que no falta la información correspondiente del SNP. Por ejemplo, un `Call.rate` del $95\,\%$ para un determinado SNP significa que el $95\,\%$ de las personas tienen datos para este SNP, por lo tanto, para filtrar por `Call.rate`, miramos la columna de Call.rate en nuestro marco de datos y elegimos los SNP que tienen un porcentaje de observaciones faltantes por debajo de nuestro umbral elegido. En esta simulación, establecemos un umbral de `Call.rate` en 0.95 y luego seleccionamos los SNP que pasan el `Call.rate`. Notar que para nuestro caso `Call.rate` es 1, esto debido a que nuestros datos son sintéticos.

La frecuencia menor de alelo (MAF) se define como la frecuencia del alelo menos común en un sitio variable. Para nuestro caso, eliminamos los SNP cuya frecuencia de alelos menores es inferior al $1\,\%$. En otros casos, particularmente cuando el tamaño de la muestra es pequeño, podemos aplicar un punto de corte del $5\,\%$. Cuando tenemos una frecuencia de alelos menores (MAF) muy pequeña, significa que la mayoría de los individuos tienen dos copias de los mismos alelos principales.

	Calls	Call.rate	Certain.calls	RAF	MAF	P.AA	P.AB	P.BB	z.HWE
rs11089130	550	1.00	1.00	0.61	0.39	0.14	0.50	0.36	1.12
rs738829	550	1.00	1.00	0.80	0.20	0.05	0.31	0.64	-0.74
rs915674	550	1.00	1.00	0.86	0.14	0.02	0.23	0.75	-1.00
rs915675	550	1.00	1.00	0.85	0.15	0.03	0.25	0.72	-0.66
rs915677	550	1.00	1.00	0.94	0.06	0.01	0.10	0.88	-3.33
rs9604721	550	1.00	1.00	0.02	0.02	0.95	0.05	0.00	0.59

Tabla 5.8: Resumen de datos.

Una vez que se filtran las muestras, volvemos al nivel de filtrado SNP y aplicamos un control del equilibrio de *Hardy-Weinberg*(HWE). Si las frecuencias de genotipo o alelo se desvían significativamente de HWE, puede indicar errores sistemáticos en el genotipado, estructura de población inesperada, presencia de regiones homólogas en el genoma, asociación con rasgo en estudios de casos y controles. Como el último de ellos es menos probable, la desviación de HWE es un indicador de que un marcador debe descartarse. Es cierto a menos que los marcadores adyacentes en LD entre sí violen HWE.Para nuestro caso, sólo eliminamos SNP con p-valores correspondientes a la estadística de prueba HWE de menos de 10^{-6}. Solo probamos HWE en controles debido a la posible violación de HWE causada por la asociación de enfermedades.

Dado que muchos SNP están extremadamente correlacionados y el tamaño de la muestra es pequeño, estos datos permiten una resolución suficientemente alta para distinguir los SNP verdaderamente importantes de sus "vecinos" más similares. La pregunta científica más convincente radica en la identificación de grupos relevantes de sitios estrechamente vinculados, en lugar de SNP individuales. Se creó entonces un dendrograma de agrupamiento jerárquico utilizando las correlaciones de muestra como medida de similitud, y luego podarlo para que no haya dos grupos con correlaciones cruzadas por encima de un valor umbral de 0.75. Aquí

obtuvimos un total de 2323 grupos, cuya media es de 1172.

Después de esto seleccionamos los grupos representativos, el podado se lleva a cabo eligiendo un SNP representativo de cada grupo. El SNP más prometedor en cada grupo se selecciona de acuerdo con los p-valores de asociación marginal (de la prueba *Cochran-Armitage*) calculados con el $20\,\%$ de los datos.

Luego de obtener las estimaciones de los parámetros obtenidas por fastPHASE, es decir, las estimaciones para nuestro modelo HMM, estamos listos para crear las variables knockoffs para el modelo. Construimos la variable knockoffs $\tilde{H}$ para H, donde $\tilde{X} = \tilde{H}^{(a)} + \tilde{H}^{(b)}$. Sea $H^{(a)}$ las filas impares de $\tilde{H}$, y $H^{(b)}$ las filas pares de $\tilde{H}$ y $X = \tilde{H}^{(a)} + \tilde{H}^{(b)}$. Como $((H, \tilde{H})_{sw(S)} \overset{d}{=} (H, \tilde{H})$, entonces

$$((H^{(a)}, \tilde{H}^{(a)})_{sw(S)} \overset{d}{=} (H^{(a)}, \tilde{H}^{(a)}) \text{ y } ((H^{(b)}, \tilde{H}^{(b)})_{sw(S)} \overset{d}{=} (H^{(b)}, \tilde{H}^{(b)}),$$

por tanto,

$$\begin{aligned}
((X, \tilde{X})_{sw(S)}) &= ((H^{(a)} + H^{(b)}, \tilde{H}^{(a)} + \tilde{H}^{(b)})_{sw(S)}) \\
&= ((H_1^{(a)} + H_1^{(b)}, \ldots, H_p^{(a)} + H_p^{(b)}, \tilde{H}_1^{(a)} + \tilde{H}_1^{(b)}, \ldots, \tilde{H}_p^{(a)} + \tilde{H}_p^{(b)}))_{sw(S)} \\
&= ((H^{(a)}, \tilde{H}^{(a)}) + (H^{(b)}, \tilde{H}^{(b)}))_{sw(S)} \\
&= ((H^{(a)}, \tilde{H}^{(a)})_{sw(S)} + (H^{(b)}, \tilde{H}^{(b)})_{sw(S)} \\
&= H^{(a)}, \tilde{H}^{(a)}) + H^{(b)}, \tilde{H}^{(b)}) \\
&= (H^{(a)} + H^{(b)}, \tilde{H}^{(a)} + \tilde{H}^{(b)}) \\
&= (X, \tilde{X})
\end{aligned}$$

esto significa que $\tilde{X}$ es una copia knockoffs de X. Al hacer la comparación de medias (figura 5.1), vemos que como la estructura de la distribución de X y $\tilde{X}$ debería de ser la misma, esperamos que $EX_i = E\tilde{X}_i$, por tanto, el diagrama de dispersión por pares debe distribuirse a lo largo de una línea recta, lo cual se da para nuestro caso.

Si hacemos un gráfico de $corr(X_j, X_{j+1})$ vs $corr(\tilde{X}_j, \tilde{X}_{j+1})$ esperamos que el diagrama de dispersión por pares se distribuya a lo largo de una línea recta, pues, para $S = \{j, j+1\}$, tenemos

$$\begin{aligned}
(X, \tilde{X})_{sw(S)} &\overset{d}{=} (X_1, \ldots, \tilde{X}_j, \tilde{X}_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, X_j, X_{j+1}, \ldots, \tilde{X}_p) \\
&\overset{d}{=} (X_1, \ldots, X_j, X_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_j, \tilde{X}_{j+1}, \ldots, \tilde{X}_p)
\end{aligned}$$

por lo cual, $corr(X_j, X_{j+1}) = corr(\tilde{X}_j, \tilde{X}_{j+1})$, esto se muestra en la figura (5.2).

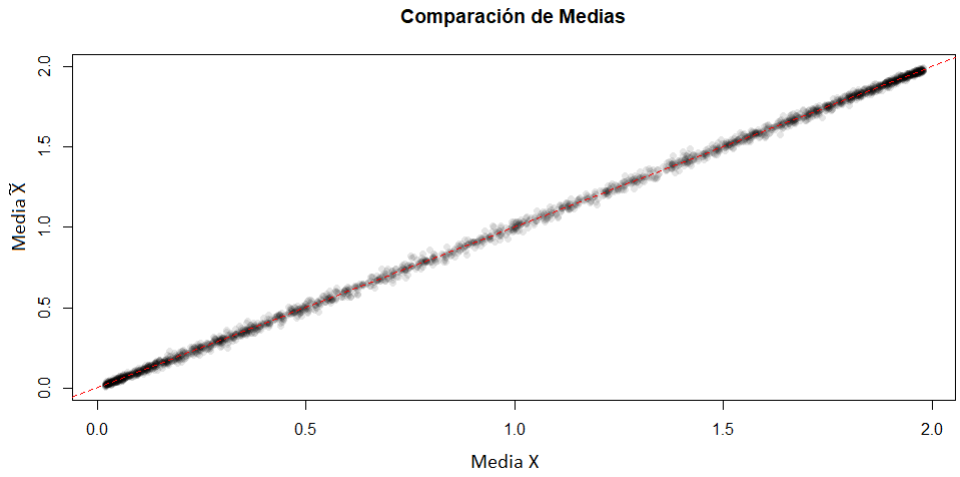


Figura 5.1: Comparación de medias entre X y $\tilde{X}$.

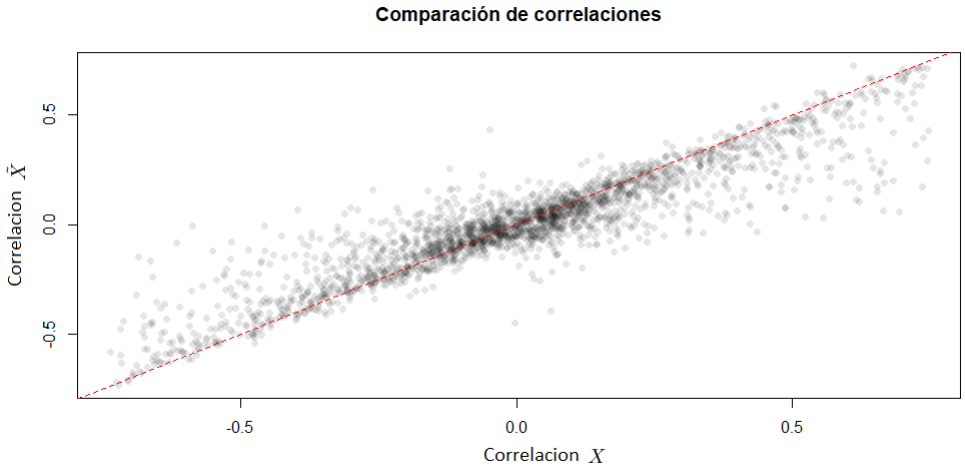


Figura 5.2: Comparación de correlaciones: $corr(X_j, X_{j+1})$ vs $corr(\tilde{X}_j, \tilde{X}_{j+1})$.

También si realizamos un gráfico de $corr(X_j, X_{j+1})$ vs $corr(X_j, \tilde{X}_{j+1})$, esperamos también obtener el mismo resultado que el caso anterior (ver figura 5.3), ya que, $S = \{j + 1\}$, tenemos

$$(X, \tilde{X})_{sw(S)} \stackrel{d}{=} (X_1, \ldots, \tilde{X}_{j+1}, \ldots, X_p, \ldots) \stackrel{d}{=} (X_1, \ldots, X_j, X_{j+1}, \ldots)$$

así $corr(X_j, \tilde{X}_{j+1}) = corr(\tilde{X}_j, \tilde{X}_{j+1})$.

Ahora bien, los genotipos knockoffs que hemos creado anteriormente se pueden usar para realizar una selección de variables controladas. Para esto, calculamos la medida de importan-

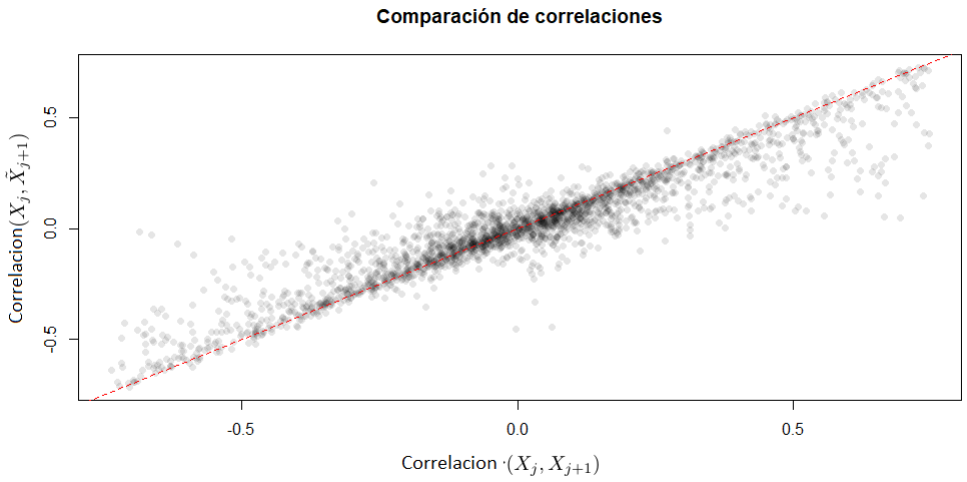


Figura 5.3: Comparación de correlaciones: $corr(X_j, X_{j+1})$ vs $corr(X_j, \tilde{X}_{j+1})$.

cia

$$W_j = |\hat{\beta}_j^{cv}| + |\hat{\beta}_{j+p}^{cv}|$$

donde $\hat{\beta}_j^{cv}$ y $\hat{\beta}_{j+p}^{cv}$ son los coeficientes estimados de regresión logística l_1-penalizada para X_j y $\tilde{X}_j$ respectivamente, calculados mediante validación cruzada, formamos entonces $W = (W_1, \ldots, W_p)$. Después, realizamos la selección de variables eligiendo un umbral adaptativo para W con el filtro Knockoffs. El umbral se elige de modo que la tasa de falsos descubrimientos objetivo sea 0.1. Para este caso se obtuvo un umbral de 0.03, esto significa que, si $W_j > 0.03$ entonces X_j es no nula. Para nuestro caso encontramos 7 SNPs, los cuales se muestran en la figura (5.4) (los X_j tal que $W_j = 0$ son incluidas en el modelo).

Los SNPs significativos encontrados con el método fueron:

rs12157341	rs5746945	rs5747338	rs451840	rs5748091	rs1210694	rs1210696
19	421	1181	1265	2008	2057	2059

Tabla 5.9: SNPs significativos encontrados con método Knockoffs.

Ahora veremos cómo fue el comportamiento de nuestro modelo, ya que como se trata de unos datos sintéticos, tenemos nuestros datos reales. La lista de los verdaderos SNP importantes también se proporciona con este conjunto de datos, la tabla (5.10) resume esto (compare con lo que obtuvimos).

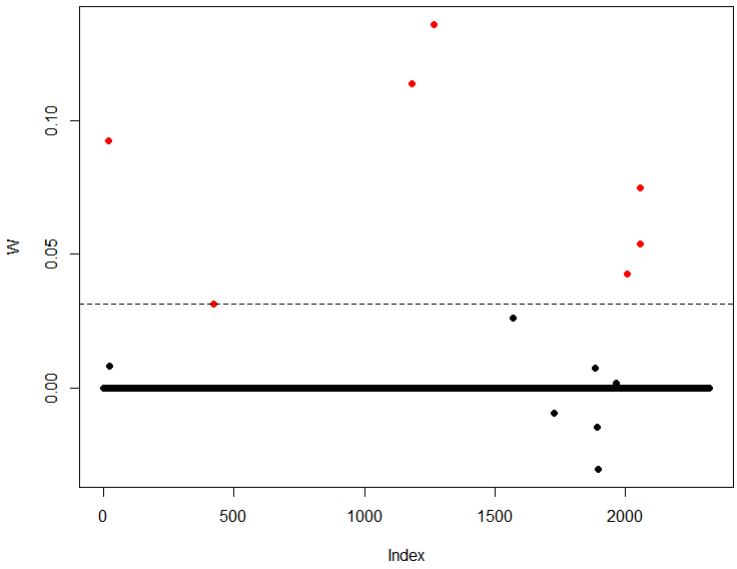


Figura 5.4: Comportamiento de W_j según el umbral.

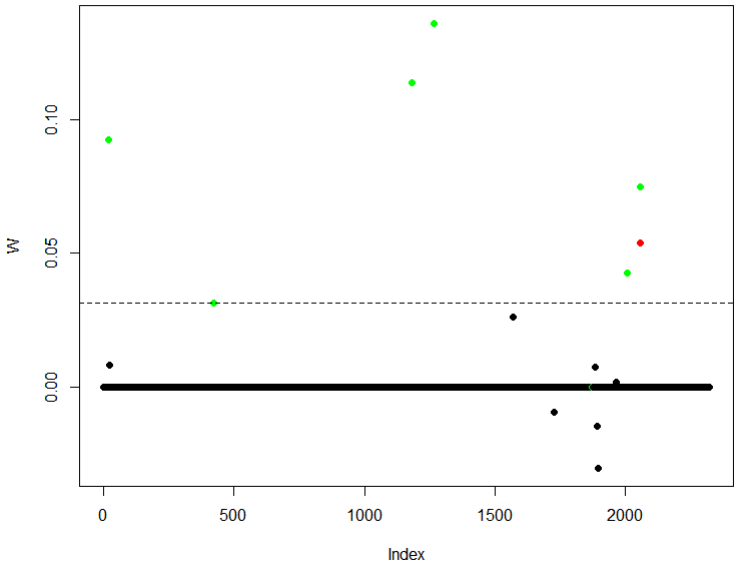


Figura 5.5: Decubrimientos vs reales

La comparación entre los descubrimientos con los verdaderos, se muestra en la figura (5.5), en esta observamos que los puntos con color rojo marcan los descubrimientos realiza-

rs6010298	rs181399	rs424962	rs455758	rs9612225	rs9620278	rs9608340	rs713998	rs9613298	rs134913
19	357	415	421	1181	1265	1454	1875	2008	2059

Tabla 5.10: Posiciones y datos reales sobre enfermedades.

dos por el algoritmo y en verde los verdaderos. Podemos entonces calcular la proporción de falsos descubrimientos (FDP) y la potencia del procedimiento en este conjunto de datos. El

$$FDP(t) = \frac{\#\{j \text{ nula: } w_j \geq t\}}{\#\{j : w_j \geq t\}} = 0.142$$

y la potencia

$$PW(t) = \frac{\#\{j \text{ no nula: } w_j \geq t\}}{\#\text{ no nulas}} = 0.6.$$

Comparación con otros Métodos

Vamos a comparar ahora con algunas pruebas marginales. Primero calculamos los p-valores de asociación marginal con todos los datos y aplicamos la prueba de tendencia de *Cochran-Armitage*. Un nivel de significancia que se ha vuelto estándar en estudios de GWAS es de 5×10^{-8}. Aquí vemos que para este nivel no se pueden hacer descubrimientos para este conjunto de datos con pruebas marginales, pues $-\log(5 \times 10^{-8}) \approx 7.3$.

Ahora, compararemos con el procedimiento de *Benjamin Hocking*. Si tratamos de controlar la tasa de falsos descubrimientos (al mismo nivel 0.1) aplicando el procedimiento a los p-valores de asociación marginal, podemos hacer algunos descubrimientos. Al aplicar el procedimiento a un nivel 0.1, obtenemos los siguientes descubrimientos:

rs6010298	rs9612225	rs9620278	rs134913	rs1210694
19	1181	1265	2059	2057

Tabla 5.11: Descubrimientos con procedimiento de Benjamin Hocking.

Aquí podemos observar que se obtuvieron 4 SNPs de los reales y uno que no corresponde. Para este caso el FDP fue de 0.2 y la potencia de 0.4. Aquí observamos el método Knockoffs tiene mejor comportamiento.

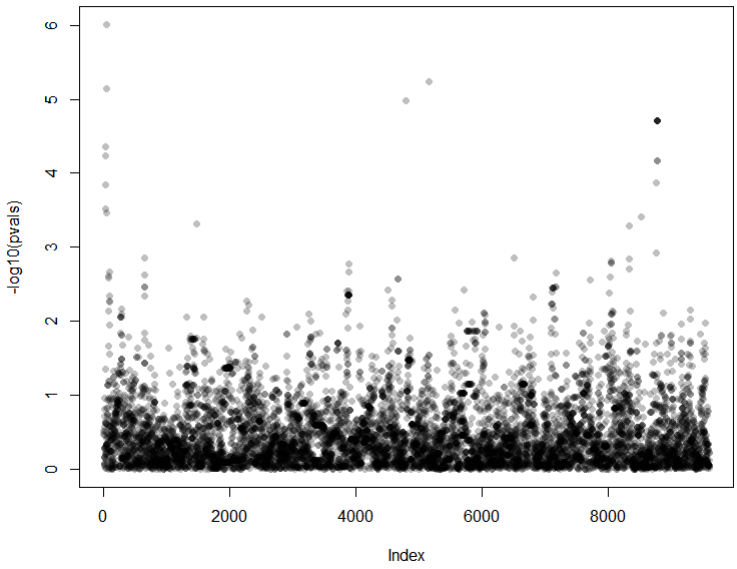


Figura 5.6: Prueba de Cochran-Armitage

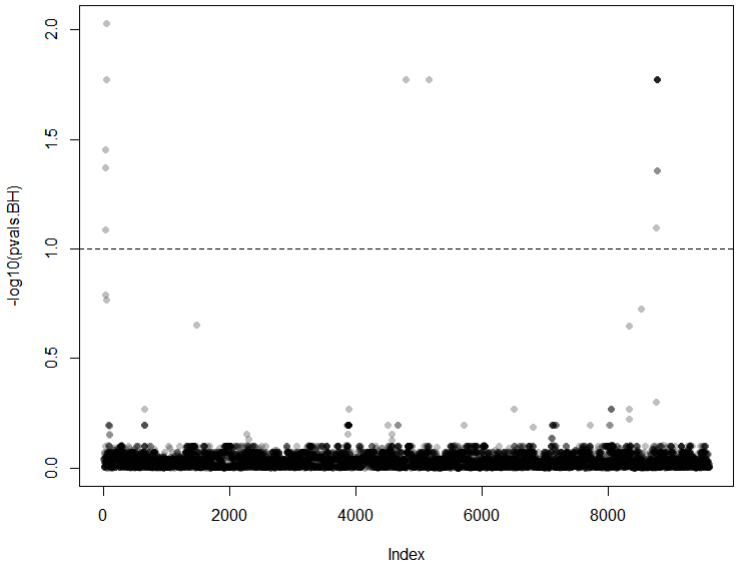


Figura 5.7: Prueba de Benjamin Hocking

CONCLUSIONES

En este trabajo estudiamos la metodología Knockoffs implementada por Candes et al. (2018). Este nuevo procedimiento busca una forma de seleccionar las covariables más importantes para una respuesta dada, controlando la tasa de falsos descubrimientos (FDR). Vimos que para una respuesta Y y covariables X, el método no se interesa en el conocimiento de la distribución $F_{Y|X}$, el único requisito necesario es el conocimiento de la distribución de las covariables X.

A pesar de que tenemos un algoritmo para generar dichas variables Knockoffs que sirven como un control negativo de las originales, no siempre es fácil aplicarlo y además tiene un gran coste computacional. Estudiamos dos aplicaciones directas, una para modelos lineales y otra para HMM. Hemos observado que para modelos lineales, la construcción es exacta, mientras que para HMM, se trata de un algoritmo iterativo. Vimos también que la forma de controlar el FDR se hace a través de un procedimiento que no es asintótico, sino probabilístico.

Estudiamos también un algoritmo que nos ayuda a obtener la fase de haplotipos, llamado fastPHASE. Este algoritimo no tiene muchas limitaciones funcionales. Su mayor inconveniente se refiere a la cantidad de tiempo que usa para procesar conjuntos de datos completos, pues la agrupación de secuencias alélicas SNP similares, que son necesarias para estimar los parámetros relevantes para el grupo antes de la imputación del genotipo y la determinación de fases de las secuencias alélicas, requiere recursos. El algoritmo fastPHASE se encuentra entre varias soluciones de imputación ampliamente utilizadas en diversos estudios, uno de ellos a GWAS, del cual una de las aplicaciones más importante radica en la detección de

enfermedades.

Dado que este algoritmo usa modelos ocultos de Markov para su construcción, el objetivo fue entonces estudiar un método en el que se pudiera construir variables Knockoffs una vez obtenido el HMM para fastPHASE. Vimos que su coste computacional es razonable y eficiente. Una vez obtenido las Knockoffs para HMM, se procede a la selección de las variables más importantes controlando el FDR. Por último vimos una aplicación donde se evidenció que la selección de variables para este tipo de aplicaciones funciona de manera efectiva.

Como se menciona en Sesia et al. (2019) se han desarrollado diferentes parametrizaciones para HMM dentro de la comunidad científica para la imputación de genotipos y el procedimiento Knockoffs puede ser aplicable fácilmente. Por ejemplo, si hay disponible una colección de haplotipos conocidos, es posible incluirlos en la descripción de F_X utilizada para generar las copias knockoffs. Sería interesante investigar desde una perspectiva aplicada las ventajas relativas de una opción sobre otra.

Dado que se ha calculado las medidas de importancia de variables basadas en modelos lineales generalizados, a pesar de que el control de la tasa de falsos descubrimientos no se basa en suposiciones de linealidad, la potencia puede verse afectada negativamente si la probabilidad real está lejos de ser lineal.

Para aprovechar al máximo la flexibilidad y la solidez de las Knockoffs, sería interesante explorar el uso de otras estadísticas que puedan capturar mejor las interacciones y las no linealidades. A la fecha, sólo se sabe cómo realizar una selección de variables controladas con knockoffs en los casos especiales en los que las variables pueden describirse mediante un modelo oculto de Markov o una distribución normal multivariada. Sería interesante extender esto a otras clases de covariables, como modelos gráficos más generales.

REFERENCIAS

Barber, R. F., Candès, E. J., et al. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Browning, S. R. and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Consortium, I. H. et al. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299.

Consortium, I. H. G. S. et al. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931.

Koski, T. (2001). *Hidden Markov models for bioinformatics*, volume 2. Springer Science & Business Media.

Lo, C. (2011). Algorithms for haplotype phasing.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644.

Sesia, M. (2008). A Tutorial for GWAS with Knockoffs. https://web.stanford.edu/group/candes/knockoffs/tutorials/gwas_tutorial.html. [Online; accedida 01-Junio-2020].

Sesia, M., Sabatti, C., and Candès, E. J. (2019). Gene hunting with hidden markov model knockoffs. *Biometrika*, 106(1):1–18.

Silva Zolezzi, I. (2011). Genómica y medicina. *Educación química*, 22(1):15–27.