EVALUACIÓN DE MÉTODOS DE IMPUTACIÓN PARA DATOS DE EXPRESIÓN GENÉTICA

Por

Sindy Díaz Hernández

Tesis sometida en cumplimiento parcial de los requerimientos para el grado de

MAESTRO EN CIENCIAS

en

MATEMÁTICAS(ESTADÍSTICA)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ

Diciembre, 2007

Aprobada por:

Edgar Acuña Fernández, Ph.D Presidente, Comite Graduado	Fecha
Tokuji Saito, Ph.D Miembro, Comite Graduado	Fecha
Julio C Quintana Díaz, Ph.D Miembro, Comite Graduado	Fecha
Mercedes Ferrer, M.E. Representante de Estudios Graduados	Fecha
Julio C. Quintana Díaz, Ph.D Director del Departmento	Fecha

Abstract of Thesis Presented to the Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master in Mathematics(Estadística)

EVALUATION OF IMPUTATION METHODS FOR GENE EXPRESSION DATA

By

Sindy Díaz Hernández

December 2007

Chair: Edgar Acuña
Major Department: Mathematical Sciences

The technology of microarrays introduced in the middle of the nineties allow the analysis of the gene expression levels of thousands of genes simultaneously. The identification of genes with an expression level very different to the others is crucial to identify the possible causes of certain illness and it permits to create a treatment for its cure. Due to many reasons related to the microarray technology is common to find missing values in the gene expression matrix. Other characteristic of the gene expression matrix is its high dimensionality. That is, it has a very large number of columns representing the genes, and few rows representing the arrays that are coming from samples taken to patients. The imputation of missing values is absolutely necessary for the application of several tasks of Data Mining and Knowledge Discovery in Bioinformatics. One of there tasks is the identification of differentially expressed genes. There are several imputation methods for this kind of data. Unfortunately, most of them have been tested in one or two datasets, and until now there is not a general evaluation of the imputation methods. In this thesis, a comparison of five methods for imputation of gene expression data is carried out.

Six well known gene expression data related to cancer are used. The comparison is done using two criterion: the normalized root mean squared error (NRMSE) and the percentage of differential expressed genes lost after the imputation. Finally, a recommendation on the use of the imputation methods is given, and an explanation of such recommendation is discussed.

Resumen de Tesis Presentado a Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestro en Ciencias

EVALUACIÓN DE MÉTODOS DE IMPUTACIÓN PARA DATOS DE EXPRESIÓN GENÉTICA

Por

Sindy Díaz Hernández

Diciembre 2007

Consejero: Edgar Acuña
Departamento: Ciencias Matemáticas

La tecnología de microarreglos, introducida en la mitad de la década de los 90, permite que se pueda analizar simultáneamente los niveles de expresión genética de miles de genes. La identificación de los genes con un nivel de expresión muy diferente a los otros genes es crucial en medicina para determinar las posibles causas de una enfermedad y poder establecer un tratamiento para su cura. Debido a varias razones relacionadas a la tecnología del microarreglo es común que haya valores faltantes en la matriz de expresión genética. Otra característica peculiar de la matriz de expresión genética es su alta dimensionalidad. Es decir, tiene un gran número de columnas, representando los genes, y pocas filas, representando los arreglos que resultan de muestras tomadas en pacientes. La imputación de los valores faltantes se hace absolutamente necesaria para la aplicación de tareas de minería de datos y descubrimiento del conocimiento en Bioinformática. Una de estas tareas es la identificación de genes diferencialmente expresados. Hay varios métodos de imputación especializados para este tipos de datos. Desafortunadamente, muchos de estos métodos han sido probados en uno o dos conjuntos de datos y hasta ahora no se ha hecho

una evaluación mas general de los mismos. En esta tesis se compararán experimentalmente cinco métodos de imputación de datos faltantes en matrices de expresión genética usando seis conjuntos de datos de expresión genética, todos ellos relacionados con cáncer y bien conocidos en la literatura genómica.

Para ello usamos dos criterios: la raíz cuadrada del cuadrado medio del error normalizada (NRMSE) y el porcentaje de genes diferencialmente expresados que dejan de ser identificados después de hacer la imputación. Al final se hace una recomendación acerca del uso de los métodos de imputación y se trata de justificar dicha recomendación.

A mi familia, en especial a dos seres que son muestra de inmenso amor por sus hijos, Julio e Imilida...

Para los que siempre seré su hermanita preferida.....Julito y Carlitos.

Aquel que ha estado a mi lado, me ha dado muestra de comprensión y amor, el que me ha acompañado a lo largo de este viaje .....Luis R Fuentes Castilla.

Todo lo puedo en Dios, esta tesis es un ejemplo de ello, gracias Padre.....

AGRADECIMIENTOS

A ese ser inexplicable de inmenso amor y que he sido testigo de su poder, a lo largo de toda mi vida...Jehová.

A mi consejero Doctor Edgar Acuña, presidente de mi comité, no existen palabras para agradecer no sólo los consejos académicos, también los personales.

A mi comité graduado Dr. Julio Quintana, Dr. Tokuji Saito y profesora Mercedes Ferrer, por sus aportes a tiempo y puntuales.

Al Departamento de Ciencias Matemáticas, sus profesores y secretarias, siempre un gran apoyo para que logremos nuestras metas.

A la profesora Damarís Santana, me enseñó más de lo que cree, no sólo lecciones académicas sino de vida.

Al Dr. Luis Daza por su ayuda en la elaboración de esta tesis.

A mi familia, a ese angelito, que seguramente se merece el paraíso, mi abuelita Cruz María, siempre estarás en mi corazón.

A Lucho, muestra de amor, compromiso y dedicación.

A mis amigos Roberto y Charlie, significados de una buena amistad, con altas y bajas como debe ser.

A las personas que hicieron de esta etapa, más dinámica y chévere: Trilce, Cata, Karen y a todos los compañeros de la oficina de estudiantes graduados 302 A.

A la oficina de Investigación Naval (ONR) por el apoyo económico recibido a través de Grant N0014-03-0359.

Al Departamento de Defensa por el apoyo económico parcial recibido a través del Grant N0014-06-1-0555.

A los miembros de CASTLE, actuales y anteriores, por compartir conmigo esta etapa.

A San Fernando, Cartagena y Colombia, se pueden alcanzar metas, ustedes son pieza importante de esta.

A todo aquél que siempre me brindó una sonrisa, me alentó a seguir y que de una u otra forma me ayudó a ser mejor persona.....

TABLA DE CONTENIDO

Página

LISTA DE TABLAS

Tabla Página

LISTA DE FIGURAS

Figura Página

LISTA DE ABREVIATURAS

ADN	Ácido Desoxirribonucleico.
ARN	Ácido Ribonucleico .
KNN	K- Vecinos más Cercanos(k-nearest neighbor).
SVD	Descomposición en Valores Singulares(Singular Value Decomposition).
LLS	Mínimos Cuadrados Locales(Local Least Squares).
gdf	Genes Diferencialmente Expresados
gdfp	Genes Diferencialmente Perdidos
PM	Perfect Match Probes
MM	Mismatch Probes

Capítulo 1
INTRODUCCIÓN

"Los avances en la computación y en las matemáticas son tan importantes y van paralelos, a la biología" J.Craig Venter

1.1. Motivación

La tecnología de microarreglos que fue introducida en la mitad de la década de los 90, permite que se pueda analizar simultáneamente los niveles de expresión genética de miles de genes. La identificación de los genes con un nivel de expresión muy diferente a los otros genes es crucial en medicina para determinar las posibles causas de una enfermedad y poder establecer un tratamiento para su cura. La matriz de expresión genética se obtiene analizando las imágenes del microarreglo, existen programas de computadora especializados para hacer esta tarea. Debido a muchas razones relacionadas a la tecnología del microarreglo es común que haya valores faltantes en la matriz de expresión genética. Otra característica peculiar de la matriz de expresión genética es su alta dimensionalidad. Es decir, tiene un gran número de columnas, representando los genes, y pocas filas, representando los arreglos que resultan de muestras tomadas en pacientes. La imputación de los valores faltantes se hace absolutamente necesaria para la aplicación de tareas de minería de datos y descubrimiento del conocimiento. Una de estas tareas es la identificación de genes diferencialmente expresados. Por esta razón desde el 2001, se comenzó a investigar en métodos para imputar eficientemente valores faltantes en matrices de expresión genética.

Desafortunadamente, muchos de estos métodos han sido probados en pocos conjuntos de datos y hasta ahora no se ha hecho una evaluación general de los mismos. En esta tesis se compararán experimentalmente 5 métodos de imputación de datos faltantes en matrices de expresión genética usando 6 conjuntos de datos bien conocidos en la literatura genómica. Para ello usamos dos criterios: la raíz cuadrada del cuadrado medio del error normalizada y el porcentaje de genes diferencialmente expresados, que dejan de ser identificados después de hacer la imputación.

1.2. Objetivo de la Tesis

El objetivo principal de esta tesis es hacer una comparación de 5 métodos de imputación de datos faltantes en matrices de expresión genética. La comparación se hace usando seis conjuntos de datos de expresión genética, todos ellos relacionados a cáncer y bien conocidos en la literatura. Al final se hace una recomendación acerca de los métodos que han tenido mejor rendimiento y se trata de explicar el resultado. Se han usado dos criterios de evaluación, uno de ellos, la raíz cuadrada del cuadrado medio del error normalizada, la cual ha sido bastante usada en la literatura, y el segundo criterio es el porcentaje de genes diferencialmente expresados, perdidos después de hacer la imputación. Este segundo criterio mide un efecto posterior de hacer la imputación en una tarea primordial en el análisis de datos de microarreglos.

1.3. Resumen de la Tesis

El capítulo 1 es Introducción, allí explicamos la motivación de esta tesis y los objetivos de la misma. También hacemos un breve resumen de los otros capítulos de la tesis. En el segundo capítulo hacemos un repaso de varios conceptos relacionados a biología molecular y que son necesarios para entender en donde se enmarca esta tesis. En el tercer capítulo, explicamos los dos tipos de microarreglos que existen.

En el cuarto capítulo se discuten algunas característica de los datos provenientes de microarreglos.
En particular, el concepto de datos faltantes que es el motivo principal de esta tesis.
En el quinto capítulo se discuten detalladamente los 5 métodos de imputación para datos de microarreglos que han sido comparados en esta tesis. Estos son: el método basado en KNN (KNNImpute), el método basado en la descomposición de una matriz en sus valores singulares (SVDImpute), el método basado en componentes principales probabilísticos (PPCA), el método basado en componentes principales bayesianos (BPCA) y el método basado en mínimos cuadrados locales(LLSImpute).

En el capítulo 6 se describe la metodología, empezando con las características de los 6 conjuntos de expresión genética usados en la tesis. Los dos criterios usados para la evaluación de los métodos de imputación, la raíz cuadrada del error cuadrático medio normalizada y el porcentaje de genes diferencialmente expresados después de imputar(PGDP). Finalmente en el capítulo 7 se detallan las conclusiones a las que se llega en esta tesis así como los posibles problemas que se podrían trabajar en un futuro.

Capítulo 2
UNA MIRADA A LA BIOLOGÍA MOLECULAR

"Por primera vez en la historia, la humanidad puede leer su propio genoma, su libro de la vida. Este libro no es como ningún otro, ya que al leerlo desveláremos una vista cada vez mas amplia de nosotros mismo."-Francis S. Collins

Director.NHGRI

El objetivo principal de este capítulo es esbozar términos de biología molecular que nos permitan plantear lo que es la expresión de un gen y su importancia.

2.1. Célula

La célula es la unidad básica esencial y fundamental de todo ser vivo. Existen dos tipos de células, las eucariotas y procariotas. Las células procariotas se supone que fueron la primera forma de ser vivo, éstas no poseen núcleo, tienen ribosoma y ADN, tema del que hablaremos más adelante, el cual se encuentra disperso en el citoplasma. Las células eucariotas son más complejas que las primeras, debido a que poseen varios organelos, cada uno con funciones específicas. Por ejemplo, poseen núcleo en el cual se halla gran parte del ADN; algunos de los organelo importantes que poseen las células eucariotas son: mitocondrias y vacuolas , entre otras. Si hablamos de las células vegetales, uno de los organelo que sólo ellas tienen son los cloroplastos.

Toda célula cumple una función específica, para que esto ocurra todo debe marchar bien dentro de ella, es decir debe haber una buena comunicación entre sus organelos.

Observemos la siguiente imagen que representa la célula:

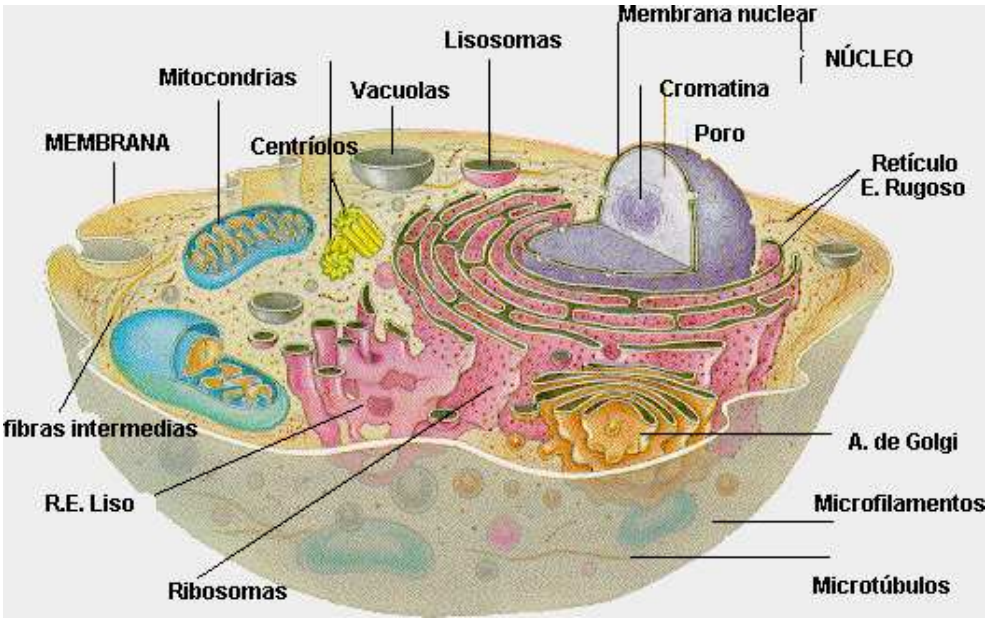

Figura 2–1: Representación de la célula, Tomada de:www.juntadeandalucia.es

Para poder cumplir con nuestros objetivos definamos primeramente lo que son núcleo y cromosomas.

2.1.1. Núcleo y Cromosomas

La célula eucariota tiene una membrana citoplasmática, la cual sirve como frontera entre la célula y su entorno, ya que es la encargada de seleccionar qué moléculas o sustancia puede entrar a la célula. Dentro de ella se halla el citoplasma, allí están los distintos organelo de la célula, siendo el más destacable es el núcleo.

El núcleo es una de las partes más importante de la célula ya que lleva a cabo muchas funciones además coordinar otras. En el núcleo se hallan los cromosomas, dentro de

los cuales están los genes, es por ello que se dice que en el núcleo es donde se halla la mayor cantidad del material genético. Veremos más adelante como los mismos orgánelos que se encargan de la herencia son los que se encargan de coordinar los procesos de la célula. En [25], podemos encontrar mucha literatura donde se refieren a la célula.

Los cromosomas están formados por ADN, están en pares. En los seres humanos hay 23 pares de cromosomas, en la cebolla 8, la mosca de la fruta tiene 4 pares y en las papas hay 24 pares. Por lo tanto el número de cromosomas no tiene relación con lo especializado que puede ser un organismo. Como los cromosomas están compuestos de proteínas y ADN, se creía que las primeras eran las encargadas de la herencia, hasta que en la década de los años cuarenta los genetistas, George Wells Breadle y Edward Lawrie por medio de un experimento sobre cepas, se dieron cuenta que en realidad la herencia era función del ADN. Pasemos pues a hablar más detalladamente de este último.

2.2. Ácidos Nucleicos

En la célula existen dos tipos de ácidos nucleicos, estos son el ácido desoxirribonucleico(ADN) y el ribonucleico(ARN). Estos constituyen parte del material genético y contribuyen en la fabricación de proteínas. A continuación detallaremos sus características.

2.2.1. Ácido Desoxirribonucleico(ADN)

Una molécula de ADN esta formada por varias moléculas pequeñas o monómeros, que son los nucleótidos, por lo tanto esta molécula es un polímero o macromolécula.

Los nucleótidos a su vez, están formados por un grupo de fosfato, bases nitrogenadas y azúcares. Las bases nitrogenadas que forman el *ADN* son cuatro: Adenina(A), Guanina(G), Citosina(C) y Tiamina(T). Estas a su vez se dividen en dos grupos, las dos primeras son púricas y las dos ultimas pirimídicas.

La estructura del *ADN* según descrito en el artículo *Molecular Structure of Nucleic Acids*, [38], es una molécula que tiene dos franjas que se entrecruzan, pareciéndose a una escalera o doble hélice, cada base nitrogenada en una de las franjas se aparean con otra base de la otra franja. Esta unión se da hacia dentro y solo lo hacen con esta condición: A con T y G con C. La gráfica 2–2 es una de las mas representativas que se hallan sobre *ADN* y se encuentra en un tutorial sobre conceptos básicos de biología molecular de David Secko, [33].

Una de las funciones principales del *ADN* es poder participar en la reproducción de un organismo con la misma características que el original. Esto sucede porque esta molécula tiene las propiedades de mutación y replicación.

2.2.2. Ácido ribonucleico *(ARN)*

El ácido ribonucleico tiene algunas similitudes con el *ADN* pero lo que los diferencia es lo siguiente:

- Es una sola cadena, de bases nitrogenadas, no es doble hélice como *ADN*.
- El azúcar es ribosa
- A pesar que tiene cuatro bases nitrogenadas, la diferencia es que tiene Uracilo(U) en lugar de T.

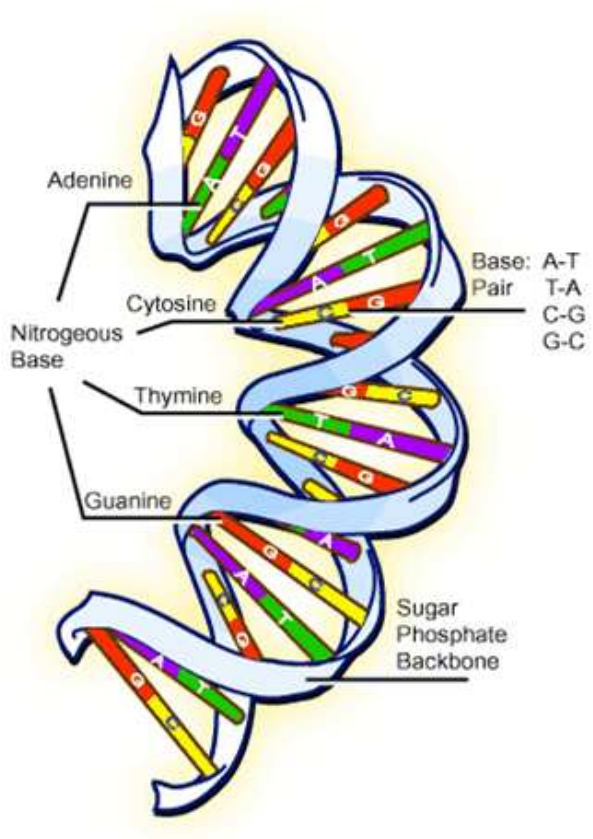

Figura 2–2: Representacion de la doble cadena de ADN , Tomada de: www.scq.ubc.ca

Hay varias clases de ADN dependiendo en donde estén situadas y de sus funciones. Esto también se aplica al ARN, pero el que juega un papel importante en la fabricación de proteína es el ARN mensajero, $mARN$.

Otros conceptos que se utilizan en esta tesis son:

Oligonucleótidos: Son secuencias cortas de nucleótidos ya sea ADN o ARN conformadas esencialmente por 20 o menos bases.

Hibridación: El proceso químico por el cual dos cadenas complementarias de ácido nucleótido se entrelazan[20].

Desnaturalización: Es el proceso inverso de hibridación.

2.3. Proteínas y Aminoácidos

Existen distintas clases de aminoácidos, sin embargo los que nos interesan son los aminoácidos proteicos, de los cuales se conocen mas de 100. La estructura en la mayoría de ellos es la misma:

- Grupo carboxilo
- Grupo amino
- Carbono α
- Cadena lateral

Los aminoácidos se clasifican en dos grupos: los esenciales- no los produce el organismo y por lo tanto deben ser ingeridos en la dieta- y los no esenciales-que pueden ser producidos por el cuerpo-. Los aminoácidos esenciales son 8, aunque hay dos adicionales que son especiales en niños.

Las proteínas son macromoléculas formadas por una secuencia de aminoácidos, por lo cual son llamadas polímeros. Existen 20 aminoácidos distintos que forman las proteínas, estos se les llama aminoácidos proteicos. Cabe señalar que para formar una sola proteína se necesitan más de cien aminoácidos.

Las proteínas son esenciales para la vida, pues cumplen funciones tales como: regulación de hormonas, cicatrización, manejo del dolor, están presentes en el crecimiento de los músculos y la transmisión de impulso nervioso, entre otros. Hay proteínas que pueden ser fabricadas por el mismo organismo, la pregunta sería ¿Cómo lo hace?.

2.4. Genes

Gregor Mendel, monje y naturalista austriaco, trabajó en plantas para formular leyes de la herencia, habló sobre factores (conocidos actualmente como genes) recesivos y

dominantes. Su trabajo estaba más inspirado hacia el fenotipo. Aunque fue ignorado en su tiempo, varios años después fue evidente la importancia de las investigaciones que él había realizado.

Podríamos destacar que una de las funciones que más se conoce sobre los genes es la herencia. Aunque gen se conoce como parte de ADN que tiene la información suficiente para la fabricación de una proteína, esta es una de sus funciones en la cual radica gran importancia. Podemos establecer varias preguntas pertinentes a la información que ha estado desarrollando hasta ahora:

1. ¿Cómo se fabrican las proteínas?
2. ¿Qué sucede si un gen da indicaciones para producir más o menos proteínas de las que la célula necesita?
3. ¿Cada célula tendrá secuencias de ADN distintas?
4. ¿Cuántos genes tiene el cuerpo humano?
5. ¿Todos los seres humanos tendremos el mismo código genético?

A estas interrogantes se le dará respuesta en las secciones posteriores.

2.5. Genoma Humano

La importancia que han adquirido términos y conceptos relacionados con la biología molecular, así como la cotidianidad de éstos va en aumento. Sólo hablar de clonación hace unos años era del manejo de expertos y la esperanza de múltiples investigadores. Aun cuando anunciaron que habían logrado leer el genoma humano, muy pocos sabíamos el significado y los beneficios que esto traería, mucho menos que la relación directa que existe entre la biología molecular y la medicina, se haría mas fuerte.

Se entiende por genoma como todo material genético de un organismo, en nuestro caso hablamos de las células eucariotas, hablamos del material genético que se halla

en el núcleo, es por lo tanto, el conjunto de todos los genes contenidos en los cromosomas.

A finales del siglo pasado y comienzo de éste se desarrolló un proyecto científico internacional, en el cual se pretendía hallar la secuencia de cada una de las bases o nucleótidos, identificar cada uno de los genes, y de ser posible hallar las funciones de cada uno.El tiempo que le fue asignado eran 15 años, con un presupuesto de tres mil millones de dólares, sin embargo se terminó en el 2001, dos años antes de lo esperado. Aunque el punto final se dio en el 2003. Ese año se cumplían 5 décadas de haber descubierto la estructura de doble hélice de parte de Watson y Crick [39].

Veamos entonces el Genoma Humano en cifras:

- Tenemos 23 pares de cromosomas(1956 por Albert Levan y Joe Him Tijo)[40]
- Si estiraramos el ADN tenemos que es aproximadamente de dos metros
- Llamemos al número de bases pares bp. Entonces hay 3000 millones de bps.
- El 99.9 % del ADN de dos seres humanos es idéntico.
- Se supone que existen entre 20,000 a 30,000 genes, mucho menos de lo que se esperaba.

El Genoma Humano, no es el único que se ha descifrado, existen otros organismos a los cuales se les conoce su genoma. Se ha encontrado que muchos guardan altos porcentajes de similaridad.

Una de las cláusulas del proyecto genoma humano, es que los resultados deben estar disponible de manera gratuita.

La interpretación del genoma humano abrió múltiples posibilidades a nivel médico y a otras aplicaciones , pero cada avance sugiere nuevos retos éticos, ya que el manejo de esta información se podría usar para crear armas u otro tipo de elementos no beneficiosos. Algunos avances que han sido de transcendencia internacional, son por ejemplo, el descubrimiento de las causas de determinadas enfermedades, como es el

caso del Alzheimer, o en la detección temprana de otras como es el caso del cáncer de mama.

Una de las metas de la era genómica, la cual estamos viviendo y que cada vez toma más auge, son los cambios de realizar diagnósticos médicos. Porque la mayoría de los diagnósticos de enfermedades actualmente se basan en los síntomas y causas. En un futuro, lo que se quiere es realizarlos con base en la información genética; no sólo para mirar las causas de determinada enfermedad sino el posible desarrollo de otra. Además, se espera modificar la farmacología. Ya que se aspira a personalizar los medicamentos y tratamientos, de acuerdo a los marcadores genéticos.

Pero lo inquietante es descubrir en que interfiere el 0.1 % de diferencia en el código genético que hay entre seres humanos. Ya que una evidencia que la similitud o diferencia entre códigos genéticos, es significativa. El código genético del mono macaco, tiene una similitud con el genoma humano del 95 %, otro es el del gorila tiene una similitud del 99 % con el ser humano.

Lo que si es seguro en palabras del genetista J.Craig Venter en una entrevista dada al periódico la Nación, el 10 de Junio del presente año[1], quién participó en la secuenciación y codificación del genoma humano, que haber secuenciado el genoma humano es solo el inicio. Hay que tratar de desarrollar el conocimiento adquirido. Un ejemplo, es que no sólo un gen influye en una enfermedad sino que es un conjunto de modificaciones genéticas. Para enfatizar esto, él asegura que hay evidencia en que 300 genes tienen implicaciones en la presión arterial. Además hizo énfasis en que se ha estudiado poco el genoma humano y se ha procurado en hallar la secuencia de otros organismos. Es decir, hay múltiples preguntas, esto solo es el inicio y hay mucho camino que recorrer.

2.6. Dogma de la Biología Molecular

En esta parte veremos como se conjugan todo lo visto para fabricar las proteínas, la magnífica ingeniería que existe a nivel celular.

El hallar la forma del *ADN* solo fue el comienzo para tratar de resolver unas cuantas preguntas, ya se sabía que el *ADN* se podía duplicar en dos moléculas mas y por lo tanto poder transmitir la información genética a las células hijas, es decir, que este debía guardar una cierta información o código genético, como razonaron Watson y Crick y basados en eso realizaron otra publicación en el *Nature*[39].

Al comprobar la relación que había entre los péptidos (que es cualquier molécula formada por la union de aminoácidos) y las bases nitrogenadas, además que los ribosomas "leen" de tres en tres bases nitrogenadas para fabricar aminoácidos, Crick lanzó una afirmación que no se pudo demostrar sino tiempo después, como los ribosomas se hallan fuera del núcleo, entonces debía haber una especie de mensajero que llevara la información desde los cromosomas hasta los ribosomas. A esta afirmación le llamó Dogma, término que todavía se usa, aunque ya se ha demostrado que sí es cierta, ese mensajero que Crick decía el candidato era el *ARN*, aunque no podía ser todo el *ARN*.

Entonces se comprobó que el *ADN* se traduce en el *ARN mensajero* y este lleva la información para la fabricación de proteínas, desde los cromosomas hasta los ribosomas. Dado que son porciones de *ADN* los cuales pueden llevar la información para la fabricación de proteína, entonces ellos son los genes, así, que cualquier modificación en los genes llevaría una diferenciación en la producción de las proteínas, pero no recíprocamente, en otras palabras los genes influyen en la síntesis de proteína mas no viceversa.

2.7. Expresión de un Gen

El mismo código genético o información genética se encuentra en todas las células del cuerpo. Es decir, tienen la misma información, sabemos que no todas cumplen la misma función dentro del cuerpo y su función esta ligada a las proteínas que en ellas están fabricadas. Todas tienen la misma cantidad de genes, pero no todas tienen activado los mismos genes, por lo anteriormente expuesto podríamos inferir que no en todas se transcribe la misma cantidad de ARNmensajero, cuando si ocurre decimos que los genes que lo forman están expresados.

El nivel de expresión de un gen es la cantidad de mARN transcrito en un momento dado, ya que por lo visto anteriormente esa cantidad es proporcional a la cantidad de proteínas fabricadas.

Pero muchas veces los genes no se expresan de la forma adecuada para que la células tenga las proteínas suficientes para poder realizar su función. Esto es, muchas veces se sobre expresan. Es decir, existe mayor número de copias de ARNmensajero del que debería haber o por el contrario no hay, esto se interpreta como que el gen esta altamente expresado o no esta expresado.
Existe gran información disponible sobre los conceptos vistos, en nuestro caso nos basamos en [2] y [43].

Como vimos tiene mucha transcendencia el poder medir la expresión de un gen, desde 1953 el crecimiento científico se intensificó en la biología molecular y a medida que pasa el tiempo así como crecen las respuestas.
También los interrogantes, uno de los mecanismo que es relevante para el estudio de

la expresión de genes o expresión génica como también se le llama, son los micro-arreglos, tema del que hablaremos en el capítulo siguiente.

La siguiente gráfica representa el dogma de la biología molecular, observemos:

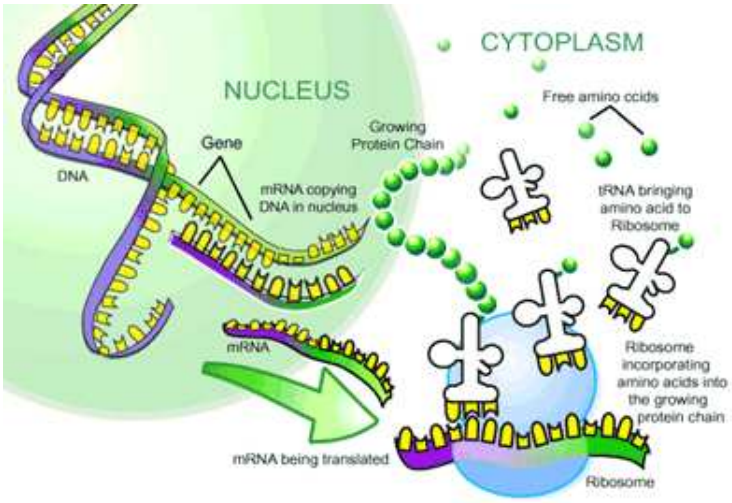

Figura 2–3: Representación del dogma de la biología molecular, Tomada de: www.scq.ubc.ca

Capítulo 3
MICROARREGLOS

"El conocimiento que tenemos sobre la biología y la genética no es buena, ni mala, solo conocimiento" -Francis S. Collins

Director.NHGRI

El capítulo anterior, acentúa la necesidad de estudiar no sólo la formación del genoma de un determinado organismo sino estudiar la expresión de los genes que lo conforman, saber qué proteínas se encargan de producir y en que funciones regulan o participan.

Por lo tanto, la búsqueda de un mecanismo que permita estudiar la expresión de varios genes al mismo tiempo se hace necesario [10]. Existían distintos métodos que nos permitían estudiar la expresión genética, pero requerían mucho tiempo. El método más usado se conoce como dejando uno afuera("knock it out"). Este consiste en aislar un gen y ver cómo afecta a los otros genes y a los procesos del organismo en estudio. Llevarlo a cabo toma mucho tiempo, así que es poco viable, costoso, expansivo y poco recomendado si queremos estudiar varios genes a la vez.

La solución a este inconveniente viene dado por los microarreglos, los cuales permiten obtener la expresión genética de cientos o miles de genes al mismo tiempo. De esta forma, podemos estudiar modelos de expresión de los genes que están en un tejido, podemos comparar los tejidos sanos y patógenos. O desarrollar pruebas que dependen de la expresión de un gen, como los marcadores genéticos. Es por esto,

que el uso de microarreglos es escogida antes que el de dejar uno afuera. Además que se puede mirar el efecto global de los genes. Aunque la técnica de dejar uno afuera no ha perdido vigencia y sirve en gran medida para estudiar en detalle un gen en particular [10].

El interés que ha despertado el uso de microarreglos ha traído como consecuencia directa una cantidad de datos, que muchas veces no se han podido analizar ya que se producen mas rápido de lo que se puede estudiar.
Y es que gran parte de la motivación que hay alrededor de los microarreglos es sobre su aplicación a áreas como la medicina. Lo que se desea es hacer instrumentos que permitan diagnosticar enfermedades, de forma certera. Es comprensible el interés que tenemos en desarrollar la teoría (si se tuviera un laboratorio adecuado, se podría hacer la práctica) sobre la manipulación de los microarreglos.

A continuación se procederá a definir algunos conceptos técnicos.
Un microarreglo ADN es un soporte sólido de plástico, silicio o vidrio. En él se fijan distintas muestras de material genético, como son oligos, cADN, o mARN, siendo los dos primeros los más usados. A estos materiales se les llama "probes". Si miramos un microarreglo lo que vemos es una serie de puntos o "spots". Se calcula que en cada microarreglo hay aproximadamente 80,000 puntos, con diámetro de 80 a 150 $m\mu$. Cada "spot" contiene millones de clones de una secuencia específica.[20], [2].
Existen dos tipos de microarreglos, muy diferentes uno del otro, no sólo en su fabricación, sino en los "probes" y en el análisis, estos son los "spotted microarrays"(microarreglos punteados) y los microarreglos oligonucleótidos de alta densidad. Los que se utilizan con mayor frecuencia son los primeros, debido al costo y de fabricación. Los segundos son más precisos, pero existen pocas compañías que los

elaboran y su precio está muy por encima de los primeros.

3.1. Microarreglos Punteados("Spotted microarrays")

Existen dos tipos de fabricación de microarreglos punteados. Se clasifican según la forma en que se colocan los "spots" en la superficie del microarreglo, estas son:

i) Distribuidor Activo: Este es donde se construyen o se fijan base por base cada uno de los "spot"[20].

ii) Distribuidor pasivo: Aquí se elabora el material genético que se va a usar como "probe" aparte y luego se va aplicando sobre el microarreglo usando un robot. En este caso se hace una solución con ADN y por medio de una aguja se va aplicando sobre la superficie del microarreglo.

Podríamos resumir algunos pasos para la elaboración de microarreglos de la siguiente forma[5]:

1. Diseño del experimento: En este punto se deben decidir muchas cosas, lo que se quiere estudiar(un tejido, gen, o todo un genoma), cómo serán colocados los "spots", la marcación, entre otros puntos, es decir fijar las bases para el microarreglos. Luego de haber decidido esto seguiría la fase de extracción del tejido.

2. Extracción del objeto a estudiar: En el punto anterior se decidió qué se quería estudiar. Se extrae el tejido, por ejemplo sangre del ser humano. Normalmente se extraen diferentes muestras ya sea del mismo organismo o de organismos distintos de la misma especie, esto se utiliza para hacer replicación del experimento o para usar como "probes" para otros experimentos [31].

3. Extracción del $mARN$: Esta cantidad se extrae del tejido y se separa de los otros tejidos. La cantidad de $mARN$ viene a ser proporcional a la expresión de los genes estudiados. Esto se realiza por medio de un proceso llamado plydeadenylation. Una vez adquirido el $mARN$ se elabora $cADN$ Recordemos que el $mARN$ se obtiene de una cadena de ADN, cambiando la Tiamina por el Uracilo, lo que hacemos en este caso es la transcripción inversa, es decir se adhiere a la cadena de $mARN$ una de ADN, fabricada en el laboratorio con anterioridad y luego esa cadena es usada en el microarreglo. Esta es llamada ADN complementario, debido a que es el complemento de $mARN$, y se escribe $cADN$. Esta última es usada en lugar de $mARN$a que es mas estable que la anterior.

4. Marcador del $cADN$: Luego debemos decidir, que cantidad de $cADN$ a usar, aunque muchas veces esto se decide antes del paso 3. Dado que lo obtenido en un microarreglo es una imagen para luego escanearla, tenemos que hacer visible las cadenas de $cADN$. Para ello se utilizan dos colores fluorescentes, uno para cada tejido, casi que por conveniencia el tejido sano o de control se marca con el color verde y el patógeno o caso con color rojo, aunque esto lo decide el fabricante al momento de elaborar el diseño. Se conocen como *Cy5* que es el rojo, su nombre es cianina5, y el verde es cianina 3 por lo tanto es *Cy3*.

5. Colocación de los *"probes"*: En la superficie del microarreglo, colocamos los *probes* que a su vez forman los "spots". Hay tres clases de "spots" en el microarreglo, existen unos que están limpios es decir no tienen material genético, otros que tienen secuencias conocidas muy parecidos a los genes que vamos agregar y otros que se preparan a partir de otros organismos. Todos ellos sirven de control para los "spots" de interés, los cuales son localizados en puntos específicos del microarreglos, recordemos que los "spots" están formados por "probes". Los "probes" que

recae interés uno a uno son fabricados con base en las secuencias de $mARN$, que se desea estudiar usando el principio descrito por Watson y Crick de complementariedad [38].

6. Hibridación: Luego de haber colocado el material genético en la superficie del microarreglo, se agrega el $cADN$, lo que se espera bajo unas condiciones adecuadas es que halla proceso de hibridación.

7. Lavado: Se retira el material sobrante, o aquello cuya hibridación no es muy fuerte. En este punto se coloca otra placa del mismo material al de la superficie del microarreglos, y se pasa a escánaer, lo proximo es analizar imágenes.

Existen ciertos puntos que hay que determinar en el diseño, esto es si colocamos "spots" consecutivos cuyos "probes" son iguales o "clones". Es decir, hacer replicas, muchas veces se extrae material genético del mismo tejido de células distintas o de la misma célula, para enfrentar los resultados y observar la diferencia. Otro caso es tomar el mismo tejido de organismos distintos y compararla, pero por lo visto en el capítulo anterior acerca del genoma humano este tendría muchas fallas.

Lo que debemos tener en cuenta en nuestra tesis, específicamente es lo minucioso que debe ser este procedimiento, aunque la mayoría de todo el proceso es realizado a través de robots o máquinas, éstas tienden a tener algunos errores sistemáticos, pero algunas partes del proceso son manipuladas por seres humanos, además que los materiales que se utilizan son altamente sensibles a un trato no muy sofisticado, y esto trae como consecuencias errores en la lectura del microarreglo. Otro de los puntos es mirar las fronteras entre "spots".

Como dijimos al principio existe otro tipo de microarreglos, que son más sofisticados y más precisos. Como vimos anteriormente los microarreglos de dos colores, no miden la expresión de un gen sino que dan una razón o comparación entre dos tejidos, eso no ocurre en los microarreglos oligonucleótidos.

3.2. Arreglos Oligonucleótidos(Oligonucleotide Microarrays)

Conocidos también como arreglos de un solo canal. En esencia estos microarreglos no usan las librerías de $cADN$. Para construirlos lo que se hace es construir los "probes" base por base, donde cada "probe" tiene aproximadamente 25 bases. Dependiendo del diseño, se construyen dos clases de "probes", uno es *perfect match probes(PM)* que no es mas que la secuencia complementaria a la parte del gen que se desea estudiar, y *mismatch probe(MM)*, en esta lo que se hace es cambiar la base del medio de un PM por su complementario. Como estamos hablando de 25 bases lo que forman cada PM entonces será la de la posición trece, es decir si tenemos que para unPM, en la posición trece la base T, lo que hacemos es cambiarla por su complemento al fabricar MM, esto es por A. Anteriormente por cada PM se usaba un MM[8], pero con el tiempo el número de MM usado por PM ha ido disminuyendo. Actualmente se utiliza cerca del 30 %, de MM en un microarreglo[20]. Estas dos pruebas se usan con el objetivo de evaluar la calidad de la hibridación.

El objetivo es que cada conjunto de "probes" estudie una secuencia específica o una parte del gen objetivo, para ello se preparan grupos de 16 a 20 "probes", para cada sector del gen. Esto es conocido como *probeset*[6], aunque por ejemplo para el microarreglo HG-U133A, se han considerado 11 *probeset*. Este último es un microarreglo que fábrica la compañía Affymetrix, que se conocen como GeneChips. D.Lockhard

y Lipshutz [24] y [23], idearon esta técnica de fabricación de microarreglos. Actualmente existen otras que también elaboran este tipo de microarreglos, como son: *GE Healthcare, Ocimum, Biosolutions y Agilent.*

A diferencia de los microarreglos de dos colores, para poder comparar dos tejidos en los microarreglos oligonucleótidos hay que elaborar un microarreglo para cada uno. Para comparar la expresión de cada gen, usando este tipo de microarreglo, se mide el nivel de expresión de cada "probe set". Para esto, se han elaborado ciertas medidas, algunas de ellas son: *Average Difference (AvgDiff), el Model Based Expression Index (MBEI)* de Li y Wong (2001)[22], *el MAS 5.0 Statistical algorithm* de Affymetrix (2001), y el *Robust Multichip Average* propuesto en Irizarry et al. (2003) .

Actualmente se esta desarrollando una tecnología híbrida de los dos tipos de microarreglos anteriormente descritos, se trata de utilizar la uniformidad de microarreglos oligonucleótidos y la especificidad de los microarreglos punteados o c*ADN*.

Luego del proceso de obtener las placas de los microarreglos, se pasa a la etapa de análisis de imágenes, en la cual usando los distintos software que existen, se pasan las intensidades de la fluorescencia a una matriz de datos. En esencia lo que se busca es leer la intensidad de cada uno de los "spots". Primero se separan cada uno de los canales los cuales se hacen visibles al ampliarla usando una computadora.

La imagen que se obtiene es una serie de puntos, que son los *"spots"*. Estos a su vez están formados por una serie de elementos gráficos, que es lo mas mínimo que se puede observar a través del monitor de una computadora, estos se llaman píxel. El valor de un píxel será el promedio de la intensidad en el área representada por el píxel. Dependiendo del software usado, el valor de cada píxel se toma como 1/10 del

diámetro de la mancha de cada “spot”. Así que cada “spot” será cubierto por 100 píxels aproximadamente(pero esto depende del escáner y de la computadora que se esté usando) lo que se recomienda es que cada “spot” tenga por lo menos 8 píxel.

El número de píxels usados por el escáner se llama resolución y este viene dado por la cantidad de líneas horizontales y verticales de píxel para cubrir la plantilla del microarreglo.

A grandes rasgos lo que se busca es obtener dos imágenes, una que proviene del tinte verde y otra del tinte rojo para luego superponer con la otra. Existen varios software algunos gratuitos, qué permiten analizar estas imágenes. Lo primero que debe hacerse es decidir cuántos sectores existen. Estos vienen dados desde el laboratorio de fabricación , las cuales usan distintas agujas para diferenciar unos de otros en el momento de la aplicación del material. Pero también puede ser a discreción del científico o la persona que esta analizando la imagen; esto se hace porque cada “spot” tiene un punto de ubicación dentro del microarreglo. Luego de ello, lo que se hace es diferenciar qué píxels forman parte del material genético a analizar o por el contrario cuál esta fuera de él mismo, es decir el fondo. Se busca extraer la intensidades de Cy3 y Cy5 de cada uno, para luego obtener una matriz de datos.

Algunos programas para análisis de imágenes de microarreglos se consiguen gratis, mientras otros son costosos y son usados en especial para investigaciones en medicina u otras ciencias o disciplinas.

A) **Programas Gratis(uso no comercial)**

- Dapple: Se encuentra disponible en http://www.cs.wustl.edu/ jbuhler/dapple/ Colocamos de entrada el par de imágenes provenientes de la hibridación, lo

que Dapple hace es hallar los “spots” individuales sobre la imagen, evalúa la calidad de la imagen, y cuantifica la intensidad de fluorescencia. Solo trabaja en microarreglos de vidrios. Dapple esta implementado en el lenguaje C++ lenguaje para UNIX como sistema operativo. Fue creado por Dr.Jeremy Buhler.

- Scanalyze: Desarrollado por Michael Eisen, es un software gráfico y fácil de manejar, además tiene un manual gráfico y puntual, se puede conseguir en la página http://rana.lbl.gov/EisenSoftware.htm, junto con otros software de análisis de microarreglos. El programa corre bajo Windows.
- MAIA: Desarrollado por Eugene Novikov at Institute Curie , en Francia, es uno de los software gratuito que recibe mantenimiento constante, va por la version 2.7, incluye un manual bastante practico y gráfico, en Windows. http://bioinfo-out.curie.fr/projects/maia/
- Spotfinder: Instituto J Craig Venter,2005,la version actual es 2.2.4,corre sobre Windows 2000. Está escrito en C++., incluye un manual pero no tan gráfico como los dos anteriores. Lo podemos hallar en la siguiente dirección ftp://occams.dfci.harvard.edu/pub/bio/Spotfinder

B) **Programas Comerciales**

- SilicoCyte, comercializado por silicoCyte, corre bajo Windows
- ImaGene comercialiazado por Biodiscovery , corre en Windows y Linux.
- QuantArray Universidad de Oregon,caprise Rosato, corre bajo Windows

La gráfica 3–2, representa un microarreglo punteado y el proceso de elaboración.

Sí usamos los microarreglos Affimetrix, al final lo que obtenemos es una matriz de expresiones, de cada uno de los genes. Pero, si usamos microarreglos punteados,

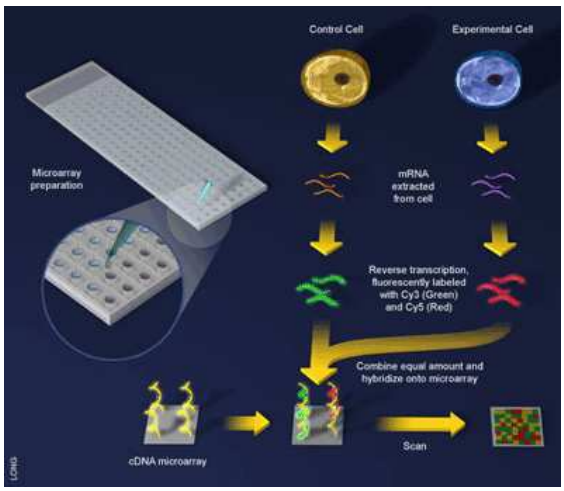

Figura 3–1: Experimento de microarreglos, fig. tomada de www.scq.ubc.ca

tenemos una matriz de razónes de las expresiones de los genes obtenidas por el color verde entre rojo.

Luego del proceso de fabricación y antes de analizar el microarreglo como imagen, lo podemos apreciar en la siguiente gráfica:

Figura 3–2: Microarreglos Affymetrix, fig. tomada de www.blep.com/journal/

Cada experimento con microarreglos depende del diseño propuesto con anterioridad, lo que hay que tener claro es cuántas muestras debemos tomar o vamos analizar. En

la literatura vemos que la mayoría de las veces usan menos de 200, ya que para cada muestra se construye un microarreglo. Obviamente en cada microarreglo estamos evaluando la expresión de los mismos genes. Un ejemplo en específico es estudiar al menos dos tipos de cáncer. Para ello lo que hacemos es tomar distintos tejidos(cada una se evalúa en su respectivo microarreglo) que presentan ese tipo de cáncer(no es necesario tomar la misma cantidad para cada uno de los cáncer (clases) que estamos evaluando) y para cada uno elaborar un microarreglo. Al final de extraer las expresiones de los genes a evaluar, se construye una matriz donde las filas son las muestras(o microarreglos) y en las columnas estarían los genes.

Por lo visto en los párrafos anteriores, es fácil predecir que esa matriz de datos, llamada matriz de expresión genética, tendrá muchísimas columnas(generalmente mas de mil) y pocas filas(normalmente menos de 200). Esto es, si queremos analizar la expresión de n genes, en m muestras obtenemos una matriz G, de orden mxn, con $n >> m$.

Normalmente la matriz que resulta no es a la que le aplican directamente los distintos métodos estadísticos, ya que puede presentar errores muy grandes, Para ello se deben preprocesar los datos.

Capítulo 4
ALGUNAS CARACTERÍSTICAS DE DATOS DE EXPRESIÓN GENÉTICA

"Necesitamos todas las formas de conocer posibles, todas las formas de decir la verdad. La ciencia es una, la fé es otra." -Francis S. Collins

Director NHGRI

En este capítulo describiremos algunos conceptos, no necesariamente estadísticos. Que son relevantes en el desarrollo posterior de la tesis.

4.1. Valores Faltantes

En el capítulo anterior hablamos un poco sobre microarreglos y su uso en ciencias médicas, farmacología, entre otras disciplinas. Por lo que hemos descrito, el proceso de fabricación y pasar de una imagen gráfica a datos, lleva cierto tipo de errores, ya sea sistemático o humano. Por tal motivo la matriz final tiene valores que no representan necesariamente la intensidad de los genes en la muestra dada o sencillamente tenemos celdas vacios en el conjunto de datos, es decir valores faltantes ("missing values").

Que ocurran valores fuera de lo esperado o que se clasifiquen de manera inadecuada algunas muestras, puede ser corregido, ya sea suprimiendo algunos genes, haciendo análisis de ruido, transformando el conjunto de datos. Se puede tener en cuenta la cantidad de ruido del conjunto de datos en el análisis posterior y poder hallar una tasa de error en nuestra interpretación. Pero cuando se presenta parte del conjunto

de datos vacío, no es tan fácil manejarlo ya que se puede presentar un alto porcentaje de datos faltantes. A pesar de la frecuencia como se presentan estos datos es sorprendente que hasta el 2001 [37], no habían métodos que al tratar de imputar o "llenar" esos valores faltantes, tuvieran en cuenta la estructura del conjunto de datos.

Cada uno de los procesos descritos anteriormente cae en la etapa del preprocesamiento de datos de microarreglos. Es importante el saber trabajar con cada uno de ellos, para tener conclusiones más cercanas a los valores reales.

Nuestro interés se centra en valores faltantes y en ruido. Del primero en la forma de cómo se puede hacer una buena imputación y del segundo es cuanto afecta el porcentaje de ruido del conjunto de datos en los métodos de imputación.

4.2. Clasificación de Valores Faltantes

Valores faltantes ("missing values") es un tema que se trabaja con mucha frecuencia en minería de datos. En el año 1987, Little y Rubin[41] subdividieron los valores faltantes en tres categorías:

1. Valores Faltantes Completamente al Azar(por sus siglas en inglés,"MCR"): Este caso es el ideal. Nos dice que la probabilidad de que un valor sea valor faltante no depende de los valores observados o de los no observados. Es decir, no se pueden predecir por los datos observados o no observados. En pocas palabras esto sería:

$$P(\text{“}missing\text{”}/observada, noobservada) = P(\text{“}missing\text{”}) \tag{4.1}$$

   Un ejemplo típico de MCR, sería cuando se les pide a las personas que contesten o no a una pregunta dependiendo del lanzamiento de una moneda.

2. Valores Faltantes al Azar(por sus siglas en inglés,"MAR"): La distribución de los valores perdidos depende de los datos observados, esto vendría dado por:

$$P(\text{“}missing\text{”}/observada) = P(\text{“}missing\text{”}) \tag{4.2}$$

   Un ejemplo donde se evidencia MAR, es cuando se le pregunta a una persona por quien votaron. Puede haber respuesta o no, pero es bastante seguro que respondan a que partido pertenece.

3. Valores Faltantes no al Azar(MNAR): En este caso los valores faltantes no son aleatorios. Es más difícil encontrarlos y de trabajar. Su distribución de los valores faltantes dependen de valores faltantes.

Este tipo de distribución de valores perdidos que hallamos con mayor frecuencia es MAR.

Los valores faltantes son frecuentes en datos de microarreglos y la importancia de poder tratarlos de forma adecuada radica en que muchos procesos para analizar datos requieren que el conjunto de datos esté completo. Hay muchos métodos que exigen de entrada el conjunto de datos completo; hay algunos, pocos que no traen esa restricción, ya que internamente algunos eliminan los valores faltantes, o lo reemplazan por cero. Ninguna de estas opciones son aconsejables. Otra alternativa es repetir el experimento bajo las mismas condiciones. A pesar que muchos han visto esto como una posibilidad, para obtener conclusiones más fiables, usando técnicas estadística cuando hay réplica. Pero en realidad desarrollar de nuevo el experimento no es viable, ya que cada elaboración del microarreglo y su lectura es costoso, por lo tanto esta opción no es la más adecuada. Es notorio que gran parte de los avances estadísticos en bioinformática están influenciados por lo que se hace en minería de datos. Lo que sorprende es que en esta última área el manejo de datos faltantes es determinante a la hora de preprocesar datos. A diferencia de bioinformática que este

tema hasta hace poco carecía de relevancia. De hecho muchos autores no incluyen el tema de valores faltantes en preprocesamiento de datos de expresión genético.

Veamos en la siguiente matriz de expresiones de 10 genes para cuatro muestras, porque no es viable, el eliminar ya sean genes o muestra, o por último hacer el reemplazo por cero o por la media de los genes:

$$\begin{pmatrix} 0{,}1 & NA & 0{,}9 & 0{,}3 & 0{,}4 & 0{,}3 & 0{,}5 & 1{,}0 & NA & 1{,}3 \\ 0{,}1 & 0{,}9 & 1{,}2 & 0{,}7 & NA & 0{,}8 & 1{,}1 & 1{,}5 & 2{,}3 & 3{,}0 \\ 0{,}2 & 0{,}9 & 2{,}1 & 1{,}0 & 0{,}4 & 1{,}2 & NA & 2{,}5 & 2{,}3 & 0{,}2 \\ 0{,}4 & 0{,}6 & 1{,}3 & 2{,}7 & 1{,}4 & NA & 1{,}6 & 1{,}6 & 2{,}8 & 4{,}0 \end{pmatrix}$$

Si observamos detalladamente la matriz de datos, podemos decir que: el 100 % de muestras, el 50 % de los genes y 12.5 % de los datos tienen valores faltantes. Esto es un ejemplo típico de los porcentajes de valores faltantes que se hallan en los conjuntos de datos reales. Por lo tanto, el manejo de este tipo de datos se debe hacer con cuidado. Por el anterior ejemplo, podemos inferir que el quitar los casos, con valores perdidos no es lo más recomendable y tampoco eliminar los genes. Una solución que parecía excelente a la hora de imputar los datos era el de transformar los datos de los valores de expresión genética por el logaritmo en base dos y los valores perdidos sustituirlos por cero. Esto fue propuesto por Alizadeh et al., 2000 [3].

En cualquier proceso de medición ocurren errores. En particular, hay presentes valores faltantes en datos de microarreglos, debido a las siguientes razones:

- Resolución insuficiente
- Error experimental

- Error durante la hibridación
- Error en la fabricación del microarreglo(como polvo en la laminilla)
- Defecto en la imagen(debido al software, manipulación no adecuada del microarreglo hibridizado)
- Error en la medida de la intensidad(como al ubicar los "spots", al marcar el límite del fondo), entre otros.

No solo tenemos valores faltantes en nuestras bases de datos, también se puede presentar ruido en las observaciones o que la complejidad del conjunto de datos sea alta. Esto influye de manera significativa en cualquier tarea que vayamos a realizar, en particular en imputación. Muchos autores una vez hallado los datos que son complejos o que no están en sintonía con los demás, sugieren el removerlos, pero ese no es nuestro objetivo (ya que hemos visto que remover muestras en nuestro caso no resulta lo mas apropiado). Lo que haremos es calcular la complejidad del conjunto de datos, y el porcentaje de ruido, y luego de realizarlo lo que miraremos es como afectan esos valores a los conjuntos de datos estudiados.

4.3. Complejidad de los datos

Si bien es cierto que muchos autores usan el término de complejidad. A veces se usa calidad del conjunto de datos, resulta un témino más apropiado. Con calidad se refiere a qué tan buena es la base de datos para realizar una tarea, o qué podemos esperar de ella al querer extraer información. Este tema es abordado ampliamente en Daza[29]. Aunque sólo fue considerado en problemas de clasificación, una de las tareas importante sería en minería de datos y en bioinformática.

Recordemos que en microarreglos podemos tomar muestras ya sea de un sólo tejido, dos o más tejidos. Sí tenemos dos o más, entonces cada uno de esos tejidos es considerado como clase. Se supone que la clasificación a priori se realiza por medio de los tejidos usados, esto es, al usar varias muestras de dos tipos de tejidos uno sano y el

otro patógeno, lo que hacemos es que toda las muestras que sean del tejido sano, lo etiquetamos con 1 y los que sean del patógeno con 0.

Pero podríamos tener casos como es ambigüedad de la clase, esto es, tenemos dos instancias cuyos valores son idénticos pero que pertenecen a distintas clases, se presume hubo una equivocación al asignarle la clase a una de las muestras.
Existen ciertas medidas que permiten poder calcular la complejidad de los datos, éstas son:

4.3.1. Razón de Fisher

Para sólo dos clases viene dada por:

$$F_1 = \frac{\mu_1 - \mu_2}{\sigma_1^2 - \sigma_2^2} \tag{4.3}$$

donde μ_1, μ_2, σ_1^2, σ_2^2 son las medias de las dos clases y sus varianzas correspondientes.
Existen conjuntos de datos que tienen más de dos clases, así se generalizó la anterior medida:

$$F_{1gen} = \frac{\sum_{i=1}^{g} n_i \delta(\mu, \mu_i)}{\sum_{i=1}^{g} \sum_{j=1}^{n_i} \delta(x_j^i, \mu_i)} \tag{4.4}$$

donde n_i denota el número de instancias en la clase i, δ es una métrica, μ es la media global, μ_i es la media de la clase i y x_j^i representa la instancia j que pertenece a la clase i.

4.3.2. Volumen de la Región de Solapo

Esta medida calcula para cada variable f_k, la longitud del solapo normalizado por la longitud del rango total en que todos los valores de ambas están distribuido:

$$Over = \sum_{(C_i, C_j)} \prod_k \frac{minmax_k - maxmin_k}{maxmax_k - minmin_k} \tag{4.5}$$

Donde (C_i, C_j) recorre por todos los pares de clases, $i, j = 1, 2, ..., g$, además:

$$minmax_k = min\{max(f_k, c_i), max(f_k, c_j)\}$$

$$maxmin_k = max\{min(f_k, c_i), min(f_k, c_j)\}$$

$$maxmax_k = max\{max(f_k, c_i), max(f_k, c_j)\}$$

$$minmin_k = min\{min(f_k, c_j), min(f_k, c_j)\}$$

4.3.3. Fracción de Puntos en la Frontera

Este método se basa en la técnica propuesta por Friedman y Rasfsky, llamado *Minimum Spanning Tree*(MST), sugiere que construimos un MST de la forma en que cada instancia se conecta con sus vecinos más cercanos, sin considerar la clase a la que pertenece. Luego se procede a contar el número de instancias conectadas a una instancia de la clase contraria por una arista en el MST. Se considera que estos puntos están cerca de la frontera de la clase. La fracción de las instancias restantes sobre el número total de instancias en el conjunto se usa como la medida de complejidad.

Daza(2007)[29] enumera varias medidas de complejidad para bases de datos, aunque las anteriores son usadas con frecuencia. El propone las siguientes medidas en su tesis, las cuales miden la calidad de los conjuntos de datos de las instancias, lo que se busca es medir la "posición" de una instancia en específico. Qué distancia tiene a la frontera de decisión.

4.3.4. Calidad de las Instancias con respecto a los Centroides

Mide la posición de la instancia dentro de la clase a la cual pertenece, ya que puede estar en el centro, en la frontera o es una instancia que cumple las condiciones para estar en otra clase(ruido). Esta viene dada por:

$$Q_i = \frac{r_i - d_i}{max(d_i, r_i)} \tag{4.6}$$

Donde d_i es la distancia métrica de la i-ésima instancia x_i al centroide de su misma clase, r_i es la mínima distancia de x_i al centroide de las otras clases, dividido entre el máximo de cada una de los valores antes señalado. Esta medida cumple ciertas propiedades:

- Q_i varía entre -1 y 1.
- Un valor cercano a 1, indica que la instancia es de buena calidad, un valor cercano a -1 indica que la instancia es de mala calidad y con posibilidad de ser ruidosa.
- Para las instancias cercanas a la frontera, Q, es cercano a 0.

Otra de las medidas desarrolladas por Daza(2007) es la siguiente:

4.3.5. Calidad de Instancias Basadas en los Vecinos Más Cercanos

La medida basada en vecinos más cercanos para medir la calidad de las instancias(QNN), viene dado por:

$$QNN_i = \frac{\tilde{r_i} - \dot{d_i}}{max(\tilde{r_i}, \dot{d_i})} \tag{4.7}$$

Donde $\dot{d_i}$ es la distancia métrica de la i-ésima instancia x_i a su vecino más cercano en su misma clase; $\tilde{r_i}$ es la distancia métrica mínima de la i-ésima instancia a sus vecinos más cercanos de cada una de las clases opuestas, normalizada por el máximo de $\tilde{r}$ y $\bar{d}$.

Para determinar la complejidad de un conjunto de datos procedemos como sigue:

- Para cada instancia se calcula la medida de calidad Q o QNN.

- Se procede a contar las instancias que tienen una medida de calidad menor que cero (negativa). Entonces para hallar el porcentaje de instancias de mala calidad, está dada por:

$$Q_{IND} = \frac{count\{Q_i < 0\}}{n} \tag{4.8}$$

Donde Q_i representa la medida de Q o QNN para x_i.

Un valor grande para las medidas como Fisher y MST menos complejidad. Es decir la relación es inversa. Mientras que para Q , QNN y Overlap , la relación es directa, esto es a mayor valor para ellas significa mayor complejidad.

Además si un conjunto de datos tiene mayor complejidad, significa que es de menor calidad.

4.4. Ruido

Hemos notado varios problemas que pueden aparecer en el conjunto de datos en estudio. Esto es, por ejemplo que haya valores muy grandes comparadas con otros, para esto se requiere que se normalice. Los valores faltantes son muy comunes e impiden muchas veces adquirir conocimiento del conjunto de datos. Pero existe otro factor que afecta la obtención y por ende la interpretación de resultados, lo más complicado es que este problema no es tan palpable como el de valores faltantes. Sugiere ciertas destrezas de parte del investigador para que pueda notar que la base de datos que está trabajando contiene cierto porcentaje de ruido.

Se dice que un conjunto de datos contiene ruido si se han presentado errores en la asignación de clases o errores introducidos en las variables, siendo estos errores no sistemáticos.

Podemos distinguir dos tipos de ruidos, veamos su clasificación:

1. Ruido en las variables: errores que se presentan al ingresar los valores de las variables. Algunos ejemplos son: Errores en los valores de las variables, variables con valores perdidos y datos redundantes.
2. Ruido en las clases: Son los errores al asignar las clases de las instancias. Esto puede deberse a:
   a) Instancias Contradictorias: Dos o más instancias que tienen valores iguales, variable por variable, pero que pertenecen a clases diferentes.
   b) Error en la clasificación: Instancias que están asignadas a una clase distinta a la que debería estar.

Lo que queremos observar es cuánto afecta el porcentaje de ruido, a los métodos de imputación. Muchas bases de datos que están disponibles gratuitamente, ya están preprocesadas y es por ello que el hallar ruido en las instancias es muy difícil. Por lo que sólo hallaremos el ruido en las clases. Esto lo haremos usando un método desarrollado por Daza[29], en el cual se propone es tomar aquellas instancias con valor de calidad Q da negativo. Luego se hallan los vecinos más cercanos a estas instancias (se recomienda usar cinco vecinos), y se cuentan cuantas pertenecen a clases distinta a la instancia considerada. Si este valor es mayor a las que están en su misma clase, se considera ruido. La cantidad de ruido del conjunto de datos se obtiene al tomar la cantidad de instancias con ruido sobre el total de instancias.

4.5. Identificación de Genes Diferencialmente Expresados

Dado un conjunto de datos de microarreglos, una de las tareas en la que recae mayor interés es en la búsqueda de los genes diferencialmente expresados. De hecho es uno de los temas de los cuales hay más literatura en bioinformática. Existen muchos métodos y medidas para hallarlos. En nuestra tesis usaremos una librería en R, la cual emplea modelos lineales para identificar los genes diferencialmente expresados.

La librería usada es llamada Limma: (*linear models for microarray data*). Como su nombre lo indica usa modelos lineales para distintas tareas como son normalización, hallar genes expresados y visualización, entre otras. Su autor es Gordon Smyth[35].

El procedimiento para detectar los genes diferencialmente expresados es como sigue:

1. Consideramos primero el modelo que vamos a usar, modelo de los mínimos cuadrados. En este se halla un modelo lineal para cada gen.

2. Se construye las matriz diseño asociada con el modelo , ya que estamos comparando grupos, por ejemplo si tenemos 5 muestras con dos casos, CT y INF, donde hay 2 CT y 5 INF, la matriz diseño nos quedaría:

CT	INF
1	0
1	0
0	1
0	1
0	1

Esto es, la matriz diseño indica las distintas clases o grupos que vamos a comparar, para compararlos se usa la matriz de contrastes, donde se específica qué conjuntos se van a comparar. Sólo es necesario si hay más de dos clases.

3. Se plantean pruebas de hipótesis, donde lo que se quiere observar los genes que difieren de los demás, esto es dependiendo de los estadíticos del modelo.
4. Luego, como estamos realizando pruebas multiples hallamos los p valores ajustados, puesto que los genes se suponen correlacionados, debido a como se elabora el microarreglo, entonces no se usan los valores de p normales sino p ajustados. Hay distintas formas de hallar los valores para los p ajustados. En nuestro caso usamos Benjamini y Hochberg ("BH"). En este punto, usted debe considerar cuantos genes diferencialmente expresados desea, en nuestro caso, requerimos 100.

En esta tesis no se discute sobre la teoría sobre genes diferencialmente expresados, ya que es muy extensa, además que no es nuestro objetivo, lo único que nos interesa es poder obtener los genes diferencialmente expresados usando las mismas condiciones para todas las bases de datos, en nuestro caso usaremos *Limma*, aunque existe otra librería en R que permite hallar genes diferencialmente expresados llamada *multtest*. La secuencia que vimos para *Limma* puede ser usada para datos de microarreglos punteados u oligonucleotidos.

Capítulo 5
MÉTODOS PARA IMPUTAR VALORES FALTANTES EN DATOS DE EXPRESIÓN GENÉTICA

"La mayoría de las enfermedades y de las características humanas están relacionadas a cientos y hasta miles de genes."-Francis S. Collins
Director.NHGRI

Como notamos en el capítulo anterior, es clara la necesidad de obtener métodos que no solo sustituyan por cero o eliminaran los valores faltantes. A pesar de la necesidad de construir métodos que tuviesen en cuenta la estructura del conjunto de datos, fue hasta el 2001[37] que aparecieron los primeros métodos elaborados para imputación de datos con valores faltantes. Esto es, hallar un valor lo más cercano posible al valor real del dato. Con el tiempo se han desarrollado nuevos métodos de imputación. Lo cual nos abre una gran posibilidad para poder preprocesar conjuntos de datos con valores faltantes para no perder muestras y genes. A menos que exista un alto número de valores faltantes(esta cantidad depende del investigador o del método a usar). En las referencias notamos que cada uno de los que elaboran estos métodos lo aplican a pocos conjuntos de datos (generalmente menos de tres), en donde el método que ellos están desarrollando se comporta mejor que los otros, sin dar razones porqué eso ocurre o qué condiciones debe tener el conjunto de datos para que un método de imputación funcione en él. Por tal razón era necesario un trabajo independiente el cual evaluara los métodos de imputación para datos faltantes, un

factor importante en esto es la calidad del conjunto de datos y la cantidad de ruido que presenta.

5.1. Métodos de Imputación para datos de microarreglos.

Aquí describiremos cada uno de los métodos que aplicaremos, además de las ventajas y desventajas según sus autores. Luego, según nuestros experimentos, evaluaremos si es cierto lo que ellos describen. Consideremos una matriz $\boldsymbol{G}$ de expresión genética, de tamaño $n \times m$, , esto es:

$$G = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{pmatrix}$$

Donde g_j, es el vector de expresiones para las distintas muestras del gen j, de dimension $1 \times n$. Consideremos x_{ij} como el valor de la i-ésima muestra para el j-ésimo gen. Rubin(1976) introdujo una matriz indicadora, $\boldsymbol{R} = (r_{ij})_{i=1,\ldots,n}^{j=1,\ldots,m}$. Donde sí x_{ij} tiene un valor faltante, entonces $r_{ij} = 0$, de lo contrario, $r_{ij} = 1$. Lo que implica que nosotros queremos aplicar los métodos de imputación al conjunto:$\{x_{ij}\colon r_{ij} = 0\}$.

En el 2001 aparece la primera publicación relacionada con la imputación de datos de microarreglos, fue desarrollada por Troyanskaya et al [37]. Los autores proponen dos métodos, aunque describen tres. El tercer método es “Gene average”, que ya era ampliamente usado en minería de datos, y utilizado por varias personas cuando los datos de microarreglos contenían valores faltantes.

"Gene average", funciona de la siguiente manera: si un gen tiene un valor faltante, lo que se hace es hallar el promedio de los datos que conocemos del gen (recordemos que este es un vector) y se coloca en lugar del valor faltante. Sí hay varios valores faltantes, ese mismo valor se coloca en cada uno de esos lugares.

Este método no es elaborado, ni es riguroso y es bastante sensible a atípico. Los otros dos métodos los discutiremos a continuación:

5.2. Imputación usando los K-vecinos más cercanos(KNN)

Este método se conoce como KNN por sus siglas en inglés (K-nearest neighbors). Fue uno de los métodos desarrollado por Troyanskaya et al.,(2001) [37].

Sin perdida de generalidad, supóngase que el gen g_j, tiene un valor faltante en la posición ν. Primero hay que hallar los posibles genes candidatos que vamos a usar para imputar. Consideremos el conjunto D_c de todos aquellos genes cuya posición ν no tienen valores faltantes, esto es $D_c = \{g_j : r_{\nu j} = 0\}$, de allí vamos a tomar los K genes cuya distancia es más pequeña con respecto a g_j. Según la literatura podríamos usar cualquier tipo de distancia, pero para este método se usa con frecuencia la distancia euclidiana. Esta es sensible a datos atípicos, pero esto algunas veces se puede corregir si se transforman los datos. Por ejemplo, usando transformación logarítmica. Si esto no aplica hay que usar métodos más rigurosos, según Elliot [12]. Medimos la distancia euclidiana entre g_j y cada uno de los genes en D_c, en particular para el gen μ, esta viene dada por:

$$d_{j\mu} \equiv d(g_j, g_\mu) = \{n_{j\mu}^{-1} \sum_{k=1}^{n} r_{k\mu} r_{kj} (x_{kj} - x_{k\mu})^2\} \tag{5.1}$$

Donde $n_{j\mu} = \sum_{i=1}^{n} r_{ij} r_{i\mu}$, son los pesos aplicados a la distancia, estos vienen dados por los lugares en común que no tienen valores faltantes los genes μ y j.

En esta parte tomamos los K genes cuya distancia sea más pequeña con g_j. Procedemos a calcular el valor con el que vamos a imputar el valor faltante que tiene en la posición ν, consideremos I el conjunto de índices de los genes seleccionados. Entonces,

$$\hat{y}_{vj} = \sum_{i \in I} w_i x_{vi} \tag{5.2}$$

Donde $w_i = 1/(d_{ji} \sum_{i \in I} d_{ji}^{-1})$, para todo $i \in I$. Esto tiene el efecto de normalizar los pesos. Claramente los pesos son inversamente proporcionales a las distancias. Es decir, queremos que entre más cerca estén al gen que estamos imputando, mayor sea su colaboración.

Existen puntos que hay que considerar como es el caso de la selección de K. Troyanskaya et al.,[37] nos sugieren que ese valor sea entre 10 y 20. Si tomamos un valor más grande para K (aunque bien se puede ver que ese valor depende en gran parte de la dimensión del conjunto de datos), se puede dar el caso que estemos escogiendo genes que no son tan "cercanos" y por lo tanto la imputación sea óptima. Además si es una matriz con mucho ruido se aumentaría la contribución de ruido de cada gen al nuevo valor. Nguyen et al.(2004), [26] asegura que usando un K entre 10 y 22 se hallan resultados óptimos, la gran dificultad para obtener los valores cercanos a los reales es el porcentaje de valores faltantes en el conjunto de datos (se comporta óptimo entre 6 al 26 %). Además se recomienda que este método sólo utilizado en conjunto de datos que tengan más de seis muestras.

Por lo visto en este método se hallan los genes más cercanos que no tengan valores faltantes en la misma posición que el gen considerado. Pero existen bases de datos cuyo porcentaje de valores faltantes es tan alto que no permiten realizar dicho cálculo. En estos casos el método que se recomienda imputación que recomienda es el "Gene average".

Vimos que el método anterior considera todo el conjunto de datos(ya que a través de ella hay que buscar los K-vecinos cercanos), pero tenemos miles de genes y cientos de muestras. Por lo tanto el uso de KNN para imputación no es recomendable. Además es sensible a datos atípicos. Aunque su ventaja es que el algoritmo computa rapidamente. Pero hay algunos investigadores que expresan que no es necesario tomar todo el conjunto de datos para realizar imputación [27]. Otro punto es que el buscar los vecinos más cercanos, no sugiere que se este considerando la multicolinealidad entre los datos. Así surgieron varios métodos que usan técnicas como SVD y PCA.

5.3. Imputación usando SVD

Primero esbozaremos algunos conceptos necesarios para luego desarrollar la teoría del uso del SVD en imputación.

5.3.1. Descomposición en Valores Singulares(SVD)

Descomposición en valores singulares o SVD, por sus siglas en inglés. SVD, es usado con frecuencia en matemáticas, álgebra numérica y en estadística multivariada. En esta última, se utiliza una extensión de SVD que recibe el nombre de componentes principales. Se utiliza con frecuencia sobre todo, para reducir la dimensionalidad del conjunto de datos, para hacer predicción y para hallar patrones en el conjunto de datos. Por lo tanto su uso en imágenes es amplia, muchas veces el uso se extiende a

ciertas áreas, ignorando que no sólo se deben cumplir las condiciones matemáticas sino que también hay que considerar los conceptos para los cuales se desea usar.

Definición 1 (**Valor y Vector Propio**). *Consideremos una matriz A, de dimensión $n \times n$, λ es un valor propio (eigenvalue) de A, si existe un vector $\boldsymbol{x}$ tal que $A\boldsymbol{x}=\lambda\boldsymbol{x}$, equivale a:*

$$(A - \lambda I)\boldsymbol{x} = 0 \tag{5.3}$$

El vector $\boldsymbol{x}$ que es solución para la ecuación 5.3, recibe el nombre de vector propio (eigenvector),se dice que λ está asociado al vector propio $\boldsymbol{x}$.

Polinomio Característico: Sea A una matriz cuadrada, el polinomio característico de A, es $P_A(\mathtt{x}) = det(A - I\mathtt{x})$
Si A es una matriz cuadrada, entonces λ es un valor propio de A si y solo si $P_A(\lambda) = 0$

Existen resultados muy interesantes con respecto a los valores propios y vectores propios, el que tiene un interés particular para nosotros, es el siguiente:

Lema 1. *Sea A una matriz simétrica,entonces existe una matriz U y una matriz diagonal D, tal que: $U^T AU = D$, donde la diagonal de D está formada por los valores propios de A, ordenados de mayor a menor, y las columnas de U son los vectores propios con respecto a los valores en D*

Así que podemos representar a A, como $A = UDU^T$, dado que D es una matriz diagonal, entonces decimos que A es similar a la matriz U.
Entonces, A se puede representar de una forma simple, ahora nuestro interés se centra en poder extender ese resultado en matrices que no son simétricas, de hecho que

ni siquiera sean cuadradas. Esta descomposición se llama descomposición en valores singulares, SVD.

Definición 2 (**Valores singulares de una matriz**). *Sea A una matriz, sus valores singulares son la raíz cuadrada de los valores propios de $A^T A$*

Definición 3 (**Vectores singulares**). *Los vectores singulares derechos de A, son los vectores propios de $A^T A$, mientras que los vectores singulares izquierdos son los vectores propios de AA^T.*

Sea A una matriz de dimensión $m \times n$, con $m \gg n$, de rango r (número de columnas linealmente independiente), entonces existen las matrices ortogonales U y V tales que: $U^T AV = D$, donde D es una matriz diagonal, formado por los valores singulares de A. U de dimensión $n \times n$ está formada por los vectores singulares izquierdos de A. V de dimensión $m \times m$, esta formado por los vectores singulares derechos. Los valores singulares en D están en orden decreciente. Si tenemos que r es el rango de A, entonces tiene r valores singulares, $d_1, d_2, ..., d_r$, así que definiríamos D como:

$$D_{ij} = \begin{cases} d_i, & \text{si } 1 \leq i = j \leq r \\ 0 & \text{de otra forma} \end{cases}$$

Donde los d_i, están ordenados en orden decreciente, algunos pueden ser muy próximos a cero. Según lo descrito anteriormente, entonces A la podemos representar como $A = UDV^T$

Existe una relación entre los valores singulares y los valores propios de una matriz, puesto que:

$$AA^T = (UDV^T)(UDV^T)^T = UDV^T VDU^T = UD^2U^T$$

$$A^T A = (UDV^T)^T(UDV^T) = VDU^T UDV^T = VD^2V^T$$

Donde se ha usado el hecho de que U y V son ortogonales. De las ecuaciones anteriores nos queda que:

$$V^T A^T A V = D^2 \tag{5.4}$$

$$U^T A A^T U^T = D^2 \tag{5.5}$$

Por lo tanto si consideramos λ_i, para i=1,...,$rango(A^T A)$, los valores propios de $A^T A$, entonces $d_i^2 = \lambda_i$. Similar para (AA^T). Esto nos da una opción si queremos hallar los valores singulares de una matriz.

5.3.2. Componentes Principales

En estadística multivariada muchas veces se trata de reducir la dimensión del conjunto de datos. Puesto que trabajar con la base de datos completo no es lo más viable. Debido a que el hacerlo envuelve usar más tiempo y por lo tanto mayor costo computacional. El método de componentes principales trata de subsanar este problema, junto al de la multicolinealidad de los datos, cuya presencia puede llevar a malas interpretaciones. A pesar que es una técnica muy usada en la actualidad, el primer trabajo que se realizó con referencia a este tema fue por Karl Pearson 1901[28]. En lugar de trabajar con el sistema de coordenadas originales, propuso trabajar con combinaciones lineales de las variables originales, pero que sean invariantes. Luego Hotelling, 1933[17], tomó la idea anterior y formuló la teoría acerca de componentes principales. En los últimos tiempos con el aumento de las dimensiones de los conjuntos de datos trabajados tanto en bioinformática como en minería de datos, este tema toma relevancia. Jollife(1972)[19], estructuró la teoría de componentes principales y le dió muchas aplicaciones . Entre las aplicaciones de PCA se encuentran el procesamiento de imágenes, compresión de datos, visualización, análisis exploratorio de datos, predicción en serie de tiempo, entre otras.

Lo que específicamente se quiere en PCA, es hallar nuevas variables que conserven la mayor información del conjunto de datos original. Para ello, se pretende es que los nuevos datos contengan la mayor variabilidad de los datos originales y poder eliminar o reducir la multicolinealidad.

Lo que se busca en componentes principales, es expresar las variables $X_1, X_2, ..., X_n$, de la matriz X de dimensión $m \times n$, en nuevas variables, $Y_1, Y_2, ...Y_p$, que sean ortogonales, $p \leq n$. Para esto queremos que Y sea una transformación de X la cual la podemos expresar de la forma:

$$Y = A^T X \tag{5.6}$$

y que además esas variables guarden la mayor información de X, la matriz X debe ser centrada. Para obtenerla, el primer paso es hallar el vector de medias $\bar{x}$ de las columnas de X, conocido como centroide. Para la i-ésima componente del centroide se la resta a cada elemento de la i-ésima columna. Esto es: $\tilde{X} = X - 1\bar{x}$ donde 1 representa un vector de 1, de dimensión $n \times 1$. Cada uno de los componentes de A, el cual expresémoslo por a_j, debemos hallarlo de tal forma que al expresar de esta manera a $Y_j = a_{1j}\tilde{X}_1 + a_{2j}\tilde{X}_2 + ... + a_{nj}\tilde{X}_n$ con a_j el vector de pesos, nos quede que $Var(Y_j)$, es máxima. Notar que, $Var(Y_j) = Var(a_j^T \tilde{X}) = a_j^T Var(\tilde{X}) a_j = a_j^T (n-1)^{-1} \tilde{X}^T \tilde{X} a_j = a_j^T C a_j$, en el cual C es la matriz de covarianzas de X.

Entonces queremos maximizar la varianza de Y_j, lo cual depende de a_j. Sí consideramos un vector a_j el cual maximiza la varianza de Y_j podríamos hallar otro valor al multiplicar a_j por una constante. Por lo tanto, hay que colocar la restricción que $a_j^T a_j = 1$, esto es para evitar la influencia en la escogencía de los pesos, sólo interesa su dirección. Ahora para maximizar la varianza de Y usamos el multiplicador de

Lagrange el cual estará definido por:

$$\Phi = a^T C a - \lambda(a^T a - 1) \text{ derivando con respecto a } a \text{ nos queda}$$
$$\frac{\partial \Phi}{\partial a} = 2Ca - 2\lambda a = 0$$
$$2(Ca - \lambda a) = 0$$
$$(Ca - \lambda a) = 0$$

Según la segunda ecuación λ es un valor propio de C asociado con el vector a. Así como los valores propios los ordenamos de manera decreciente nos queda que el mayor valor propio λ maximiza $Var(Xa)$. Esto plantea una relación entre el mayor valor propio y la primera componente principal. Esto es, la primera componente es la multiplicación de X con el vector propio asociado al mayor valor propio, sabemos que si el rango de A es p entonces existen p valores propios y ese es el número máximo de componentes principales para X.

Entonces, $Y = A^T \tilde{X} = (a_1, a_2, ...a_p)\tilde{X} = (a_1\tilde{X}, a_2\tilde{X}, ...a_p\tilde{X})$.

La ortogonalidad de los vectores propios garantiza que las componentes principales lo sean. Siempre y cuando los vectores propios de C sean ortogonales esto es:

$$Cov(Y_i, Y_j) = Cov(a_i^T \tilde{X}, a_j^T \tilde{X}) \tag{5.7}$$
$$= E[a_i^T (\bar{\tilde{X}} \bar{\tilde{X}}^T) a_j] \tag{5.8}$$
$$= a_i^T E[\tilde{X}\tilde{X}^T] a_j = a_i^T S a_j \text{ , donde S es la varianza de} \tilde{X} \tag{5.9}$$
$$= a_i \lambda_j a_j = 0 \tag{5.10}$$

Por lo visto en 5.7, tenemos que $C = A\Lambda A^T$, donde Λ es una matriz diagonal, de esta forma:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

Como $Y = A^T\tilde{X}$, hallando la varianza de Y nos quedaría,$Var(Y) = Var(A^T\tilde{X}) = A^TVar(\tilde{X})A = A^TCA = A^TA\Lambda AA^T = \Lambda$, por lo tanto, los Y son no correlacionados.

El primer objetivo era transformar X en una matriz de menor dimensión pero que conservara la variabilidad. Puede notarse que Λ es una matriz diagonal lo que significa que la matriz de la varianza de Y también lo es. Usando ese hecho, tenemos: $\sum_{i=1}^{p} Var(Y_i) = traza(\Lambda) = traza(A^TCA) = traza(A^TA)traza(C) = \sum_{i=1}^{p} Var(\tilde{X}_i)$

Esto implica que logramos el objetivo, que las componentes principales conserven la varianza de las originales.

Las componentes principales no están relacionadas a ningún modelo probabilístico.

5.3.3. Uso de SVD para el modelo de regresión lineal multiple

El modelo de regresión lineal permite encontrar una variable dependiente(cuantitativa) con una o más variables predictoras o independientes. El objetivo es hallar un modelo que relacione la variable respuesta con las predictoras.

Dado el modelo: $Y = X\beta + e$, con las suposiciones $E(e) = 0$ y $Var(\mathtt{e}) = \sigma^2 I_n$, donde I_n es la matriz identidad de orden n, y $\mathtt{e}$ es un vector aleatorio de dimensión n.

Queremos hallar el valor de β que minimiza $\|X\beta - Y\|^2$, el cual viene dado por:

$$\hat{\beta} = (X^TX)^{-1}X^TY \tag{5.11}$$

Pero el cálculo de $(X^TX)^{-1}$, no es tan sencillo, ya que X puede ser una matriz cualquiera, en particular, ser singular. Así que no se puede sencillamente despejar para β. Lo que implica que hay que usar métodos más especializados para hallar esa estimación. Una alternativa es usar *SVD*. Consideremos U y V cuyas columnas están formadas por los vectores singulares izquierdo y derechos respectivamente, esto es $X = USV^T$. Dado que se busca minimizar.

$$\begin{aligned}
min\|X\beta - Y\|^2 &\quad \text{usando la transformación de X} \\
\|USV^T\beta - Y\|^2 &\quad \text{dado que U es ortogonal} \\
\|U(SV^T\beta - U^TY)\|^2 &\quad \text{por propiedades de la norma} \\
\|U\|\|SV^T\beta - U^TY\|^2 &\quad \text{como U es unitaria, nos quedaría} \\
\|SV^T\beta - U^TY\|^2 &
\end{aligned} \tag{5.12}$$

Tenemos una equivalencia para nuestra norma anterior, nos quedamos con matrices mas fáciles de manejar, las cuales sabemos que tienen inversa. Hagamos:

$$\tilde{\beta} = V^T\beta \tag{5.13}$$

$$\tilde{Y} = U^TY \tag{5.14}$$

Usando componentes principales, tenemos una transformación para X, esta es: $Z = US$, donde U y S son los que se obtienen al aplicar *SVD* dado 5.12, nos queda: $\|S\tilde{\beta} - \tilde{Y}\|$, así que al usar 5.11,tendríamos:

$$\hat{\tilde{\beta}} = (S^T S)^{-1} S^T \tilde{Y} \tag{5.15}$$

$$= (S^2)^{-1} (S)^T \tilde{Y} \tag{5.16}$$

$$= S^{-2} (S) \tilde{Y} \tag{5.17}$$

$$= S^{-1} \tilde{Y} \tag{5.18}$$

Y esa sería un solución equivalente a nuestro modelo original, para 5.11. La solución para 5.18, con base en la teoría de componentes principales, sería la inversa generalizada para los pesos, ya que no necesariamente es singular(originalmente es no singular), pero se reduce su dimensión al hacer *SVD*, véase apéndice A.

5.3.4. Descripción del modelo de Imputación SVD (SVDImpute)

Este método utiliza la relación de multicolinealidad que hay entre las columnas de una matriz, para ello consideremos G una matriz de expresión genética, de dimensión $m \times n$, con $m >> n$, la cual tiene valores faltantes, como emplearemos descomposición en valores singulares, el procedimiento sería el siguiente:

1. Hallar la media por columna(por muestra), de los valores presentes y restar a cada uno de los valores de la columna, es decir$X_i - \bar{x}_i.$, llamémosla $\tilde{G}$
2. Guardar los índices de los valores faltantes, y sustituirlos por cero(sino se centra se sustituyen los valores faltantes por la media de los genes).
3. Hallar la descomposición en valores singulares de la matriz. $\tilde{G} = U \sum V^T$, a ésta considerémosla $\tilde{G}_{old}$, V^T, contiene los eigengenes (vectores singulares) que están relacionados con los eigenvalues que forman la diagonal de $\sum$.
4. Identificar los eigengenes más significantes, usando componentes principales, este valor es conocido como nPcs. Es decisión del investigador fijar cuántas usará. Luego

se divide la matriz, (según lo que teníamos originalmente) en G^C, cuyas columnas no tienen valores faltantes y G^m,las columnas tienen valores faltantes. Obsérvese tenemos el conjunto de datos completos, realizamos SVD y hallamos los mayores valores de U(según el número de componentes principales que consideramos). Dividimos los valores de los pesos en los que son relativos a valores faltantes y los que no, llamémosle A y B respectivamente, usando 5.11 $X = A^+G^C$, ahora los valores de X son considerados para hallar $EST = B^+X$.

5. Ahora en $G^i = (G^C, EST)$, esto significa que obtenemos una nueva matriz, donde los valores de los genes originales no se cambian. Pero en donde tenemos los índices de los valores faltantes, los reemplazamos por los correspondientes EST.
6. SVDImpute es un método iterativo. Es decir, que la matriz en el paso i, al cual reemplazamos los valores faltantes, con ella volvemos hacer los pasos anteriores y obtenemos G^{i+1}, y luego se halla:

$$dif = \sqrt{\sum (\tilde{G}_{Old} - \tilde{G}^i)^2 / \sum \tilde{G}^2_{Old}} \tag{5.19}$$

7. De antemano, se ha considerado una cota. Usualmente es 0.01, esto es, si $dif > cota$,entonces $\tilde{G}_{Old} = G^i$, y retorna al paso tres hasta que se halla G^{i+1}, hasta que dif no supere la cota

Supóngase que obtenemos la matriz $\tilde{G}_{Imp}$, la cual esta centrada, para obtener los datos originales, se suma la media que le habíamos restado en el primer paso.

La parte teórica de este método es tratado en [15], pero su aplicación y difusión fue en el articulo de Troyanskaya et al.,[37], en donde su enfoque es más práctico que teórico.

Una vez se reduce la dimensión, las columnas de la matriz obtenida son conocidas como eigengenes y las filas por eigenarrays. Los autores recomiendan que el número de componentes principales (nPcs) entre 10 y 20. Sin embargo, aseguran que en *SVD*, se afectan los valores de la imputación al hacer el menor cambio de nPcs. Un punto significativo que influye a la hora de usar *SVD* es el hecho de que se debe realizar "gene average", en el primer paso, que es bastante sensible a valores atípicos y a ruidos. Muchas veces, el querer obtener la matriz original a partir de las componentes principales, no es certero. Como dijimos anteriormente la *SVD* es usada con frecuencia para hallar modelos en los conjuntos de datos o explorarla, así que al usar un modelo de regresión se debe esperar que para valores mayores de nPcs se obtiene una buena estimación. Existen otros puntos que requieren atención, entre ellos es que se comporta mejor para los conjuntos de datos de serie de tiempo. Otro punto que no existe un método adecuado para escoger el nPcs óptimo, no existe un método adecuado para ello.

5.4. Imputación usando Componentes Principales Probabilístico

Primero describiremos lo que es análisis factorial, para desarrollar el método *PPCA*.

5.4.1. Análisis Factorial

Como vimos en el modelo de la sección anterior, al reescribir el modelo usando componentes principales, no deja en evidencia que exista ningún error. Por ello muchos autores manifiestan su inconformidad con este tipo de métodos. Así que surge otra clase de modelo en el cual debemos representar factores observables, en función de los no observables, además se supone que siguen una distribución, adicionalmente se agrega un error, este modelo es conocido como análisis factorial, y es el eje principal para los dos métodos siguientes.

La idea básica es el siguiente modelo:

$$t = Wx + \mu + e \tag{5.20}$$

Donde t=Vector q-dimensional de datos observados

W=Una matriz $d \times q$, $d < q$

μ= Es un vector de medias

e=Un vector q-dimensional de errores

Este modelo asume algunas suposiciones como es:

$$x \sim N(0, I)$$

$$e \sim N(0, \psi)$$

Donde ψ es diagonal. Si observamos en 5.20, tenemos que t, tiene un distribución gaussiana, donde:

$$E[t] = E[Wx + \mu + e] = \mu \tag{5.21}$$

$$Var[t] = Var[Wx + \mu + e] = Var[Wx] + Var[e] = WW^T + \psi \tag{5.22}$$

por lo tanto, $t \sim N(\mu, C)$, con $C = WW^T + \psi$

5.4.2. Descripción del Método de Imputación *PPCA*

Como podemos notar el modelo de análisis factorial difiere del PCA, pero algunos autores han destacado ciertas condiciones para hallar una relación entre estos. Lo primero es hacer $\psi_i = \sigma^2$. Young-While consideran σ^2 conocida, así nos queda un modelo para el cual necesitamos determinar dos parámetros, μ y W, lo cual se puede hacer usando mínimos cuadrados. Pero Tipping y Bishop [36] consideran que podemos estimar los tres parámetros, para esto utilizan el algoritmo EM.

Consideremos el siguiente modelo:

$$t|x \sim N(Wx + \mu, \sigma^2 I) \tag{5.23}$$

Según la ecuación anterior y la distribución para $x|t$, usando la regla de Bayes nos quedaría:

$$x|t \sim N(M^{-1}W^T(t-\mu), \sigma^2 M^{-1}) \tag{5.24}$$

donde $M = W^T W + \sigma^2 I$. Tenemos según las dimensiones vista que M es de dimensión $q \times q$.

Un resultado que usaremos con posterioridad es el esperado de la función de máxima verosimilitud para t, la cual viene dada por

$$L = \sum_{n=1}^{N} E[ln(p(t_n))] = \sum_{n=1}^{N} E[-\frac{d}{2}ln(2\pi) - \frac{1}{2}ln|C| - \frac{1}{2}(t_n - \mu)^T C^{-1}(t_n - \mu)] \tag{5.25}$$

$$= \sum_{n=1}^{N} \{-\frac{d}{2} - \frac{1}{2}ln|C| - \frac{1}{2}E[(t_n - \mu)^T C^{-1}(t_n - \mu)]\} \tag{5.26}$$

$$= -\frac{N}{2}dln(2\pi) - \frac{N}{2}ln|C| - \frac{N}{2}tr(C^{-1}S) \tag{5.27}$$

$$= -\frac{N}{2}\{dln(2\pi) - ln|C| - tr(C^{-1}S)\} \tag{5.28}$$

Propiedades de los Estimadores

Ahora debemos estimar los parámetros que no conocemos, que son μ, W y σ^2. Para ello se utiliza el algoritmo EM, el cual es un algoritmo iterativo para maximizar el logaritmo de la función máxima verosimilitud, $L(\mu, W, \sigma^2)$. Este algoritmo se divide en dos partes:

1. **Paso E**:Determinar $E_{x_n|t_n}[lnP(X|\theta)]$ donde $\theta = (\mu, W, \sigma^2)$
2. **Paso M**: Maximizar la expresión anterior con respecto a θ

Estimamos μ de los datos, mediante el cálculo del promedio, esto es: $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} t_n$. Usamos el algoritmo EM para estimar σ^2, W.

La función de máxima verosimilitud, estaría dada por: $L_C = \sum_{n=1}^{N} lnp(t_n, x_n)$
Según nuestro modelo nos quedaría:

$$p(t_n, x_n) = (2\pi\sigma^2)^{-d/2} exp\{-\frac{\|t_n - Wx_n - \mu\|^2}{2\sigma^2}\}(2\pi)^{-q/2} exp\{-\frac{\|x_n\|^2}{2}\} \tag{5.29}$$

La función máxima verosimilitud viene dada por:

$$L_C = \sum_{n=1}^{N} lnp(x_n, t_n) = \sum_{n=1}^{N} (-d/2)ln(2\pi\sigma^2) - \frac{\|t_n - Wx_n - \mu\|^2}{2\sigma^2} - (q/2)ln2\pi - \frac{\|x_n\|^2}{2} \tag{5.30}$$

Para aplicar el paso **E**, según PPCA, hay que hallar el valor esperado a la función anterior y los valores de los parámetros que maximicen la función de máxima verosimilitud. Como queremos maximizar descartamos los valores constantes, entonces

$$E[L_C] = -\sum_{n=1}^{N} E\{(d/2)ln(\sigma^2) + \frac{\|t_n - Wx_n - \mu\|^2}{2\sigma^2} + \frac{\|x_n\|^2}{2}\} \tag{5.31}$$

Vamos a reescribir ese valor esperado usando la siguiente propiedad: Sea Y un valor aleatorio dimensión k tal que $e[Y] = \mu$ y $Var[Y] = V$,entonces $E(Y^T AY) = traza(AV) + \mu^T A\mu = traza(A(V + \mu^T \mu))$. Haciendo el cálculo del valor esperado de la ecuación 5.31, nos quedaría:

$$\begin{aligned} E[L_C] = -\sum_{n=1}^{N} \{&(d/2)ln(\sigma^2) + 1/2tr(\langle x_n x_n^T \rangle) + \frac{1}{2\sigma^2}(t_n - \mu)^T(t_n - \mu) \\ &- \frac{1}{\sigma^2}(\langle x_n \rangle)^T W^T (t_n - \mu) + \frac{1}{2\sigma^2} tr(W^T W \langle x_n x_n^T \rangle)\} \end{aligned}$$

La demostración del valor esperado anterior se puede ver en el apéndice B.

La notación usada en la ecuación anterior sería:

$$\langle x_n \rangle = M^{-1} W^T (t_n - \mu)$$
$$\langle x_n x_n^T \rangle = \sigma^2 M^{-1} + \langle x_n \rangle \langle x_n \rangle^T$$
$$M = W^T W + \sigma^2 I$$

Como estamos usando el algoritmo EM lo que debemos hacer es hallar los estimadores para cada uno de los parámetros, los cuales maximizan la función logaritmo del valor esperado de la máxima verosimilitud, vienen dados por:

$$\tilde{W} = [\sum_n (t_n - \mu)(\langle x_n \rangle)^T][\sum_{n=1}^{N} \{\| \langle x_n x_n^T \rangle \|\}]^{-1} \tag{5.32}$$

$$\tilde{\sigma^2} = \frac{1}{Nd} \sum_{n=1}^{N} \{\|t_n - \mu\|^2 - 2(\langle x_n \rangle)^T \tilde{W^T} (t_n - \mu) + tr(\langle x_n x_n^T \rangle \tilde{W^T} \tilde{W})\} \tag{5.33}$$

La solución para las derivadas anteriores lo podemos hallar en el apéndice B.

Dada las dos soluciones anteriores la podemos reescribir usando:

$NS = \sum_{n=1}^{N} (t_n - \mu)(t_n - \mu)^T$, y $(t - \mu)^T \cdot V \cdot (t - \mu) = tr(V \cdot W)$ donde $W = (t - \mu)(t - \mu)^T$

Consideremos 5.32, nos queda haciendo uso de las ecuaciones anteriores:

$$\begin{aligned}
\tilde{W} &= [\sum_n (t_n - \mu)(\langle x_n \rangle)^T][\sum_{n=1}^{N} \{\|\langle x_n x_n^T \rangle\|\}]^{-1} && (5.34)\\
&= [\sum_n (t_n - \mu)(t_n - \mu)^T W M^{-1}][\sum_n (\sigma^2 M^{-1} + M^{-1} W^T (t_n - \mu)(t_n - \mu)^T W M^{-1})]^{-1} \\
&= [NSWM^{-1}][N\sigma^2 M^{-1} + M^{-1} W^T NSWM^{-1}]^{-1} \\
&= [SWM^{-1}][\sigma^2 M^{-1} + M^{-1} W^T SWM^{-1}]^{-1} \\
\tilde{W} &= [SW][\sigma^2 I + M^{-1} W^T SW]^{-1} && (5.35)
\end{aligned}$$

Lo que nos indica esta ecuación, es que $\tilde{W}$, es un nuevo valor que estimamos a partir del valor anterior, W.

Reescribamos σ^2, tenemos:

$$\begin{aligned}
\tilde{\sigma^2} &= \frac{1}{Nd} \sum_{n=1}^{N} \{\|t_n - \mu\|^2 - 2(\langle x_n \rangle)^T \tilde{W}^T (t_n - \mu) + tr(\langle x_n x_n^T \rangle \tilde{W}^T \tilde{W})\} && (5.36)\\
&\quad \text{Hagamoslo parte por parte} \\
&\longrightarrow \sum_{n=1}^{N} \|t_n - \mu\|^2 = \sum_{n=1}^{N} (t_n - \mu)^T (t_n - \mu) = \sum_{n=1}^{N} tr((t_n - \mu)(t_n - \mu)^T = \\
&\quad = NS && (5.37)
\end{aligned}$$

$$\longrightarrow \quad \sum_{n=1}^{N} \langle x_n \rangle^T \widetilde{W}^T (t_n - \mu) = \sum_{n=1}^{N} (t_n - \mu)^T W M^{-1} \widetilde{W}^T (t_n - \mu) =$$
$$= Ntr(WM^{-1}\widetilde{W}^T S) \tag{5.38}$$
$$\longrightarrow \quad \sum_{n=1}^{N} \langle x_n x_n^T \rangle \widetilde{W}^T \widetilde{W} =$$
$$= \sum_{n=1}^{N} ((\sigma^2 M^{-1} + M^{-1} W^T (t_n - \mu)(t_n - \mu)^T W M^{-1}) \widetilde{W}^T \widetilde{W})$$
$$= tr((N\sigma^2 I + M^{-1} W^T N S W) M^{-1} \widetilde{W}^T \widetilde{W})$$
$$= Ntr((\sigma^2 I + M^{-1} W^T S W) M^{-1} \widetilde{W}^T \widetilde{W}) \quad \text{usando el resultado anterior}$$
$$= tr(WSM^{-1}\widetilde{W}^T) \tag{5.39}$$

Usando 5.37, 5.38 y 5.39, tenemos que:

$$\widetilde{\sigma^2} = \frac{1}{Nd}\{Ntr(S) - 2Ntr(WSM^{-1}\widetilde{W}^T) + Ntr(WSM^{-1}\widetilde{W}^T)\}$$
$$\widetilde{\sigma^2} = \frac{1}{d}\{tr(S - WSM^{-1}\widetilde{W}^T)\}$$

Hallamos los estimadores que maximizan la función de máxima verosimilitud, hasta una cota señalada haciendo iteración, es decir este sería el paso **M**. Esto sería si nuestros datos estuviesen completos, pero como estamos en el problema de imputación, lo que los autores de este método tratan es de alejarse de hallar simplemente máxima verosimilitud para datos incompletos, ya que esto en esencia lo que hace es hallar los parámetros de la distribución que se supone(lo que la mayoría de las veces es la gaussiana). En cambio Tipping y Bishop lo que buscan es hallar la relación con PPCA, y no solo de estimar los parámetros, sino de reducir la dimensión del conjunto de datos.

Lo primero es mirar esa relación, para ello tenemos que C es nuestra matriz de covarianza del modelo y S es la matriz de covarianza muestral, como reduciremos la dimensión del conjunto de datos consideraremos que $C \neq S$.

Supóngase la descomposición en valores singulares: $W = ULV^T$, donde U es una

matriz ortogonal $d \times q$, $L = diag(l_1, l_2, ..., l_q)$, es la matriz diagonal de los valores singulares y V una matriz ortogonal $q \times q$.

$$
\begin{aligned}
C^{-1}W &= (\sigma^2 I + WW^T)^{-1}W = [W^{-1}(\sigma^2 I + WW^T)]^{-1} \\
&= [W^{-1}(\sigma^2 WW^{-1} + WW^T)]^{-1} = [W^{-1}W(\sigma^2 W^{-1} + W^T)]^{-1} \\
&= [(\sigma^2 I + W^T W)W^{-1}]^{-1} = W(\sigma^2 I + W^T W)^{-1} \\
&= ULV^T(\sigma^2 I + VLU^T ULV^T)^{-1} = ULV^T(\sigma^2 VV^T + VL^2 V^T)^{-1} \\
&= ULV^T(V(\sigma^2 I + L^2)V^T)^{-1} = ULV^T V(\sigma^2 I + L^2)V^{-1}
\end{aligned}
$$

Por lo tanto,

$$C^{-1}W = UL(\sigma^2 I + L^2)^{-1}V^T \tag{5.40}$$

Debería cumplirse: $SC^{-1}W = W$, lo que implica:

$$
\begin{aligned}
SUL(\sigma^2 I + L^2)^{-1}V^T &= ULV^T \\
SUL &= ULV^T V(\sigma^2 I + L^2) \\
SUL &= UL(\sigma^2 LL^{-1} + L^2) \\
SUL &= UL(\sigma^2 LL^{-1} + L^2) \\
SUL &= U(\sigma^2 I + L^2)L \\
SU &= U(\sigma^2 I + L^2)
\end{aligned}
$$

Esto implica que $SU = U(\sigma^2 I + L^2)$, donde las columnas de U son los vectores propios de S, para los vectores propios $\lambda_j, j = 1, ..., q$. por consiguiente:

$\sigma^2 + l_j^2 = \lambda_j$, despejando nos quedaría:

$l_j = (\sigma^2 - \lambda_j)^{\frac{1}{2}}$

La descomposición de W puede ser escrito como:

$W = U_q(\Lambda_q - \sigma^2 I)^{\frac{1}{2}}R$, donde R es una matriz ortogonal y Λ_q la definimos como:

$$\Lambda_n = \begin{cases} \lambda_j & \text{el correspondiente valor propio de } u_j \\ \sigma^2 & \text{Si } \lambda_j \text{ es cercano a cero} \end{cases}$$

Conseguimos una relación entre PPCA y análisis factorial.

Luego de sustituir los valores de W , hallamos una función para hallar σ^2, la cual viene dada por:

$\sigma^2 = \frac{1}{q'} \sum_{n=q'+1}^{d} \lambda_j$, donde $l_{q+1} = 0$ y $\lambda_{q+1}, \ldots \lambda_d$, son los valores mas pequeños de S.

Ahora hallamos la relación con SVD, es claro que reducimos la matriz original, de aquí en adelante se trabaja igual que SVD.

5.5. Imputación usando Componentes Principales Bayesianos

Este método fue introducido por Oba et al., en el 2003[27]. Sus siglas son BPCA por su nombre en inglés Bayesian Probabilistic Components Analysis. La idea es introducir métodos bayesianos al modelo anterior, Usaremos análisis factorial, para esto lo que hacemos es suponer que existen factores x_n tal que cada una de las columnas del conjunto de datos, es: $t_n = Wx_n + \mu + \epsilon$, Donde $\epsilon \sim N(0, \frac{1}{\tau}I)$. Vamos a usar el algoritmo EM (Se puede mirar con detenimiento en el apéndice C), entonces debemos calcular la distribución condicional, veamos:

$$P(t|x, W, \tau, \mu) \equiv \frac{P(t, x, W, \tau, \mu)}{P(x, W, \tau, \mu)} = \frac{P(t, W, \tau, \mu)P(x)}{P(W, \tau, \mu)P(x)} = \frac{P(t, W, \tau, \mu)}{P(W, \tau'\mu)} = P(t|W, \tau, \mu)$$

$P(t|x, W, \tau, \mu) \equiv P(t|W, \tau, \mu)$ La ecuación anterior nos dice que maximizar $P(t|x, W, \tau, \mu)$ equivale a maximizar $P(t|W, \tau, \mu)$, pero según esta última la probabilidad buscada no depende de los factores supuestos en el módelo por lo tanto esto podría causar muchas soluciones, alguna sobre estimada. La estadística bayesiana plantea el caso de valores faltantes común problema de estimación y para evitar "overfitting", lo que nos recomiendan es incluir un hiperparámetro, en nuestro caso $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_q\}$ donde cada una de sus componentes es la inversa de las varianzas de las columnas de

W y debe cumplir: $P(W|\alpha) \equiv \Pi_{i=1}^{q}(\frac{\alpha_i}{2\pi})^{-\frac{1}{2}}exp\{-\frac{\alpha}{2}\|w_i\|^2\}$ Hagamos $\theta = \{W, \tau, \mu\}$, nos interesa hallar la probabilidad que relaciona a ambos hiperparámetros θ y α, entonces tenemos:

$$P(\theta|\alpha) = P(W, \mu, \tau|\alpha) \equiv \frac{P(W, \mu, \tau, \alpha)}{P(\alpha)} = \frac{P(W, \mu, \tau, \alpha)P(\mu|\tau)P(\tau)}{P(\alpha)P(\mu, \tau)}$$
$$= \frac{P(\mu)P(W, \tau, \alpha)P(\mu|\tau)P(\tau)}{P(\alpha)P(\mu)P(\tau)} = \frac{P(W, \tau, \alpha)P(\mu|\tau)P(\tau)}{P(\alpha)P(\tau)} = \frac{P(W, \tau, \alpha)P(\mu|\tau)P(\tau)}{P(\tau, \alpha)}$$
$$= P(\mu|\tau)P(\tau)\Pi_{i=1}^{q}P(w_i|\alpha_i, \tau)$$
$$P(\theta|\alpha) = P(\mu|\tau)P(\tau)\Pi_{i=1}^{q}P(w_i|\alpha_i, \tau)$$

Siguiendo con la teoría bayesiana, entonces debemos definir las a priori, esto es:

$$P(\mu|\tau) = N(\mu|\bar{\mu_0}(\gamma_{\mu_0}\tau)^{-1}I_q) \tag{5.41}$$

$$P(w_i|\tau, \alpha_i) = N(w_i|0, (\alpha_i, \tau)^{-1}I_q) \tag{5.42}$$

$$P(\tau) = g(\tau|\gamma_0, \gamma_{\tau_0}) \tag{5.43}$$

Donde $g(\tau|\bar{\tau}, \tau_\gamma)$, denota una distribución gamma con hiperparámetros $\bar{\tau}$ y γ_τ, definido como:

$g(\tau|\bar{\tau}, \gamma_\tau) \equiv \frac{(\gamma_\tau\tau^{-1})^{\gamma_\tau}}{\Gamma(\gamma_\tau)}exp[-\gamma_\tau\bar{\tau}^{-1} + (\gamma_\tau - 1)ln\tau]$

Donde $\Gamma(\cdot)$ es una función gamma.

Los valores usados en las a priori arriba deben considerarse valores iniciales, para que sean funciones a priori no informativa, estos son: $\gamma_{\mu_0} = \gamma_{\tau_0} = 10^{-10}, \bar{\mu_0} = 0 \text{y} \tau_0 = 1$.

La idea es calcular $P(\theta, \alpha, X|T)$, donde T representa el conjunto de datos, el problema es que esa distribución no puede ser simplificada, también hay que estimar X, α y θ, pero de las dos últimas hallamos uno de los hiperparámetros, θ_{ML} haciendo máxima verosimilitud. Luego nos falta hallar los demás, como este problema no es fácil y nos encontramos con integrales que no son fáciles de resolver, por lo tanto usaremos el método Bayes variacional, es decir definimos una distribución q que cumple las propiedades que nos interesa, esto es: $q(X, \theta, \alpha) = q(X)q(\theta)q(\alpha)$, además

dividimos la matriz de datos en T^{obs} y T^{miss}, parte completa y con valores faltantes respectivamente.

Asi, lo que queremos calcular es: $q(X, \theta, \alpha) = P(\theta, \alpha, X|T)$ Entonces, es claro que: $q(T^{miss}) = \int d\theta q(\theta) p(T^{miss}|T^{obs}, \theta)$

El primer paso es hallar la distribución $q(T^{miss})$, el valor inicial de la distribución usando el promedio de las columnas. Luego lo que hacemos es hallar la posterior de θ, $q(\theta)$ usando T^{obs} y $q(T^{miss})$. El proceso se repite hasta que alcance una cota.

La estimación de los valores faltantes se hace, como: $T^{miss} = \int T^{miss} q(T^{miss}) dT^{miss}$.

5.6. Imputación usando Mínimos Cuadrados Locales(LLS)

Este es uno de los métodos más recientes para imputación de datos, publicado en el año 2005 por Kim et al.[14]. En comparación con los demás métodos su parte teórica resulta mas sencilla. Lo primero que debemos considerar es una matrix de expresión de m genes dcon n muestras, llamémosla $\mathtt{G} \in \Re^{mxn}$, por lo dicho anteriormente m$\gg$ n. Esto quiere decir que G se puede representar así:

$$G = \begin{pmatrix} g_1^T \\ g_2^T \\ \vdots \\ g_m^T \end{pmatrix}$$

Además supóngase que existe un valor faltante en el i-ésimo gen, al hacer el j-ésimo experimento, esto convendríamos en denotarlo de la siguiente manera:

$G(i, l) = g_i(l) = \alpha$

Por ejemplo sí el segundo gen tiene un valor faltante en el segundo experimento o muestra, podríamos representarlo: $G(2, 2) = g_2(2) = \alpha$

La manera en que usamos este método se puede dividir en dos partes:

1. Dividir los genes en dos grupos en genes con valor faltante y genes completos.
2. Seleccionar k genes similares al gen que tiene valores faltantes.
3. Usar regresión utilizando los genes hallados en la parte dos para reemplazar los valores perdidos.

Definición 4 (k-Genes Similares)**.** *Los k genes similares de g_j^t , son aquellos cuyo valor absoluto de la correlación con g_j^t es mas alta.La correlación puede ser la de Pearson, que es la usada en nuestros experimentos, pero se puede cambiar por otra correlación.*

Veamos como queda este método paso a paso, si consideramos que tenemos un valor faltante en la primera posición, entonces $G(1,1) = G_i(l) = \alpha$, hallamos los genes mas similares, llamémosle $g_{s_i}^T \in \Re^{1\times n}, 1 \leq i \leq k$.

Si consideramos que tenemos seis experimentos, con un valor faltante en la primera posición nos quedaría:

$$G = \begin{pmatrix} g_1^T \\ g_{s_1}^T \\ \vdots \\ g_{s_k}^T \end{pmatrix} = \begin{pmatrix} \alpha & \mathbf{W}^T \\ \mathbf{b} & A \end{pmatrix} = \begin{pmatrix} \alpha & \mathbf{w}_1 & \mathbf{w}_2 & \mathbf{w}_3 & \mathbf{w}_4 & \mathbf{w}_5 \\ \mathbf{b}_1 & A_{1,1} & A1,2 & A1,3 & A_{1,4} & A_{1,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{b}_k & A_{k,1} & A_{k,2} & A_{k,3} & Ak,4 & Ak,5 \end{pmatrix}$$

Donde α es el valor faltante que queremos imputar y $g_{s_1}^T, \cdots, g_{s_k}^T$ son los genes similares para g_1^T. Ahora queremos imputar el valor faltante, nosotros podemos estimar el cualquier valor de la primera fila de la segunda matriz utilizando regresión lineal esto es de la forma:

$$\min_{\mathbf{x}} = \|A^T\mathbf{x} - \mathbf{w}\|^2 \tag{5.44}$$

y según la ecuación 5.18, y el apéndice A, tenemos que la ecuación anterior tiene solución y esta dada por:

$$\alpha = \mathbf{b}^T\mathbf{x} = \mathbf{b}^T\mathbf{x} = \mathbf{b}^T(A^T)^+\mathbf{w} \tag{5.45}$$

Si quisiéramos estimar los valores que están completos en el gen g_1^T, lo que haríamos es hallar el siguiente modelo lineal $w \simeq x_1a_1 + x_2a_2 + \cdots + x_ka_k$, donde los x_i son los coeficientes del modelo lineal hallados al usar 5.45, así que el valor de α se puede estimar usando: $\alpha = \mathbf{b}^T\mathbf{x} = b_1x_1 + b_2x_2 + \cdots + b_kx_k$.

Hay veces que tenemos mas de un valor faltante para un gen, si consideramos que el gen g_1 tiene q valores faltantes, debemos tomar los K-genes similares para g_1, llamémosle g_{s_i}, para $1 \leq i \leq k$, los cuales deben cumplir las condiciones antes citadas de no tener valores faltantes en los lugares en que g_1 lo tiene. Entonces nos quedarían A una matriz $\Re^{k\times(n-q)}$, B es una matriz de dimensión $k \times q$ y $\mathbf{w}$ un vector $n - q \times 1$, tenemos el mismo caso anterior lo único es que queremos hallar un vector $\mathbf{u}$, de $1 \times q$, par poder sustituir en los valores faltantes, de acuerdo a la ecuación 5.18 y 5.45, nos quedaría:

$$\mathbf{u} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_q \end{pmatrix} = B^T\mathbf{x} = B^T(A^T)^+\mathbf{w}$$

Así que hay que calcular el modelo de regresión y luego estimar los valores para imputar, el problema está en hallar la seudoinversa, pero según lo que vimos en secciones anteriores, ella se calcula usando descomposición en valores singulares.

Los autores de este método Kim et al.,[14] no dan un valor específico para el número de genes similares, como habíamos visto con métodos anteriores sino que consideran que eso depende de los datos. Por regresión sabemos que hay factores que pueden perjudicar la estimación de la variables respuesta, en este caso el vector de imputación, pero sabemos que la gran mayoría de las veces al agregar mas variable en

nuestro modelo podemos tener una sobre estimación de la variable respuesta, conocido como “overfitting”, basándonos en ese principio, se esperaría que a medida que aumenta el K el error nos de mas pequeño.

No hay teoría acerca del mejor valor para K, los autores recomiendan en ese caso es que cambiemos el k hasta que se obtengan resultados óptimos, aseguran además que este método es mejor que los anteriormente mostrados, midiendo NRMSE.

Capítulo 6
METODOLOGÍA Y RESULTADOS

"El poder estudiar, por primera vez en la historia de la humanidad, las tres mil millones de letras del ADN humano- que considero el lenguaje de Dios nos permite vislumbrar el imenso poder creador de su mente "-Francis S. Collins

En este capítulo, hallaremos la metodología que usaremos los conceptos teóricos antes planteados.

6.1. Bases de Datos

Como dijimos anteriormente la mayoría de los artículos publicados dan sus demostraciones con una o dos conjuntos de datos, con características similares. Nosotros trabajamos con seis conjuntos de datos. Se anotarán algunos aspectos característicos de cada una de ellas. Cinco de estas bases de datos son utilizado por Marcel Dettling y Perter Bühlmann en sus papers,[9] y [42] están disponible en su página de internet
http://stat.ethz.ch/ dettling/bagboost.html, se encuentran preprocesados, y son usados ampliamente en investigaciones, las cuales por distintos métodos la han preprocesados, cabe señalar que ellas están completas, no tienen ningún valor faltante. la única que no se halla en la página anteriormente mencionada es Breast Cancer.

6.1.1. Breast Cancer

Este conjunto de datos, contiene la expresión genética de 3226 genes para 22 muestras, tiene tres distintas clases BRCA1 y BRCA 2 y Sporadic. Fue publicada por Hendenfalk et al.,[16].

6.1.2. Colon Cancer

Este conjunto de datos es trabajada ampliamente en el paper de Alon et al.[4]. Originalmente este conjunto de datos que proviene de experimentos con microarreglos de Affymetrix. Contiene 6500 genes humanos, los cuales se han tomado 62 muestras, 40 de ellos tienen tumor y las 22 restantes están sanos. Pero en el articulo anteriormente citado lo utilizan en la evaluación de clustering en datos de Affymetrix, así que lo primero que hacen es preprocesarlo, para ello refiérase a [4]. Al final lo que se obtiene es un conjunto de datos con 2000 genes para 62 muestras. Luego de escoger los genes lo que se hace es una transformación logarítmica y estandarización, para obtener por muestras media cero y varianza uno.

6.1.3. Leukemia

Este conjunto de datos es preprocesado usando técnicas en el siguiente orden: considerando cotas, filtrando, haciendo transformación logarítmica y estandarización. Después de tomarle muestras a 72 pacientes, donde se halla la expresión a 7129 genes, nos quedamos con la expresión de 3571 para 72 pacientes, los cuales presentaban alguno de estos dos tipos de cáncer, leucemia linfomática aguda (acute lymphocytic leukemia, ALL). Los cuales se tienen 47 casos y la otra clase es leucemia mielogena aguda (acute myelogenous leukemia, AML), de las que tenemos 25 casos. Es utilizada ampliamente para clasificación por Golub et al.,[13], muchas veces es conocida como *Golub dataset*. El procedimiento para preprocesarlo aparece en Dudoit et al.,[11].

6.1.4. Lymphoma

Este conjunto de datos contiene 62 muestras de humanos que presentan la enfermedad de linfoma, pero con tres clases. Hay 42 muestras con *diffuse large B-cell lymphoma* (DLBCL), 9 de *follicular lymphoma* y 11 casos de *lymphocytic leukemia* (CLL), en el que se estudian 4026 genes en las diferentes muestras. Este conjunto de datos contiene valores faltantes, pero en la página en que se hallan ya esta completa, debido a que fue preprocesada. Es un resultado de microarreglos punteados. Es una de las primeras bases de datos que se imputaron, asignando cero por los valores perdidos[3].

6.1.5. Prostate

Esta base de datos es ampliamente estudiada por Singh et al.[34]. Se obtienen muchas muestras ya que cáncer de prostata es uno de mas frecuentes en Estados Unidos. Hay 102 muestras dividida en dos clases, sana y con tumor, de la primera hay 50, por lo tanto existen 52 muestras con tumor. Para este tipo de tejido se estudian 6033 genes. Originalmente se consideran 6533 genes, pero variaban con respecto a los demás por lo tanto fueron quitados. Luego de ello se aplicó normalización. Fue diseñada por los laboratorios Affymetrix.

6.1.6. SRBCT

El tipo de cáncer que representa esta base de datos es interesante, ya que según [21], nos dice que hay evidencia que es uno de los tipos de cánceres mas difícil de distinguir. Consta de cuatro clases distintas. Los autores nos dicen que los tipos de cánceres SRBCT, por su nombre en inglés, “small round blue cell sarcomas”, (en español diríamos como sarcomas de pequeñas células redondas pequeñas y azules). No se pueden identificar fácilmente por los medios normalmente usados como es bajo

el microscopio o por pruebas. Es importante el poder distinguir cada uno de los posibles casos ya que dependiendo de ello escoge el tratamiento a seguir y que actúan de forma diversa. La clasificación según la clase del tumor sería: *Ewing family of tumors*(EWS) 23 ,*rhabdomyosarcoma*(RMS) 20, *neuroblasta*(NB) 12 y *non-Hodgkin lymphoma*(NHL) 8. Este tipo de datos es obtenido por microarreglos punteados o cDNA. Originalmente eran 6767 genes para las 63 muestras, al final luego de transformarse quedaron 2308.

Consideramos distintos conjuntos de datos, aunque cabe anotar que todas están preprocesadas, por lo tanto esperamos que tengan poca complejidad y poco porcentaje de ruido.

6.2. Evaluación de Métodos de Imputación

Nuestro objetivo es evaluar los métodos de imputación. El sentido común nos llevaría a usar conjunto de datos con valores faltantes. Pero como queremos es comparar resultados, lo que nos aconseja la literatura es utilizar el conjunto de datos completos. Para comparar resultados entre los completos y la resultante luego de imputar.

1. Tomar un conjunto de datos completo.

2. Medir la calidad del conjunto de datos y la cantidad de ruido

3. Insertar valores perdidos, los porcentajes que consideremos son: 5, 10, 15, 20, 25, 30 %. El valor a eliminar del conjunto de datos, se obtiene al multiplicar el número de genes por el de muestras y obtenemos el % de ese valor a eliminar. Esto es, por ejemplo si tenemos 20 genes y 10 muestras y queremos incluirle el 5 % de valores

faltantes. Hacemos $0{,}05 \times (10 \times 20) = 10$, entonces eliminamos aleatoriamente 10 datos de la matriz de expresiones.

4. Aplicar el método de imputación

5. Utilizar una medida estadística de error la cual permita
   medir cuan cerca estan los datos imputados de los completos el conjunto de datos imputado al real.

El insertar valores faltantes en el conjunto de datos completos. Esto se hace para poder determinar que tan certeros son los métodos de imputación o cuánto nos afectaría en las conclusiones al usar uno u otro método de imputación.

6.3. Raíz Cuadrada Normalizada del Error Medio Cuadrático Normalizado

Por sus siglas en inglés, NRMSE. El error cuadrático medio (MSE)de un estimador $W(x_1, x_2, ...x_n)$ de un parámetro θ, se define como $MSE = E[(W - \theta)^2]$, entonces $RMSD = \sqrt{MSE}$, aquí tenemos la raíz del error cuadrático medio. Por último lo que hacemos es normalizar ese error. Basándonos en que sea ,Y_r y Y_i la primera sea la matriz de expresiones de los valores reales y Y_i la matriz de expresiones de los valores imputada entonces podemos decir que el *NRMSE* vendría dado por:

$$NRMSE = \frac{\sqrt{mean[(Y_r - Y_i)^2]}}{sd[Y_r]} \tag{6.1}$$

Donde Y_r: La matriz original

Y_i: La matriz imputada

Pero este error lo único que hace es mirar cuan alejados están los datos imputados de los datos reales. Sin importar cuan alejados están los resultados si usamos una u otra.

Una de las tareas más importantes para la cual se usa la bioinformática es para hallar genes diferencialmente expresados. La importancia de estos lo vimos anteriormente. Por lo tanto lo que queremos mirar es si usamos un determinado método de imputación cuánto se afectan los genes diferencialmente expresados.

Para ello, veamos el próximo tema:

6.4. Porcentaje de Genes Diferencialmente Perdidos después de la Imputación(PGDP)

Este criterio mide el porcentaje de genes que se pierden al hallarlos con el conjunto de datos imputados con referencia a los genes que salen diferencialmente expresados con el conjunto de datos completo[30]. Veamos pues como debemos hallar este valor:

1. Consideramos el conjunto de datos completos y le insertaremos valores faltantes(le quitamos valores) de forma aleatoria
2. Le aplicamos el método de imputación que queremos evaluar
3. Hallamos los *gdf* , del conjunto de datos completo y del conjunto de datos imputado.
4. Calculamos *PGDP*.

Es decir que ahora necesitamos saber como calcular *PGDP*, lo primero es mirar el número de *gdf* que queremos comparar que estén en los conjuntos, supóngase que sea k, ese número. Es decir miramos los k primeros genes diferencialmente perdidos del conjunto de datos completos *gdfR* y el de los conjuntos da datos imputados *gdfI*, entonces:

$$PGDP = [k - card((gdfR \cap gdfI))]/k \tag{6.2}$$

donde card(A) significa la cardinalidad del conjunto A.

Este criterio refleja lo que hemos anotado desde el principio. Que no sólo hay que tener en cuenta a la hora de imputar datos el error normalizado, sino que tantos

genes se pierden lo que interesa al final.

Notamos en los métodos anteriores que el escoger el parámetro ya lo hallan definido sea K o nPcs, no es tan sencillo. Desde KNN que ha sido estudiado por varias personas y han revelado cual es el mejor intervalo para escoger ese K hasta en LLS que no nos aseguran cuantos genes similares sea mejor, sino que nos dice que cambiando el número de los genes similares se puede hallar el mas óptimo para cada conjunto de datos.

Una de las formas que no hemos descrito con anterioridad de escoger el valor del parámetro, es considerando el plot de los valores imputado versus el valor verdadero y hallando su correlación, debido a que NRMSE, no es tan fiable, así que esa es nuestra primera recomendación.

Lo que nosotros realizamos se puede describir de la siguiente forma, simulamos los valores faltantes, y para cada una de los conjuntos de datos con valores faltantes cambiamos los valores de los parámetros, para poder escoger el K o nPcs. Una vez escogido el valor de K se simulaban valores faltantes, generabamos para cada base de datos 5 conjuntos de datos con valores faltantes en distintos lugares. Se calcula NRMSE y PGDP, al igual que las gráficas de el valor original versus el valor imputado, dada esas condiciones. Dado K promediar los valores para NRMSE y PGDP, y ese es el valor que vamos hallar en cada tabla. El que acompaña al error NRMSE entre paréntesis es el valor que se ha escogido como el mas óptimo para cada uno de los métodos.

Para algunas de las bases de datos aquí descrita se habían considerado en Acuña E y Díaz S.[44]. En esta tesis se han extendido la cantidad de conjuntos de datos y las medidas de calidad.

6.4.1. Resultados de las Medidas de Calidad

Según lo descrito en la metodología, lo que haremos es mostrar en la siguiente tabla las medidas de calidad de cada uno de los conjuntos de datos trabajados, esto es:

Tabla 6–1: Medidas de Calidad de los Conjuntos de Datos

	MEDIDAS DE CALIDAD				
Datos	**Overlap**	**Quality**	**QNN**	**DCFisher**	**MST1**
SRBCT	2.142(5)	0.016(5)	0.063(4)	0.514(5)	0.936(4)
Lymphoma	1.422(6)	0.016(4)	0.016(5)	0.594(6)	0.984(5)
Colon	3.381(2)	0.113(2)	0.258(2)	0.266(2)	0.742(2)
Prostate	3.783(1)	0.362(1)	0.157(3)	0.256(1)	0.843(3)
Breast	2.820(3)	0.045(3)	0.410(1)	0.415(4)	0.59(1)
Leukemia	2.685(4)	0.0149(6)	0.014(6)	0.330(3)	0.986(6)

Los valores en la tabla 6–1, indican la calidad de ruido de cada uno de los conjuntos de datos, entre paréntesis tenemos la posición que ocupa para la medida dada, el último conjunto de datos que le hallamos la medida de calidad es el de Breast. Las que tenían mayor complejidad era Prostate y Colon, mientras que los otros conjuntos, SRBCT, lymphoma,leukemia, tienen poca complejidad. Así al momento de incluir la calidad de Breast ya no se nota la subdivision tan clara, puesto que para algunas está entre las mejores mientras que para otras están entre los peores.

Por lo tanto, de menor a mayor complejidad nos quedarían los conjuntos de datos: Leukemia, Lymphoma, SRBCT, Colon, Prostate. Colocamos a Breast aparte ya que según nuestras medidas no podríamos definirlo en los dos grupos. Una razón para ello es la pocas observaciones por la cantidad de clases que tiene.

6.4.2. Resultados para Porcentajes de Ruido en los Conjuntos de Datos

Por lo que vimos anteriormente, la complejidad esta dado por las instancias que quedan en la frontera o son mas cercanos al centro de otra clase que a la que

está clasificada. Existen algunas instancias que a pesar que cumplen las condiciones para estar clasificadas en una clase, por error humano se clasifican en otra. Veamos el porcentaje de ruido de los conjuntos de datos:

Tabla 6–2: Porcentaje de Ruido

Datos	**% de Ruido**
SRBCT	0
Leukemia	1.38
Lymphoma	1.61
Breast	4.54
Colon	4.83
Prostate	8.82

Los conjuntos de datos en 6–2, están en orden creciente según su porcentaje de ruido. Tenemos que SRBCT no presenta ruido mientras que para Prostate el 8,82 % de las instancias tiene ruido.

El porcentaje de ruido no es muy alto para los conjuntos de datos, ya que están preprocesados.

Describiremos cada uno de los resultados de NRMSE y PGDP para cada conjunto de datos. Se presentará una tabla para cada una de las medidas anteriores, y va acompañado de una gráfica para poder apreciar mejor como cambian las medidas a través de los porcentajes.

El resultado para cada medida del porcentaje dado, va acompañado de un número entre paréntesis lo que indica el valor de K (ya sean genes similares, vecinos más cercanos y nPcs), para el cual se obtiene NRMSE y PGDP mas pequeño.

Recuérdese que los valores de NRMSE y PGDP, en la tabla es el promedio de 5 corridas hechas para el K dado.

Para hallar los valores dados usamos las librerías limma, pcaMethods e impute de R.

Otro punto es que los conjuntos de datos están completos y lo que hacemos es simular valores faltantes, para luego imputar y hallar NRMSE y PGDP. Comencemos a estudiar cada uno de los conjuntos de datos, según lo indicado anteriormente.

6.4.3. Resultados para SRBCT

Los resultados para NRMSE, se hallan el siguiente tabla:

Tabla 6–3: NRMSE para SRBCT

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	.449(55)	.4537(30)	.512(30)	.4678(30)	.5401(25)	.5064(50)
PPCA	.442(55)	.494(30)	.462(30)	.468(30)	.477(25)	.472(50)
SVD	.474(55)	.599(30)	.6418(30)	.6145(30)	.644(25)	.697(50)
KNN	.5425(10)	.5504(10)	.557(10)	.566(25)	.5765(15)	.5851(25)
LLS	.5008(8)	.5301(7)	.5371(7)	.5465(6)	.5634(6)	.5731 (6)

Para poder apreciar mejor los datos de la tabla anterior veamos la gráfica 6–1

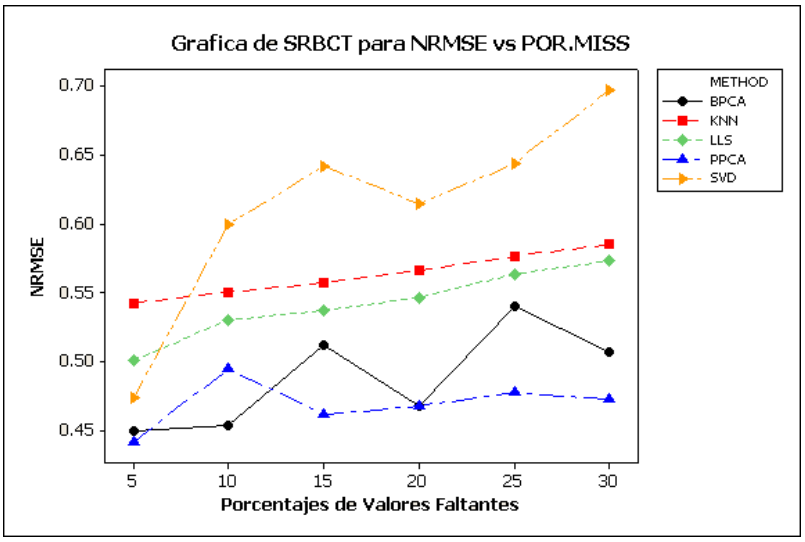

Figura 6–1: NRMSE para SRBCT

Recordemos que SRBCT no tiene ruido. Y es uno de los que presenta menor complejidad. En la gráfica 6–1 notamos que PPCA es el que mejor resultado da para

NRMSE. Sin embargo BPCA lo sigue muy cerca, podemos ver que para 10 % da mejor resultado que PPCA. Mientras que LLS es el que peor resultado da al hallar NRMSE de SRBCT.

A continuación hallamos los resultados al hallar el porcentaje de genes diferencialmente expresados, luego de aplicar cada uno de lo métodos de imputación para SRBCT. En la siguiente tabla:

Tabla 6–4: PGDP para SRBCT

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	5.6	8.2	10.4	13.2	15.0	31.2
PPCA	5.6	8.2	10.4	13.2	15.0	31.2
SVD	11.2	9.8	13.6	15.0	19.4	30.2
KNN	8.0	11.8	16.6	20.6	24.4	27.0
LLS	8.0	8.0	13.0	16.0	22.0	20.0

El gráfico correspondiente a los datos planteados en los valores anteriores es 6–2.

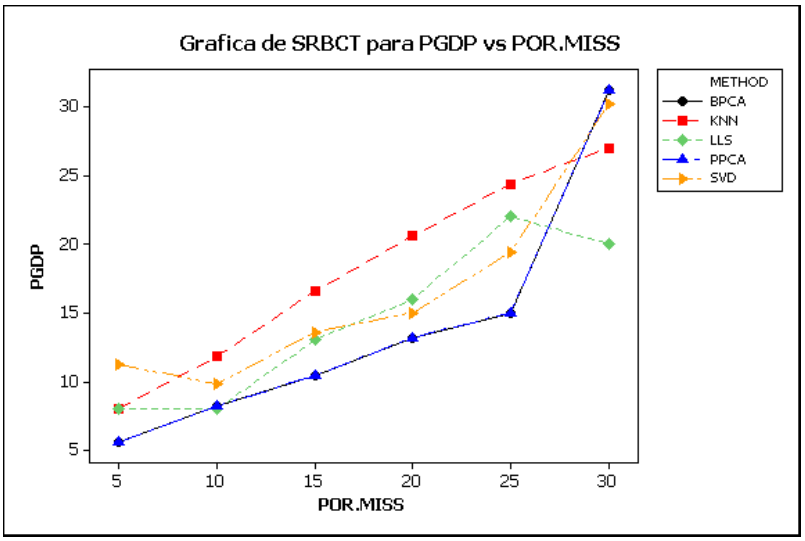

Figura 6–2: PGDP para SRBCT

Tenemos que el BPCA y PPCA, son los que pierden menor porcentaje de genes diferencialmente expresados. Aunque en el 30 % son los que peor resultados dan. En términos generales, KNN es el que peor resultados da para esta medida.

6.4.4. Resultados para Leukemia

En la primera parte tenemos la tabla y la gráfica con respecto al error NRMSE, analicémoslas:

Tabla 6–5: NRMSE para Leukemia

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	.4292(71)	.4733(25)	.4840(20)	.4982(20)	.5116(20)	.5130(35)
PPCA	.4304(71)	.4367(25)	.4436(20)	.4476(20)	.4672(20)	.4685(35)
SVD	.4597(71)	.5031(25)	.492(20)	.5029(20)	.5154(20)	.5644(35)
KNN	.4858(7)	.4896(10)	.4955(12)	.5005(12)	.5055(15)	.5117(20)
LLS	.5020(4)	.4941(6)	.4987(5)	.5084(5)	.5140(5)	.5233(4)

Veamos ahora la gráfica 6–3 que tiene relacionado los valores de NRMSE para leukemia.

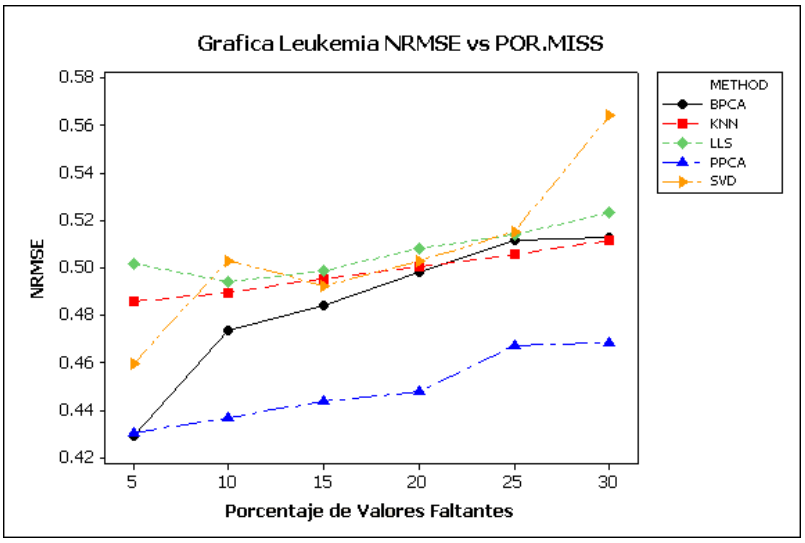

Figura 6–3: NRMSE para Leukemia

Según la gráfica 6–3, es claro que los métodos para imputación PPCA es el que tiene valores mas pequeños de NRMSE. Por lo tanto es el que mejor se comporta.

Mientras que LLS, KNN y SVD, se comportan similares y son con los que peor resultados se obtiene para NRMSE.

La siguiente tabla representa la tabla de GDFP para leukemia, vendría dada por:

Tabla 6–6: PGDP para Leukemia

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	4.4	8.0	10.4	13.8	17.4	20.2
PPCA	4.2	8.4	11.6	12.6	17.0	28.6
SVD	4.6	9.0	10.6	14.2	16.0	24.0
KNN	5.2	1.0	12.4	14.4	17.6	20.2
LLS	7.0	9.0	11.0	14.0	23.0	23.0

La gráfica 6–4 representa los valores de la tabla anterior, veamosla y análicemos.

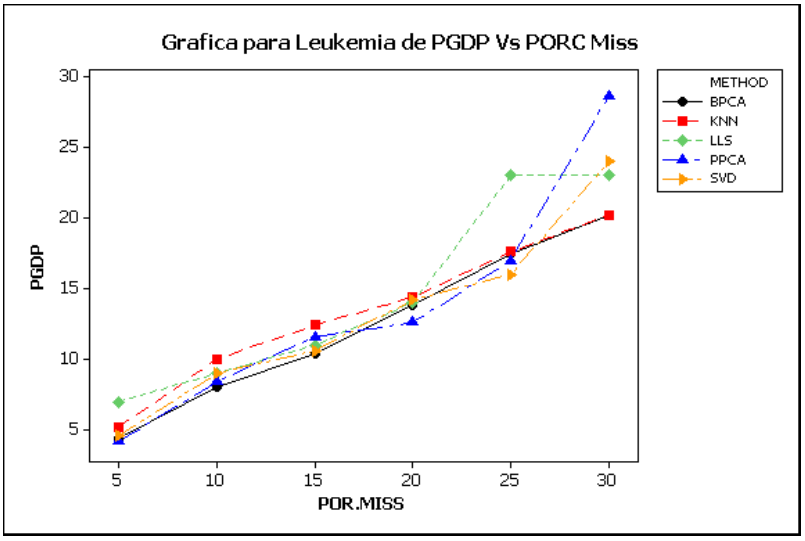

Figura 6–4: PGDP para Leukemia

Salvo algunos valores particulares como en el caso de LLS para 25 % que es el mas alto podríamos decir que el porcentaje de genes diferencialmente expresados es el mismo para cualquier método.

6.4.5. Resultados para Lymphoma

Lo primero que vamos a ver es la tabla referente a Lymphoma para NRMSE, 6–7:

Tabla 6–7: NRMSE para Lymphoma

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	.689(40)	.7073(35)	.6993(45)	.7438(30)	.7524(30)	.778(25)
PPCA	.658(40)	.6684(40)	.6837(45)	.6897(30)	.7621(30)	.782(30)
SVD	.659(40)	.689(35)	.695(45)	.7063(30)	.749(30)	.746(25)
KNN	.683(5)	.6945(3)	.7109(5)	.7182(5)	.7377(5)	.7499(7)
LLS	.6000(7)	.6146(8)	.6338(8)	.6704(7)	.6916(7)	.7196 (7)

La gráfica correspondiente a los valores anteriores, vendría dada por 6–5

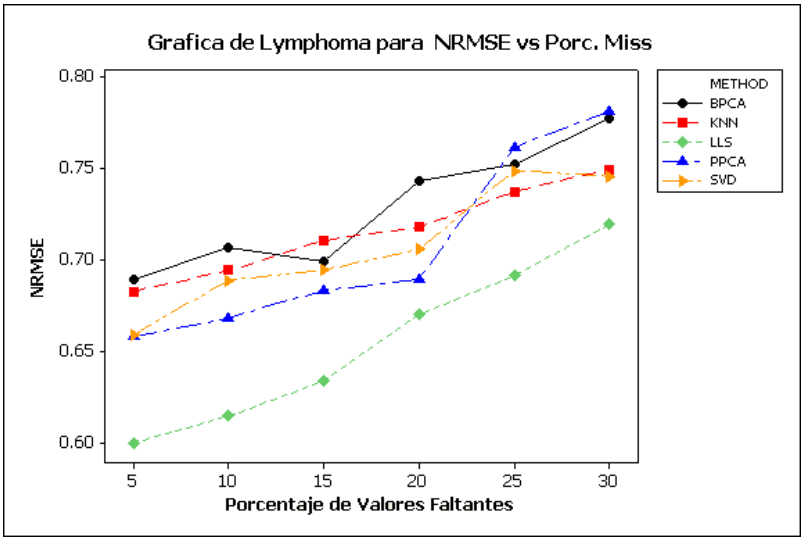

Figura 6–5: NRMSE para Lymphoma

Podemos concluir que LLS es el que mejor resultado da para lymphoma de NRMSE. Mientras que PPCA da buenos resultados para un porcentaje de valores faltantes, del 20 % o menos. Mientras que BPCA en términos generales es el que peor comportamiento tiene. Los otros métodos, tienen un comportamiento similar.

Veamos la tabla 6–8, se hace referencia a los valores obtenidos PGDP después de la imputación:

Tabla 6–8: PGDP para Lymphoma

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	7.8	11.6	14.0	19.6	23.2	28.2
PPCA	7.6	10.4	14.0	16.8	25.8	28.0
SVD	8.2	12.0	22.0	19.8	22.8	27.0
KNN	8.4	12.2	15.2	17.6	21.2	25.8
LLS	9.0	9.0	13.0	18	23.0	24.0

Ahora, para poder observar mejor los datos en la tabla anterior, haremos la siguiente gráfica 6–6:

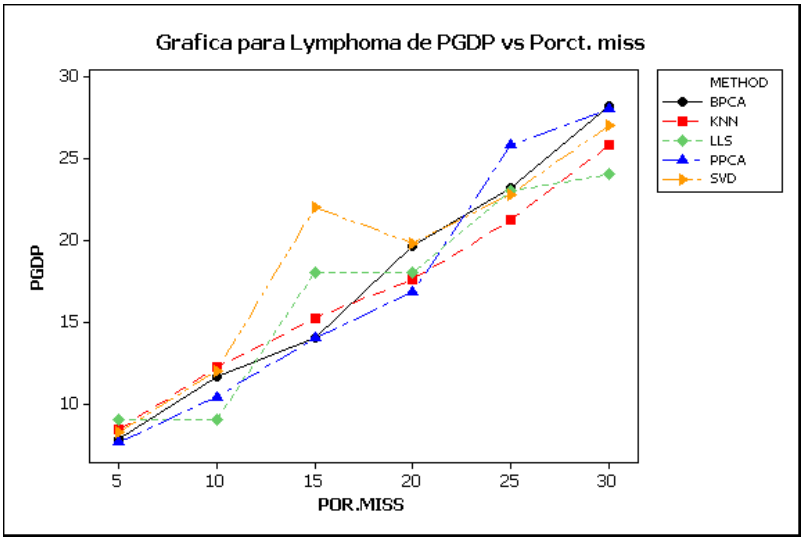

Figura 6–6: PGDP para Lymphoma

Según la gráfica 6–6, los métodos se comportan relativamente igual. Esto es, si nuestro objetivo es hallar los genes diferencialmente expresados de un conjunto de datos con las mismas características que lymphoma. No hay gran diferencia en los resultados al usar cualquier método de imputación.

6.4.6. Resultados para Breast

La tabla 6–9, son los valores de NRMSE para breast:

Tabla 6–9: NRMSE para Breast

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	.593(5)	.652(15)	.716(7)	.664(20)	.7385(20)	.6565(7)
PPCA	.593(5)	.652(15)	.716(7)	.664(20)	.7385(20)	.656(7)
SVD	.593(5)	.652(15)	.716(7)	.672(15)	.7385(20)	.657(7)
KNN	.733(10)	.659(15)	.664(12)	.6795(15)	.760(15)	.714(20)
LLS	.627(15)	.865(12)	.687(5)	.664(10)	.664(10)	.6446(5)

Los valores de la tabla 6–9 están representados en la gráfica 6–7, veamos:

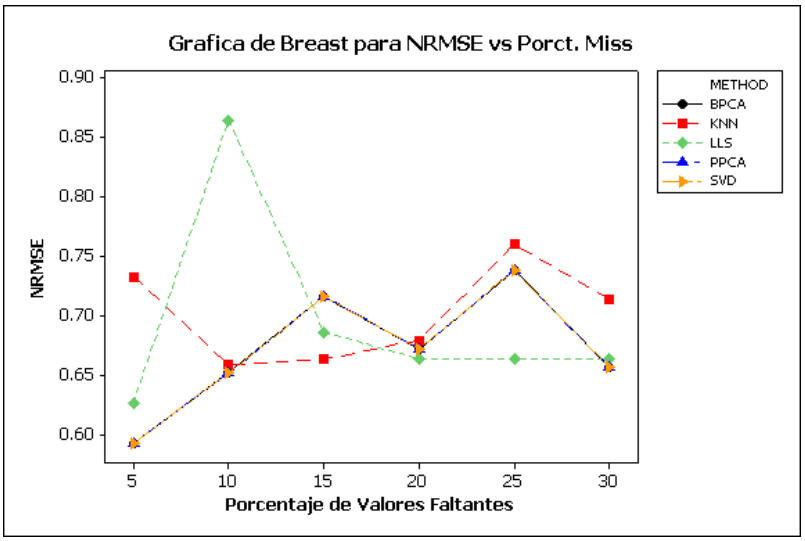

Figura 6–7: NRMSE para Breast

Recordemos que para breast los métodos de calidad no lo pudieron clasificar como de buena o mala. Algo parecido sucede con los métodos de imputación al evaluar NRMSE de este conjunto de datos y es que no podemos determinar cual es el mejor método de imputación.

Ahora miremos que sucede cuando comparamos los genes diferencialmente expresados, al comparar con los hallados con el conjunto de datos original y el conjunto de datos imputados.

La tabla con referencia a PDGP vendría dada por:

Tabla 6–10: PGDP para Breast

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	17.0	23.0	30.0	36.0	38.0	39.0
PPCA	17.0	23.0	30.0	36.0	38.0	39.0
SVD	10.7	23.0	30.0	35.0	38.0	39.0
KNN	15.0	22.0	24.0	32.0	44.0	46.0
LLS	14.0	22.0	34.0	31.0	31.0	46.0

Para poder visualizar mejor los datos de la tabla 6–10, observemos la gráfica 6–8:

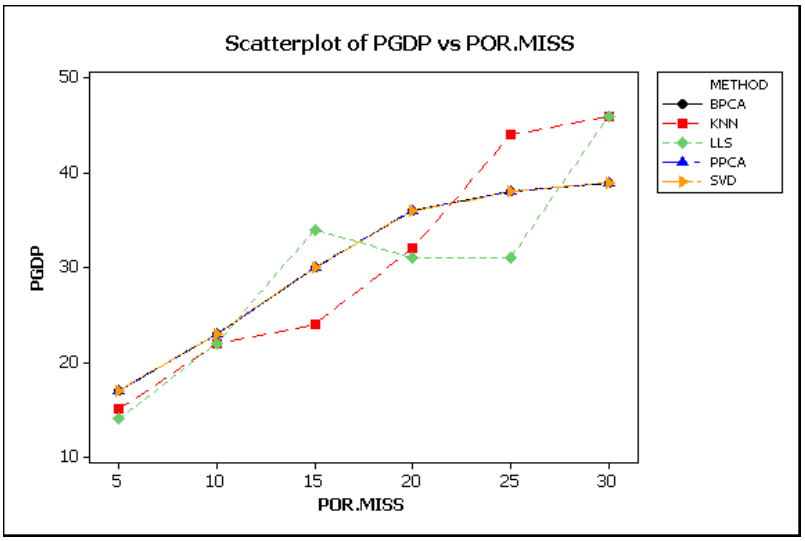

Figura 6–8: PGDP para Breast

Según la gráfica 6–8, no podemos concluir cual de los método es el óptimo.

6.4.7. Resultados para Colon

En las siguientes tablas y gráficas describimos los resultados para colon de NMRSE:

Tabla 6–11: NRMSE para Colon

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	.683(3)	.686(3)	.687(5)	.703(5)	.708(5)	.708(5)
PPCA	.677(3)	.682(3)	.683(5)	.682(5)	.683(5)	.684(5)
SVD	.6812(3)	.687(3)	.704(5)	.705(5)	.708(5)	.711(5)
KNN	.686(15)	.699(20)	.692(20)	.695(25)	.698(25)	.701(25)
LLS	.692(5)	.707(4)	.709(5)	.719(5)	.725(4)	.737(3)

El gráfico 6–9 corresponde a los valores de la tabla anterior.

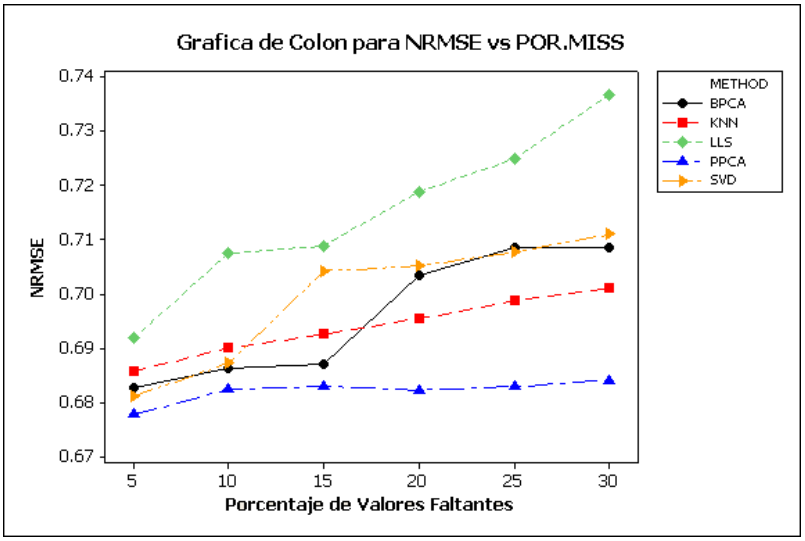

Figura 6–9: NRMSE para Colon

Si observamos detalladamente la gráfica 6–9, tenemos que PPCA es el mejor. Mientras que LLS es el que peor resultados nos da.

Pasemos ahora a PGDP, los valores de dicha cantidad para cada uno de los métodos aplicados a los distintos porcentajes de valores faltantes, sería:

Tabla 6–12: PGDP para Colon

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	6.2	8.4	12.0	14.0	16.6	19.2
PPCA	6.4	9.2	11.6	13.6	15.8	16.6
SVD	6.6	9.2	13.4	15.8	18.8	22.0
KNN	6.4	8.8	10.8	12.6	16.0	16.4
LLS	5.0	8.0	13.0	14.0	10.0	19.0

La gráfica 6–10, tiene valores de PGDP para colon, veamosla y miremos las conclusiones:

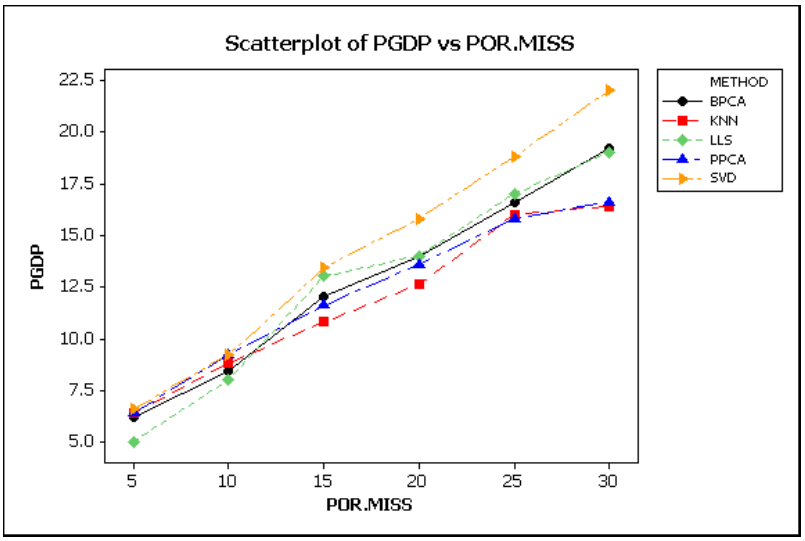

Figura 6–10: PGDP para Colon

En la gráfica, 6–10 no podemos hacer una separación entre los métodos que se comportan mejor y cuales no. Es decir, que cualquier método se comporta similar en cuanto a la obtención de genes diferencialmente expresados.

6.4.8. Resultados para Prostate

Lo primero que veremos es la tabla 6–13 de NRMSE para las distintas imputaciones correspondiente a este conjunto de datos:

Tabla 6–13: NRMSE para Prostate

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	.3497(40)	.3483(80)	.3485(101)	.3494(101)	.3534(101)	.3570(101)
PPCA	.3497(40)	.3483(80)	.3485(101)	.3494(101)	.3534(101)	.3570(101)
SVD	.3497(40)	.3483(80)	.3485(101)	.3494(101)	.3534(101)	.3570(101)
KNN	.3725(12)	.3726(12)	.3755(15)	.3737(12)	.3786(15)	.3815(15)
LLS	.333(5)	.335(80)	.3370(70)	.3405(50)	.3529(25)	.3744(10)

La gráfica 6–11, para estos valores vendrían dado por:

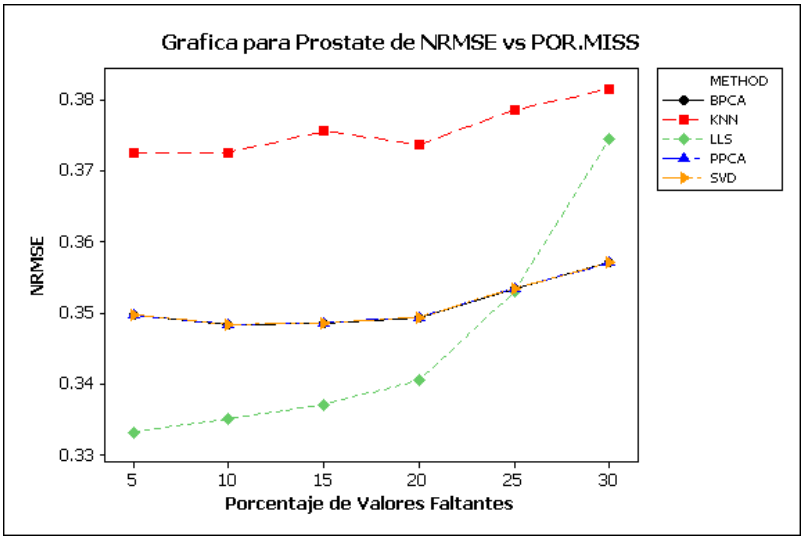

Figura 6–11: NRMSE para Prostate

Según lo que podemos observar el método KNN, es el que peor valores presenta para NRMSE, y que sucede lo mismo que en breast ya que PPCA, BPCA y SVD coinciden. El que mejor se comporta es LLS.

Pasemos a observar la tabla y gráfica correspondiente a PGDP, la tabla 6–12, representa los valores de los porcentaje de genes diferencialmente expresados perdidos después de la imputación, vemos que coinciden muchos valores, como no es tan fácil observar con una tabla, miramos la correspondiente gráfica.

Tabla 6–14: PGDP para Prostate

	PORCENTAJES DE VALORES FALTANTES					
Met	**5**	**10**	**15**	**20**	**25**	**30**
BPCA	5.0	8.0	9.0	19.0	19.0	19.0
PPCA	5.0	8.0	9.0	19.0	19.0	19.0
SVD	5.0	8.0	9.0	19.0	19.0	19.0
KNN	8.0	8.0	12.0	20.0	23.0	22.0
LLS	5.0	7.0	7.0	17.0	16.0	21.0

La gráfica 6–12 corresponde a los datos de la tabla 6–14:

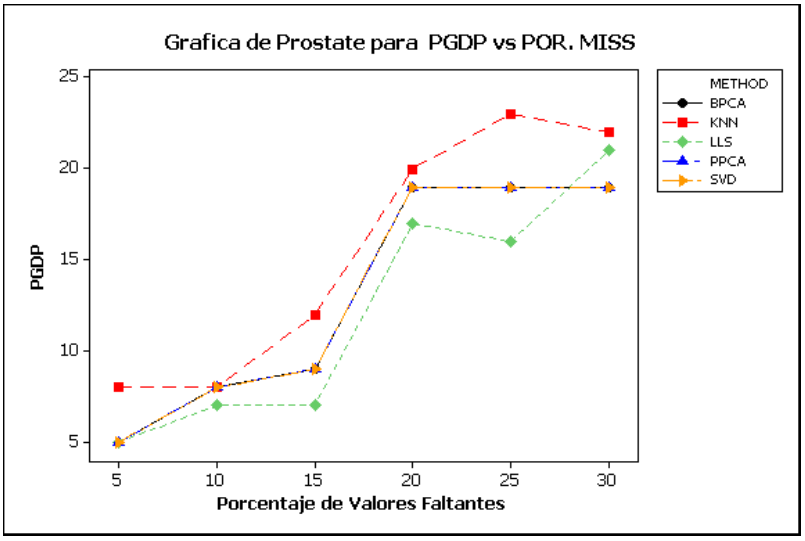

Figura 6–12: PGDP para Prostate

En él miramos que los valores de BPCA, PPCA, y SVD coinciden. Además KNN es el que tienen PGDP mayor, por lo tanto es el que peor resultado da. Mientras, que con LLS es el que mejor resultados da para PGDP.

Nuestra intuición y lo que esperaríamos, es que a medida que aumenta el porcentaje de valores faltantes, NRMSE aumente. Ya que perdería calidad. Esto no sucede. Un ejemplo es en breast en el cual no se nota que a medida que aumentamos los valores en X (porcentaje de valores faltantes)aumente en el eje Y(NRMSE).

Lo anterior se cumple en la otra medida PGDP, ya que en términos generales, a medida que aumenta el porcentaje de valores faltantes,PGDP aumenta.

1. PPCA da óptimos resultados, excepto para lymphoma y prostate.
2. BPCA algunas veces se mantiene cercano a PPCA, aunque para colon y lymphoma, está entre los de peor rendimiento.
3. LLS para lymphoma y prostate da buenos resultados.
4. KNN este método que es ampliamente usado, tiene un rendimiento regular, muy pocas veces está entre los mejores.
5. Interesante que para dos de los conjuntos de mayor complejidad y más ruidosos, colon y breast, todos los métodos tienen prácticamente comportamiento similar.
6. NRMSE Da como resultado en general, que no importa que métodos se utilice, se obtendrían similares resultados para esta medida

Notamos que en breast PGDP es mayor que para todos los demás conjuntos de datos.

6.4.9. Relación de cada método con todos los Conjuntos de Datos

Queremos observar como se comportan los método según el porcentaje de ruido de los conjuntos de datos, para ello graficaremos los valores de NRMSE para los distintos porcentajes de valores faltantes, a lo largo de todos los conjuntos de datos. Pasemos a observar cada una de las gráficas y analizar lo que sucede en cada una de ellas.

En la siguiente gráfica notamos los resultados de NRMSE al aplicar PPCA a todos los conjuntos, tenemos:

En la gráfica 6–13, estan los resultados de todos los conjuntos al usar PPCA.

Vemos que se subdivide en tres grupos, el que mejor se comporta es prostate, mientras los siguientes son srbct y leukemia, y los que tienen NRMSE, mas alto son colon, lymphoma y breast.

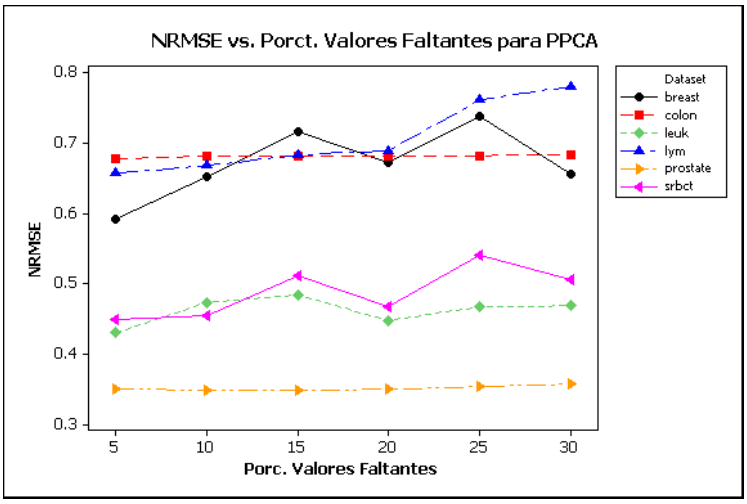

Figura 6–13: NRMSE al usar PPCA para todos los conjuntos de Datos

Tenemos que el comportamiento no parece estar ligado al ruido. Prostate que tiene mayor porcentaje de ruido, para PPCA los valores de NRMSE son menores. Mientras que para lymphoma ocurre lo contrario, este conjunto de datos tiene poco ruido, pero tiene NRMSE más alto.

Casos similares si miramos la gráfica 6–14, correspondiente a LLS para cada uno de los conjuntos de datos:

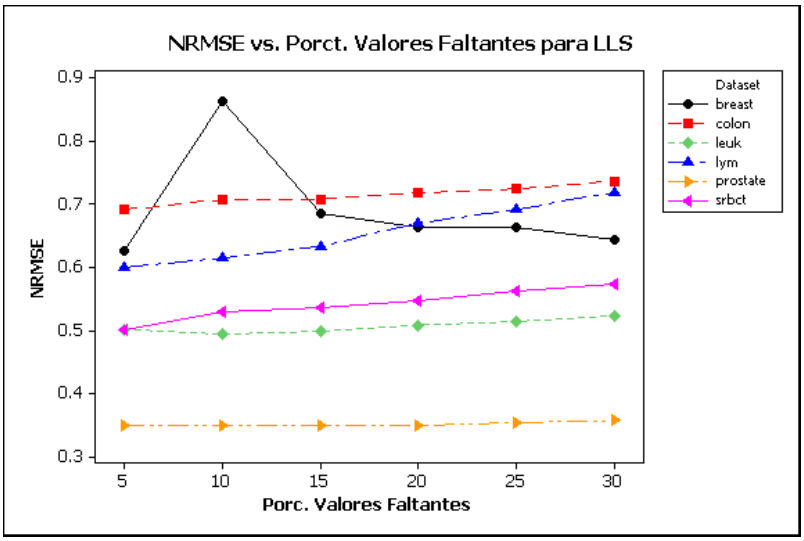

Figura 6–14: NRMSE al usar LLS para todos los conjuntos de Datos

LLS da resultados similares para lymphoma, colon y breast. Aunque entre los conjuntos al aplicarle LLS y obtener NRMSE, no se tiene muchas diferencia.

Observemos los resultados se obtienen al gráficar los valores de NRMSE al aplicar BPCA para cada uno de los conjuntos de datos:

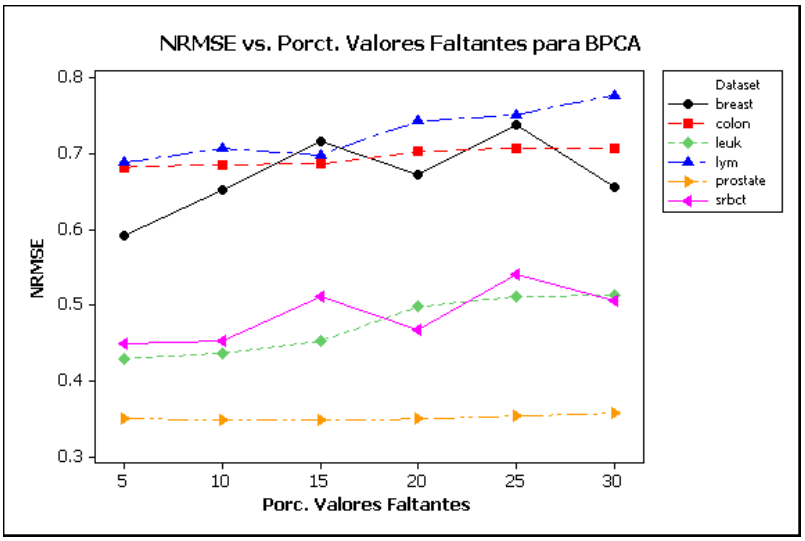

Figura 6–15: NRMSE al usar BPCA para todos los conjuntos de Datos

En la gráfica 6–15, tenemos la misma subdivision que en los otros dos métodos, aunque un poco mas clara que en LLS. También es fácil darse cuenta que lymphoma sigue entre los que peor valor dan NRMSE.

Pasemos a mostrar la gráfica 6–16 correspondiente a SVD:

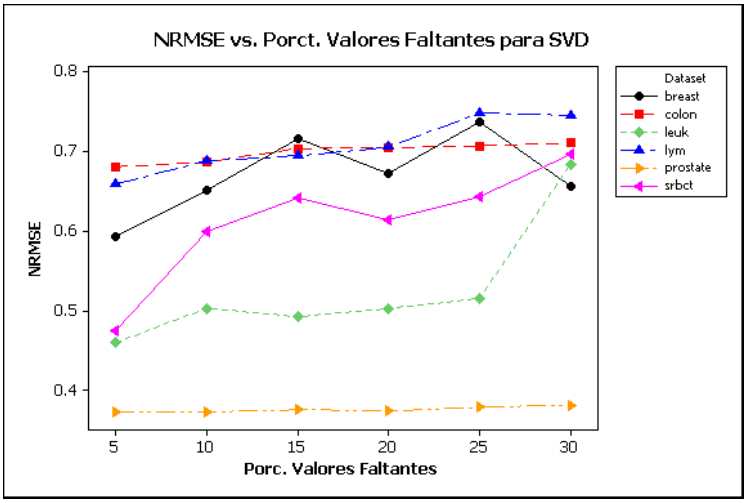

Figura 6–16: NRMSE al usar SVD para todos los conjuntos de Datos

Podemos mirar el cambio entre SRBCT y leukemia,permanecían muy juntos, acá vemos como SRBCT se separa de leukemia, aunque no llega al punto donde están los ruidosos. El patrón que se sigue es el mismo a los datos anteriores.

Por último es la gráfica referente a KNN, veamos que conclusiones tenemos al observar la gráfica 6–17

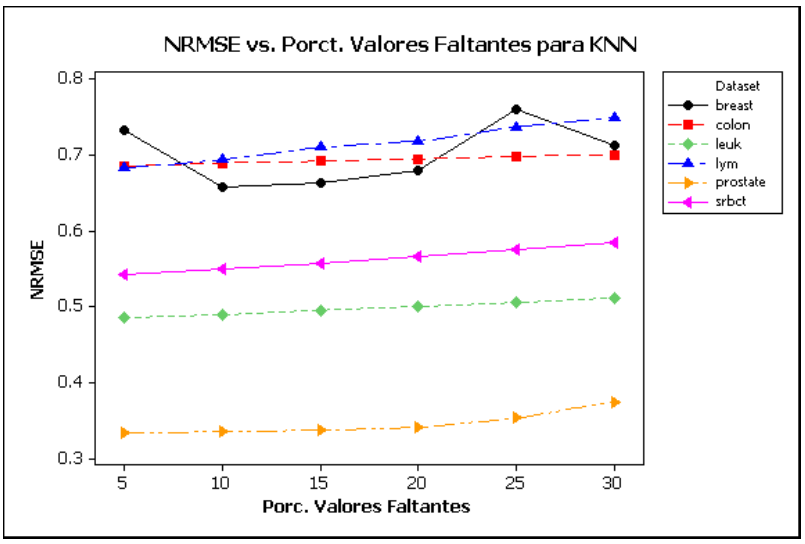

Figura 6–17: NRMSE al usar KNN para todos los conjuntos de Datos

En la gráfica 6–17, también queda claro los mismos resultados que para los métodos anteriores, es así como breast, colon y lymphoma tienen mayor valor para NRMSE, mientras que leukemia y SRBCT están en el medio y prostate es el que tiene menor valor para NRMSE. Si no fuese por lymphoma, tendríamos que existe una relación entre NRMSE y el porcentaje de ruido. Así que no solamente el porcentaje de ruido es el que debe afectar, sino otras características del conjunto de datos, como pueden ser el número de muestras por cada clase o la correlación entre las variables.

Para observar si esta última tiene influencia en los resultados anteriores, realizaremos forma gráfica la matriz de las correlaciones. Antes observaremos las gráficas

relativa a PGDP, para ver si hallamos una relación entre el porcentaje de ruido que presentan los conjuntos de datos y el comportamiento de los métodos de imputación. En la gráfica 6–18 presentamos los resultados con referencia a PPCA.

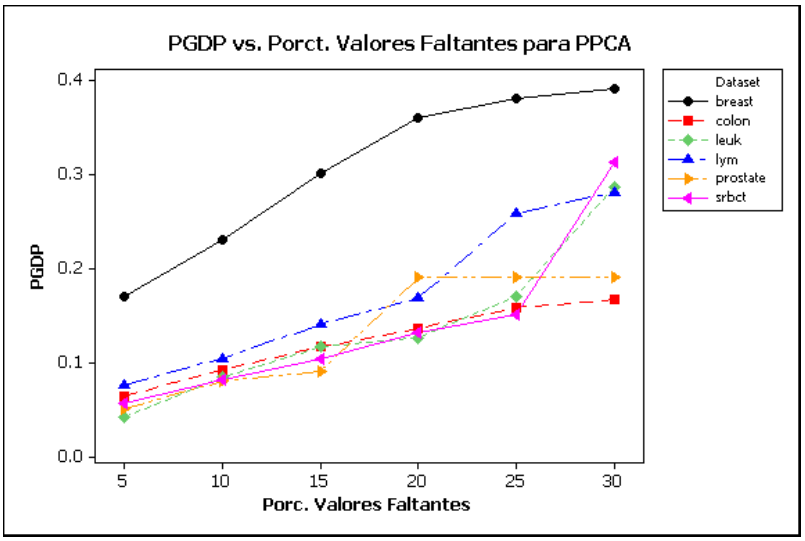

Figura 6–18: PGDP al usar PPCA para todos los conjuntos de Datos

Según esta gráfica el conjunto que mayor número de genes diferencialmente expresados pierde al imputarse usando PPCA, es breast, los otros tienen unos porcentajes similares.

Según la gráfica para LLS, 6–19. El conjunto que tiene mayor PGDP es breast:

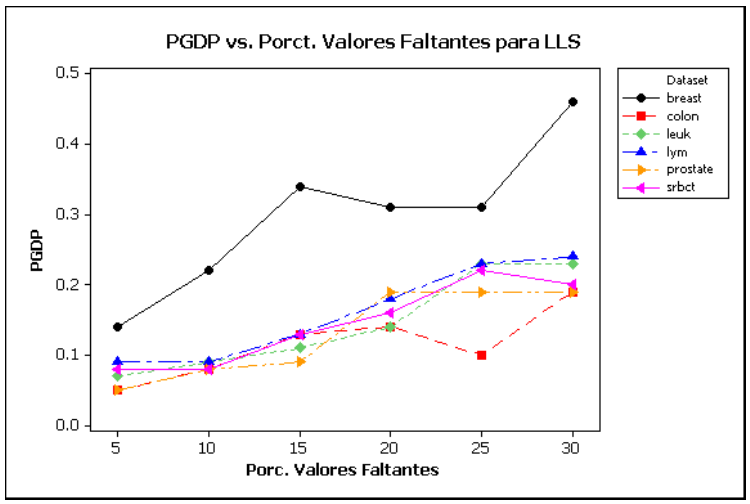

Figura 6–19: NRMSE al usar LLS para todos los conjuntos de Datos

Ahora, según la tablas los valores de BPCA y PPCA son bastantes similares, por lo tanto la gráfica con respecto a BPCA debe comportarse de manera similar a PPCA, observemos en 6–20.

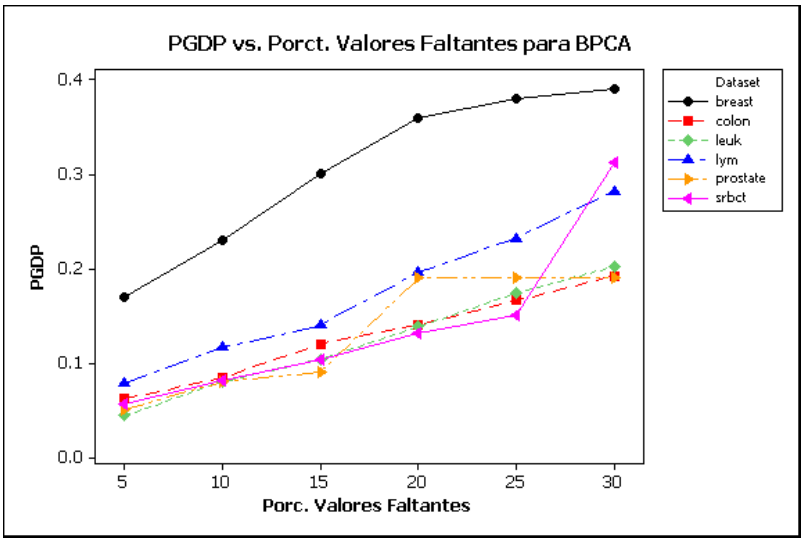

Figura 6–20: PGDP al usar BPCA para todos los conjuntos de Datos

Al observar la gáfica 6–20, nos podemos percatar que tenemos lo presupuestado, es bastante similar a la PPCA y los conjuntos se comportan similar a los anteriores. Luego de esto, miremos la gráfica correspondiente a SVD:

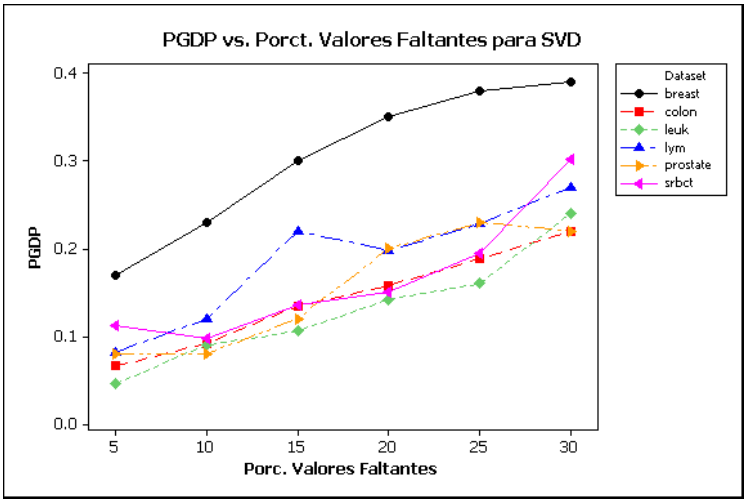

Figura 6–21: PGDP al usar SVD para todos los conjuntos de Datos

En ella notamos que breast está por encima de los demás, es por lo tanto el que mayor número de genes diferencialmente expresados deja por fuera luego de la imputación. Por último tenemos a KNN, para ello miremos la grafica siguiente:

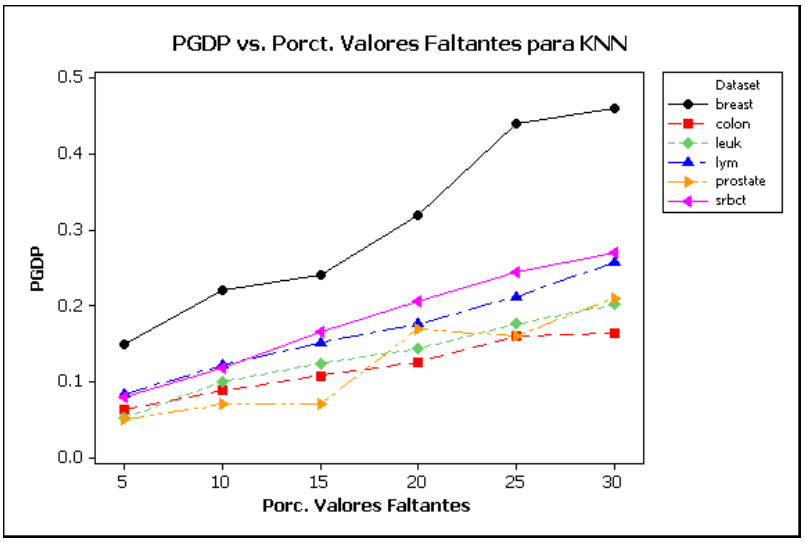

Figura 6–22: PGDP al usar KNN para todos los conjuntos de Datos

Según la gráfica 6–22, tenemos un comportamiento similar al que tenemos para los otros métodos, el porcentaje de genes diferencialmente expresados perdidos luego de la imputación es similar para todos los conjuntos, salvo para breast.

Una de las razones por las cuales podría pasar ello es que breast tiene pocas muestras 22, para tres clases, tiene pocas muestras comprado con los demás conjuntos y la matriz diseño requiere el uso de las clases para poder hallar los genes diferencialmente expresados, según lo hace limma, que es la librería de R usada en esta tesis.

Retornando a NRMSE, las razones no suelen ser tan fáciles para notar el comportamiento de ciertos conjuntos al aplicarle los métodos como lo es para PGDP, así que debemos hallar otras razones, sobre todo para el comportamiento de lymphoma y de prostate, ya que en el primero deberíamos esperar valores pequeños y por el

contrario, grandes valores para NRMSE en el segundo. Así que según lo anterior miraremos la correlación de los conjuntos de datos.

6.4.10. Correlación de los Conjuntos de Datos

Como no es fácil observar la matriz de correlaciones, debido a las dimensiones de los conjuntos de datos que estamos manejando. Las distintas variaciones del color significaran alta correlación, que no existe o es negativa.

Según lo que indica las siguientes gráficas en el margen izquierdo. La gama de azul indica que la correlación es negativa, la de rojo es correlación es positiva y en blanco es que no hay correlación. Los primeros que miraremos son lymphoma y prostate, ya que según lo anteriormente expresados en ellos recae nuestro interés.

Correlación de Prostate

La siguiente gráfica representa la matriz correlación entre las variables de prostate:

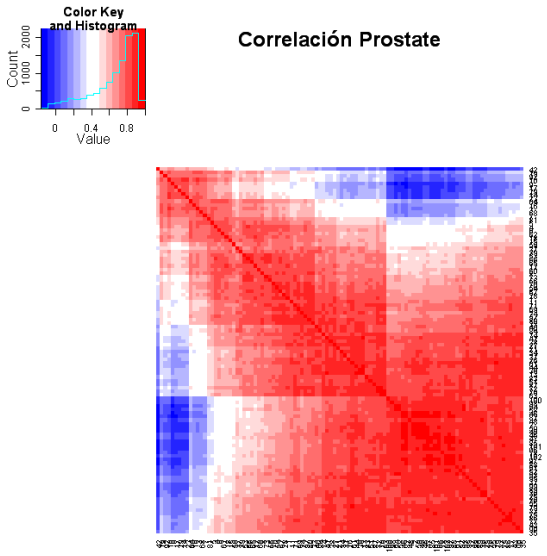

Figura 6–23: Gráfica de la matriz de correlación para Prostate

En la gráfica 6–23, es claro por el rojo intenso que las variables tienen una correlación positiva alta, lo que indica si notamos la tabla de los valores de prostate 6–11, que LLS es el mejor método y esta gráfica parece indicar que es debido a su alta correlación. Ya que LLS usa regresión local y entre mas correlacionadas las variables, mejor su estimación.

Correlación de Lymphoma

Ahora, pasemos a mirar la gráfica de la matriz de correlación correspondiente a lymphoma, en ella tenemos:

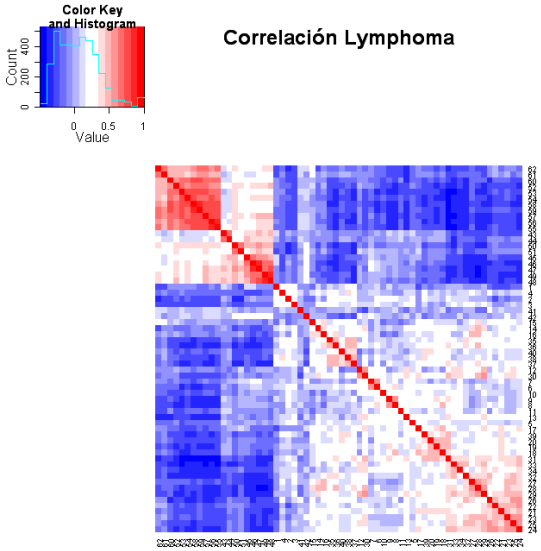

Figura 6–24: Gráfica de la matriz de correlación para Lymphoma

Las variables de lymphoma tienen una alta correlación negativa, siguiendo las razones dadas para prostate, parece explicar porqué LLS, es el de mejor comportamiento para este conjunto de datos.

En cuanto a correlación y ruido las diferencias están dadas, en lymphoma existen bastantes variables con correlación negativa y es de buena calidad. Mientras que para prostate la correlación es positiva, alta y es de mala calidad.

Correlación de Breast

La gráfia 6–25 corresponde a la matriz de correlaciones para breast. Las variables en su mayoría no tienen correlación o con correlación negativa.

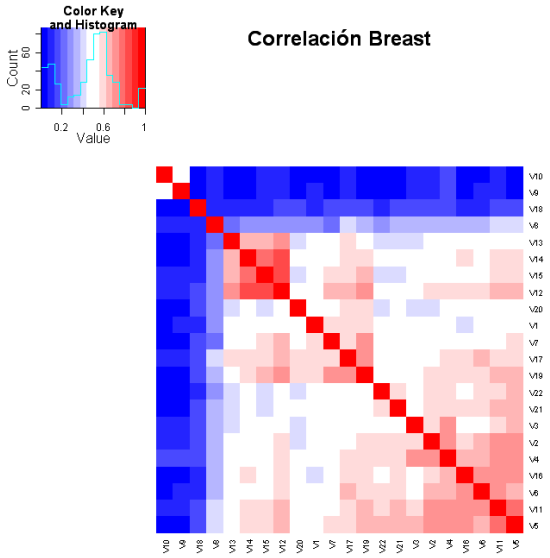

Figura 6–25: Gráfica de la matriz de correlación para breast

Correlación de Colon

Colon tiene mayor complejidad y alto porcentaje de ruido.

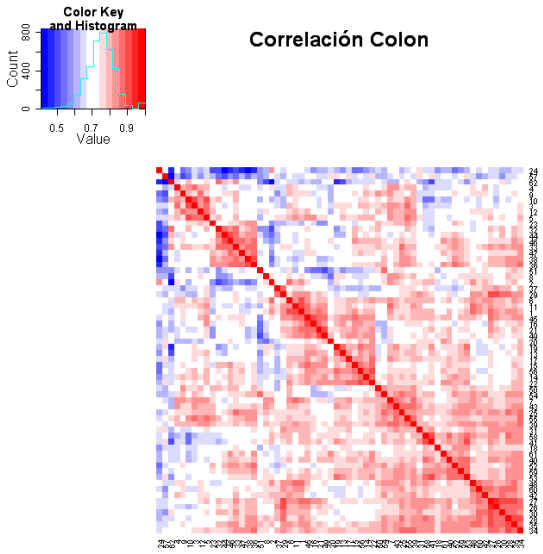

Figura 6–26: Gráfica de la matriz de correlación para colon

Podríamos decir que salvo algunas, no existe correlación entre las variables y por lo tanto si notamos el gráfico 6–9, tenemos que LLS es el que peor resultados obtiene.

Correlación de Leukemia

Miremos el gráfico de correlación para leukemia, veamos:

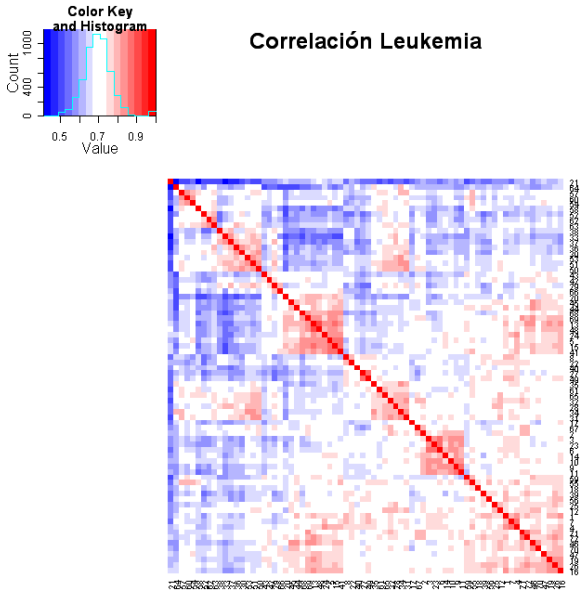

Figura 6–27: Gráfica de la matriz de correlación para leukemia

Según percibimos, están en el mismo porcentaje las variables con correlación negativa y las que no tienen correlación.

Correlación de SRBCT

Recordemos que SRBCT no tiene ruido y es la de mejor calidad. La gráfica correspondiente a su matriz de correlación es 6–28. El porcentaje de las no correlacionadas y las que tienen correlación negativa es similar.

Por esta última gráfica, 6–28 miramos la grandes similitudes que existen entre leukemia y srbct, ya que ambas tienen buena calidad, poco ruido y sus matrices de

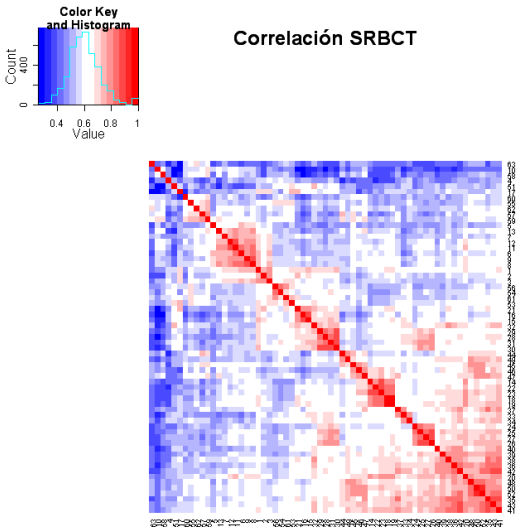

Figura 6–28: Gráfica de la matriz de correlación para srbct

correlaciones son similares, de ahí el comportamiento de los métodos con ella. Luego esta última razón es la que separa a lymphoma de ellas, y por eso es que obtenemos los resultados tan particulares para ese conjunto de datos.

Capítulo 7
CONCLUSIONES Y TRABAJOS FUTUROS

7.1. Conclusiones

Los resultados de nuestro estudio experimental nos conducen a las siguientes conclusiones, estas son:

1. Se utilizó una medida alternativa a NRMSE, la cual es PGDP.
2. El PGDP [30] parece ser un mejor criterio que el NMRSE para evaluar métodos de imputación en datos de microarreglos.
3. PPCA es el mejor método de imputación, ya que resultó dar buenos resultados para cuatro de seis conjuntos de datos.
4. LLS parece no estar influenciada por el porcentaje de ruido y es el que mejor resultados da al calcular NRMSE siempre y cuando las variables del conjunto de datos estén fuertemente correlacionadas.
5. Si el conjunto de datos tiene un alto porcentaje de ruido y mala calidad, PPCA, SVD y BPCA dan los mismos resultados para NRMSE.

Teniendo en cuenta los resultados de NRMSE, podemos decir que recomendaríamos PPCA. Si hay alta correlación entre los genes usaríamos LLS.

Pero, si lo que queremos hallar son los genes diferencialmente expresados de un conjunto de datos que tiene valores faltantes, usaríamos KNN. Esto, se debe a que demora menos que los demás métodos y notamos que para PGDP, todos los métodos tienen similares resultados.

Siempre queda mucho por evaluar, en nuestro caso estamos conscientes que la labor apenas comienza. El comparar métodos descritos por otros autores es importante, ya que podemos determinar cual es la mejor alternativa a usar, dependiendo de la estructura del conjunto de datos. Cuando hablamos de estructura del conjunto de datos, nos referimos a su dimensión, correlación, cantidad de observaciones por clases y porcentaje de ruido. Aunque pueden haber otros factores que se pueden medir.

7.2. Trabajos Futuros

Existe una gama de trabajos posteriores que se pueden realizar, veamos a continuación lo que a nuestro criterio se puede hacer a corto plazo:

- Aumentar la cantidad de conjunto de datos usados para hacer la comparación. Si queremos realizar una evaluación más optima.
- Hacer un análisis computacional de los algoritmos de los métodos de imputación. Sería otra manera de evaluar los métodos. Por ejemplo, para cierto conjuntos de datos KNN se demora minutos o algunas horas. BPCA se demora días.
- Ante los buenos resultados que se notó con el uso de PGDP. Podría agregarse medidas similares evaluando otras tareas. Un ejemplo, sería al hacer clasificación y mirar si cambian de clases las muestras después de realizar la imputación.

APÉNDICES

Apéndice A
MATRIZ INVERSA GENERALIZADA

Consideremos una matriz A, no singular, si la transformamos usando SVD, nos quedaría, $A = U \sum V^T$, su inversa sería: $A^{-1} = V \sum^{-1} U^T$, pero como vimos en las secciones anteriores, nosotros trabajamos con matrices no cuadrada, supóngase que A es dimensión $m \times n$,con $m >> n$, además consideremos que rank r, la matriz $\sum$ sería de orden $m \times n$,de la forma:

$$\sum = \left(\begin{array}{c|c} \sum_1 & 0 \\ \hline 0 & 0 \end{array}\right) = \left(\begin{array}{ccc|c} \sigma_1 & & & 0 \\ & \ddots & & \vdots \\ & & \sigma_r & 0 \\ \hline 0 & \cdots & 0 & 0 \end{array}\right)$$

por lo tanto A no tiene inversa, la solución vino de parte de Moore en 1920 y Perouse en 1955, la matriz es conocida como la seudoinversa de Moore-Perouse, A^+, y se define como:

$$A^+ = V \overset{+}{\sum} V^T donde \overset{+}{\sum} = \left(\begin{array}{c|c} \sum_1^{-1} & 0 \\ \hline 0 & 0 \end{array}\right) \tag{A.1}$$

Para que A^+ definida anteriormente pueda ser llamada seudoinversa debe cumplir ciertas propiedades, veamos:

Sea A una matriz de dimensión $n \times m$, con rank r, X es la seudoinversa de A si:

$$AXA = A \tag{A.2}$$

$$XAX = X \tag{A.3}$$

$$(AX)^T = AX \tag{A.4}$$

$$(XA)^T = XA \tag{A.5}$$

Claramente $X = V \sum^+ U^T = A^+$ cumple con las condiciones anteriores demostremos una de ellas:

$$AX = (U \sum V^T)(V \sum^+ U^T) \tag{A.6}$$

$$= U \sum (V^T V)^T \sum^+ U^T \tag{A.7}$$

$$= U (V^T V \sum^T)^T \sum^+ U^T \tag{A.8}$$

$$= (v^T V \sum U^T)^T \sum^+ U^T \tag{A.9}$$

$$= (U \sum^+ V^T V \sum U^T)^T \tag{A.10}$$

$$= [(U \sum^+ V^T)(V \sum U^T)]^T \tag{A.11}$$

$$= (AX)^T \tag{A.12}$$

similarmente se obtiene que A^+, cumple con las propiedades, por lo tanto es candidata a la inversa generalizada de A, ahora lo que nos falta mostrar es que es única, para ello consideremos X y Y inversas generalizadas de A, así que ambas cumplen

las propiedades de la 1 a 4, por lo tanto:

$$X = XAX = (XA)^T X = A^T X^T X \tag{A.13}$$

$$= (AYA)^T X^T X = A^T Y^T A^T X^T X \tag{A.14}$$

$$= (A^T Y^T)(A^T X^T) X = (YA)^T (XA)^T X \tag{A.15}$$

$$= YAXAX = YAX = YAYAX \tag{A.16}$$

$$= Y(AY)^T (AX)^T = YY^T A^T X^T A^T \tag{A.17}$$

$$= YY^T (AXA)^T = YY^T A^T \tag{A.18}$$

$$= Y(AY)^T = YAY = Y \tag{A.19}$$

comprobamos la unicidad de A^+, esta matriz es interesante ya que la podemos hallar para soluciones que encierran matrices para las cuales es difícil hallar la inversa.

Podemos hallar una lista larga sobre propiedades, existencia y utilidad de este tipo de matrices para mas detalle ,vea [18]

Apéndice B
DEMOSTRACIONES CON RELACIÓN A PPCA.

La primera demostración que tenemos en base a PPCA es la concerniente a la reescritura de el valor esperado de la función de máxima verosimilitud, para ello consideremos:

$$L_C = \sum_{n=1}^{N} lnp(x_n, t_n) = \sum_{n=1}^{N} (-d/2)ln(2\pi\sigma^2) - \frac{\|t_n - Wx_n - \mu\|^2}{2\sigma^2} - (q/2)ln2\pi - \frac{\|x_n\|^2}{2} \tag{B.1}$$

Tenemos que al aplicarle valor esperado nos quedaría:

$$E[L_C] = \sum_{n=1}^{N} \{(\frac{d}{2})ln(2\pi\sigma^2) - E[\frac{\|t_n - Wx_n - \mu\|^2}{2\sigma^2}] - E[\frac{\|x_n\|^2}{2}]\} \tag{B.2}$$

Donde no hemos tenido en cuenta las constantes, teniendo en cuenta que la variable es x que depende de t, las otras se consideran fijas:

$$E[\|x_n\|^2] = E[x_n^T x_n] = tr(\sigma^2 M^{-1} + (t_n - \mu)^T W M^{-1} M^{-1} W^T (t_n - \mu)) \tag{B.3}$$

$$= tr(\sigma^2 M^{-1} + (\langle x_n \rangle)^T \langle x_n \rangle) \tag{B.4}$$

$$= tr(\langle x_n x_n^T \rangle) \tag{B.5}$$

Ahora hallemos:

$$E[(t_n - Wx_n - \mu)^T (t_n - Wx_n - \mu)] \tag{B.6}$$

$$= E[(Wx_n)^T (Wx_n) - (Wx_n)^T (t_n - \mu) - (t_n - \mu)^T Wx_n + (t_n - \mu)(t_n - \mu)] \tag{B.7}$$

entonces si miramos el primer valor esperado es similar a la de B.3, por lo tanto nos queda:

$$E[(Wx_n)^T(Wx_n)] = tr(W^TW\langle x_n x_n^T\rangle) \tag{B.8}$$

utilizando la segunda parte del valor esperado en la ecuación B.6

$$-E[(Wx_n)^T(t_n-\mu)] = -E[x_n^T]W^T(t_n-\mu) = -M^{-1}W^TW^T(t_n-\mu)(t_n-\mu) = -\langle x_n\rangle W^T(t_n-\mu) \tag{B.9}$$

la tercera parte nos quedaría:

$$-E[(t_n-\mu)^TWx_n] = -E[(x_nW(t_n-\mu))^T] = -E[x_n]W^T(t_n-\mu) \tag{B.10}$$

$$= -M^{-1}W^T(t_n-\mu)W^T(t_n-\mu) = \tag{B.11}$$

$$-\langle x_n\rangle W^T(t_n-\mu) \tag{B.12}$$

Podemos sustituir los valores en B.2, por lo tanto nos queda:

$$E[L_C] = -\sum_{n=1}^{N}\{(d/2)ln(\sigma^2) + 1/2tr(\langle x_n x_n^T\rangle) + \frac{1}{2\sigma^2}(t_n-\mu)^T(t_n-\mu)$$
$$-\frac{1}{\sigma^2}(\langle x_n\rangle)^TW^T(t_n-\mu) + \frac{1}{2\sigma^2}tr(W^TW\langle x_n x_n^T\rangle)\}$$

B.0.1. Derivadas de los parámetros del modelo

Según lo visto en la sección donde se desarrolló PPCA, tenemos que:

$$E[L_C] = -\sum_{n=1}^{N}\{(d/2)ln(\sigma^2) + 1/2tr(\langle x_n x_n^T\rangle) + \frac{1}{2\sigma^2}(t_n-\mu)^T(t_n-\mu)$$
$$-\frac{1}{\sigma^2}(\langle x_n\rangle)^TW^T(t_n-\mu) + \frac{1}{2\sigma^2}tr(W^TW\langle x_n x_n^T\rangle)\}$$

Para hallar el valor que maximiza la función anterior debemos derivar, la expresión anterior:

$$
\begin{aligned}
\tfrac{\delta}{\delta\sigma^2}(\tfrac{d}{2}ln\sigma^2) &= \frac{d}{2\sigma^2} \\
\tfrac{\delta}{\delta\sigma^2}(\tfrac{1}{2}tr\langle x_n x_n^T\rangle) &= \frac{2}{2\sigma^2}(\frac{1}{2}tr(\sigma^2 M^{-1} + \langle x_n\rangle\langle x_n^T\rangle)) = \frac{1}{2}tr(M^{-1}) \\
\tfrac{\delta}{\delta\sigma^2}(\tfrac{(t_n-\mu)^T(t_n-\mu)}{2\sigma^2}) &= -\frac{1}{2\sigma^4}(t_n-\mu)^T(t_n-\mu) \\
\tfrac{\delta}{\delta\sigma^2}(-\tfrac{\langle x_n\rangle^T W^T(t_n-\mu)}{2\sigma^2}) &= -\frac{1}{\sigma^4}\langle x_n\rangle^T W^T(t_n-\mu) \\
\tfrac{\delta}{\delta\sigma^2}(\tfrac{1}{2\sigma^2}tr(W^T W\langle x_n x_n^T\rangle)) &= \frac{1}{2\sigma^2}(tr(W^T W(\sigma^2 M^{-1} + \langle x_n\rangle\langle x_n\rangle^T))) \\
= \quad \tfrac{1}{2\sigma^4}tr(W^T W\langle x_n\rangle\langle x_n\rangle^T) &+ \frac{1}{2\sigma^2}tr(W^T W M^{-1})
\end{aligned}
$$

Sumando las derivadas anterior e igualando a cero, nos quedaría:

$0 = -\frac{Nd}{2\sigma^2} - \frac{1}{2}\sum_{i=1}^{N} tr(I) + \frac{1}{2\sigma^4}\|t_n-\mu\| + \frac{1}{\sigma^4}\langle x_n\rangle^T W^T(t_n-\mu) + \frac{1}{2\sigma^4}tr(\langle x_n x_n^T\rangle W^T W)$

Despejando tenemos:

$\frac{Nd}{\sigma^2} = \frac{1}{\sigma^4}(\sum_{i=1}^{N}\|t_n-\mu\|^2 - 2\langle x_n\rangle^T W^T(t_n-\mu) + tr(\langle x_n x_n^T\rangle W^T W))$

La estimación de σ, nos queda:

$\hat{\sigma}^2 = \frac{1}{\sigma^4}\sum_{i=1}^{N}\{\|t_n-\mu\|^2 - 2\langle x_n\rangle^T W^T(t_n-\mu) + tr(\langle x_n x_n^T\rangle)W^T W\}$

Para W se procede de manera similar.

Apéndice C
ALGORITMO EM

Este algoritmo es usado para distintas tareas en estadística multivariada, en particular para imputación de datos, como se utiliza en dos métodos descritos, razones suficientes para dar énfasis en como se utiliza. Lo aquí descrito se halla en [32] y [7]

Describiremos el algoritmo EM para ello consideremos un conjunto de datos $X \in \Re^{n\times p}$, deseamos estimar su media $\mu \in \Re^{p\times p}$,y la matriz de covarianza $\Sigma \in \Re^{p\times p}$ para luego imputar los valores faltantes. Por lo anterior, la matriz X tiene n genes y p muestras, esto es $n \gg p$. Consideremos $x = X_i$ con valores faltantes. Entonces, dividimos la matriz en X_m la parte de X con valores faltantes en el mismo lugar que lo presenta x y X_a formada por los genes completos, sin valores faltantes. Así, consideremos $x_a \in X_a$ donde $x_a \in \Re^{1\times p_a}$ y $x_m \in X_m$ con $x_m \in \Re^{1\times p_m}$.

Ahora, hallamos μ, el cual dividimos en μ_ay μ_m, según la notación anterior. Para cada $x = X_i, (i = 1, ...n)$ con valores faltantes, tenemos el siguiente modelo que relaciones las completas con la que tiene valores faltantes:

$$x_m = \mu_m + (x_a - \mu_a)B + \epsilon \tag{C.1}$$

Donde $B \in \Re^{p_a\times p_m}$, viene dado por los coeficientes de regresión y ϵ $N(0, C)$, con $C \in \Re^{p_m\times p_m}$. Lo primero que debemos hacer es estimar los parámetros para ello usamos maxima verosimilitud, luego se halla B y C; y por último usamos el modelo de regresión anterior para hallar los valores de x_m que faltan.

El algoritmo EM es iterativo, por lo cual lo que hacemos son los pasos anteriores, y

se fija una cota para los valores, el cual una vez la alcanza(ya sea menor o igual que la cota) se suspende el algoritmo.

En el primer paso se calcula μ y Σ, usando los datos de la muestra. Como es iterativo, supóngase vamos en el paso l, tenemos: $\hat{\mu}^l$ y $\hat{\Sigma}^l$ los parámetros, los cuales lo hemos estimado usando máxima verosimilitud. Ahora particionamos la matriz de covarianza $\hat{\Sigma}^l$ en dos submatrices: $\hat{\Sigma}^l_{aa}$ correspondiente a la covarianza de los datos completos y $\hat{\Sigma}^l_{mm}$ para las muestras que tienen datos faltantes. Además, $\hat{\Sigma}^l_{am} = \hat{\Sigma}^{l^T}_{ma}$ La cuales consisten en la estimación de las covarianzas cruzadas de las variable observadas y las que tienen valores faltantes.

Así que la matriz de coeficientes vendría dada por:

$$\hat{B} = \hat{\Sigma}^{-1}_{aa}\hat{\Sigma}_{am} \tag{C.2}$$

De la estructura del modelo de regresión, tenemos:

$$\begin{aligned}
\hat{C} &= \hat{\Sigma}_{mm} + \hat{B}^T\hat{\Sigma}_{aa}\hat{B} - \hat{B}^T\hat{\Sigma}_{am} - \hat{\Sigma}_{ma}\hat{B} && \text{(C.3)}\\
&= \hat{\Sigma}_{mm} + \hat{\Sigma}^T_{am}\hat{\Sigma}^{-1}_{aa}\hat{\Sigma}_{aa}\hat{\Sigma}^{-1}_{aa}\hat{\Sigma}_{am} - \hat{\Sigma}^T_{am}\hat{\Sigma}^{-1}_{aa}\hat{\Sigma}_{am} - \hat{\Sigma}_{am}\hat{\Sigma}^{-1}_{aa}\hat{\Sigma}_{am} && \text{(C.4)}\\
&= \hat{\Sigma}_{mm} + \hat{\Sigma}_{ma}\hat{\Sigma}^{-1}_{aa}\hat{\Sigma}_{am} - \hat{\Sigma}_{ma}\hat{\Sigma}^{-1}_{aa}\hat{\Sigma}_{am} - \hat{\Sigma}_{am}\hat{\Sigma}^{-1}_{aa}\hat{\Sigma}_{am} && \text{(C.5)}\\
&= \hat{\Sigma}_{mm} - \hat{\Sigma}_{ma}\hat{\Sigma}^{-1}_{aa}\hat{\Sigma}_{am} && \text{(C.6)}
\end{aligned}$$

Así tenemos los datos preliminares para cada estimación, por lo dado anteriormente , tenemos que: $x_m \equiv E[x_m|x_a; \hat{\mu}^{(t)}, \hat{\Sigma}^{(t)}]$, considerando el modelo antes descrito y las estimaciones hechas obtenemos: $\hat{x}_m = \hat{\mu}_m + (x_a - \hat{\mu}_a)\hat{B}$ Así con base en lo anterior llenamos los valores faltantes, y miramos si la diferencia entre los datos anteriores y el de paso actual no supere la cota.

Entonces una vez llenado los valores de $x = X_i, i = 1, ..., n$, nos quedaría:

$\hat{\mu}^{t+1} = \frac{1}{n}\sum_{i=1}^{n} X_i$ Esta seria la estimación de la media en el paso $t+1$, la matriz de covarianza, nos queda: $\hat{\Sigma}^{t+1} = \frac{1}{n}\{\hat{S}^{(t)}_i - [\hat{\mu}^{t+1}]^T\hat{\mu}^{t+1}\}$ Es la matriz de covarianza anterior.

Luego que estimamos μ y Σ, estimamos los valores faltantes, usando el modelo de regresión dado. Este es un proceso iterativo, el cual se realiza hasta que los valores imputados y los anteriores no superen una cota dada.[32]

REFERENCIAS BIBLIOGRÁFICAS

[1] *Periodico la Nacion de Costa Rica*, 10 de Junio 2007. Entrevista Realizada a J. Craig Venter.

[2] E. Acuna. Statistical methods for microarray data. Technical report, Universidad de Puerto Rico-mayaguez, http://math.uprm.edu/ edgar, Agosto 2006.

[3] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson Jr, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, L.M. Brown y P.O. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–11, 2000.

[4] U. Alon, N. Barkai, D. Notterdam, K. Gish, S. Ybarra, A. Mack y D. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed oligonucleotide arrays. *Proc Natl Acad Sci*, 96:6745–6750, 1999.

[5] C. Auesukaree. cdna microarray techonology for the analysis of gene expression. *KMITL Sci. Tech. J.*, 6(1):29–34, Jan-Jun 2006.

[6] B. M. Bolstad, R. A. Irizarry, M. Astrand y T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, January 2003.

[7] S. Borman. *The Expectation Maximization Algoritm A short tutorial*, July-October 2006.

[8] C. Chen y J. Chen. *Encyclopedia of Biopharmaceutical Statistics*, chapter Microarrays Gene Expression, pages 599–613. M. Dekker, 2003.

[9] M. Dettling. Bagboosting for tumor classification with gena expression data. *Bioinformatics*, 20:3583–3593, December, 2004.

[10] S. Draghici. *Data Analysis Tools for DNA Microarrays.* Chapman and Hall/CRC, 2000.

[11] S. Dudoit, T. Fridlyand y J. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am stat Assoc*, 97:77–87, 2002.

[12] M. Elliott. Multiple imputation in the presence of outliers. *The University of Michigan Department of Biostatistics Working Paper Series*, pages 1–31, may 2006.

[13] T.R. Golub, P. Slonim, D.K. Tamayo, C. Huard, J.P Gaasenbeek, M. Mesirov, H. Coller, M.L. Lohl, J.R. Downing, M.A. Caligiuri, E.S. Bloomfield y C.D. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–538, 1999.

[14] Kim H., G. H. Park y H. Golub. Missing value estimation for dna microarray gene expression data: local least squares imputation. *Bioinformatics*, 21:187–198, 2005.

[15] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Botstein y Brown. Imputing missing data for gene expression arrays. Technical report, Stanford University, 1999.

[16] D. Hedenfalk, I. Duggan, M. Radmacher, M. Bittner, P. Simon, R. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Loman, N. Pittaluga, S. Gruvberger, O. Johannsson, H Olsson, B. Wilfond, G. Sauter, O. kallioniemi, J. Borg, y P. Trent. Gene-expression profiles in hereditary breast cancer. *Journal of*

Medicine, 344:539–548, 2001.

[17] H. Hoteling. Analysis of a complex of statistical variable into principal components. *Educ. Physhol*, 24:417–441, 1933.

[18] S. J.Leon. *Linear Algebra with Applications*, volume 1. Prentice Hall, 2002.

[19] I.T. Jollife. Discarding variables in a principal components analysis. *Appl. Statist*, 21:160–173, 1972.

[20] D. M. Kelmansky. *Analisis Exploratorio y Confirmatorio de Datos de Experimetos de Microarrays*, 2006.

[21] J. Khan, Wei J., R. Markus, L. Saal, M. Ladany, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, P. Peterson y C. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *nature medicine*, 7:673–679, 2001.

[22] C. Li y W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS*, 98(1):31–36, January 2001.

[23] R. J. Lipshutz, S. P. Fodor, D. J. Gingeras y T. R. Lockhart. High density synthetic oligonucleotide arrays. *Nat Genet*, 21(1 Suppl):20–24, Jan 1999.

[24] D. Lockhard, E. Brown, G. Wong, T. Chee y Markand Gingeras. Expression monitoring by hibridation high density oligonucleotide arrays. *United States Patent*, 1996.

[25] C. D. Medan. *Celula*. Universidad Nacional de Tres de Febrero, Argentina, 2001.

[26] N. Carroll, R. J. Nguyen y D. V. Wang. Evaluation of missing value estimation for microarray data. *Journal of Data Science*, 2:347–370, 2004.

[27] S. Oba, M. Sato, I. Takemasa, M. Monden, S. Matsubara y K. Ishii. A bayesian missing value estimation method for gene expression profile. *Bioinformatics*, 19:2088–2096, 2003.

[28] K. Pearson. On lines and planes of closed fit to system of print in space. *Phil. Mag*, 6:559–572, 1901.

[29] L. A. Daza Portocarrero. *Metodos para Mejorar la Calidad de un Conjunto de Datos para Descubrir Conocimiento*. PhD thesis, University of Puerto Rico Mayaguez, 2007.

[30] Ida Scheel, Magne Aldrin, Ingrid K. Glad, Ragnhild Sorum, Heidi Lyng y Arnoldo Frigessi. The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics*, 21(23):4272–4279, 2005.

[31] M. Schena, D. Shalon, P.O. Davis y R.W. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[32] Tapio Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Atmospheric Environment*, 38:2895–2907, June 2004.

[33] D. Secko. A monk's flourishing garden: The basic or molecular biology explained. *The Science Creative Quarterly*, 2003.

[34] D. Singh, P. Febbo, K. Ross, J. Jackson, C. Manola, J. andLadd, P. Tamayo, A. Renshaw, A. Damico, J. Richie, E. Lander, L. Massimo, T. Sellers W. kantoff y P. Golub. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.

[35] G. Smyth, M. with contributions from Ritchie, J. Silver, J. Wettenhall, N. Thorne, M. Langaas, E. Ferkingstad, M. Davy, F. Dongseok y C. Pepin. *Linear Models for Microarray Data*, 2.10.5 edition, 05 2007.

[36] M. Tipping y C. Bishop. Probabilistic principal component analysis. *Journal of the Royal statistical Society*, B:611–622, 1999.

[37] O. Troyanskaya, M.l Cantor, G. Sherlock, B. Pat, T. Hastie, R. Tibshirani, R. B. Botstein y D. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17:520–525, 2001.

[38] J. D. Watson y F. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *nature*, 17:737–738, Apr 25 de 1953.

[39] Watson JD. y Crick F.HC. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171:964–967, 1953.

[40] JH. Tjio y A. Levan. The chromosome number of man. *Am J Obstet Gynecol*, 130:723–724, 1956.

[41] R. J. A. Little y D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

[42] M. Dettling y Peter Buhlman. Supervised clustering genes. *Genome Biology*, 3:RESEARCH0069, 2002.

[43] S. Dudoit y R. Gentleman. *Introduction to Genome Biology and DNA Microarray Experiments*. Bioconductor, http://www.bioconductor.org, January 2002. Havard School of Public Health.

[44] E. Acuna y S. Diaz. Evaluation of imputation methods for gene expression data. *Proceedings of 56th Session of the International Statistical Institute*, 1, Agosto 2007.