UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Tesis de Licenciatura

Transformaciones estabilizadoras de la varianza para datos de experimentos de microarreglos

Pablo Delieutraz

Directora: Dra. Diana Kelmansky

Octubre de 2008

Índice general

Introducción

El objetivo de la mayoría de los experimentos con microarreglos es analizar los niveles de expresión de cientos a decenas de miles de genes en un solo ensayo. Muchos de estos experimentos investigan relaciones entre muestras biológicas buscando genes que estén expresados diferencialmente, es decir, genes que presenten cambios significativos en los niveles de expresión en las distintas muestras.

En la medición de la expresión de un gen lo que se trata de medir es que tan activo está ese gen. Como es difícil identificar una escala absoluta para medir y por varios motivos más, se suele utilizar una referencia para obtener una escala relativa. Incluso genes de una misma muestra no son directamente comparables por lo cual cada gen tiene su propia referencia. De esta forma, se pueden obtener relaciones de expresión (*gene expression ratios*) que pueden ser utilizadas para testear si un gen (en la muestra testeada) está diferencialmente expresado (comparado con el gen de la muestra de referencia) o no.

Los valores de expresión de genes obtenidos en los experimentos de microarreglos presentan varianzas que dependen de la media. Por este motivo muchos métodos estadísticos tradicionales no pueden aplicarse directamente a los datos obtenidos. Una solución a este problema es aplicar una transformación a los datos de forma tal que los datos transformados tengan varianza aproximadamente independiente de la media.

En el capítulo 1 se describe en qué consiste la tecnología de microarreglos y cómo se realiza un experimento típico.

En el capítulo 2 presentamos el modelo utilizado para describir las relaciones que existen entre los valores medidos en un experimento de microarreglos, los verdaderos valores de expresión y la influencia de otros factores (ruido y sesgos), incorporando parámetros de normalización. Este modelo permite explicar la relación entre la varianza y la media de las intensidades de los spots de un microarreglo y de esta forma se puede deducir una transformación que estabiliza aproximadamente la varianza, la cual se presenta en el capítulo 3.

El tema principal de esta tesis es la estimación de los parámetros de normalización, el cual se desarrolla en el capítulo 4. En primer lugar, en la sección 4.4.1, se supone que las intensidades transformadas tienen distribución normal y no hay genes expresados diferencialmente y se obtienen los estimadores de máxima verosimilitud de dichos parámetros. Luego, en la sección 4.4.2, se

considera la presencia de genes expresados diferencialmente y distribuciones simétricas aproximadamente normales y se describe un método de estimación robusto bajo estas condiciones. Por último, en el capítulo 5 se muestran los resultados de la aplicación de dicho método a datos de microarreglos simulados, donde se analiza como afectan distintas variables a la estimación.

Capítulo 1

Experimentos de microarreglos

En este capítulo daremos una visión general de la tecnología de microarreglos y los experimentos que producen datos de expresión de genes.

1.1. Tecnología de microarreglos

El microarreglo es un soporte sólido, generalmente de vidrio o silicio, al que se le ha adherido cierto tipo de material genético formando una matriz de miles de puntos equiespaciados. Cada punto contiene millones de clones de una secuencia específica asociada a un gen. El número de puntos (también llamados *spots*) en un microarreglo puede llegar a alcanzar varias decenas de miles.

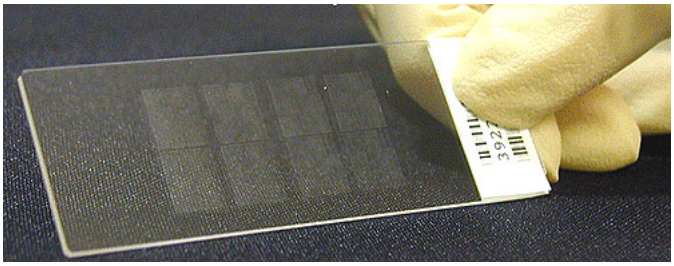

Figura 1.1: Imagen de un microarreglo de vidrio.

Llamaremos sonda (*probe*) al material inmovilizado sobre el vidrio y target al material que se agrega luego sobre el arreglo. Esta es la nomenclatura que se usa en la actualidad, aunque inicialmente se utilizaba la nomenclatura contraria.

Existen distintos tipos de microarreglos dependiendo del material que se utilice como sonda. Los tipos más comunes son:

1. **Microarreglos de ADNc**.

El ADNc (ADN complementario) es una molécula de ácido nucleico derivada del ARNm (ARN mensajero). Los niveles de ARNm en una célula reflejan la actividad metabólica del gen particular del cual fue transcripto.

Los experimentos con este tipo de microarreglos permiten realizar análisis cuantitativos del ARN transcripto desde un gen específico y la principal ventaja sobre otros métodos utilizados para medir niveles de expresión de genes, es que los microarreglos permiten analizar miles de genes simultáneamente. Entre los problemas que se presentan en los microarreglos de ADNc está la posibilidad de hibridación cruzada entre ARNm y otros elementos no específicos de los clones de ADNc. Otro de los problemas que surgen es el gran esfuerzo que requiere mantener las grandes librerías de ADNc. Estos problemas pueden ser eludidos en gran parte mediante el uso de microarreglos de oligonucléotidos.

2. **Microarreglos de oligonucleótidos**. Los oligonucleótidos son secuencias cortas de nucleótidos de ARN o ADN. Estas secuencias pueden tener unas 20 o menos bases o pares de bases. Muchas veces los oligonucleótidos son referidos simplemente como oligos. Cuando la secuencias son de 50-70 nucleótidos hablamos de oligonucleótidos largos.

   El uso de oligonucleótidos tiene la ventaja de no requerir grandes librerías de ADNc para su fabricación; sin embargo su precio es más elevado.

   En este tipo de microarreglo las sondas se sintetizan directamente sobre el chip en vez de sintetizarlas *in vitro* y adherirlas luego al chip. Cada gen está representado por un grupo de sondas (*probe set*) que investigan distintas partes de la secuencia de un mismo gen.

   Se utilizan dos tipos de sondas: PM (Perfect Match) y MM (Miss Match). Las sondas PM son secuencias de 25 bases perfectamente complementarias a una región específica de un gen. Las sondas MM coinciden con una PM salvo en una única base (la central).

   Affymetrix es la compañía líder en la fabricación de este tipo de chips.

3. **Microarreglos de proteínas**. Estos microarreglos consisten en chips de proteínas inmovilizadas en una posición concreta sobre una superficie sólida, dispuestas en una forma similar a como se disponen las sondas en los microarreglos de ADN.

   El desarrollo de microarreglos de proteínas ha sido técnicamente complicado debido a la complejidad y diversidad estructural de las proteínas. Al contrario de los ácidos nucleicos, las proteínas no tienen una estructura homogénea ni un patrón de unión específico, sino que cada proteína tiene unas características bioquímicas particulares. Además el

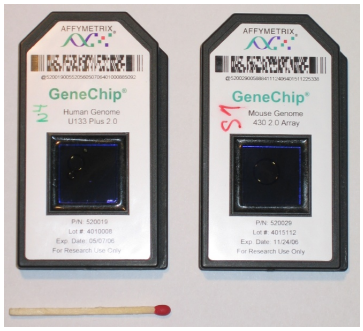

Figura 1.2: Microarreglos de Affymetrix.

plegado tridimensional de las secuencias de aminoácidos hace que sea difícil preservarlas sobre superficies planas.

4. **Microarreglos de tejidos**. Los microarreglos de tejidos (TMAs, Tissue Microarrays) consisten en colecciones miniaturizadas de hasta 1000 muestras de tejidos inmovilizadas sobre un soporte que permiten el análisis de ADN, ARN y proteínas. Prácticamente la mayoría de los artículos publicados sobre microarrays de tejidos están relacionados con el análisis de tumores, aunque se ha utilizado con éxito en otro tipo de patologías relacionadas con el sistema nervioso.

1.2. Descripción de experimento de microarray de ADN

En esta sección haremos una breve descripción de como se realiza un experimento típico de microarray de ADN.

De acuerdo con el tipo de experimento los microarreglos de ADN se clasifican en dos grandes grupos:

- **Arreglos de dos canales (o dos colores).** Los arreglos de dos canales permiten comparar, en la misma laminilla, material proveniente de dos tejidos bajo distintas condiciones. Por ejemplo, se puede comparar un tejido sano con uno enfermo, o tejidos sometidos a distintos tratamientos, etc.

- **Arreglos de un canal.** Affymetrix es la compañía que fabrica este tipo de microarreglos usando una combinación de fotolitografía y reacciones químicas para sintetizar los oligonucleótidos *in situ*.

A continuación se mencionan los pasos a seguir en un experimento de microarreglos.

1. Diseño del microarreglo.

   En esta etapa se selecciona el tipo y cantidad de material biológico que se va a inmovilizar sobre la superficie, que variará en función del tipo de experimento que se desee llevar a cabo. Se determina también la densidad de integración, es decir, el número de sondas que se desean inmovilizar sobre la superficie del chip, que se verá limitada por el método de fabricación que se desee emplear.

2. Fabricación del arreglo.

   Este paso está muy diversificado como consecuencia de la gran cantidad de soluciones tecnológicas presentes en el mercado. En general las grandes empresas que comercializan los microarreglos ya listos son capaces de ofrecer mayores densidades de integración que las que se pueden alcanzar empleando los robots para la fabricación de microarreglos personalizados en laboratorio.

3. Extracción y marcado del ARN de cada una de las muestras.

   Los procesos a seguir son la extracción y purificación del material a analizar, un proceso de amplificación en el caso de tratarse de material genético y por último el etiquetado de la muestra para permitir su detección en el proceso de revelado. Los marcadores más comúnmente empleados son los fluorescentes.

4. Hibridación, en el arreglo, de los ADNc marcados.

   Resulta un paso clave ya que en él se produce la reacción de afinidad en la que se hibridan las hebras de ADN de las muestras marcadas para permitir su posterior identificación, con sus complementarias inmovilizadas en la superficie del microarreglo. Según las condiciones en las que se produzca esta reacción de afinidad se obtendrán, posteriormente, mejores o peores resultados en el proceso de revelado. El lavado se realiza para eliminar las interacciones no específicas que se dan entre la muestra y el material inmovilizado o la superficie del arreglo.

5. Lectura del arreglo.

   Es un proceso que viene condicionado por la gran variedad de alternativas tecnológicas diseñadas para esta función. Entre estas soluciones las más comunes son la utilización de scanners láser y cámaras CCD para la detección de marcadores fluorescentes con los que se ha marcado la muestra.

6. Cuantificación de la imagen.

   Consiste en la localización de los puntos en la imagen y la obtención de sus intensidades de fluorescencia. En este paso se analizan las imágenes

de 16 bits obtenidas para cada una de las longitudes de onda en la cuales se pueden apreciar los puntos en los que la reacción de hibridación ha sido positiva y los puntos en los que no ha habido tal hibridación.

7. Análisis estadístico de los datos obtenidos.

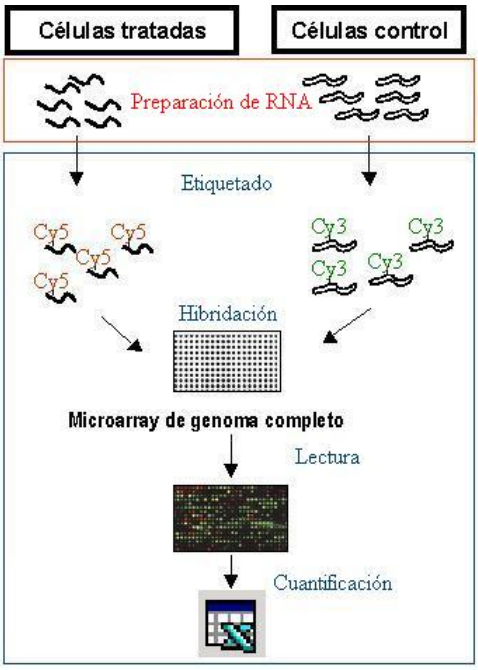

Figura 1.3: Pasos de un experimento típico de análisis comparativo de expresión de genes utilizando microarreglo de ADN.

Capítulo 2

Modelo para el error de medición de expresión de genes en microarrays

La construcción de un *modelo* matemático permite describir la influencia de las diferentes fuentes de variabilidad que intervienen en una medición y explicar la relación entre el valor verdadero que se desea medir y los resultados obtenidos en las mediciones. En el caso de los experimentos con microarreglos, la magnitud que se desea medir es la abundancia de moléculas específicas presentes en un conjunto de muestras biológicas y las cantidades que se miden son las intensidades de fluorescencia de los diferentes spots del arreglo (Huber et al., 2004).

El proceso de medición consiste en una serie de reacciones bioquímicas y un sistema de detección óptica con un scanner láser o una cámara CCD. En dicho proceso intervienen una gran cantidad de factores que afectan a los valores obtenidos.

Hay muchos tipos de ruido que pueden afectar la señal final producida por el scanner que pueden ser clasificados en dos categorías: ruido de fuente y ruido de detección.

- Ruidos de la fuente: ruido de fotones, polvo en los vidrios, tratamiento de los vidrios.
- Ruidos de detección: están vinculados al proceso de amplificación y digitalización.

Una imagen perfecta debería reflejar únicamente medidas de la intensidad de fluorescencia de los tintes de interés. Sin embargo en la práctica el sistema es imperfecto y las imágenes son combinaciones de señales no deseadas (tales como ruido fotónico, ruido electrónico, luz láser reflejada y fluorescencia de fondo) con las señales de fluorescencia deseadas.

Además del ruido, que es la componente aleatoria de la variabilidad, la señal está afectada por errores sistemáticos provenientes de las diferencias en ciertas propiedades de los tintes utilizados:

- El tamaño diferente de las moléculas de Cy3 y Cy5.
- La eficiencia de la emisión de fotones (quantum yield) en el proceso de fluorescencia.
- El blanqueado por la luz (photo bleaching).

La importancia de construir un modelo para los errores de medición en experimentos de microarrays se debe a las siguientes cuestiones (Huber et al., 2004):

1. Es necesario conocer como es la distribución de los posibles resultados de las mediciones para poder realizar inferencias basadas en un número pequeño de mediciones.

   Si tuviéramos una cantidad suficiente de mediciones podríamos utilizar las distribuciones empíricas para comparar el nivel de expresión de los genes en estudio. Desgraciadamente, obtener varias mediciones de todos los genes bajo todas las condiciones de interés no siempre es posible o resulta muy costoso. Pero si tenemos confianza en un modelo para el error, entonces estamos en condiciones de sacar conclusiones significativas con unas pocas replicaciones.

2. Un modelo del error es una herramienta eficiente para el resumen y reporte de los resultados experimentales. Si tenemos razones para creer que los resultados de las mediciones tienen una cierta distribución, entonces los parámetros de la distribución (por ejemplo la media y el desvío) describen bastante bien a los valores obtenidos en los experimentos.

3. Un modelo del error es un resumen de la experiencia y de nuestro entendimiento del sistema de medición. También puede ser utilizado para **control de calidad**: si la distribución de un nuevo conjunto de datos se desvía considerablemente del modelo, entonces podríamos cuestionar la calidad de estos datos.

Se han propuesto varios modelos para explicar la relación que existe entre los valores obtenidos en experimentos de microarreglos y los diversos factores involucrados. En la sección 2.1 mostramos una deducción teórica del modelo aditivo-multiplicativo. Este modelo fue introducido en el contexto de los datos de microarreglos en dos versiones distintas, una propuesta por Ideker et al. (2000) y otra por Rocke y Durbin (2001).

En la sección 2.2 presentamos la versión del modelo aditivo-multiplicativo en la que está basado el resto de este trabajo.

2.1. Deducción teórica del modelo aditivo - multiplicativo

Consideremos una observación genérica y llamemos y al valor obtenido en la medición, x a la cantidad que se desea medir y representemos todos los otros parámetros de los cuales puede depender el valor medido por medio del vector $r = (r_1, \ldots, r_n)$. Podemos describir la relación que existe entre estos valores por medio de una función (Huber et al., 2004):

$$y = f(x, r) \tag{2.1}$$

Si el aparato de medición está bien construido entonces podemos suponer que f es una función suave y escribir la ecuación (2.1) en la siguiente forma:

$$y = f(0, r) + f'(0, r)x + O(x^2) \tag{2.2}$$

donde f' es la derivada de f respecto de x y $O(x^2)$ representa los efectos no lineales.

Idealmente, los parámetros r podrían estar fijos en un valor $\overline{r} = (\overline{r}_1, \ldots, \overline{r}_n)$. En la práctica estos valores fluctúan alrededor de $\overline{r}$ en distintas repeticiones del experimento. Si las fluctuaciones no son muy grandes, podemos aproximar los valores de $f(0, r)$ y $f'(0, r)$ por medio del polinomio de Taylor de orden 1:

$$f(0, r) \approx f(0, \overline{r}) + \sum_{j=1}^{n} \frac{\partial f(0, \overline{r})}{\partial r_j} (r_j - \overline{r}_j) \tag{2.3}$$

$$f'(0, r) \approx f'(0, \overline{r}) + \sum_{j=1}^{n} \frac{\partial f'(0, \overline{r})}{\partial r_j} (r_j - \overline{r}_j) \tag{2.4}$$

Las sumas en las ecuaciones (2.3) y (2.4) son combinaciones de una gran cantidad de variables aleatorias con media cero. Por lo tanto, es una aproximación razonable suponer que $f(0, r)$ y $f'(0, r)$ son variables aleatorias normalmente distribuidas con medias $a = f(0, \overline{r})$ y $b = f'(0, \overline{r})$ y varianzas σ_a^2 y σ_b^2 respectivamente. Entonces, omitiendo el término no lineal en la ecuación (2.2), tenemos el siguiente modelo:

$$Y = \alpha + \beta X \tag{2.5}$$

con $\alpha \sim N(a, \sigma_a^2)$ y $\beta \sim N(b, \sigma_b^2)$.

Si escribimos

$$\begin{aligned} \alpha &= a + \varepsilon \\ \beta &= b + \zeta = b\,(1 + \eta) \end{aligned}$$

con $\varepsilon \sim N(0, \sigma_a^2)$, $\zeta \sim N(0, \sigma_b^2)$ y $\eta = \frac{\zeta}{b} \sim N(0, \sigma_b^2/b^2)$, el modelo (2.5) quedaría expresado de la siguiente forma:

$$y = a + \varepsilon + b\,x\,(1 + \eta) \tag{2.6}$$

con $\varepsilon \sim N(0, \sigma_a^2)$ y $\eta \sim N(0, \sigma_b^2/b^2)$. Este es el **modelo aditivo-multiplicativo** propuesto por Ideker et al. (2001) para las mediciones en experimentos de microarreglos.Rocke y Durbin (2001) propusieron el siguiente modelo:

$$y = a + \varepsilon + b\,x\,e^{\eta} \tag{2.7}$$

que es equivalente a (2.6) hasta los términos de primer orden en η. Los modelos (2.6) y (2.7) difieren significativamente sólo si el coeficiente de variación σ_b/b es grande. Para los datos de microarreglos este coeficiente típicamente es menor que 0.2, con lo cual la diferencia entre los dos modelos es de poca relevancia práctica.

2.2. Modelo aditivo-multiplicativo

En esta sección presentaremos el modelo propuesto por Huber et al. (2003), en el cual está basado el resto de este trabajo.

Un microarray consiste de un conjunto de sondas inmovilizadas sobre un soporte sólido. Las sondas se eligen de forma tal que se unan a moléculas específicas. La muestra biológica de interés se prepara en solución, se etiqueta con tintura fluorescente y se aplica sobre el arreglo permitiendo la hibridación con las sondas ubicadas sobre el mismo. La abundancia de moléculas de la muestra puede ser comparada a través de la comparación de la intensidad de fluorescencia en los sitios donde se encuentran las sondas correspondientes.

La intensidad medida y_{ki} de la sonda $k = 1, \ldots, n$ para la muestra $i = 1, \ldots, d$ puede ser descompuesta en una parte específica y una parte no específica:

$$y_{ki} = \alpha_{ki} + \beta_{ki} x_{ki} \tag{2.8}$$

donde x_{ki} es la abundancia de transcripción representada por la sonda k en la muestra i, β_{ki} es un factor de proporcionalidad y α_{ki} son las contribuciones de señal no específica que pueden ser causadas por efectos tales como hibridación no específica, hibridación cruzada o fluorescencia del background.

Generalmente los valores de α_{ki} y β_{ki} no son conocidos, pero la tecnología de microarreglos está diseñada de tal forma que estos valores están relacionados para distintos k e i. Esto hace posible que se puedan hacer inferencias acerca de las concentraciones x_{ki} a partir de los datos y_{ki} obtenidos en las mediciones. Las relaciones entre los valores de α_{ki} y β_{ki} para diferentes k e i pueden ser expresados en términos de otra descomposición:

$$\begin{aligned} \beta_{ki} &= \beta_i \gamma_k e^{\eta_{ki}} \\ \alpha_{ki} &= a_i + \overline{\nu}_{ki} \end{aligned}$$

Por lo tanto, el factor β_{ki} es el producto de una afinidad de la sonda γ_k, que es la misma para todas las mediciones que involucran a las sondas del tipo k, multiplicado por un factor de normalización β_i, el cual se aplica a todas las mediciones de la muestra i.

El resto $\beta_{ki}/(\beta_i\gamma_k)$ se explica por medio de $e^{\eta_{ki}}$. Se pueden elegir las unidades de β_i y γ_k de forma tal que $\sum_k \eta_{ki} = \sum_i \eta_{ki} = 0$.

La contribución de señal no específica α_{ki} puede ser descompuesta en una *contribución* por muestra a_i y un resto $\overline{\nu}_{ki}$ con $\sum_k \overline{\nu}_{ki} = 0$.

La afinidad de la sonda γ_k puede depender, por ejemplo, de la secuencia de la sonda, de la estructura secundaria y de la abundancia de moléculas de la sonda en el arreglo. El factor de normalización puede depender, por ejemplo, de la cantidad de ARNm en la muestra, de la eficiencia del etiquetado y del rendimiento cuántico del tinte.

La idea detrás de esta descomposición es que mientras que los valores individuales de η_{ki} y $\overline{\nu}_{ki}$ pueden fluctuar alrededor de cero, lo hacen en una forma no sistemática y aleatoria. Así, por ejemplo, suponemos que no hay efectos sistemáticos no lineales, lo cual podría implicar tendencias en los valores de η_{ki} o $\overline{\nu}_{ki}$ dependiendo de los valores de x_{ki}.

Con esta nueva descomposición el modelo (2.8) queda de la siguiente forma:

$$y_{ki} = a_i + \overline{\nu}_{ki} + \beta_i\gamma_k e^{\eta_{ki}} x_{ki} \tag{2.9}$$

Ahora uno puede reducir la complejidad de los parámetros de esta ecuación a través de los siguientes 3 pasos de modelado:

1. No tratar de determinar explícitamente la afinidad de las sondas γ_k. Estas pueden ser absorbidas en $m_{ki} = \gamma_k x_{ki}$, que se puede considerar como una medida de la abundancia de la transcripción k en la muestra i en unidades específicas de la sonda.

2. Tratar a η_{ki} y $\overline{\nu}_{ki}$ como "términos de ruido" provenientes de distribuciones de probabilidad apropiadas.

3. Estimar los valores de β_i y a_i, así como los parámetros de la distribución de probabilidad de los datos.

Así, la ecuación (2.9) da lugar al siguiente modelo estocástico:

$$\frac{Y_{ki} - a_i}{\beta_i} = m_{ki}e^{\eta_{ki}} + \nu_{ki} \tag{2.10}$$

con $\eta_{ki} \overset{\text{iid}}{\sim} \mathcal{L}_\eta$, $\nu_{ki} \overset{\text{iid}}{\sim} \mathcal{L}_\nu$. Aquí, $\nu_{ki} = \overline{\nu}_{ki}/\beta_i$ es el ruido aditivo dividido por el factor de normalización β_i.

El miembro derecho de la ecuación (2.10) es una combinación de dos términos de error, uno aditivo y otro multiplicativo. Este modelo fue propuesto por Rocke y Durbin(2001), usando distribuciones normales $\mathcal{L}_\eta = N(0, \sigma_\eta^2)$ y

$\mathcal{L}_\nu = N(0, \sigma_\nu^2)$. En lo que sigue, consideraremos distribuciones $\mathcal{L}_\eta$ y $\mathcal{L}_\nu$ unimodales, aproximadamente simétricas y que tienen media cero y varianzas σ_η^2 y σ_ν^2 respectivamente, pero no supondremos por ahora distribución normal.

El miembro izquierdo describe la calibración de las intensidades del microarreglo Y_{ki}.

2.3. Dependencia varianza vs. media

Según el modelo (2.10) la varianza de las intensidades normalizadas depende de la media de la siguiente forma:

$$\mathrm{Var}\Big(\frac{Y_{ki} - a_i}{\beta_i}\Big) = c^2 \mathrm{E}^2\Big(\frac{Y_{ki} - a_i}{\beta_i}\Big) + \sigma_\nu^2$$

donde $c^2 = \frac{\mathrm{Var}(e^\eta)}{\mathrm{E}^2(e^\eta)}$ es un parámetro de la distribución de $\eta \sim \mathcal{L}_\eta$. En efecto:

$$\mathrm{E}\Big(\frac{Y_{ki} - a_i}{\beta_i}\Big) = m_{ki}\,\mathrm{E}(e^\eta) \tag{2.11}$$

Como η_{ki} y ν_{ki} son independientes:

$$\begin{aligned}
\mathrm{Var}\Big(\frac{Y_{ki} - a_i}{\beta_i}\Big) &= m_{ki}^2\,\mathrm{Var}(e^{\eta_{ki}}) + \sigma_\nu^2 \\
&= \Big[m_{ki}\,\mathrm{E}(e^\eta)\Big]^2 \frac{\mathrm{Var}(e^\eta)}{\mathrm{E}^2(e^\eta)} + \sigma_\nu^2 \\
&= c^2 \mathrm{E}^2\Big(\frac{Y_{ki} - a_i}{\beta_i}\Big) + \sigma_\nu^2
\end{aligned}$$

En el capítulo siguiente presentaremos algunas transformaciones estabilizadoras de la varianza que surgieron en el contexto de los microarreglos.

Capítulo 3

Transformaciones que estabilizan la varianza

Varias metodologías estadísticas tradicionales están basadas en el supuesto de que los datos tienen varianza constante independiente de la media. En el caso que este supuesto no se cumpla, una solución posible es aplicar una transformación adecuada a los datos de forma tal que la varianza de los datos transformados sea aproximadamente independiente de la media.

En el capítulo anterior vimos que las intensidades medidas en un experimento de microarreglo no cumplen con dicho supuesto. Además vimos, basándonos en el modelo aditivo-multiplicativo, como depende de la varianza de la media. Esto sugiere que debemos aplicarle una transformación a los datos obtenidos en experimentos de microarreglos para estabilizar la varianza.

En la sección 3.1 presentamos un método general para encontrar transformaciones que estabilizan la varianza. En la sección 3.2 obtenemos la transformación que estabiliza la varianza para las variables aleatorias del modelo aditivo-multiplicativo (2.10) y en la sección 3.3 mostramos que dicha transformación efectivamente estabiliza la varianza.

En la sección 3.4 mostramos otras transformaciones que se han propuesto en el contexto de microarreglos.

3.1. Método general para encontrar transformaciones que estabilizan la varianza

Consideremos una variable aleatoria X con $\mathrm{E}(X) = \mu$ y sea $h : I \to \mathbb{R}$ una función diferenciable, donde I es un intervalo que contiene al rango de X. Haciendo el desarrollo de Taylor de h alrededor de μ tenemos:

$$h(X) = h(\mu) + h'(\mu)\big(X - \mu\big) + \frac{h''(\xi)}{2}\big(X - \mu\big)^2$$

Resulta entonces que la varianza de $h(X)$ es:

$$\mathrm{Var}(h(X)) = h'(\mu)^2\,\mathrm{Var}\big(X\big) + \mathrm{Var}\big(\frac{h''(\xi)}{2}(X-\mu)^2\big) + 2\,\mathrm{Cov}\Big(h'(\mu)(X-\mu)\,;\,\frac{h''(\xi)}{2}(X-\mu)^2\Big)$$

Como $\mathrm{E}(X-\mu)=0$, entonces

$$\mathrm{Cov}\big(h'(\mu)(X-\mu)\,;\,\frac{h''(\xi)}{2}(X-\mu)^2\big) = h'(\mu)\,\mathrm{E}\big(\,\frac{h''(\xi)}{2}(X-\mu)^3\big)$$

y la ecuación anterior queda:

$$\mathrm{Var}(h(X)) = h'(\mu)^2\,\mathrm{Var}\big(X\big) + \mathrm{Var}\big(\frac{h''(\xi)}{2}(X-\mu)^2\big) + 2\,h'(\mu)\mathrm{E}\big(\,\frac{h''(\xi)}{2}(X-\mu)^3\big)$$

Si dentro de los valores típicos de X, h es una función aproximadamente lineal, entonces el término de segundo orden del desarrollo de Taylor es pequeño y los términos que lo involucran en la ecuación anterior son despreciables.

Luego, tenemos que:

$$\mathrm{Var}(h(X)) \approx h'(\mu)^2\,\mathrm{Var}\big(X\big) \tag{3.1}$$

Supongamos ahora que $\{Y_\mu\}$ es una familia de variables aleatorias tales que $\mathrm{E}(Y_\mu)=\mu$ y cuyas varianzas dependen de la media, es decir $\mathrm{Var}(Y_\mu)=g(\mu)$.

Si queremos encontrar una función h tal que $\mathrm{Var}(h(Y_\mu))$ sea aproximadamente independiente de μ y suponemos que vale la siguiente aproximación

$$\mathrm{Var}(h(Y_\mu)) \approx h'(\mu)^2\,g(\mu)\,, \tag{3.2}$$

podríamos pedir que h verifique

$$h'(\mu)^2\,g(\mu) = \mathrm{cte}$$

Una función h que tenga esta propiedad se llama función estabilizadora de varianza.

Notar que si h estabiliza la varianza y $\gamma_1, \gamma_2 \in \mathbb{R}$ entonces $\gamma_1 h + \gamma_2$ también es una función estabilizadora. Por lo tanto, podemos obtener una transformación h que estabilice la varianza integrando $h'(\mu) = g(\mu)^{-1/2}$:

$$h(y) = \int^y \frac{1}{\sqrt{g(\mu)}}\,d\mu \tag{3.3}$$

Si h es una transformación obtenida de esta forma se tiene que $\mathrm{Var}(h(Y_\mu)) \approx 1$.

Notar que este método se basa fuertemente en la aproximación (3.2). En la sección (3.2) aplicaremos este método para deducir una transformación que estabiliza aproximadamente la varianza de las intensidades de microarreglos. Luego estudiaremos la validez de la aproximación (3.2) para la transformación hallada.

Ejemplo 1

Si $\{Y_\mu\}$ es una familia de variables aleatorias tales que $\mathrm{E}(Y_\mu) = \mu$ y $\mathrm{Var}(Y_\mu) = k\mu^\alpha$, con $k, \alpha \in \mathbb{R}_{>0}$, entonces obtenemos una transformación que estabiliza la varianza de la siguiente forma:

$$h(y) = \int^y \frac{1}{\sqrt{k\mu^\alpha}}\, d\mu = \begin{cases} \frac{1}{\sqrt{k}} \frac{y^{1-\alpha/2}}{1-\alpha/2} & \text{si} \quad \alpha \neq 2 \\ \\ \frac{1}{\sqrt{k}} \ln(y) & \text{si} \quad \alpha = 2 \end{cases}$$

Por simplicidad podríamos considerar la siguiente transformación:

$$f(y) = \begin{cases} y^{1-\alpha/2} & \text{si} \quad \alpha \neq 2 \\ \\ \ln(y) & \text{si} \quad \alpha = 2 \end{cases}$$

ya que al ser un múltiplo escalar de h también estabiliza la varianza, es decir la varianza de $f(Y_\mu)$ es aproximadamente independiente de μ. Más precisamente:

$$\mathrm{Var}(f(Y_\mu)) \approx \begin{cases} k(1-\alpha/2)^2 & \text{si} \quad \alpha \neq 2 \\ \\ k & \text{si} \quad \alpha = 2 \end{cases}$$

Ejemplo 2

Si $X \sim \mathcal{P}(\lambda)$ entonces $\mathrm{E}(X) = \lambda$ y $\mathrm{Var}(X) = \lambda$. Este es un caso particular del ejemplo anterior ($k = \alpha = 1$) , por lo tanto una transformación que estabiliza aproximadamente la varianza para una familia de variables aleatorias con distribución Poisson es:

$$h(y) = \sqrt{y}$$

Ejemplo 3

Si $X \sim \varepsilon(\lambda)$ entonces $\mathrm{E}(X) = \frac{1}{\lambda}$ y $\mathrm{Var}(X) = \frac{1}{\lambda^2}$. Este es otro caso particular del ejemplo 1 ($k = 1$, $\alpha = 2$), por lo tanto una función estabilizadora de varianzas para una familia de variables aleatorias con distribución exponencial es:

$$h(y) = \ln(y)$$

Ejemplo 4

Si $X \sim Bi(n, p)$ entonces $\mathrm{E}(X) = np$ y $\mathrm{Var}(X) = np(1-p)$. En este caso la varianza depende de la media en la forma: Var(X)=E(X)(1-E(X)/n). Luego una transformación estabilizadora de varianza en el caso binomial será:

$$h(y) = \int^y \frac{1}{\sqrt{\mu(1-\mu/n)}}\, d\mu$$

Haciendo la sustitución $z = \sqrt{\mu/n}$ tenemos:

$$h(y) = 2\sqrt{n} \int^{\sqrt{y/n}} \frac{1}{\sqrt{1-z^2}}\, dz = 2\sqrt{n}\, \operatorname{arsin}\left(\sqrt{\frac{y}{n}}\right)$$

3.2. Transformación arco-seno hiperbólico

En el capítulo anterior vimos que la varianza de las intensidades normalizadas de los spots de microarreglos $Z_{ki} = \frac{Y_{ki}-a_i}{\beta_i}$ dependen de su media de la siguiente forma:

$$\operatorname{Var}(Z_{ki}) = c^2 \mathrm{E}^2(Z_{ki}) + \sigma_\nu^2$$

Si aplicamos el método descripto en la sección anterior para estabilizar la varianza, en este caso obtenemos la transformación h integrando:

$$h(z) = \int^z \frac{1}{\sqrt{c^2\,\mu^2 + \sigma_\nu^2}}\, d\mu$$

donde $c^2 = \frac{\operatorname{Var}(e^\eta)}{\mathrm{E}^2(e^\eta)}$ y σ_ν^2 son parámetros de las distribuciones $\mathcal{L}_\eta$ y $\mathcal{L}_\nu$ respectivamente.

Por lo tanto:

$$h(z) \;=\; \frac{1}{\sigma_\nu} \int^z \frac{1}{\sqrt{\left(\frac{c\,\mu}{\sigma_\nu}\right)^2 + 1}}\, d\mu$$

Luego de hacer la sustitución $v = \frac{c\,\mu}{\sigma_\nu}$ queda:

$$\begin{aligned} h(z) &= \frac{1}{c} \int^{\frac{c\,z}{\sigma_\nu}} \frac{1}{\sqrt{v^2+1}}\, dv \\ &= \frac{1}{c}\, \operatorname{arsinh}\Big(\frac{c\,z}{\sigma_\nu}\Big) \end{aligned}$$

Tenemos entonces que a las intensidades obtenidas en experimentos de microarreglos debemos aplicarles la siguiente transformación para normalizar y estabilizar la varianza:

$$h_i(y) = \operatorname{arsinh}\Big(\frac{y-a_i}{b_i}\Big) \tag{3.4}$$

con $b_i = \beta_i \sigma_\nu / c$. Así, el modelo (2.10) en escala transformada toma la siguiente forma:

$$\operatorname{arsinh}\Big(\frac{Y_{ki}-a_i}{b_i}\Big) = \mu_{ki} + \varepsilon_{ki}\,, \quad \varepsilon_{ki} \overset{\text{iid}}{\sim} \mathcal{L}_\varepsilon \tag{3.5}$$

donde μ_{ki} representa el verdadero nivel de expresión del gen k en la muestra i en escala transformada y $\mathcal{L}_\varepsilon$ es una distribución con media cero y varianza c^2.

Recordando el modelo (2.10) se puede ver que μ_{ki} y m_{ki} están relacionados de la siguiente forma:

$$\mu_{ki} = \mathrm{E}\Big(\operatorname{arsinh}\big(\frac{c}{\sigma_\nu}(m_{ki}e^{\eta_{ki}} + \nu_{ki})\big)\Big)$$

Además se puede ver que $\mathrm{E}\Big(\operatorname{arsinh}\big(\frac{c}{\sigma_\nu}(m_{ki}e^{\eta_{ki}} + \nu_{ki})\big)\Big) \approx \operatorname{arsinh}(\frac{c}{\sigma_\nu}m_{ki})$. En efecto, sean $\eta \sim \mathcal{L}_\eta$, $\nu \sim \mathcal{L}_\nu$ y

$$f(\eta, \nu) = \operatorname{arsinh}\left(\frac{c}{\sigma_\nu}(me^\eta + \nu)\right)$$

Haciendo el desarrollo de Taylor de orden 1 de esta función alrededor del $(0, 0)$, obtenemos el siguiente polinomio:

$$P(\eta, \nu) = \operatorname{arsinh}(\frac{c}{\sigma_\nu}m) + \frac{c}{\sqrt{\sigma_\nu^2 + c^2m^2}}\,\nu + \frac{c\,m}{\sqrt{\sigma_\nu^2 + c^2m^2}}\,\eta$$

Como $f(\eta, \nu) \approx P(\eta, \nu)$ y además η y ν tienen media cero, entonces:

$$\begin{aligned}
\mathrm{E}(f(\eta, \nu)) &\approx \mathrm{E}(P(\eta, \nu)) \\
\mathrm{E}\Big(\operatorname{arsinh}\big(\frac{c}{\sigma_\nu}(me^\eta + \nu)\big)\Big) &\approx \mathrm{E}\Big(\operatorname{arsinh}(\frac{c}{\sigma_\nu}m) + \frac{c}{\sqrt{\sigma_\nu^2 + c^2m^2}}\,\nu + \frac{c\,m}{\sqrt{\sigma_\nu^2 + c^2m^2}}\,\eta\Big) \\
\mathrm{E}\Big(\operatorname{arsinh}\big(\frac{c}{\sigma_\nu}(me^\eta + \nu)\big)\Big) &\approx \operatorname{arsinh}(\frac{c}{\sigma_\nu}m)
\end{aligned}$$

3.3. Propiedades de la transformación arco-seno hiperbólico

Para hallar la transformación estabilizadora de la varianza $h(y) = \operatorname{arsinh}\Big(\frac{y-a}{b}\Big)$ nos basamos en la aproximación (3.1). En esta sección veremos que dicha transformación efectivamente estabiliza la varianza para una familia Y_m de variables aleatorias distribuidas de acuerdo a:

$$Y_m = me^\eta + \nu \qquad \eta \sim N(0, \sigma_\eta^2), \nu \sim N(0, 1) \tag{3.6}$$

con $m > 0$. Esto corresponde al modelo (2.10) con $a_i = 0$ y $\beta_i = 1$.

Recordando que $b_i = \beta_i\sigma_\nu/c$, en este caso la transformación estabilizadora será $h(y) = \operatorname{arsinh}(\frac{y}{b}) = \operatorname{arsinh}(cy)$, donde $c^2 = \frac{\mathrm{Var}(e^\eta)}{\mathrm{E}^2(e^\eta)}$. Como e^η tiene distribución log-normal tenemos que $c^2 = e^{\sigma_\eta^2} - 1$ (ver apéndice).

Para valores grandes de m, Y_m está dominada por el término me^η y además la función arsinh es parecida al logaritmo, por lo tanto $h(Y_m) \approx \log(Y_m) + \log(c) \approx \log(Y_m)$. Tenemos entonces que:

$$\mathrm{Var}(h(Y_m)) \quad \approx \quad \mathrm{Var}(\log(me^{\eta})) = \mathrm{Var}(\log(m) + \eta)) = \sigma_\eta^2 \qquad (3.7)$$

Como $\mathrm{E}(Y_m) = me^{\sigma_\eta^2/2}$, comparamos para distintos valores de m el desvío de $h(Y_m)$ con el desvío asintótico σ_η. Para esto se generó una muestra de Y_m de tamaño 10^5 y se calculó es desvío muestral de $h(Y_m)$ para cada $m = 1, 2, \ldots, 60$.

Los valores de Y_m se generaron de la siguiente forma: primero se generó una muestra de ν con una distribución $N(0, 1)$ y 4 muestras de η con distribuciones $N(0, \sigma_\eta^2)$ para σ_η =0.05, 0.1, 0.2 y 0.4. Luego obtenemos los valores de $Y_m = m\, e^{\eta} + \nu$ para cada m entre 1 y 60.

En la figura 3.1 se puede observar que el cociente entre el desvío de $h(Y_m)$ y σ_η no es superior a 1.035.

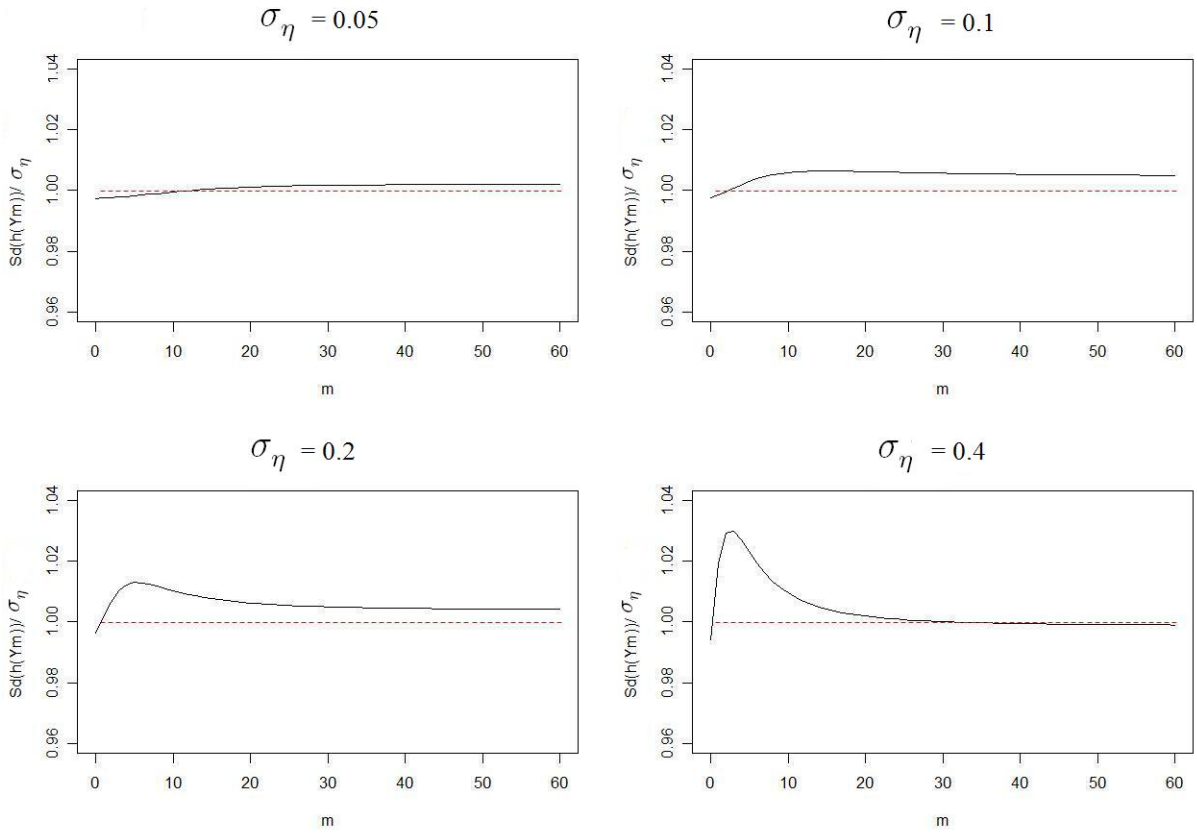

Figura 3.1: Validación de la transformación estabilizadora.

3.4. Otras transformaciones

Las funciones arco-seno hiperbólico y logaritmo están relacionadas de la siguiente forma:

$$\begin{aligned} \operatorname{arsinh}(x) &= \ln(x + \sqrt{x^2 + 1}) \\ \ln(x) &= \operatorname{arsinh}\left(\frac{1}{2}\left(x - \frac{1}{x}\right)\right) \end{aligned}$$

Además:

$$\lim_{x \to \infty} (\operatorname{arsinh}(x) - \ln(x) - \ln(2)) = 0$$

En la figura 3.2 se muestra el gráfico de la función $\operatorname{arsinh}((y - a_i)/b_i)$ para $a_i = 0$ y tres valores distintos de b_i.

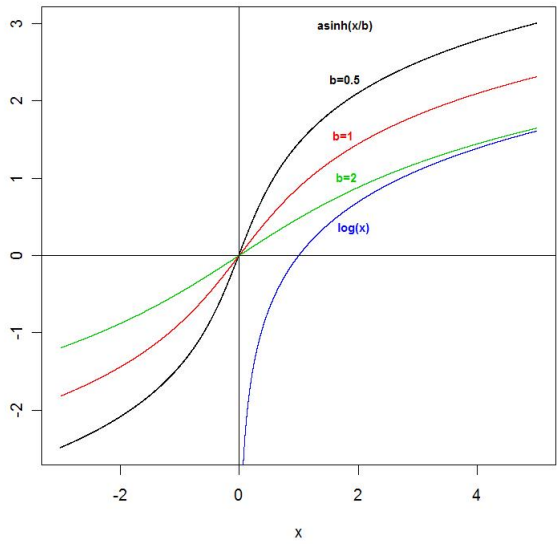

Figura 3.2: Gráfico de la función (3.4) para $a_i = 0$ y tres valores distintos de b_i. También se muestra el gráfico de la función logaritmo para comparar con estas curvas.

En el contexto de los microarreglos se han propuesto otras funciones que, dentro del rango de los datos, tienen gráficos similares a $\operatorname{arsinh}((y - a_i)/b_i)$, entre ellas están las siguientes:

$$\widetilde{h}(y) = \log\left(\frac{y-\widetilde{a}}{\widetilde{b}}\right) \tag{3.8}$$

$$\bar{h}(y) = \begin{cases} \log(y/\bar{b}) & \text{si} \quad y \geq \bar{a} \\ y/\bar{a} + \log(\bar{a}/\bar{b}) - 1 & \text{si} \quad y < \bar{a} \end{cases} \tag{3.9}$$

Mientras que la transformación (3.4) corresponde a una dependencia varianza-media de la forma $v(\mu) \propto (\mu - a)^2$+const., donde $v(\mu)$ es la varianza como función de la media, las dos transformaciones (3.8) y (3.9) corresponden a dependencias varianza-media de la forma

$$\widetilde{v}(\mu) \propto (\mu - \widetilde{a})^2$$

$$\bar{v}(\mu) \propto \begin{cases} \mu^2 & \text{si} \quad \mu \geq \bar{a} \\ \bar{a}^2 & \text{si} \quad \mu < \bar{a} \end{cases}$$

respectivamente. Nosotros utilizaremos, como Huber et al. (2003), la transformación (3.4) por conveniencia computacional y por su interpretabilidad en términos del modelo (2.10).

Capítulo 4

Estimación de parámetros

El modelo (3.5) relaciona las intensidades medidas Y_{ki} con los valores de expresión verdaderos (en escala transformada) μ_{ki} en términos de los parámetros de calibración y estabilización de la varianza a_i y b_i y de la distribución $\mathcal{L}_\epsilon$. Nuestro objetivo será estimar dichos paráme-tros. Primero describiremos los métodos de estimación de máxima verosimilitud y LTS (Least Trimmed Squares) para luego aplicar estas ideas al caso que nos interesa. En la sección 4.4 mostraremos un método resistente propuesto por Huber et al. (2003), basado en máxima verosimilitud y LTS, para la estimación de los parámetros de calibración y estabilización de la varianza de las intensidades de microarreglos.

4.1. Método de máxima verosimilitud

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $\mathcal{D}_\theta$, donde $\theta \in \Theta \subset \mathbb{R}^k$ es un vector de parámetros desconocidos que se desea estimar y sea f_θ la función de densidad ó probabilidad asociada.

Para $x_1, \ldots, x_n$ fijos la función $f_\theta(x_1, \ldots, x_n, \theta)$ depende sólo de θ. Esta función se llama función de verosimilitud y la notaremos $\ell(\theta) = f_\theta(x_1, \ldots, x_n, \theta)$.

En el caso discreto $f_\theta(x_1, \ldots, x_n, \theta)$ representa la probabilidad de observar el vector $(x_1, \ldots, x_n)$ cuando el valor del parámetro es θ.

Si $X_1, \ldots, X_n$ son independientes e idénticamente distribuidos, la función de verosimilitud se puede escribir como producto de n funciones de densidad ó probabilidad univariadas:

$$\ell(\theta) = \prod_{i=1}^{n} f_\theta(x_i, \theta)$$

Definición 4.1 *Diremos que $\hat{\theta}$ es un estimador de máxima verosimilitud de θ si se cumple:*

$$\ell(\hat{\theta}) = \max_{\theta \in \Theta} \ \ell(\theta)$$

Ejemplo 5

Sea $x_1, \ldots, x_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$, donde los parámetros μ y σ^2 son desconocidos. En este caso el vector de parámetros a estimar es $\theta = (\mu, \sigma^2)$. La función de densidad de cada variable es:

$$f(x_i; \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

Luego tenemos que:

$$\ell(\theta) = f_\theta(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

Como la función $\ln(x)$ (logaritmo natural) es monótona creciente, maximizar $\ell(\theta)$ será equivalente a maximizar $\ln(\ell(\theta))$. En este caso la función $\ln(\ell(\theta))$ es diferenciable, por lo tanto el estimador de máxima verosimilitud $(\hat{\mu}, \hat{\sigma}^2)$ debe verificar:

$$\begin{aligned} \frac{\partial \ln(\ell(\hat{\mu}, \hat{\sigma}^2))}{\partial \mu} &= 0 \\ \frac{\partial \ln(\ell(\hat{\mu}, \hat{\sigma}^2))}{\partial \sigma^2} &= 0 \end{aligned}$$

Luego, $(\hat{\mu}, \hat{\sigma}^2)$ debe ser solución del sistema de ecuaciones:

$$\begin{aligned} \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})}{\hat{\sigma}^2} &= 0 \\ \sum_{i=1}^{n} -\frac{1}{2\hat{\sigma}^2} + \frac{(x_i - \hat{\mu})^2}{2(\hat{\sigma}^2)^2} &= 0 \end{aligned}$$

Este sistema tiene la siguiente solución:

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{n} \frac{x_i}{n} = \overline{x} \\ \hat{\sigma}^2 &= \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n} \end{aligned}$$

Este es el único punto crítico de $\ln(\ell(\theta))$ pero falta ver que efectivamente es un máximo.

La matriz jacobiana de la función $\ln(\ell(\theta))$ es:

$$-\begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) \\ \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) & \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{\sigma^4} \end{pmatrix}$$

Además se puede verificar que esta matriz evaluada en $(\hat{\mu}, \hat{\sigma}^2)$ es definida negativa y de esta forma queda demostrado que $(\hat{\mu}, \hat{\sigma}^2)$ es el estimador de máxima verosimilitud de (μ, σ^2).

4.2. Least Trimmed Squares

Least Trimmed Squares (LTS) es un método de estimación de parámetros propuesto por Rousseeuw (1985) como una alternativa robusta al método clásico de mínimos cuadrados. A continuación definimos el estimador LTS para el caso general.

Definición 4.2 *Consideremos el modelo:*

$$y_i = h(x_i, \theta) + \varepsilon_i, \qquad 1, \ldots, n \tag{4.1}$$

donde y_i representa la variable dependiente y $h(x_i, \theta)$ es una función de los datos $x_i \in \mathbb{R}^p$ y del vector de parámetros $\theta \in \mathbb{R}^p$. Dada una muestra (y_i, x_i), el estimador LTS está definido por:

$$\hat{\theta}_{(LTS,k)} = \arg\min_{\theta \in B} \sum_{i=1}^{k} r^2_{[i]}(\theta) \tag{4.2}$$

donde $B \in \mathbb{R}^p$ es el espacio de parámetros, $r^2_{[i]}(\theta)$ representa la muestra de los cuadrados de los residuos $r_i^2(\theta) = (y_i - h(x_i, \theta))^2$ ordenados y k es una constante que satisface $\frac{n}{2} < k \leq n$.

Una definición equivalente es:

$$\hat{\theta}_{(LTS,k)} = \arg\min_{\theta \in B} \min_{K} \sum_{i \in K} r_i^2(\theta) \tag{4.3}$$

donde la minimización sobre K recorre los subconjuntos de $\{1, \ldots, n\}$ de k elementos, con $\frac{n}{2} < k \leq n$.

La constante k determina la robustez del estimador LTS, ya que la igualdad (4.2) implica que las $n - k$ observaciones con mayores residuos no tienen influencia directa sobre el estimador. El mayor nivel de robustez se alcanza para $k = [n/2] + [(p+1)/2]$ (Rousseeuw y Leroy, 1987), mientras que para $k = n$ el nivel de robustez es mínimo. En este caso el estimador LTS es equivalente al estimador de mínimos cuadrados.

Rousseeuw y Van Driessen (1999) proponen un algoritmo llamado FAST-LTS que mejora el tiempo computacional del cálculo del estimador LTS para modelos lineales.

4.3. C-step y la idea básica del algoritmo FAST-LTS

El propósito de esta sección es dar una idea del algoritmo para calcular el estimador LTS en modelos lineales. Los detalles pueden consultarse en Rousseeuw y Van Driessen (1999).

Consideremos el modelo lineal

$$y_i = x_{i1}\theta_1 + \cdots + x_{ip}\theta_p + e_i \quad i = 1, \ldots, n$$

donde los datos son de la forma $(\boldsymbol{x_i}, y_i) = (x_{i1}, \ldots, x_{ip}, y_i)$ con $x_{ip} = 1$ para regresión con término de intercepción.

Una de las claves del algoritmo propuesto por Rousseeuw es el hecho de que empezando desde cualquier estimación de los coeficientes de regresión, es posible calcular otra estimación de los parámetros que dan un valor de la función objetivo aún menor.

Proposición 4.1 *Consideremos un conjunto de datos* $(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_n}, y_n)$.

Sea $K_1 \subset \{1, \ldots, n\}$ *con* $|K_1| = k$, *sean* $\hat{\boldsymbol{\theta}}_{\mathbf{1}} = (\hat{\theta}_1^1, \hat{\theta}_2^1, \ldots, \hat{\theta}_p^1) \in \mathbb{R}^p$ *y* $r_1(i)$ *el residuo del* i*-ésimo dato, i.e.* $r_1(i) = y_i - (\hat{\theta}_1^1 x_{i1} + \hat{\theta}_2^1 x_{i2} + \ldots + \hat{\theta}_p^1 x_{ip})$. *Llamemos* $Q_1 := \sum_{i \in K_1} \big(r_1(i)\big)^2$.

Sea K_2 *el conjunto de índices de aquellos* k *residuos con menores valores absolutos. Calculemos ahora el estimador de mínimos cuadrados* $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$ *correspondiente a las* k *observaciones cuyo índice pertenece a* K_2. *Esto da lugar a nuevos residuos* $r_2(i)$ *para* $i = 1, \ldots, n$ *y* $Q_2 := \sum_{i \in K_2} \big(r_2(i)\big)^2$. *Entonces se tiene que*

$$Q_2 \leq Q_1$$

Demostración. Como K_2 corresponde a los k residuos de menor valor absoluto, tenemos que $\sum_{i \in K_2} \big(r_1(i)\big)^2 \leq \sum_{i \in K_1} \big(r_1(i)\big)^2 = Q_1$. Como $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$ es el estimador de mínimos cuadrados correspondiente a las k observaciones en K_2, se tiene que:

$$Q_2 = \sum_{i \in K_2} \big(r_2(i)\big)^2 \leq \sum_{i \in K_2} \big(r_1(i)\big)^2 \leq Q_1 \quad \blacksquare$$

Aplicando esta proposición a un subconjunto de índices K_1 se obtiene K_2 con $Q_2 \leq Q_1$. En su algoritmo, Rousseeuw llama a este paso el C-step, donde C se refiere a *concentración* ya que K_2 está mas concentrado (la suma de

los cuadrados de los residuos es menor) que K_1. En términos algorítmicos, el C-step se puede describir de la siguiente forma:

Dado un k-subconjunto H_{old}:

- calcular $\hat{\theta}_{old}$:= el estimador de mínimos cuadrados basado en H_{old}
- calcular los residuos $r_{old}(i)$ para $i = 1, \ldots, n$
- ordenar los valores absolutos de estos residuos, o equivalentemente, buscar una permutación π para la cual $|r_{old}(\pi(1))| \leq |r_{old}(\pi(2))| \leq \cdots \leq |r_{old}(\pi(n))|$
- definir $H_{new} := \{\pi(1), \pi(2), \ldots, \pi(k)\}$
- calcular $\hat{\theta}_{new}$:= el estimador de mínimos cuadrados basado en H_{new}

Repeticiones del C-step generan un proceso iterativo. Si $Q_2 = Q_1$ paramos, sino aplicamos otra vez este paso, obteniendo Q_3, Q_4, etc. La sucesión $Q_1 \geq Q_2 \geq Q_3 \geq \ldots$ es no negativa y por lo tanto converge. Mas aún, como hay una cantidad finita de k-subconjuntos existe un índice m para el cual $Q_m = Q_{m-1}$ (en la práctica m suele ser menor que 10), por lo tanto, la convergencia se alcanza siempre después de una cantidad finita de iteraciones. Esta no es una condición suficiente para que Q_m sea el mínimo global de la función objetivo del estimador LTS, pero sí es una condición necesaria.

Esto provee una idea parcial para un algoritmo:

Tomar varias conjuntos iniciales H_1 y aplicar el C-step a cada uno hasta alcanzar la convergencia y quedarse con la solución con menor valor de $\sum_{i=1}^{k} r_{[i]}^2$

Para poner en práctica esta idea, hay varios puntos a tener en cuenta: cómo generar los conjuntos iniciales H_1, cuántos de estos conjuntos se necesitan, cómo evitar la duplicación de trabajo ya que varios H_1 podrían llevar a la misma solución, etc. Estos temas son tratados por Rousseeuw y van Driessen (1999).

4.4. Regresión resistente VSN

Para estimar los parámetros a_i y b_i del modelo (3.5) supondremos en primer lugar que no hay genes expresados diferencialmente, es decir, el nivel de expresión de cada gen es el mismo para todas las muestras analizadas. En este caso tendremos que $\mu_{ki} = \mu_k$ para todo i. Además supondremos que la distribución $\mathcal{L}_\varepsilon$ es normal, por lo tanto tenemos que:

$$\operatorname{arsinh}\Big(\frac{Y_{ki} - a_i}{b_i}\Big) = \mu_k + \varepsilon_{ki}\,, \quad \varepsilon_{ki} \overset{\text{iid}}{\sim} N(0, c^2) \tag{4.4}$$

Bajo estas hipótesis, en 4.4.1, aplicaremos el método de máxima verosimilitud para estimar los parámetros a_i y b_i. Luego, en 4.4.2, presentaremos el

método robusto propuesto por Huber et al. (2003) para el caso en el que haya una fracción minoritaria de genes expresados diferencialmente y la distribución $\mathcal{L}_\varepsilon$ sea aproximadamente normal en el centro.

En 4.4.1 además de suponer que ε_{ki} tiene distribución normal, hacemos uso de los siguientes teoremas para deducir la función de densidad de Y_{ki}:

Teorema 4.1 *Sea $g : \mathbb{R} \to \mathbb{R}$ una función estrictamente creciente y sea Y una variable aleatoria. Entonces si $Z = g(Y)$ tiene función de distribución F_Z, la función de distribución de Y será:*

$$F_Y(y) = F_Z(g(y)) \tag{4.5}$$

Demostración. Como g es estrictamente creciente se tendrá

$$F_Y(y) = P\big(Y \leq y\big) = P\big(g(Y) \leq g(y)\big) = P\big(Z \leq g(y)\big) = F_Z\big(g(y)\big) \quad \blacksquare$$

Teorema 4.2 *Sea $g : \mathbb{R} \to \mathbb{R}$ una función derivable con $g'(y) > 0$ y sea Y una variable aleatoria. Si $Z = g(Y)$ es absolutamente continua con función de densidad f_Z, entonces la función de densidad de Y será:*

$$f_Y(y) = f_Z(g(y))\, g'(y)$$

Demostración. Se deduce derivando (4.5) $\blacksquare$

4.4.1. Estimación de máxima verosimilitud

Supongamos que $h_i(Y_{ki}) = \operatorname{arsinh}\left(\frac{Y_{ki}-a_i}{b_i}\right)$ tiene distribución normal con media μ_k independiente de i y desvío c. Los parámetros de este modelo son $\{a_i, b_i\}_{i=1,\ldots,d}$, $\{\mu_k\}_{k=1,\ldots,n}$ y c. Los valores que se observan son las intensidades Y_{ki} y queremos estimar a_i y b_i $(i = 1, \ldots, d)$.

Si $g(Y_{ki}) = \frac{h_i(y_{ki})-\mu_k}{c}$, entonces $g(Y_{ki})$ tiene distribución normal estándar y de acuerdo con el teorema 4.2 la función de densidad de Y_{ki} es:

$$\phi\left(\frac{h_i(y_{ki}) - \mu_k}{c}\right) \frac{h_i'(y_{ki})}{c} \tag{4.6}$$

donde ϕ es la densidad de una variable aleatoria con distribución normal estándar.

Por lo tanto, los estimadores de máxima verosimilitud de los parámetros del modelo ($\{a_i\}$ $\{b_i\}$, $\{\mu_k\}$ y c) son los que maximizan la siguiente función:

$$\ell(a_1, b_1, \ldots, a_d, b_d, \mu_1, \ldots, \mu_n, c) = \prod_{k=1}^{n} \prod_{i=1}^{d} \phi\left(\frac{h_i(y_{ki}) - \mu_k}{c}\right) \frac{h_i'(y_{ki})}{c} \tag{4.7}$$

$$= \prod_{k=1}^{n}\prod_{i=1}^{d} \frac{1}{\sqrt{2\pi c^2}} \exp\left(-\frac{(h_i(y_{ki}) - \mu_k)^2}{2c^2}\right) h_i'(y_{ki}) \tag{4.8}$$

Los parámetros a_i y b_i son parámetros de las transformaciones h_i. Si fijamos estos valores y hacemos una cuenta similar a la del ejemplo 5 obtenemos los valores de $\{\mu_k\}_{k=1,\ldots,n}$ y c que maximizan (4.7):

$$\hat{\mu}_k = \frac{1}{d}\sum_{i=1}^{d} h_i(y_{ki}) \quad (k = 1, \ldots, n) \tag{4.9}$$

$$\hat{c}^2 = \frac{1}{nd}\sum_{k=1}^{n}\sum_{i=1}^{d}(h_i(y_{ki}) - \hat{\mu}_k)^2 \tag{4.10}$$

Definimos:

$$\begin{aligned} p\ell(a_1, b_1, \ldots, a_d, b_d) &= \max_{(\mu_1,\ldots,\mu_n,c)} \ell(a_1, b_1, \ldots, a_d, b_d, \mu_1, \ldots, \mu_n, c) \\ &= \ell(a_1, b_1, \ldots, a_d, b_d, \hat{\mu}_1, \ldots, \hat{\mu}_n, \hat{c}) \end{aligned}$$

Ahora los estimadores de a_i y b_i son aquellos que maximizan la denominada "profile likelihood"$p\ell(a_1, b_1, \ldots, a_d, b_d)$ que se obtiene reemplazando μ_k y c^2 por sus estimadores $\hat{\mu}_k$ y $\hat{c}^2$.

Reemplazando $\hat{\mu}_k$ y $\hat{c}$ en ℓ tenemos:

$$p\ell(a_1, b_1, \ldots, a_d, b_d) = \prod_{k=1}^{n}\prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\hat{c}} \exp\left(\frac{-(h_i(y_{ki}) - \hat{\mu}_k)^2}{\frac{2}{nd}\sum_{k=1}^{n}\sum_{i=1}^{d}(h_i(y_{ki}) - \hat{\mu}_k)^2}\right) h_i'(y_{ki})$$

$$= \frac{1}{(2\pi)^{nd/2}\,\hat{c}^{nd}} \exp\left[\sum_{k=1}^{n}\sum_{i=1}^{d}\left(\frac{-(h_i(y_{ki}) - \hat{\mu}_k)^2}{\frac{2}{nd}\sum_{k=1}^{n}\sum_{i=1}^{d}(h_i(y_{ki}) - \hat{\mu}_k)^2}\right)\right] \prod_{k=1}^{n}\prod_{i=1}^{d} h_i'(y_{ki})$$

$$= \frac{e^{-nd/2}}{(2\pi)^{nd/2}\,\hat{c}^{nd}} \prod_{k=1}^{n}\prod_{i=1}^{d} h_i'(y_{ki})$$

El logaritmo de esta expresión (profile log-likelihood) está dado por:

$$p\ell\ell(a_1, b_1, \ldots, a_d, b_d) = -nd \log(\hat{c}) + \sum_{k=1}^{n} \sum_{i=1}^{d} \log(h_i'(y_{ki})) + \text{Cte}$$

$$= -\frac{nd}{2} \log\left(\sum_{k=1}^{n} \sum_{i=1}^{d} (h_i(y_{ki}) - \hat{\mu}_k)^2 \right) + \sum_{k=1}^{n} \sum_{i=1}^{d} \log(h_i'(y_{ki})) + \text{Cte} \qquad (4.11)$$

donde $\text{Cte} = -\frac{nd}{2}(1 + \log(2\pi))$.

Los estimadores de máxima verosimilitud de $(a_1, b_1, \ldots, a_d, b_d)$ pueden hallarse maximizando esta última expresión. Esta función puede maximizarse numéricamente con la restricción de que $b_i > 0$.

4.4.2. Estimación resistente

El estimador de máxima verosimilitud que se obtiene maximizando (4.11) es sensible a desviaciones de la normalidad y a la presencia de genes expresados diferencialmente, para los cuales no vale $\mu_{ki} = \mu_k$.

Huber et al. (2003) propuso un procedimiento heurístico donde utiliza el método de regresión LTS para hacer mas robusta la estimación bajo la presencia de genes expresados diferencialmente y con $\mathcal{L}_\varepsilon$ unimodal y simétrica pero no necesariamente normal. Básicamente, el procedimiento consiste en reemplazar en la expresión (4.11) las sumas sobre $k = 1, \ldots, n$ por sumas sobre $k \in K$, eligiendo en cada iteración el subconjunto K de tamaño $\lceil nq_{lts} \rceil$ de forma tal que los residuos mas chicos sean aquellos r_k con $k \in K$, donde $0{,}5 \leq q_{lts} \leq 1$ y $\lceil x \rceil$ es el menor entero mayor o igual que x. En realidad, en la elección de K se tiene en cuenta el contexto y no se descartan exactamente los valores con mayores residuos. Aquí se supone que la fracción de outliers debería ser aproximadamente la misma a lo largo de todo el rango de intensidades promedio $\hat{\mu}_k$ y por lo tanto se particiona el conjunto $\{1, \ldots, n\}$ en 10 partes de forma tal que la partición 1 corresponde a los datos para los cuales $\hat{\mu}_k$ es menor que el 10 %-cuantil de los $\hat{\mu}_k$, la partición 2 corresponde a los datos para los cuales $\hat{\mu}_k$ está entre el cuantil del 10 % y el del 20 %, etc. Luego se eligen dentro de cada partición los datos con menores residuos.

A continuación describimos el procedimiento en términos algorítmicos.

En lo que sigue utilizaremos la siguiente notación:

$$p\ell\ell_K(a_1, b_1, \ldots, a_d, b_d) = -\frac{nd}{2} \log\left(\sum_{k \in K} \sum_{i=1}^{d} (h_i(y_{ki}) - \hat{\mu}_k)^2 \right) + \sum_{k \in K} \sum_{i=1}^{d} \log(h_i'(y_{ki}))$$

Algoritmo:

1. Hacer $K = \{1, \ldots, n\}$
2. Hallar $\hat{\theta} = (\hat{a}_1, \hat{b}_1, \ldots, \hat{a}_d, \hat{b}_d)$ que maximiza $p\ell\ell_K(\theta)$, es decir:
$$\hat{\theta} = \arg\max \ \ p\ell\ell_K(\theta)$$
3. Para $k = 1, \ldots, n$ calcular y guardar los residuos $r_k = \sum_{i=1}^{d} \left(\hat{h}_i(y_{ki}) - \hat{\mu}_k\right)^2$
4. Ordenar los valores de $\hat{\mu}_k$, es decir, hallar una permutación $\Pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ tal que:
$$\hat{\mu}_{\Pi(1)} \leq \hat{\mu}_{\Pi(2)} \leq \ldots \leq \hat{\mu}_{\Pi(n)}$$
5. Particionar el conjunto $\{1, \ldots, n\}$ en diez partes $T_1, \ldots, T_{10}$ de forma tal que T_1 contenga aquellos k para los cuales $\hat{\mu}_k$ es menor que el 10 %-cuantil de los $\hat{\mu}_k$, T_2 contenga aquellos k para los cuales $\hat{\mu}_k$ está entre el cuantil del 10 % y el del 20 %, etc. Es decir:
$$T_j = \{\ \Pi(i) \,/\, (j-1)\frac{n}{10} < i \leq j\frac{n}{10}\} \qquad\qquad j = 1, \ldots, 10$$
6. Definir $Q_s = q_{lts}$-cuantil de $\{r_k \,/\, k \in T_s\}$ $\qquad\qquad s = 1, \ldots, 10$
7. Elegir $K = \bigcup_{s=1}^{10} \{k \in T_s \,/\, r_k \leq Q_s\}$
8. Mientras no se alcance la cantidad máxima de iteraciones predeterminada volver a 2

Para lograr un algoritmo más eficiente el criterio de finalización podría depender también de algún criterio de convergencia.

Capítulo 5

Simulaciones

Para analizar el comportamiento del algoritmo en distintas situaciones hicimos varias simulaciones generando datos a partir del modelo (2.10) para diferentes valores de parámetros tales como cantidad de sondas, cantidad de arreglos, proporción de genes expresados diferencialmente, etc.

Para generar los datos y estimar los parámetros del modelo en base a dichos datos se utilizó el software R, el cual se puede descargar gratuitamente de la página www.r-project.org y el paquete *vsn* versión 3.2.1 desarrollado por el grupo de Huber, disponible en www.bioconductor.org.

En la sección 5.1 se describe la forma en que se generaron los datos. En el resto del capítulo aplicamos el algoritmo a los datos generados y analizamos su performance.

5.1. Generación de datos

El propósito de la simulación es generar datos que reflejen ciertas propiedades de las intensidades de las sondas de microarreglos obtenidas experimentalmente. En nuestro caso, las simulaciones fueron realizadas según el modelo:

$$\operatorname{arsinh}\Big(\frac{Y_{ki} - a_i}{b_i}\Big) = \mu_{ki} + \varepsilon_{ki}\,, \quad \varepsilon_{ki} \overset{\text{iid}}{\sim} \mathcal{L}_\varepsilon \tag{5.1}$$

definido en los capítulos anteriores, suponiendo $\mathcal{L}_\varepsilon = N(0, c^2)$.

Según Newton et al. (2001) los valores recíprocos de los niveles de expresión se pueden modelar con una distribución gamma. Por lo tanto, los valores de los parámetros μ_{ki} fueron generados de la siguiente forma. En primer lugar, para cada gen k se generó un valor μ_k de acuerdo a:

$$\mu_k = \operatorname{arsinh}(m_k), \quad 1/m_k \sim \Gamma(1,1) \tag{5.2}$$

Para modelar la mezcla de genes expresados diferencialmente y genes no expresados diferencialmente, se generan indicadores $p_k \in \{0,1\}$ con $P(p_k = 1) = p_{dif}$, donde p_{dif} es la proporción de genes expresados diferencialmnete que

se quiere simular. Luego, para cada gen expresado diferencialmente (aquellos con $p_k = 1$) y para cada muestra $i \geq 2$ se genera un factor $s_{ki} \in \{-1, 1\}$ para simular genes sub-regulados y sobre-regulados, con $P(s_{ki} = 1) = p_{up}$, donde p_{up} es la proporción de genes expresados diferencialmente que están sobre-regulados ($p_{up} = 0{,}5$ indica que hay la misma cantidad de genes sobre-regulados y sub-regulados). También se simula la amplitud de la expresión diferencial en estos genes z_{ki}, generada a partir de una distribución $U(0, z_{max})$.

Tenemos entonces que p_k indica si el gen k se expresa diferencialmente en las distintas muestras; s_{ki} indica si está sub-regulado o sobre-regulado en la muestra i comparado con la muestra 1; z_{ki} indica la diferencia de expresión entre la muestra i y la muestra 1 en valor absoluto. Combinando ésto obtenemos:

$$\begin{aligned} \mu_{k1} &= \mu_k \\ \mu_{ki} &= \mu_k + p_k\, s_{ki}\, z_{ki} \quad (i \geq 2) \end{aligned}$$

El paquete *vsn* trabaja con el siguiente modelo equivalente a (5.1):

$$\operatorname{arsinh}(A_i + e^{B_i} Y_{ki}) = \mu_{ki} + \varepsilon_{ki} \tag{5.3}$$

Aquí los parámetros A_i y B_i no tienen restricciones mientras que en el modelo (5.1) b_i debía ser positivo.

La relación entre los parámetros de calibración de (5.1) y los de (5.3) es la siguiente:

$$\begin{aligned} a_i &= -A_i e^{-B_i} \\ b_i &= e^{-B_i} \end{aligned}$$

Para la simulación se generaron los valores de A_i y B_i a través de una distribución uniforme en (-2,2):

$$A_i, B_i \sim U(-2, 2)$$

Con los valores de μ_{ki}, A_i y B_i, obtenemos los datos simulados de las intensidades por medio de:

$$y_{ki} = -A_i e^{-B_i} + e^{-B_i} \sinh(\mu_{ki} + \varepsilon_{ki}) \qquad \varepsilon_{ki} \sim N(0, c^2) \tag{5.4}$$

Con los datos simulados y_{ki} se estimaron los parámetros mediante la función `vsn2` del paquete *vsn*, obteniendo de esta forma una estimación de los datos transformados:

$$\hat{h}_{ki} = \operatorname{arsinh}\left(e^{\hat{B}_i} y_{ki} + \hat{A}_i\right) \tag{5.5}$$

Nota: la función `vsn2` devuelve, entre otras cosas, los parámetros estimados $\hat{A}_i$, $\hat{B}_i$ y los datos transformados, pero los datos devueltos están sujetos a la transformación:

$$\begin{aligned}\tilde{h}_{ki} &= \frac{\operatorname{arsinh}\left(e^{\hat{B}_i} y_{ki} + \hat{A}_i\right)}{\log(2)} + c \\ &= \frac{\log\left(e^{\hat{B}_i} y_{ki} + \hat{A}_i + \sqrt{1 + (e^{\hat{B}_i} y_{ki} + \hat{A}_i)^2}\right)}{\log(2)} + c\end{aligned}$$

donde c es una constante calculada a partir de los valores de B_i de forma tal que para valores grandes de y_{ki} la transformación corresponda aproximadamente a la función logaritmo en base 2. Esta constante se debe a que a muchos usuarios les gusta ver los datos en un rango de valores con los cuales estén familiarizados y la función logaritmo en base 2 es utilizada frecuentemente en el contexto de los microarreglos.

Más precisamente,

$$c = -\log_2(2) - B\log_2(e)$$

donde $B = \frac{1}{d}\sum_{i=1}^{d} B_i$. Efectivamente, para valores grandes de y_{ki} tenemos:

$$\begin{aligned}\tilde{h}_{ki} &= \log_2\left(e^{\hat{B}_i} y_{ki} + \hat{A}_i + \sqrt{1 + (e^{\hat{B}_i} y_{ki} + \hat{A}_i)^2}\right) + c \\ &= \log_2\left(e^{\hat{B}_i} y_{ki}\left[1 + \frac{\hat{A}_i}{e^{\hat{B}_i} y_{ki}} + \sqrt{\frac{1}{(e^{\hat{B}_i} y_{ki})^2} + (1 + \frac{\hat{A}_i}{e^{\hat{B}_i} y_{ki}})^2}\right]\right) + c \\ &= \log_2(y_{ki}) + \hat{B}_i\log_2(e) + \log_2\left(1 + \frac{\hat{A}_i}{e^{\hat{B}_i} y_{ki}} + \sqrt{\frac{1}{(e^{\hat{B}_i} y_{ki})^2} + \left(1 + \frac{\hat{A}_i}{e^{\hat{B}_i} y_{ki}}\right)^2}\right) + c \\ &\approx \log_2(y_{ki}) + \hat{B}_i\log_2(e) + \log_2(2) + c \\ &\approx \log_2(y_{ki})\end{aligned}$$

La comparación entre los verdaderos datos transformados y las estimaciones de éstos se realizó por medio de:

$$\delta = \sqrt{\frac{1}{d\,|\kappa|}\sum_{i=1}^{d}\sum_{k\in\kappa}\left(\Delta\hat{h}_{ki} - \Delta h_{ki}\right)^2} \tag{5.6}$$

donde κ es el conjunto de los k para los cuales $p_k = 0$, es decir, se tienen en cuenta solamente los genes no expresados diferencialmente, $\Delta\hat{h}_{ki}$ está definido por:

$$\Delta\hat{h}_{ki} = \hat{h}_{ki} - \frac{1}{d}\sum_{j=1}^{d}\hat{h}_{kj} \tag{5.7}$$

y Δh_{ki} son los valores verdaderos:

$$\Delta h_{ki} = h_{ki} - \frac{1}{d}\sum_{j=1}^{d} h_{kj} \tag{5.8}$$

En las próximas secciones analizamos los resultados obtenidos en las simulaciones, donde para cada conjunto de datos simulado se estimaron los parámetros por medio de `vsn2` y se calculó de valor de δ.

5.2. Número de sondas n

Para analizar como influye el número de sondas n en la estimación de los parámetros de calibración y estabilización de la varianza se simularon las intensidades de los spots de dos microarreglos ($d = 2$) sin genes expresados diferencialmente ($p_{dif} = 0$). Los valores de n utilizados son 250, 500, 1000, 2000, 4000, 8000, 16000 y 32000.

Para cada valor de n se generaron 30 conjuntos de datos para los cuales se estimaron los parámetros utilizando distintos valores de q_{lts}, los resultados obtenidos se muestran en la figura 5.1.

Para comparar la eficacia del método utilizando distintos valores de q_{lts} se realizó un gráfico que muestra el δ promedio para los distintos valores de n en escala semi-logarítmica, figura 5.2.

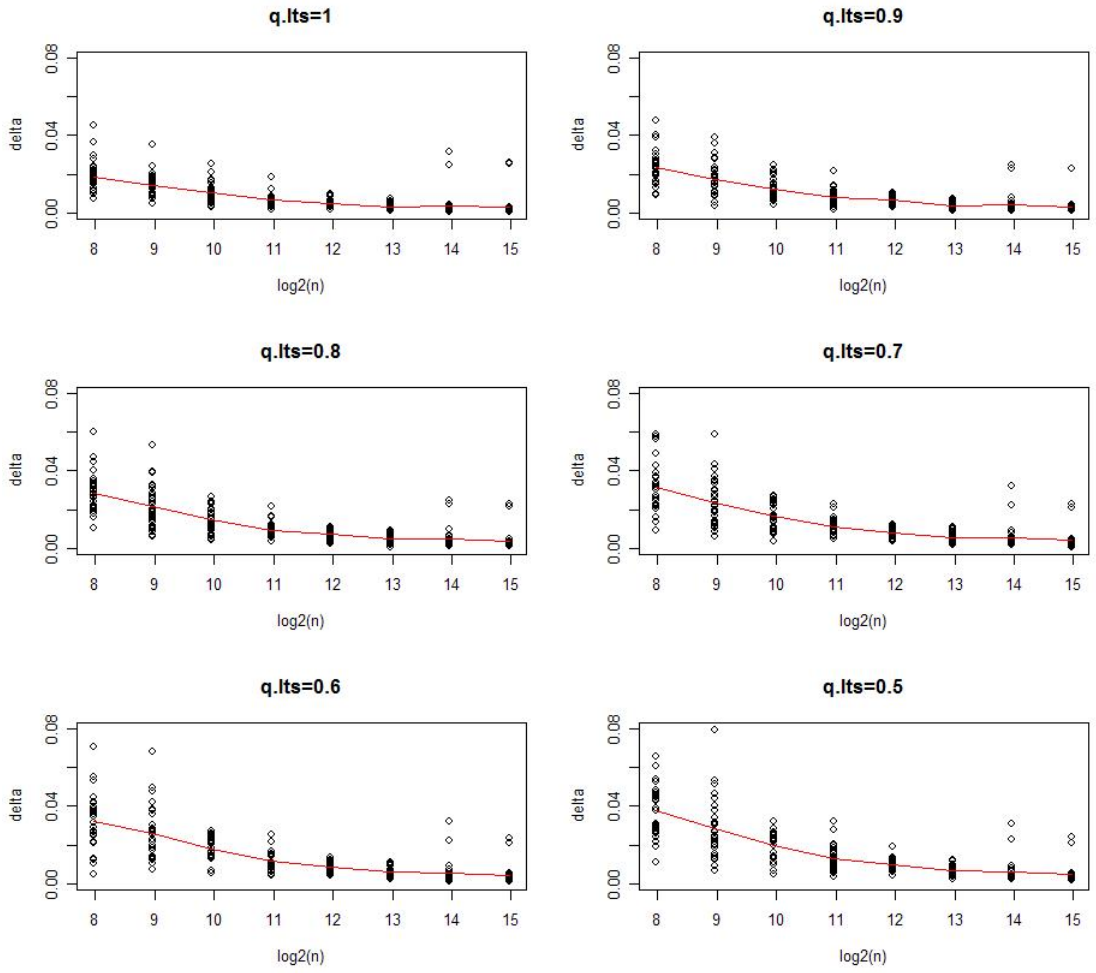

Figura 5.1: Los datos fueron generados para distintos valores de n, con $d = 2$ y $p_{dif} = 0$. Para cada n se generaron 30 conjuntos de datos y se calculó el valor de δ, la línea pasa por los valores promedio de δ para cada n.

En la figura 5.2 se observa que a medida que la cantidad de spots n aumenta las estimaciones son mejores. En este caso al no haber genes expresados diferencialmente, las estimaciones mejoran al aumentar q_{lts}. Por otro lado, para valores grandes de n la diferencia entre las estimaciones con distintos q_{lts} no es tan grande.

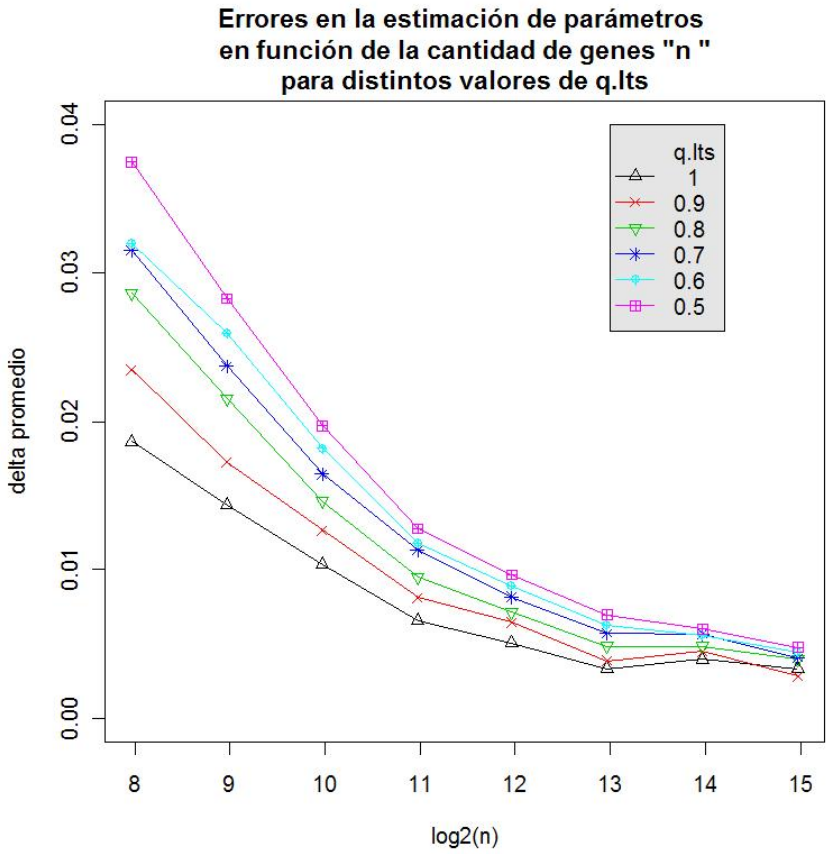

Figura 5.2: Comparación de resultados para distintos valores de q_{lts}.

5.3. Número de arreglos d

Para analizar como influye el número de arreglos d en la estimación de los parámetros se realizaron simulaciones de las intensidades de 8064 spots sin genes expresados diferencialmente ($p_{dif} = 0$). Los valores de d utilizados son 2, 4, 8, 16 y 32.

Para cada valor de d se generaron 30 conjuntos de datos para los cuales se estimaron los parámetros utilizando distintos valores de q_{lts}, los resultados obtenidos se muestran en la figura 5.3.

Para comparar la eficacia del método utilizando distintos valores de q_{lts} se realizó un gráfico que muestra el δ promedio para los distintos valores de d en escala semi-logarítmica, figura 5.4.

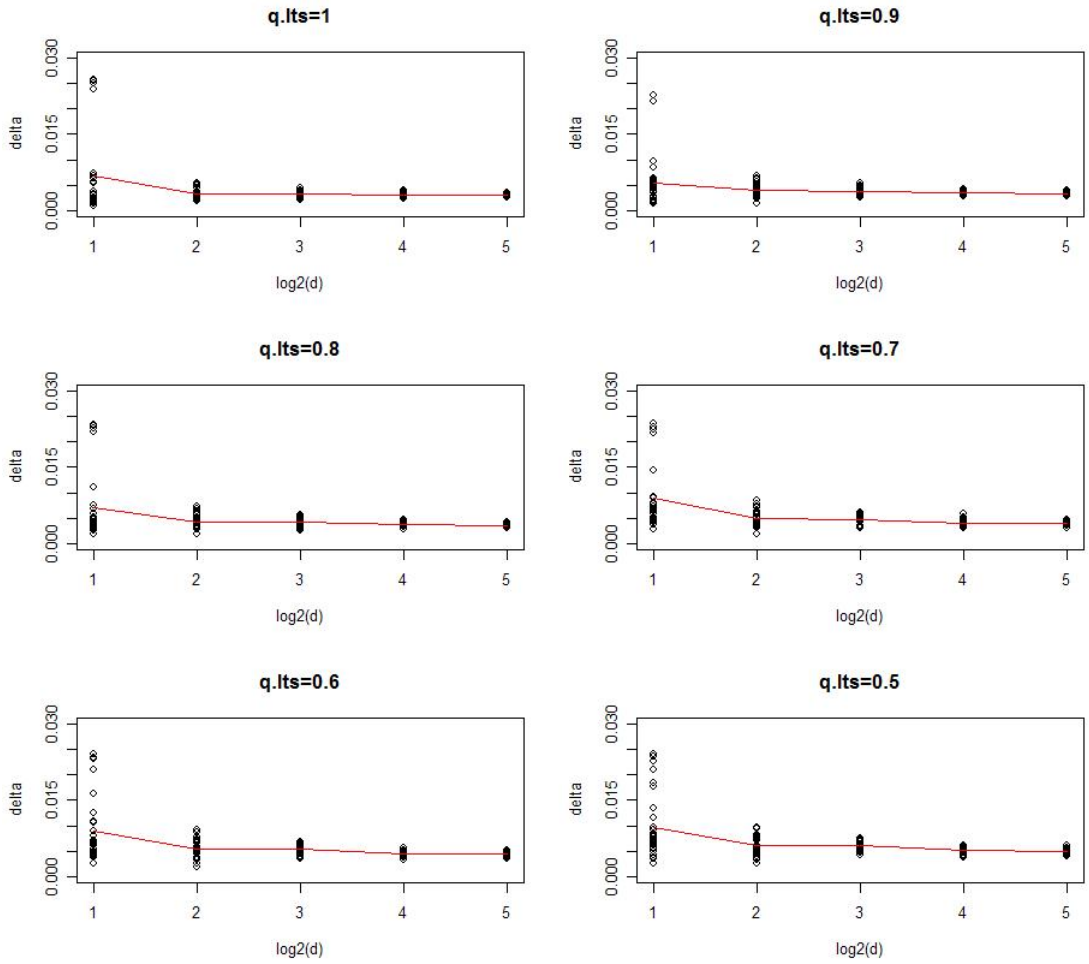

Figura 5.3: Los datos fueron generados para distintos valores de d, con $n = 8064$ y $p_{dif} = 0$. Para cada d se generaron 30 conjuntos de datos y se calculó el valor de δ, la línea pasa por los valores promedio de δ para cada d.

Se puede observar que los errores en la estimación permanecen aproximadamente constantes para los distintos valores de d, esto se debe a que la cantidad de parámetros a estimar es proporcional a la cantidad de arreglos.

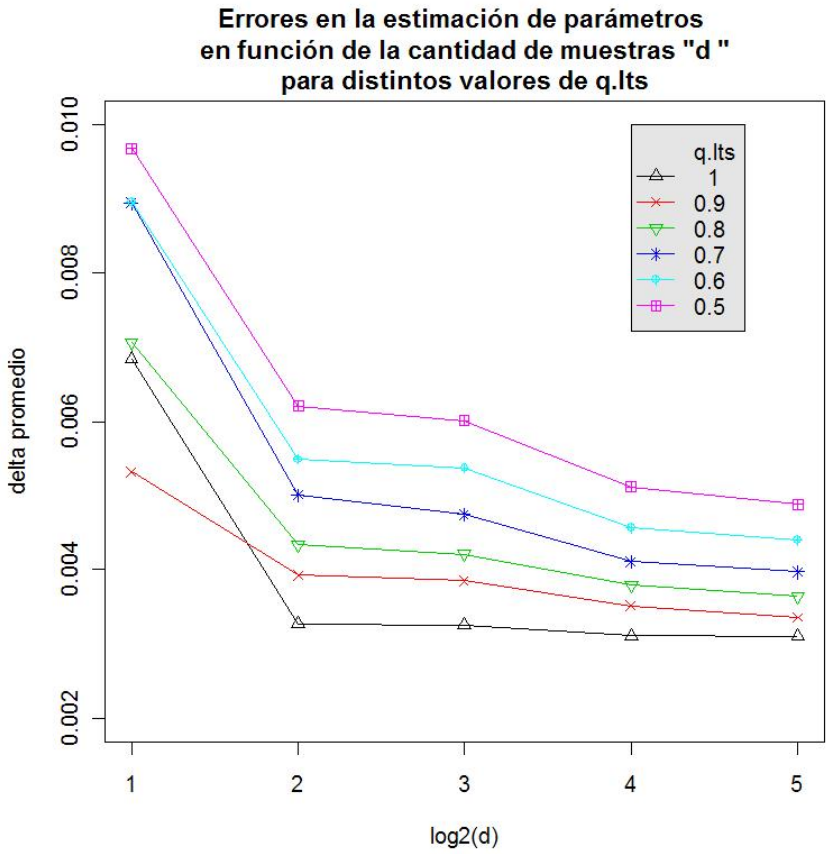

Figura 5.4: Comparación de resultados para distintos valores de q_{lts}.

5.4. Genes expresados diferencialmente.

Para estudiar las estimaciones bajo la presencia de genes expresados diferencialmente se simularon las intensidades de 8064 spots de 2 microarreglos con un valor de $z_{max} = 2$, donde z_{max} es el valor máximo de expresión diferencial en escala transformada. El paquete *vsn* permite simular datos de microarreglos a través de la función `sagmbSimulateData` pudiendo variar distintos parámetros. El valor de z_{max} no es un parámetro que se pueda modificar en esta función y su valor está fijo en 2. En la sección siguiente mostraremos los resultados que obtuvimos para distintos valores de z_{max}.

Las proporciones de genes expresados diferencialmente que se analizaron son 0, 0.1, 0.2, 0.3, 0.4 y 0.5.

Cuando la proporción de genes expresados diferencialmente aumenta las estimaciones mejoran con valores de q_{lts} cercanos a 0.5.

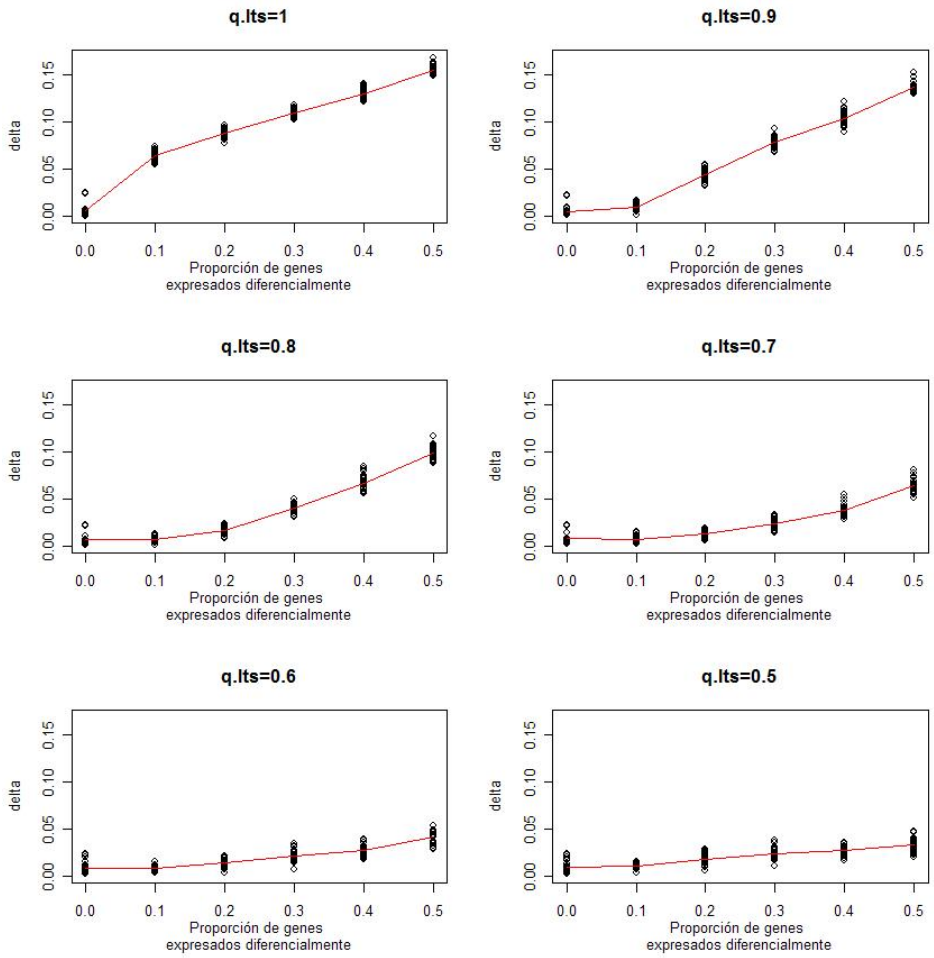

Figura 5.5: Para cada valor de p_{dif} se generaron 30 conjuntos de datos.

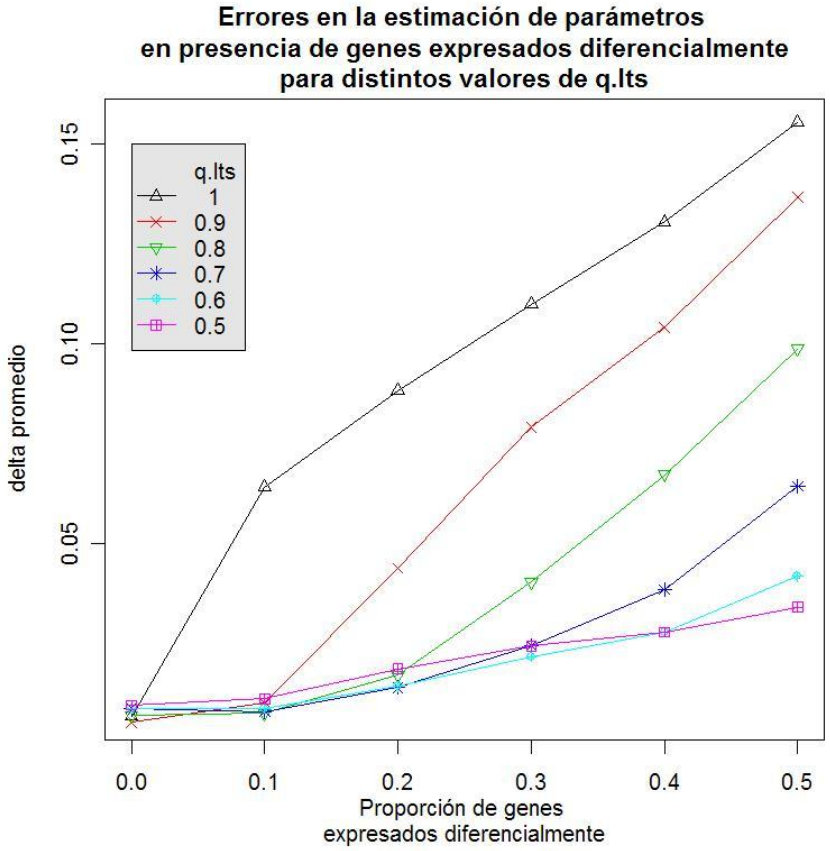

Figura 5.6: Comparación de las estimaciones con distintos valores de q_{lts}.

5.5. Nivel máximo de expresión diferencial

Para simular distintos valores del nivel máximo de expresión diferencial se generaron las intensidades de 8064 spots de 2 microarreglos con una proporción de 30 % de genes expresados diferencialmente. Los valores de z_{max} que se utilizaron son 2, 2.5, 3.5, 4.5, 5.5 y 6. Las estimaciones se realizaron con un valor de q_{lts} =0.7.

Los resultados obtenidos se muestran en la figura 5.7.

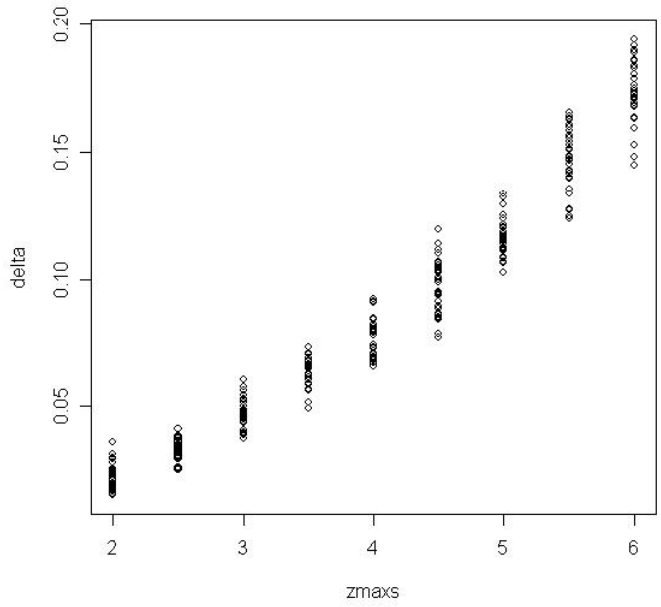

Figura 5.7: Para cada valor de z_{max} se generaron 30 conjuntos de datos.

Como era de esperar, a medida que aumenta el valor de z_{max} los errores en la estimación también aumentan.

Apéndice A

Distribución log-normal

Definición A.1 *Si X es una variable aleatoria tal que* $\ln(X) \sim N(\mu, \sigma^2)$ *entonces diremos que X tiene distribución* lognormal *y escribiremos* $X \sim LN(\mu, \sigma^2)$.

Proposición A.1 *Si* $X \sim LN(\mu, \sigma^2)$ *entonces la función de densidad de X es:*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\Big(\frac{\ln(x) - \mu}{\sigma}\Big)^2 \right) I_{(0,\infty)}(x)$$

Demostración. Como $Y = \ln(X) \sim N(\mu, \sigma^2)$ tenemos que:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\Big(\frac{y - \mu}{\sigma}\Big)^2 \right)$$

Entonces la función de distribución de X resulta:

$$F_X(x) = P\big(X \leq x\big) = P\big(e^Y \leq x\big) = \begin{cases} 0 & \text{si} \quad x \leq 0 \\ \\ P\big(Y \leq \ln(x)\big) & \text{si} \quad x > 0 \end{cases}$$

Es decir:

$$F_X(x) = \begin{cases} 0 & \text{si} \quad x \leq 0 \\ \\ \displaystyle\int_{-\infty}^{\ln(x)} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\Big(\frac{y - \mu}{\sigma}\Big)^2 \right) dy & \text{si} \quad x > 0 \end{cases}$$

Derivando esta última expresión obtenemos la función de densidad de X:

$$F'_X(x) = f_X(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\Big(\frac{\ln(x) - \mu}{\sigma}\Big)^2 \right) I_{(0,\infty)}(x). \quad \blacksquare$$

Proposición A.2 *Si* $X \sim LN(\mu, \sigma^2)$ *entonces el momento de orden* n *es:*

$$E(X^n) = e^{(n\mu + n^2\sigma^2/2)}$$

Demostración.

$$E(X^n) = \int_0^{+\infty} \frac{x^n}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\Big(\frac{\ln(x) - \mu}{\sigma}\Big)^2 \right) dx$$

Haciendo el cambio de variable $t = \ln(x)$ nos queda:

$$E(X^n) = \int_{-\infty}^{+\infty} \frac{\exp(nt)}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\Big(\frac{t - \mu}{\sigma}\Big)^2 \right) dt$$

$$E(X^n) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\Big(t^2 - 2t(\mu + n\sigma^2) + \mu^2\Big) \right) dt$$

$$E(X^n) = \exp\big(n\mu + n^2\sigma^2/2\big) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\Big(t - (\mu + n\sigma^2)\Big)^2 \right) dt$$

El integrando en la última expresión es la función de densidad de una variable aleatoria con distribución normal de media $\mu + n\sigma^2$ y desvío σ, por lo tanto el valor de la integral es 1 y la proposición queda demostrada. ∎

Corolario A.1 *Si* $X \sim LN(\mu, \sigma^2)$ *entonces:*

$$E(X) = e^{(\mu + \sigma^2/2)} \quad \text{y} \quad \mathrm{Var}(X) = e^{(2\mu + \sigma^2)}\Big(e^{\sigma^2} - 1\Big)$$

Demostración. Usando la proposición anterior con $n = 1, 2$ y teniendo en cuenta que $\mathrm{Var}(X) = E(X^2) - E^2(X)$, el resultado es inmediato. ∎

Apéndice B

Implementación del método

A continuación se muestra el código de la implementación del método en lenguaje R y una aplicación al dataset `kidney` incluído en el paquete `vsn`.

```
library(vsn) #Carga el paquete vsn

##Ejemplo "Kidney"
data(kidney)
A<-getIntensityMatrix (kidney)
##matriz con las intensidades (n filas correspondientes a n genes
##y d columnas correspondientes a d arreglos)
n<-dim(A)[1]
d<-dim(A)[2]

####Funciones de los parámetros "a" y "b" ###############
####
#### Los parámetros "b_i" del texto aquí los cambiamos por
#### exp(b_i) para asegurar que sean positivos
############     Transformacion afin       ###############
afin<-function(a,b){
B<-matrix(NA, dim(A)[1] , dim(A)[2] )
for(i in 1:d){
  B[,i]<-(A[,i]-a[i])/exp(b[i])
##Con exp me aseguro que el denominador es distinto de cero
  }
return(B)
 }

#############################################################

############     Funcion h (asinh)       ##################
h<-function(a,b){
```

```
return( asinh(afin(a,b)) )
}

############################################################

############   mu sombreros (en funcion de "a" y"b")  #####
mu<-function(a,b){
return( apply(h(a,b),1,mean) ) #es el promedio de hi(yki)
}  #tomado por fila

############################################################

############         Residuos
res<-function(a,b){
return( h(a,b)-mu(a,b) )
}

############################################################

############   Derivadas de hi respecto a yki
dh<-function(a,b){
#matriz donde guardo los resultados
deriv<-matrix(NA , dim(A)[1] , dim(A)[2] )
#aplico la derivada a cada columna con sus resp. parametros
for(i in 1:d){
deriv[,i]<-1/sqrt( (A[,i]-a[i])^2 + exp(b[i])^2 )
}
return(deriv)
}

############################################################

#############  Profile log-likelihood     #################
pll<-function(a,b){
return( -n*d/2*log(sum(res(a,b)^2))+sum(log(dh(a,b))) )
}

############################################################

############  Función a minimizar   ######################
f<-function(x){   # x es un vector donde las primeras
a<-x[1:d] # d coordenadas corresponden a parámetros "a"
b<-x[(d+1):(2*d)] # y las restantes corresponden a parámetros "b"
```

```
return(-pll(a,b))
}


###### Cuadrados de residuos
rk<-function(a,b){
return( apply(res(a,b)^2,1,sum) )
}

############################################################

####### Iteraciones

estimacion_robusta<-function(Datos = A , par_iniciales,
 iter = 10 , qlts= 0.9){
T<-vector("numeric",n)
A<-Datos
a<-par_iniciales[1:d]
b<-par_iniciales[(d+1):(2*d)]

#iteraciones
for(j in 1:iter){

for(i in 1:10){  ##Partición del conjunto 1,...,n
Ti<-(1:n)[(0.1*(i-1)*n)<sort.list(sort.list(mu(a,b))) &
 sort.list(sort.list(mu(a,b)))<=(0.1*n*i)]
T[Ti]<-i
}


T1<-(1:n)[T==1]
Q1<-quantile(rk(a,b)[T1],qlts)
ind1<-T1[rk(a,b)[T1]<=Q1] #indices

T2<-(1:n)[T==2]
Q2<-quantile(rk(a,b)[T2],qlts)
ind2<-T2[rk(a,b)[T2]<=Q2]

T3<-(1:n)[T==3]
Q3<-quantile(rk(a,b)[T3],qlts)
ind3<-T3[rk(a,b)[T3]<=Q3]

T4<-(1:n)[T==4]
```

```
Q4<-quantile(rk(a,b)[T4],qlts)
ind4<-T4[rk(a,b)[T4]<=Q4]

T5<-(1:n)[T==5]
Q5<-quantile(rk(a,b)[T5],qlts)
ind5<-T5[rk(a,b)[T5]<=Q5]

T6<-(1:n)[T==6]
Q6<-quantile(rk(a,b)[T6],qlts)
ind6<-T6[rk(a,b)[T6]<=Q6]

T7<-(1:n)[T==7]
Q7<-quantile(rk(a,b)[T7],qlts)
ind7<-T7[rk(a,b)[T7]<=Q7]

T8<-(1:n)[T==8]
Q8<-quantile(rk(a,b)[T8],qlts)
ind8<-T8[rk(a,b)[T8]<=Q8]

T9<-(1:n)[T==9]
Q9<-quantile(rk(a,b)[T9],qlts)
ind9<-T9[rk(a,b)[T9]<=Q9]

T10<-(1:n)[T==10]
Q10<-quantile(rk(a,b)[T10],qlts)
ind10<-T10[rk(a,b)[T10]<=Q10]

## nuevo conjunto K
K<-c(ind1,ind2,ind3,ind4,ind5,ind6,ind7,ind8,ind9,ind10)

A<-Datos[K,] #nuevos datos para la iteracion
opt<-optim(x,f,method="L-BFGS-B")

a<-(opt$par)[1:d]
b<-(opt$par)[(d+1):(2*d)]
x<-c(a,b)
A<-Datos
}
return( x )
}
```

B.1. Ejemplo

Aplicación a `kidney`:

Datos de intensidades de un slide de ADNc con dos muestras de tejidos adyacentes obtenidos de una nefrectomía (extirpación quirúrgica del riñón).

El chip fue producido en el año 2001 por Holger Sueltmann en la División de Análisis Molecular del Genoma del Centro Alemán de Investigación del Cáncer en Heidelberg.

```
> ####  Parámetros iniciales aleatorios
> a<-runif(d)
> b<-runif(d)
> x<-c(a,b)

> ######## Optimización
> opt<-optim(x,f,method="L-BFGS-B")

> parametros<-estimacion_robusta(A,par_iniciales = opt$par,
+  iter = 10 , qlts= 0.9)

> par_vsn<-vsn2(A,lts.quantile=0.9)
vsn: 8704 x 2 matrix (1 stratum). 100% done.
> ##Datos crudos
> canal1<-A[,1]
> canal2<-A[,2]
> ##Datos transformados con mi implementación del método
> v<-h(parametros[1:2],parametros[3:4])[,1]
> r<-h(parametros[1:2],parametros[3:4])[,2]
> ##Datos transformados por función vsn2 del paquete vsn
> v.vsn<-par_vsn@hx[,1] #datos del canal 1 transformados por vsn
> r.vsn<-par_vsn@hx[,2] #datos del canal 2 transformados por vsn
> par(mfrow=c(3,1))
> plot(rank(canal1+canal2),canal1-canal2,ylim=c(-1000,1000),pch=".")
> plot(rank(v+r),r-v,ylim=c(-4,4),pch=".")
> plot(rank(v.vsn+r.vsn),r.vsn-v.vsn,ylim=c(-4,4),pch=".")
```

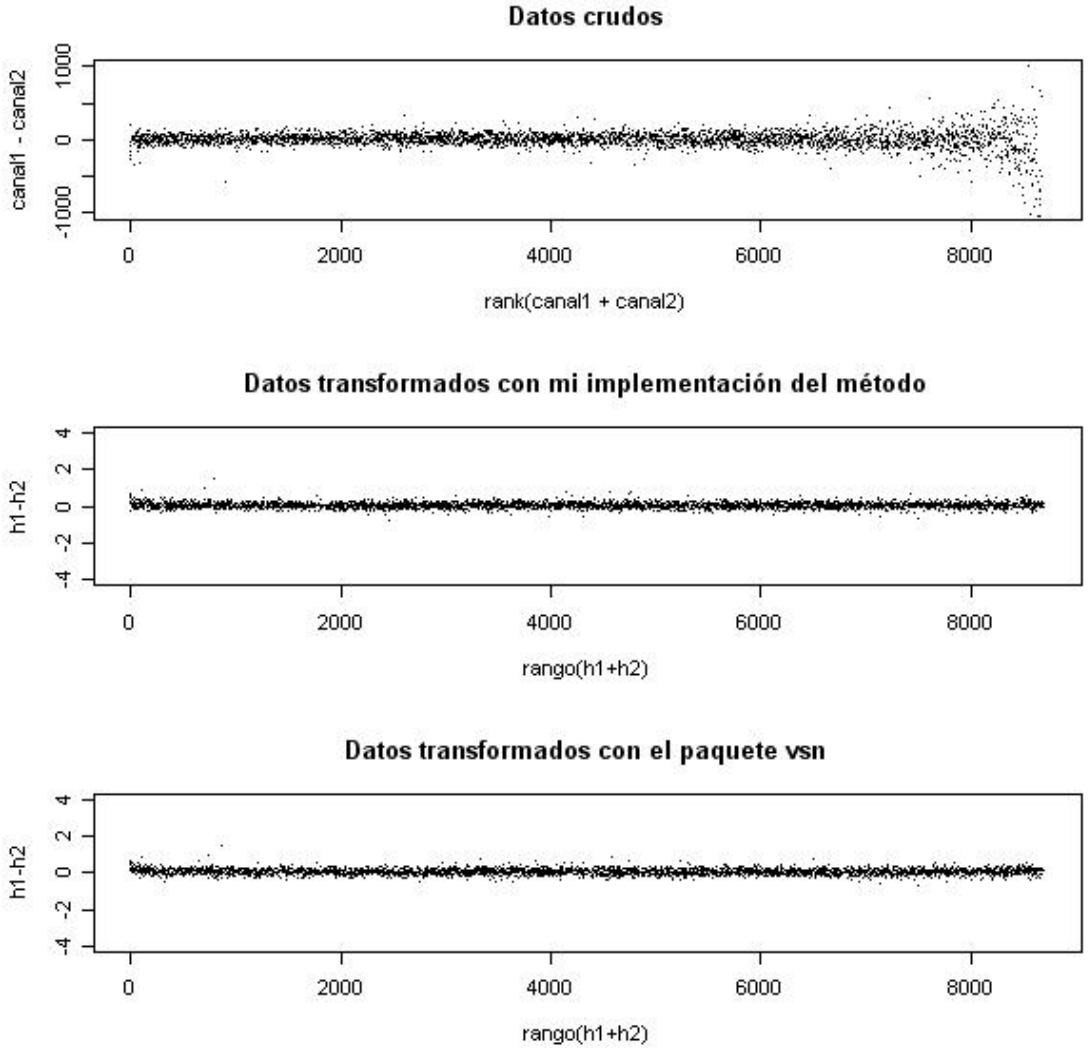

Figura B.1: La figura de arriba muestra en el eje y la diferencia de intensidades entre ambos canales y en el eje x el rango de su suma. El gráfico del medio muestra la diferencia de los datos transformados con nuestra implementación versus el rango de $h(y_{k1}) + h(y_{k2})$. El último gráfico muestra la diferencia de los datos transformados con el paquete `vsn` versus el rango de $h(y_{k1}) + h(y_{k2})$.

Apéndice C

Código para la simulación de intensidades de microarreglos.

```
##Funcion similar a sagmbSimulateData del paquete vsn
##pero con la posibilidad de cambiar zmax
SimulacionDatos<-function (n = 8064, d = 2, de = 0, up = 0.5, zmax=2) {
    sigsq = 0.04
#Genero expresiones en escala transformada
    mu = asinh(1/rgamma(n, shape = 1, scale = 1))
#Genero los coeficientes
    coeficientes = array(runif(d * 2, min = -2, max = +2), dim = c(d , 2))
#Decido cuales son los genes expresados diferencialmente
    is.de <- (runif(n) < de)
#Matriz de datos transformados
#primer columna generada segun el modelo,
#a las demás les agrego un término para la expresión diferencial
    hy <- matrix(as.numeric(NA), nrow = n, ncol = d)
    hy[, 1] <- mu + rnorm(n, sd = sqrt(sigsq))
    for (j in 2:d) {
        s <- 2 * as.numeric(runif(n) < up) - 1
        hy[, j] <- mu + as.numeric(is.de) * s * runif(n, min = 0,
            max = zmax) + rnorm(n, sd = sqrt(sigsq))
     }
    offs <- coeficientes[ , 1]
    facs <- coeficientes[ , 2]
#Genero datos "crudos"
    y = (sinh(hy) - offs)/exp(facs)
    return(list(y = y, hy = hy, mu = mu, sigsq = sigsq,
    coeficientes = coeficientes,
        is.de = is.de))
}
```

Bibliografía

[1] Xiangqin Cui, M. Kathleen Kerr, and Gary A. Churchill. *Data transformations for cDNA microarray data.* Technical report, The Jackson Laboratory, http://www.jax.org/research/churchill, 2002.

[2] Sandrine Dudoit, Yee Hwa Yang, Terence P. Speed, and Matthew J. Callow. *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.* Statistica Sinica, 2002.

[3] Blythe P. Durbin, Johanna S. Hardin, Douglas M. Hawkins, and David M. Rocke. *A variance-stabilizing transformation for gene-expression microarray data.* Bioinformatics, 18 Suppl. 1:S105-S110, 2002. ISMB 2002.

[4] W. Huber. *Vignette: Robust calibration and variance stabilization with vsn, 2002* The bioconductor project, http://www.bioconductor.org

[5] Wolfgang Huber, Anja von Heydebreck, Holger Sültmann, Annmarie Poustka, and Martin Vingron. *Variance stabilization applied to microarray data calibration and to the quantification of differential expression.* Bioinformatics, 2002.

[6] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka y M. Vingron. *Parameter estimation for the calibration and variance stabilization of microarray data.* Statistical Applications in Genetics and Molecular Biology, Vol. 2: No. 1, Article 3, 2003. http://www.bepress.com/sagmb/vol2/iss1/art3

[7] W. Huber, A. von Heydebreck y M. Vingron. *Error models for microarray intensities.* Bioinformatics, 2004.

[8] T. Ideker, V. Thorsson, A.F. Siegel, and L.E. Hood. *Testing for differentially expressed genes by maximum-likelihood analysis of microarray data.* Journal of Computational Biology, 2000.

[9] Ross Ihaka and Robert Gentleman. *R: A language for data analysis and graphics.* Journal of Computational and Graphical Statistics, 1996. http://www.bioconductor.org.

[10] Peter Munson. *A consistency test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations.* Genelogic Workshop on Low Level Analysis of Affymetrix Genechip data, http://stat-www.berkeley.edu/users/terry/zarray/Affy/GL_Workshop/genelogic2001.html, 2001.

[11] Susan A. Murphy and A. W. van der Vaart. *On profile likelihood.* Journal of the American Statistical Association, 2000.

[12] M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner y K.W. Tsui. *On differential variability of expression ratios: improving statistical inference abaout gene expression changes from microarray data.* Journal of Computational Biology, 8(1):37-52, 2001.

[13] David M. Rocke y Blythe Durbin. *A model for measurement error for gene expression analysis.* Journal of Computational Biology, 8:557-569, 2001.

[14] David M. Rocke y Blythe Durbin. *Approximate variance-stabilizing transformations for gene-expression microarray data.* Bioinformatics, 2002.

[15] Peter J. Rousseeuw *Multivariate estimation with high breakdown point.* Mathematical Statistics and Applications, Vol. B, Reidel, Dordrecht, The Netherlands, 283-297, 1985.

[16] Peter J. Rousseeuw y Annick M. Leroy. *Robust Regression and Outlier Detection.* John Wiley and Sons, 1987.

[17] Peter J. Rousseeuw y Katrien van Driessen. *Computing LTS regression for large data sets.* Technical report, Antwerp Group on Robust & Applied Statistics, 1999.