Departamento de Física Aplicada
Universidad de Granada

Métodos Bayesianos para la estimación de redes reguladoras de genes y de perfiles de proteínas a partir de microarrays de expresión genética

Tesis para la obtención del título de doctor por la Universidad de Granada
dentro del programa oficial de *Doctorado en Física y Ciecias del Espacio*

Autor:
Manuel Sánchez Castillo

Directores:
María del Carmen Carrión Pérez[†]
Isabel MaríaTienda Luna[††]
David Blanco Navarro[†]

[†] Departamento de Física Aplicada, Universidad de Granada
[††] Departamento de Electrónica y Tecnología de Computadores, Universidad de Granada

27 de septiembre de 2012

Editor: Editorial de la Universidad de Granada
Autor: Manuel Sánchez Castillo
D.L.: GR 999-2013
ISBN: 978-84-9028-501-5

Department of Applied Physics
University of Granada

Bayesian methods for inferring gene regulatory networks and protein profiles from gene expression microarrays data

Thesis for obtaining the PhD mention at the Universisty of Granada
within the official program titled *Doctorado en Física y Ciencias del Espacio*

Author:
Manuel Sánchez Castillo

Supervisors:
María del Carmen Carrión Pérez[†]
Isabel MaríaTienda Luna[††]
David Blanco Navarro[†]

[†] Department of Applied Physisc, University of Granada
[††] Department of Electronics and Computer Technology, University of Granada

September 27, 2012

Abstract

The genetic inheritance that a living being transmits to its offspring is stored in DNA macromolecules inside the nucleus of prokaryote cells, or in form of RNA in the cytoplasm of eukaryotic organisms. The nucleotide sequence of these nucleic acids encodes the characteristics of each individual of a species and potentially controls its cell development. The part of the code that regulates a feature completely is called a gene and is the hereditary information storage unit. The central dogma of Molecular Biology describes by a unidirectional flowchart the way the genetic information is encoded, stored or transmitted from a living being to its offspring. In prokaryotes, unlike in eukaryotes, DNA is not directly responsible for the cellular development. First, the information stored in DNA is transcribed into a RNA molecule and subsequently it is translated into a protein, that is actively involved in cell metabolism. When the information stored within a gene is translated into a functional protein, it is said that the gene is expressed or activated.

Microarray experiments is an experimental procedure for quantifying the expression of thousand of genes simultaneously. With this technique, gene expression can be massively profiled, leading to huge data sets that are widely available in public databases. Additionally, chromatin inmunoprecipitation as well as other novel techniques combined with gene sequencing allows to identifying gene-protein interactions and performing TF binding site prediction to estimate the topology of the transcriptional network. Whilst most recent experimental and computational techniques allows to measure gene expression and to predict the transcriptional regulatory structure, other kind of biological features of interest such as the gene regulatory network or protein abundance are difficult to estimate. Uncovering the GRN is interesting form the point of view of Systems Biology to understand how genes compete and are associated to produce complex responses and co-operative effects. On the other hand, TF profiles may dissect diseases with characteristic molecular signatures allowing its diagnostic. All these information is extremely important in many fields such as disease treatment and new drug design and it analysis demands help from the Computer Science community.

El doctorando Manuel SÁNCHEZ CASTILLO y los directores de la tesis: María del Carmen CARRIÓN PÉREZ, Isabel María TIENDA LUNA y David BLANCO NAVARRO; al firmar esta tesis doctoral

GARANTIZAMOS:

que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y en la realización del trabajo, hasta donde nuestro conocimiento alcanza, se han respetado los derechos de otros autores a ser citados cuando se han utilizado sus resultados o publicaciones.

Granada, septiembre de 2012.

Directores de la Tesis:

María del Carmen CARRIÓN PÉREZ

.....

Isabel María TIENDA LUNA

.....

David BLANCO NAVARRO

.....

Doctorando:

Manuel SÁNCHEZ CASTILLO

.....

Financiación

El doctorando agradece a los siguientes organismos la financiación que ha permitido llevar a cabo esta Tesis Doctoral:

- Consejería de General de Universidades, Investigación y Tecnología de la Junta de Andalucía. Beca y contrato predoctoral asociado al Proyecto de Investigación de Excelencia con referencia P07-TIC-02589 durante el periodo desde el 01/11/2008 hasta el 31/07/2012. Resolución del 04/11/2008 que modifica el BOJA número 12, de 17/01/2008.
- Consejería de Economía, Innovación y Ciencia de la Junta de Andalucía. Incentivos de carácter científico y técnico individuales de las universidades y organismos de investigación de Andalucía para las modalidades ed estancias de excelencia. Convocatoria 3/2010.
- Valero Energy Corporation. Foreign visiting PhD student award 2010.
- Valero Energy Corporation. Foreign visiting PhD student award 2011.

Figura 1: Viñeta que caricaturiza a William de Ockham, fraile franciscano y filósofo inglés del siglo XIV, a quien se le atribuye el principio de parsimonia también conocido como *la navaja de Ockhan*. Fuente: `http://www.chrismadden.co.uk/`.

Acrónimos

BEFM *Bayesian Expansion Factor Model*, modelo Bayesiano de factores expandido

ChIP *Chromatin Inmuno-Precipitation*, inmunoprecipitación de cromatina

FT factor de transcripción

IID Independiente e Idénticamente Distribuido

MAP Máxima probabilidad A Posteriori

ML *Maximum Likelihood*, máxima verosimilitud

TF *Transcription Factor*, FT

RRG Red Reguladora de Genes

RRT Red de Regulación Transcripcional

TRN *Transcriptional Regulatory Network*, RRT

VBEM Variacional Bayesiano de Esperanza y Maximización

Índice general

Capítulo 1

Introducción

La herencia genética que un ser vivo transmite a su descendencia está almacenada en el interior celular, en macromoléculas de ácidos nucleicos. En concreto, los organismos procariotas almacenan dicha información en moléculas de ácido desoxirribonucleico (ADN) que codifican la síntesis de proteínas con funciones reguladoras del metabolismo celular. La parte del código que controla por completo la síntesis de una proteína se denomina gen y constituye la unidad de almacenamiento de información hereditaria. Sin embargo, el ADN no es el responsable directo de la síntesis proteica. La información codificada en un gen se transcribe a una molécula funcional de ácido ribonucleico (ARN) mensajero que participa activamente en el metabolismo celular. Cuando la información almacenada en un gen es transcrita y finalmente traducida a una proteína, se dice que el gen se ha expresado. Este mecanismo de codificación, transmisión y traducción de la herencia genética se conoce como dogma central de la Biología Molecular.

Todas las células de un organismo contienen la misma información genética. Por tanto, la diferenciación y el metabolismo celular quedan controlados por la actuación conjunta de los genes expresados. La expresión genética es un proceso complejo en el que, a través de distintos mecanismos reguladores, una gran diversidad biomoléculas se asocian e interaccionan para producir respuestas diferentes. Entre estos mecanismos reguladores se encuentran: la iniciación de la transcripción del ADN, la maduración del ARN, modificaciones postranslacionales y la degradación proteica.

Durante la última década del siglo XX la Genética ha avanzado enormemente gracias al desarrollo del microarray, una técnica multiplex que permite cuantificar la expresión genética. Mientras que otros procedimientos experimentales clásicos permiten cuantificar el estado de expresión individualmente o para un numero limitado de genes, los microarrays son capaces de obtener el perfil de expresión de un genóma completo. A este gran avance se suman las técnicas de secuenciación de

nueva generación, que han permitido secuenciar el ADN al completo de organismos tan complejos como el ser humano con más de 20000 genes. Estas tecnologías abren nuevas perspectivas en el estudio y entendimiento del proceso de regulación genética.

Inicialmente, el análisis de micorarrays se limitaba a derivar estadísticos descriptivos y a la clasificación de genes por grupos con patrones de expresión similares [6]. Se ha demostrado que este tipo de estudios resulta útil para identificar qué genes comparten funciones metabólicas. Sin embargo, un análisis más ambicioso trata de entender e identificar el propio mecanismo de regulación. Para ello, es necesario el desarrollo de metodologías, capaces de procesar este tipo de datos, con una base matemática robusta a la vez que tratable desde un punto de vista computacional. Por tanto, el análisis de este tipo de experimentos demandan la ayuda de otras áreas afines a las ciencias de la información, como: Matemáticas, Estadística, Física e Informática.

1.1 Intereses y motivaciones

El gran interés de la técnica del microarray reside en su potencial para cuantificar el estado de expresión del genoma completo de un organismo en un tipo de tejido concreto. Este perfil de expresión constituye una firma biomolecular característica del estado celular de un individuo. Este tipo de información es un candidato perfecto para analizar los orígenes genéticos de la diferenciación y el metabolismo celular [89] [40]. Por otro lado, resulta de especial interés el estudio de una patología a niveles moleculares mediante un análisis comparativo, entre el perfil de expresión de un paciente sano frente y otro que padece una enfermedad concreta [55].

En Biología de sistemas, resulta interesante estudiar los mecanismos de regulación genética desde un punto de vista fenomenológico. Este enfoque integra la regulación genética en el funcionamiento de un proceso biológico desde una perspectiva funcional. Un modelo muy común usado en este tipo de descripciones sistemáticas del metabolismo celular son las redes de regulación [91]. Una red reguladora es un modelo abstracto en el que sus elementos interaccionan para producir diferentes respuestas. Esta descripción de un proceso biológico permite a los investigadores estudiar y describir los orígenes de los distintos procesos metabólicos. Por otro lado, desde el punto de vista de la Biología molecular es interesante conocer el mecanismo de regulación a un nivel más detallado. En lugar de analizar su funcionamiento, esta perspectiva analiza los origines del desarrollo celular a niveles moleculares [54]. Bajo este enfoque, se estudia cómo la celula es capaz de integrar diferentes señales químicas y el modo en el que interaccionan para producir diferentes respuestas.

Conocer la maquinaria celular a diferentes niveles, funcionales y moleculares, es de enorme interés en la investigación médica y en la industria farmacológica [94] [72]. Por un lado, esta información podría usarse para identificar una patología a partir de una firma molecular concreta. Por otro lado, conocer las respuestas celulares ante diferentes señales químicas permitiría diseñar medicamentos y terapias específicas, más eficientes en el tratamiento de enfermedades.

1.2 Trabajos previos

El análisis de datos de microarray ha avanzado relativamente poco frente a la gran difusión y las mejoras conseguidas en esta técnica experimental [40]. Entre los primeros análisis, destacan los problemas de ingeniería inversa para la estimación de las redes de regulación genética [80]. En estos métodos, los diferentes mecanismos de regulación a niveles transcripcional, translacional y postransduccional se proyectan al espacio de interacción genética. Este modelo abstracto simplifica el fenómeno de regulación genética a una serie de relaciones causales entre los genes que componen la red. La primera aproximación de la que se tiene constancia en la literatura es el modelo de las redes Booleanas probabilistas. Basándose en el modelo de las redes Boolenas [43], en su extensión probabilista, la expresión genética se cuantifican con dos estados: activación e inhibición [84]. En este modelo, el mecanismo de regulación se describe mediante una serie de reglas lógicas que van acompañadas de una medida de incertidumbre. Una aproximación posterior son las redes Bayesianas [32]. Este modelo proporciona una representación gráfica al modelado de las redes de regulación, acompañado de un formalismo matemático adecuado. En una red Bayesiana, el estado de expresión de un gen se adapta a las medidas proporcionadas en los experimentos de microarray y permiten describir las relaciones de independencia y regulación entre los genes que componen la red [38]. Además, este método goza del formalismo Bayesiano que permite modelar el ruido experimental y complementar los datos de expresión con otro tipo de información biológica relevante [89]. Otras aproximaciones importantes prescinden de una descripción de la red de regulación y tratan de explicar la variabilidad de los datos mediante modelos cinéticos [23] [22]. Estas aproximaciones resultan de interés en la simulación de datos de expresión sintéticos. Los modelos presentados anteriormente se han aplicado con éxito en datos sintéticos y en datos reales de microarrays para la estimación de la red de regulación genética en sistemas biológicos modelos [90] [28].

Con la aparición de bases de datos con otro tipo de conocimiento biológico, surge la necesidad de modelar esta información e incluirla en el análisis de datos de microarray. En concreto, las bases de datos de interacciones entre genes y proteínas permiten modelar la regulación genética a un nivel molecular más detallado

desde un enfoque transcripcional [53]. La transcripción genética es uno de los primeros procesos regulación de la expresión. Este fenómeno está mediado por una serie de proteínas funcionales que interaccionan directamente con el ADN, favoreciendo o inhibiendo el ensamblaje del complejo proteico encargado de transcribir el ADN a ARN mensajero. La descripción de las redes reguladoras transcripcionales comparte algunas propiedades de las redes genéticas pero posee una formulación propia e independiente. La estructura de esta red se estima a través de algoritmos de predicción de alineamiento de las proteínas y secuencias genéticas [44]. Una vez aprendida, la red transcripcional puede incorporarse al modelado de microarrays para estimar otro tipo de información biológica de interés, como la abundancia de las proteínas resultantes de las expresión genética. La mayoría de los análisis de datos microarray que integran redes transcripcionales consideran un modelo de factores ocultos. Las diferencias entre estas metodologías reside principalmente en el modelado de la red y el método de resolución. Por un lado, las técnicas clásicas como el análisis de componentes independientes [47] y el análisis de componentes principales [37] permiten descomponer los datos de expresión en un espacio condensado con diferentes contribuciones a la regulación genética. Otro tipo de métodos similares, como el análisis de componentes de redes [46] y la factorización no negativa [26] imponen una serie de restricciones al modelo, que tratan de resolver problemas de indeterminación inherentes al análisis de factores latentes, que descomponen los datos de expresión en un espacio de metagenes [12]. Esta aproximaciones, aunque aplicables en algunos casos concretos, no consiguen describir eficientemente la red en un contexto de regulación transcripcional realista.

1.3 Dificultades y objetivos

El modelado y la inferencia de los mecanismos de regulación mediante las técnicas clásicas de procesado de señal es un problema difícil. En primer lugar, los datos de microarray contienen una gran cantidad de ruido experimental. Además, existen otras fuentes de ruido, procedentes de variaciones estocásticas inherentes a los procesos biológicos, que no corresponden con el mecanismo de regulación real. Por otro lado, los datos de microarray proporcionan una visión parcial del fenómeno de regulación al completo. Finalmente, el número de muestras disponibles es muy limitado en comparación al número de genes considerados en estos tipos de problemas. Debido a este tipo de dificultades, el análisis de datos microarray apenas ha avanzado, en comparación a su difusión y al gran desarrollo de la técnica experimental.

Tanto en el problema de ingeniería inversa como en la estimación de los productos de la expresión genética, uno de las dificultades principales es el coste computacional. Para un conjunto con un gran número de genes, el tiempo de cálculo puede

llegar a hacer prohibitivo el método de inferencia. Este tipo de problemas afectan acusadamente a métodos basados en la simulación secuencial de las incógnitas del problema, como el muestreo de Gibbs. Las soluciones propuestas pasan por reducir el número de variables a estimar apoyándose en información a priori sobre la estructura de la red. La metodología Bayesiana ofrece un formalismo óptimo par modelar e incorporar este tipo de conocimiento biológico a priori. Además, este marco de trabajo permite enriquecer los datos de microarrays con otro tipo de información a priori que permitan adaptar el problema a diferentes aplicaciones. Conocida la estructura de la red, otra aproximación puede ser la transformación de los datos a un espacio de dimensiones reducidas que permita recuperar las estimaciones en un contexto biológico real. Como alternativa a la reducción de variables, se propone una metodología subóptima basada en reglas de aprendizaje de bajo coste computacional.

Expuestas la dificultades y posibles soluciones a los problemas de modelado de datos de microarray para la inferencia de la red de regulación genética y la estimación de perfiles de proteínas, se propone una serie de aproximaciones que cumpla con los siguientes objetivos:

- Modelar los datos microarrays eficientemente, con un modelo capaz de diferenciar entre ruido y las diferentes contribuciones del efecto de regulación real. En el caso de series temporales de microarrays, con un modelo que describa la red de regulación genética y parametrice adecuadamente el ruido. En el caso de datos estáticos de microarrays, con un modelo que no imponga ligaduras estrictas en la red transcripcional para conservar el contexto biológico de las interacciones.

- Implementar un método de inferencia de la red reguladora genética de bajo coste computacional basado en el algoritmo variacional Bayesiano para el aprendizaje de series cortas de microarrays temporales.

- Desarrollar un marco de trabajo Bayesiano para modelar la red transcripcional y datos estáticos de microarray para la estimación de la actividad proteica en un espacio de dimesnioes reducidas que permitan aplicar las metodologías de muestreo.

1.4 Estructura de la memoria

En el Capítulo 2 se presenta a modo descriptivo, para detalles y definiciones formales se reseñará a la bibliografía, una serie de conceptos fundamentales de genética y biología molecular; imprescindibles para comprender los mecanismos de codificación, interacción y cuantificación del material genético que condicionarán el posterior modelado de la información.

En el Capítulo 3 se introducen los fundamentos y algunos métodos de la estadística inferencial con el fin de repasar algunos conceptos importantes, los más básicos se suponen conocidos, y presentar al lector la notación usada en el resto del texto.

En el capítulo 4 se propone una aproximación novedosa en modelado de la red reguladora de genes y su aprendizaje a partir de series de microarrays temporales. La metodología presentada enmarca un modelo autoregresivo de media móvil en el formalismo Bayesiano y hace uso del método variacional para la estimación de la red de regulación genética. El método propuesto se valida con datos artificiales mediante simulación y con un generador de datos sintéticos. Además, los resultados se comparan con otras metodologías para este tipo de problemas de ingeniería inversa.

En el Capítulo 5 se aplica el método variacional desarrollado en el Capítulo 4 a un conjunto de datos reales. En concreto, se consideran datos de un organismo modelo como la levadura, para la cual se conoce la red de regulación en diferentes procesos metabólicos. Las estimaciones obtenidas muestran resultados satisfactorios en datos reales.

En el capítulo 6 se presenta un modelo Bayesiano de factores latentes expandidos que integra datos a priori de la red transcripcional para el análisis de datos de microarrays y la estimación de perfiles de proteínas en un espacio de dimensiones reducidas. El método propuesto incorpora eficientemente información a priori de la red trasncripcional y prescinde de ligaduras rígidas que permitan estimar las incógnitas mediante la técnica del muestreo de Gibbs. Esta aproximación se valida mediante simulación de datos sintéticos de microarray.

En el Capítulo 7 se aplica el método de factores expandidos desarrollado en el Capítulo 6 a un conjunto de datos reales. En concreto, se consideran datos de pacientes afectados de tumores de mama para estimar los perfiles de proteínas y clasificar la patología en subtipos con implicaciones relevantes en su diagnostico y prognosis.

Finalmente, en el Capítulo 8 se repasan los principales avances presentados en esta tesis en el campo de análisis de datos de microarrays y las principales contribuciones en revistas y congresos.

Capítulo 2

Fundamentos de genética y biología molecular

La herencia biológica que un individuo transmite a su descendencia se denomina genotipo y está codificada por una secuencia de nuclétotidos, en una macromolécula de ácido desoxirribonucleico (ADN). El ADN almacena dicha información en una serie de subunidades elementales de información, denominadas genes. El conjunto de todos los genes que constituyen el genotipo de un ser vivo concreto se denomina genoma. El estudio del origen, funcionamiento y evolución del genoma, mediante diferentes técnicas y disciplinas se conoce como genómica [21].

El genotipo almacenado en el ADN contiene instrucciones para la síntesis de proteínas, moléculas funcionales encargas de controlar el metabolismo celular. Sin embargo, la información almacenada en el ADN suele permanecer intacta en el núcleo en las células. La síntesis proteica precisa de una molécula intermedia a la cual se transfiere la información contenida en el ADN. Esta molécula en la que se transcribe la información genética es el ácido ribonucleico (ARN), que posteriormente se traduce en una proteína. Cuando la infomación alamacenada en un gen es finalmente transcria a una molécula de ARN, se dice que el gen se ha expresado. El fenotipo es el resultado de la expresión del genotipo. Este mecanismo de almacenaje, transcripción y traducción de la información genética se conoce como dogma central de la biología molecular.

Los avances tecnológicos de los últimos veinte años han permitido desarrollar métodos y técnicas experimentales capaces de cuantificar la expresión genética. En concreto la técnica del microarray permite determinar simultáneamente el nivel de expresión genético del genoma completo de un organismo tan complejo como el ser humano, con más de 20000 genes. Esta tecnología ha supuesto un avance enorme en el estudio de la Genética y la Biología Molecular [48] [1].

2.1 Conceptos de biología molecular

Los monómeros son moléculas de pequeña masa molecular, simples y estables que constituyen unidades básicas de construcción en Biología Molecular, por ejemplo: monosacáridos, ácidos grasos, nucleótidos o aminoácidos. Los monómeros suelen aparecer unidos entre si, mediante enlaces covalentes, formando grandes estructuras llamadas polímeros. Si el polímero está formado por un único tipo de monómero, se denomina homopolimero. Si la estructura posee diferentes tipos de moléculas, se llama heteropolímero o copolímero.

Los nucleótidos son polímeros formados por la unión de una base nitrogenada, un monosacárido y ácido fosfórico. La unión entre la base nitrogenada y la pentosa se hace a través del grupo hidroxilo del carbono 1', formando una estructura que se conoce como nucleósido. Este nucleósido se une a un grupo fosfato a través del carbono 5' de la pentosa y se conoce cómo nucleótido-monofosfato (NMP), si existen dos grupos fosfatos se denomina nucleótido-difosfato (NDP) y para tres grupos fosfatos se llama nucleótido-trifosfato (NTP). El ácido nucleico es un copolímero formado por diferentes nucleótidos unidos entre si mediante enlaces fosfodiéster (enlace entre el carbono 5' de la base sacárida a un ion fosfato y éste, a su vez, al grupo hidroxilo del carbono 3' de otro monosacárido).

2.1.1 Ácidos nucleicos

Los ácidos nucleicos son cadenas polinucleótidas de diferente longitud, que adoptan diferentes estructuras. Existen dos tipos de ácidos nucleicos: ácido desoxirribonucleico (ADN) y ácido ribonucleico (ARN). Aparte de la composición sacárida y otras diferencias estructurales, la característica principal que distingue a ambos ácidos nucleicos es el conjunto de bases nitrogenadas que los componen.

2.1.1.1 ADN

El ADN está compuesto por dos cadenas de nucleótidos, constituidas por largas secuencias de cuatro bases nitrogendas: Adenina (A), Guanina (G), Citosina (C) y Timina (T). Las dos hebras están unidas entre si por puentes de hidrógeno entre los nucleótidos, con una peculiaridad: la Adenina se empareja con Timina a través de dos enlaces, mientras que la Guanina lo hace con Citosina mediante tres enlaces. Los pares A=T y G≡C poseen el mismo tamaño y la molécula de ADN se enrolla adquiriendo la estructura de doble hélice que la caracteriza. Sin embargo, el ADN se organiza en una serie de estructuras de mayor nivel que compactan dicha molécula en el interior de la célula. La unidad estructural inmediatamente superior a la cadena de doble hélice es la cromatina, compuesta por una serie de complejos proteicos entorno a los que se enrolla la cadena bicatenaria de ADN. La

cromatina se organiza en estructuras más complejas hasta compactar la molécula de ADN en una estructura superior denominada cromosoma.

La secuencia de nucleótidos a lo largo de la cadena de ADN codifica la información sobre una proteína. Sin embargo, no es la responsable de su síntesis, simplemente almacena la información necesaria sobre la cantidad y momento de su producción. Esta parte del código que regula por completo la síntesis de una proteína se denomina gen y es la unidad de almacenamiento de información hereditaria. La información total codificada por toda la secuencia del ADN se denomina genoma.

2.1.1.2 ARN

El ARN está compuesto por una cadena monocatenária de cuatro nucleótidos: A, G, C y Uracilo (U). El ARN, se diferencia del ADN no sólo en su composición química (el ARN se basa en ribosa en lugar de desoxirribosa y la base complementaria a la Adenina es el Uracilo, en lugar de Timina) sino también en su estructura. El ARN aparece como una cadena simple que puede plegarse sobre si misma adoptando diversas formas. En comparación con el ADN, las secuencias de nucleótidos que lo componen son mucho más cortas. Sin embargo, la característica principal de esta molécula es su gran actividad en el desarrollo celular que, tras diferentes procesos de maduración, es la responsable final de la síntesis proteica.

2.1.2 Dogma central de la biología molecular

El ADN almacena la información necesaria para la producción de una proteína, pero no participa directamente en su síntesis. Para ello, la información almacenada en el ADN se transcribe a otro tipo de ácido: ARN mensajero (mARN) que, posteriormente, se encargará de sintetizar la proteína en otro proceso denominado traducción. Cuando un gen se ha transcrito a ARN, se dice que se ha expresado. Este mecanismo de codificación, transcripción y traducción de la información genética se conoce como el dogma central de la biología molecular.

2.1.2.1 Transcripción del ADN: mARN

El proceso de transcripción genética es muy complejo y pueden distinguirse hasta cinco etapas diferentes:

1. *Preiniciación*
   El primer paso consiste en localizar la región en la que debe comenzar la transcripción. Por ello, a cada gen le precede una secuencia de ADN no codificante (UTR, acrónimo del inglés *untranslated region*) conocida como región promotora y que identifica el inicio del gen. Para marcar este punto, existen

Figura 2.1: Diagrama de flujo de la información genética que representa el mecanismo de codificación y transmisión de la herencia biológica. El dogma central de la Biología Molecular establece un mecanismo unidireccional, en el cual, la información hereditaria codificada en el ADN se replica y se transmite a la descendencia. La transcripción del ADN a ARN es la primera fase del proceso de expresión genética que finaliza con la traducción del transcrito a una proteína funcional. Por otro lado, existen procesos especiales que introducen relaciones inversas como la síntesis de ADN complementario a partir de ARN.

un conjunto de proteínas funcionales llamadas factores de transcripción (FT), que localizan y reconocen estas secuencias UTR y se adhieren al carbono 5' inmediatamente anterior al gen, formando un complejo de preiniciación que establece el punto de inicio de la transcripción genética. Las secuencias UTR constituyen elementos reguladores fundamentales en una primera etapa del proceso de transcripción. La longitud de estas regiones depende del organismo y del gen considerado, oscilando entre las 500 bases (para organismos más simples, como la levadura) hasta 2000 bases (para especies más complejas, como el ser humano).

2. *Iniciación*
   Una vez localizado el promotor, el complejo de preiniciación se cierra con la unión de la ARN polimerasa (ARNp): una enzima que cataliza la polimerización de ribonucleótidos, uniéndolos mediante enlaces fosfodiéster. En este momento, uno de los TF (la helicasa, una enzima capaz de romper puentes de hidrógeno) desnaturaliza parcialmente el ADN, separando la doble hélice a lo largo de dieciocho pares de bases nitrogenadas, formando un complejo abierto conocido como burbuja de transcripción.

3. *Disgregación del promotor*
   Una vez sintetizado el primer enlace fosfodiéster, el complejo de preiniciación se desprende del promotor. La separación se debe a otro TF, la quinasa, una enzima que fosforilíza la ARNp.

4. *Elongación*
   Los ribonucleótidos se aparean complementariamente a la secuencia de ADN, mediante enlaces por puente de hidrógeno. Una vez apareados, el centro ac-

tivo de la ARNp sintetiza el enlace fosfodiéster entre dos ribonucleótidos, formando la cadena de ARN complementaria, proceso conocido como elongación.

5. *Terminación*
   El proceso de elongación continúa hasta que una determinada secuencia de ARN sea sintetizada. Esta secuencia esta situada en el extremo del gen y es rica en Adenina y Citosina seguida de Timina. Una vez sintetizada, esta estructura desestabiliza el complejo ADN-ARN, separándolas, liberando la ARNp y renaturalizando la molécula de ADN.

2.1.2.2 Postprocesamiento del mARN: maduración

El proceso de transcripción de ADN a mARN se produce en el núcleo celular y es exclusivo de organismos eucariotas. Hay otra característica importante que diferencia el mARN procariota del eucariota: el mARN eucariota contiene secuencias que no codifican información (secuencias no codificantes) llamadas intrones, que separan las secuencias que si contienen información genética, llamadas exones. El mARN procariota solo contiene secuencias codificantes. Por ello, el mARN eucariota sufre un postprocesado del tránscrito primario, que se denomina maduración del mARN.

Antes de ser transportado al citoplasma (en ocasiones, incluso antes de haber transcrito por completo la cadena de mARN) se añade al extremo 5' de la secuencia un nucleótido modificado: la 7-metil-guanina (metilG). Esta molécula, denominada caperuza (CAP) aporta estabilidad al tránscrito. Una vez finalizada la transcripción, se añade al extremo 3' una cola de poliadenilato (poli-A, secuencia de ARN basada en Adenina) que protege el mARN de la degradación y ayudará a su transporte. Existe una secuencia de poliadenilación (AAUAAA) que se sintetiza unos veinte nucleótidos antes de la secuencia de terminación.

Tras el encapuchado y poliadenilizado del transcrito primario, se procede a la eliminación de secuencias no codificantes, proceso conocido como empalme o splicing. El proceso se lleva a cabo por un complejo (conocido como spliceosoma) formado por pequeñas ribonucleo-proteínas nucleares (snRNP), que eliminan los intrones y unen las secuencias de mARN codificantes, dando como resultado una molécula de mARN funcional.

2.1.2.3 Codificación del código genético: codones

El genotipo codificado en cada gen del ADN consiste en una serie de instrucciones que sintetizan una proteína. Cada proteína queda determinada por una secuencia concreta de aminoácidos. El mARN almacena esta información agrupando los ribonucleótidos en tripletes, llamados codones, lo que le permite codificar hasta 64

aminoácidos diferentes. Además de codificar un aminoácido, ciertos codones poseen otras funciones; por ejemplo, el codón AUG codifica la metionina y además, actúa de punto de inicio para la traducción genética.

2.1.2.4 Traducción del mARN: síntesis proteica

El mARN es transportado a los ribosomas (en el caso de eucariotas, atravesando los poros del núcleo celular hasta llegar al citoplasma) donde se lleva a cabo su traducción. El ribosoma es un orgánulo celular compuesto por un armazón de ácido ribonucleico, conocido como ARN ribosomal (rARN) y por proteínas que se ensamblan en dos estructuras: subunidad mayor y subunidad menor. Su función es sintetizar las proteínas codificadas por el mARN usando un adaptador: otro ácido ribonucleico presente en el citoplasma, conocido como ARN transferente (tARN).

El tARN es una cadena (de alrededor de 80 ribonucleótidos) que aparece plegada con algunos de sus nucleótidos apareados, formando una estructura concreta, en la que se diferencia un triplete desapareado denominado anticodón que se apareará con su codón complementario de la secuencia de mARN. En el otro extremo del tARN, los terminales 3' y 5' se pliegan formando un brazo aceptor con un grupo carboxilo que puede ligarse a un aminoácido concreto, en una reacción catalizada por unas enzimas específicas denominadas aminoasil-tARN-sintetasas. Además, aparecen dos bucles (o brazos) laterales: el bucle T, que reconoce al ribosoma y lo une al tARN durante el proceso de síntesis; y el bucle D, que identifica el tARN ante la enzima aminoasil-tARN-sintetasas, la cual unirá mediante un enlace covalente un aminoácido determinado al brazo aceptor del tARN.

El proceso de síntesis comienza con las subunidades ribosomales desacopladas. La subunidad menor, posee tres sitios: sitio-A es el lugar de entrada del aminoacil-tARN, el sitio-P es el lugar en el que se produce el ensamblado del aminoacil-tARN con la cadena polipéptida (formando el complejo peptidil-tARN), el sitio-E por donde saldrá el tARN tras ceder su aminoácido. Se puede distinguir tres fases:

1. *Iniciación*
   El primer paso es la adición de una serie de factores de iniciación (FI) a la subunidad menor, que bloquean el sitio-A, el sitio-E y facilita el acoplo del tARN de iniciación, que en células eucariotas es el anticodón UAC que sintetiza la metionina pero en procariotas es una variación de éste que sintetiza la formilmetionina (fmet-tARN); este aminoácido inicial suele eliminarse de la cadena proteica tras su formación. Posteriormente, la subunidad menor engancha el extremo 5' de la cadena de mARN (con ayuda del CAP, en eucariotas) y comienza a desplazarse hasta encontrar el codón de iniciación AUG (en organismos procariotas la secuencia de reconocimiento está localizada a pocos aminoácidos del enganche y se conoce como secuencia de

Shine-Dalgarno: AGGAGG). Una vez localizado, los factores de iniciación son liberados, permitiendo el acoplamiento de la subunidad mayor ribosómica y liberando el sitio-A y el sitio-E.

2. *Elongación*
   El sitio-P se encuentra ocupado por un peptidil-tARN (que en un primer momento será el tARN de iniciación). El sitio-A está esperando a ser ocupado por un aminoacil-tARN que se apareará complementariamente al codón del mARN. Una vez formado el nuevo par codón-anticodón en el sitio-A, la cadena polipéptida del sistio-P posee un nuevo aminoácido para su crecimiento; el polipéptido es transferido al sitio-A (reacción catalizada por la peptidil-transferasa, una ribozima del rARN) y el tARN del sitio-P queda descargado. Finalmente el ribosoma se desplaza un codón, trasladando el tARN descargado al sitio-E y el nuevo petidil-tARN ocupa el sitio-P. Este proceso se repite contínuamente, aumentando la cadena polipéptida, hasta que se alcance un codón de parada.

3. *Terminación*
   Cuando se alcanza un codón de terminación, determinados factores de terminación actúan bloqueando el sitio-A y provocan la hidrólisis del peptil-tARN, liberando la proteína del complejo ribosoma-mARN-tARN.

2.2 Microarrays

En los últimos veinte años, la técnica del microarray se ha consolidado como un método de análisis fundamental de la Genómica. Esta tecnología permite cuantificar simultáneamente numerosas interacciones biológicas a nivel molecular. Por ejemplo: (*i*) microarrays de expresión genética (chips de ADN), interacciones gen-proteina (ChIP-on-chip), (*ii*) variaciones en el número de copias cromosómicas (CGH arrays), (*iii*) polimorfismos de nucleótido simple (SNP arrays) o (*iv*) el estado de metilación del ADN (arrays de metilación).

Un microarray consiste en una colección de biomoléculas activas, denominadas sondas, ordenadas e inmovilizadas sobre un sustrato sólido en regiones micrométricas denominadas spots. El material sobre el que se fijan las sondas puede ser muy variado. En los chips porosos, las sondas se adhieren mediante enlaces covalentes a pequeñas porciones de geles, membranas de nylon o nitrocelulosa depositadas sobre un portaobjetos de cristal. Por otro lado, en los arrays no porosos se fijan directamente a la superficie mediante enlaces covalentes. En este caso la superficie usada suele ser un sustrato de silicio, plástico, oro o un recubrimiento de agarosa. Estas superficies sobre las que se depositan las sondas forman matrices bidimensionales molécularmente activas [48].

La técnica del microarray es un procedimiento estándar que permite detectar, cuantificar y analizar simultáneamente grandes cantidades de material biológico. Las biomoléculas contenidas en una muestra de interés se denominan dianas. La detección estas dianas requiere un marcaje previo del material sometido a análisis. Este marcaje se consigue mediante diferentes indicadores luminiscentes, radiactivos o enzimáticos. Los métodos más usados son los enzimáticos y luminiscentes, en los que se utilizan tintes fluoróforos. El material marcado se distribuye sobre la superficie activa del microarray, donde cada diana se alinea complementariamente a una sonda específica, en un proceso denominado hibridación. El postprocesado de este microarray permite detectar e identificar las sondas hibridadas.

2.2.1 Cuantificación de la expresión genética: chips de ADN

En un microarray de expresión o chip de ADN, el material biológico considerado es una una colección de moléculas de ADN de una sola hebra. Estas sondas pueden ser: (*i*) secuencias cortas (entre 20 y 100 nucleótidos) denominadas oligonucleótidos o (*ii*) fragmentos constituidos por varios miles de bases que forman genes o fragmentos de ellos. Cada sonda puede extraerse directamente de una muestra o a partir de la clonación de ADN, procedente de librerías génicas. La fabricación de chips de ADN se realiza de forma automática, en un proceso robotizado, con una duración inferior a las 16 horas. Existen diferentes casas comerciales que se dedican a su diseño y construcción. Entre las más conocidas se encuentran: Agilent, Ilumina y Affymetrix.

Una vez construido el microarray, es necesario obtener el material genético de estudio. En el caso de arrays de expresión, el material de interés es el mARN transcrito. Cada marca comercial especifica un protocolo de extracción específico para cada chip. Una vez extraido el mARN, se sintetiza su ADN complementario (cADN) con ayuda de una enzima denominada retrotranscriptasa. Esta encima sintetiza un complejo cADN-ARN a partir de una hebra de ARN, proceso inverso al producido por la ARNp. Posteriormente, este complejo se desnaturaliza y se separan las dos hebras, obteniendo el cADN del ADN transcrito y libre de partes no codificantes. Las hebras de cADN se marcan y se colocan sobre la superficie del microarray donde se hibrida con su ADN sonda complementario. Finalmente, el microarray se lava y se elimina el resto de cADN diana no fijado.

El cADN hibridado se detecta en un proceso acorde con el marcaje del material genético. En el caso de indicadores fluoróforos, el revelado consiste en detectar la fluorescencia de las dianas fijadas en el microarray, cuando este es iluminado con determinada longitud de onda. Este tipo de marcaje posee una ventaja: usando diferentes indicadores, con diferentes respuestas espectrales, se pueden analizar varias muestras en un mismo microarray. Por lo general, uno de los genomas analizados se utiliza como muestra de control, mientras que la otra es una muestra

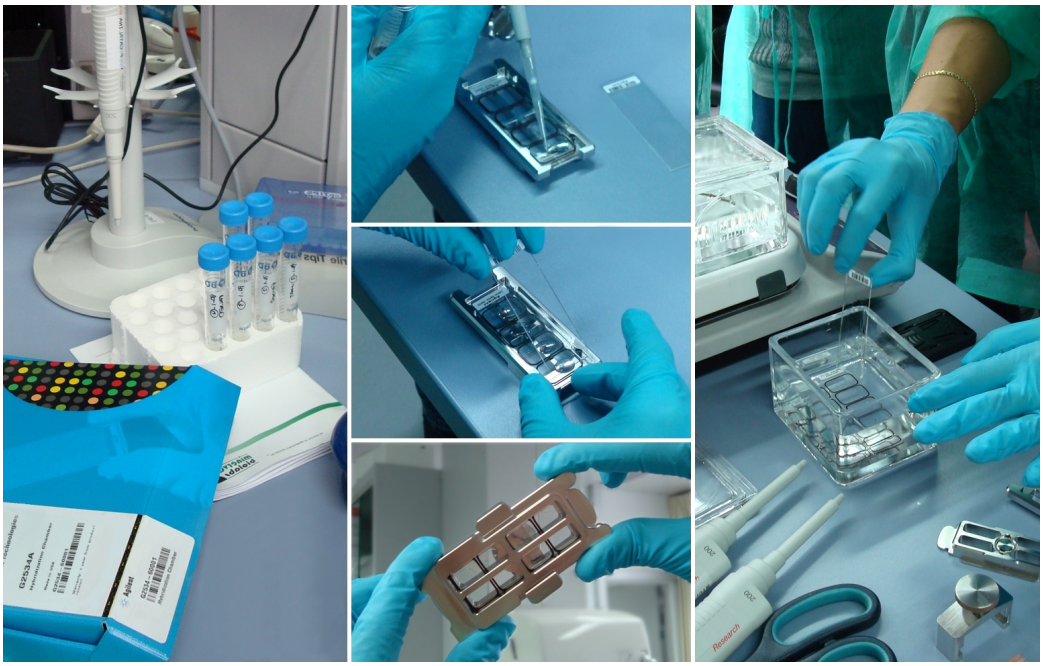


Figura 2.2: Preparación de un microarray de expresión, modelo G4112F de la casa comercial Agilent. Cada portaobjeto contiene 4 microarrays, cada uno con 44000 spots cubriendo el genoma humano al completo. El paquete comercial incluye el material necesario para la extracción de la muestra. Tras ser extraido y marcado, el material genetico se deposita sobre los arrays. Posteriormente, se cubre y se sella en una camara de hibridación siguiendo el protocolo indicado. Posteriormente, tras retirar el banco de trabajo, se elimina el material genetico sobrante y el chip queda listo para ser escaneado y procesado.

experimental en la cual se quiere observar una expresión diferencial respecto a la de referencia. Unos indicadores usados tradicionalmente son las cianinas $Cy3$ y $Cy5$ que emiten luz verde y roja cuando se iluminan con longitudes de onda $\lambda_{Cy3} \approx 550$ (nm) y $\lambda_{Cy5} \approx 660$ (nm) respectivamente [25].

El revelado de un microarray de dos canales consiste en la obtención de dos imágenes que muestren el nivel de intensidad emitido por cada uno de los indicadores. Para ello, el chip se escanea cuando se ilumina con la longitud de onda que activa un marcador y posteriormente con la otra. La información recogida se resume en una razón de intensidades entre la muestra de control y la experimental. Este cociente se considera proporcional a la diferencia en la concentración de mARN entre ambas muestras. Para un gen en el que se observa un cociente superior a la unidad se tiene una concentración de mARN mayor en la muestra de control que en la experimental, es decir, el gen se ha reprimido en la muestra de estudio; se dice entonces que el gen se ha inhibido. Por el contrario, si el cociente es menor que uno, el gen se ha activado en la muestra experimental. Con el fin de facilitar el análisis de esta información, se suele hacer uso de una transformación logarítmica que permite comparar diferentes cocientes de intensidades. Por ejemplo, razones de intensidades 4 y $0'25$ indicarían la activación e inhibición de un gen en la misma proporción. Al tomar el logaritmo en base dos, esta abundancia relativa de mARN se expresaría como 2 y -2 respectivamente. De este modo, se consigue una escala proporcional, en la que se indica la activación o inhibición de un gen con un signo positivo o negativo respectivamente. A esta magnitud, el logaritmo en base dos de la razón de intesidades entre una muestra de control y otra experimental, se le conoce cómo nivel de expresión.

2.2.2 Análisis de interacciones a nivel gen-proteína: ChIP-on-chip

La inmunoprecipitación de cromatina (ChIP) es una técnica experimental que permite analizar interacciones entre genes y proteínas [97]. Este método bioquímico consiste en el uso de un anticuerpo específico que reconoce y se adhiere a la proteína de interés, incluso cuando esta aparece ligada al ADN. Una vez formado, el complejo gen-proteína-anticuerpo es selectivamente inmunoprecipitado. Posteriormente, el compuesto es disociado y el fragmento de ADN al que estaba ligada la proteína es purificado y amplificado. Una vez obtenido el producto de la ChIP, el ADN puede ser analizado mediante un microarray de expresión específico que cubra la región genómica de interés. Este tipo de microarrays, conocidos cómo ChIP-on-chip, proporcionan información directa sobre la actividad especifica de una proteína y una secuencia de ADN concreta. En particular, resulta de especial interés el análisis de proteínas que actúan como FT durante el proceso de expresión genética [17].

La ChIP es un procedimiento laborioso e individualizado que precisa de mi-

croarrays específicos para su análisis. Sin embargo, las técnicas de secuenciación de nueva generación han encontrado una aplicación inmediata en el análisis de los productos de la ChIP, mejorando el rendimiento y la resolución con la que se identifican las regiones de interacción de un gen y una proteína. Esta metodología, que combina la técnica ChIP con la secuenciación en paralelo, se conoce como ChIP-sequencing (ChIP-seq). Tanto el análisis mediante ChIP-on-chip y ChIP-seq dan como resultado una secuencia de varias centenas de nucleótidos en la que una proteina de interés posee un sitio de enlace. Esta información se procesa para caracterizar los sitos de enlaces mediante estructuras menores, denominadas *motifs*, constituidas por un patrón de pocos nucleotidos que se conservan en diferentes experimentos [53] [69].

2.3 Red reguladora de genes

El dogma central de la biología molecular explica el proceso de expresión genética mediante un mecanismo de regulación a nivel transcripcional. Sin embargo, el metabolismo celular de un organismo depende de numerosos procesos subyacentes, entre ellos la expresión genética, que interaccionan entre si de forma compleja para producir respuestas diferentes. Por ejemplo: (i) la transducción de señales extracelulares que modifican la función de ciertas proteínas, alterando el proceso de regulación o (ii) modificaciones del estado de metilación del ADN que, sin producir mutaciones, alteran su estructura espacial e impiden la formación del complejo de preiniciación transcripcional e incluso (iii) la actividad de un gen puede estar controlada por la expresión de otros que codifican la síntesis de proteínas funcionales para su transcripción. El fenómeno de expresión genética es, por tanto, un proceso dinámico y complejo que integra diferentes elementos reguladores, internos y externos al sistema biológico considerado. Este mecanismo de adaptación funcional del proceso de expresión es lo que se conoce como epigenética [42].

Un modelo que trata de simplificar los mecanismos de regulación desde un punto de vista fenomenológico, a niveles de interacción gen-gen, es la red de regulación genética (RRG) [19]. Este modelo proyecta todos los procesos reguladores, a diferentes niveles metabólicos, en el espacio de la actividad genética. Las RRG establecen relaciones causales directas entre un conjunto de genes, para ser exactos, entre su estado de expresión [24]. En una RRG se considera que la expresión de un gen, al que se denomina hijo, depende del estado de expresión y la actividad de otros, denominados padres [40].

Las RRG heredan características propias de la Teoría de Grafos así como sus propiedades [70]. Gráficamente, una RRG se caracteriza mediante un conjunto de nodos que representan los genes[1]. Por otro lado, la estrutura topológica de la red

[1]Generalmente, se dice que los nodos de una RRG simbolizan genes, cuando en realidad re-

representa las relaciones causales entre los genes. En la Figura 2.3 se representa un ejemplo de una RRG. Cada gen está representado por un nodo mientras que las relaciones de parentesco se representan con vertices dirigidos, desde los padres a sus hijos. Además, este esquema permite representar el efecto regulador de activación o inhibición.

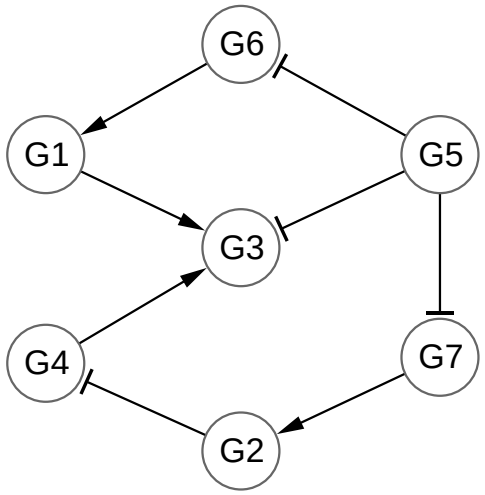


Figura 2.3: Representación grafica de una RRG con $G = 7$ genes. Los nodos representan genes y los vértices las relaciones de parentesco entre ellos. El tipo de regulación se expresa con una punta en forma de flecha y con un extremo en forma de $\ll$T$\gg$, para efectos de activación e inhibición respectivamente. Por ejemplo, el gen G3 posee tres padres: G1 y G4 que activan su expresión y G5 que la inhibe. Por otro lado, el gen G1 es hijo de G6 que activa su expresión.

Cualitativamente, las RRG permiten analizar ciertas propiedades del genoma, tales como su funcionalidad y evolución; información de vital interés en la industria famacológica y en la investigación médica. En la actualidad se desconocen la mayoría de estas estructuras, salvo para determinadas rutas metabólicas de algunas especies usadas como sistemas biológicos modelos. Por ejemplo, la levadura *Saccharomyces cerevisiae*, uno de los organismos eucariotas más simples, pero con más de $G = 6600$ genes, para el que se ha estudiado la RRG en diferentes procesos celulares y metabólicos [57] [87] [45].

presentan su expresión. Sin embargo, este modelo trata de establecer relaciones causales entre sus elementos por lo que, sin perdida de generalidad, se puede hablar de genes o de su actividad indistintamente.

2.4 Red de regulación transcripcional

Las RRG tratan de explicar la expresión genética, desde un punto de vista funcional, mediante un mecanismo de regulación a nivel global. Sin embargo, la expresión de un gen es un proceso complejo que, en una primer etapa, está regulado mediante el fenómeno de regulación transcripcional. Los FT son proteínas funcionales que reconocen y se adhieren a la región promotora de un gen específico, facilitando o impidiendo el acoplamiento del complejo transcripcional. Este mecanismo describe la regulación genética desde un punto de vista de interacciones a nivel molecular, mientras que en una RRG se toma una perspectiva propia de la Biología de sistemas.

De forma similar a las RRG, una red de regulación transcripcional (RRT) es un modelo que describe el mecanismo de regulación transcripcional. Sin embargo, a diferencia de las RRG, en una RRT las interacciones son exclusivamente a nivel gen-proteína. Por tanto, en su representación gráfica, se distinguen entre dos tipos de nodos: genes y FT. Por otro lado, la topología de la red describe la interacción[2] entre pares gen-proteína y el efecto de regulación ejercido. En consecuencia, una RRT es una red de vértices dirigidos, donde los TF siempre son nodos fuente y los genes nodos destino. En la literatura, se pueden encontrar RRT en la que todos los nodos son genes. Este tipo de estructura, favorecida por la falta de consenso en la nomenclatura, sustituye cada FT por el gen que lo codifica. Sin embargo, este tipo de RRT posee una estructura similar, en la que los genes con funciones reguladoras siempre son los nodos fuentes y no poseen ningún vértice entrante. En la Figura 2.4 se representa un ejemplo de una RRT. Los genes y los FT están representados por nodos mientras que la interacción entre ellos se establece mediante vértices.

La mayor parte de la información de interacciones entre FT y genes proceden de datos de ChIP-on-chip u otro tipo de experiencias similares [27]. Aunque también es posible representar en una RRT el efecto de regulación, en la mayoría de los casos, los métodos usados sólo permiten constatar la existencia de dichas interacciones y no su efecto en el proceso de transcripción. Estas relaciones son directamente observadas o estimadas mediante técnicas computacionales que permiten predecir los FT que potencialmente regulan un gen [44] [69]. Es por ello que, en las RRT, se suelen representar sólo las relaciones causales, es decir la estructura topológica de la red, y no el efecto de regulación ejercido.

[2]En una RRT, la interacción se produce entre un FT y la región promotora de un gen, en la cual dicha proteína posee un sitio de enlace. Sin embargo, a efectos causales, se puede hablar de interacciones gen-proteína.

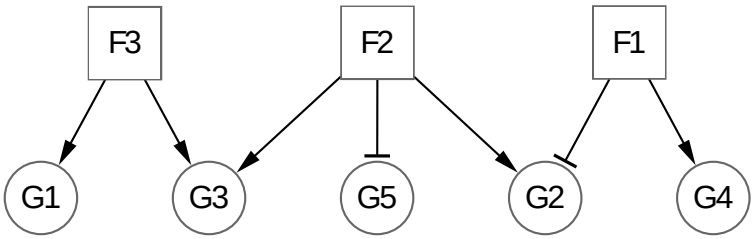


Figura 2.4: Representación gráfica de una RRT con $F = 3$ factores de transcripción y $G = 5$ genes. Cada gen está representado por un nodo circular mientras que los FT se representan mediante nodos rectangulares. El tipo de regulación se describe mediante un extremo en forma de flecha o en forma de ≪T≫, si el FT efectos de activa o inhibe la transcripción del gen. Por ejemplo, el factor F1 posee un sitio de enlace en la región promotora de los genes: G2 y G4 impidiendo y favoreciendo su transcripción respectivamente. Además, el gen G2 también es regulado por F2 que favorece su transcripción.

Capítulo 3

Paradigma Bayesiano

En numerosas ocasiones, las observaciones de un fenómeno físico no pueden modelarse haciendo uso de leyes deterministas. La variabilidad de las propiedades observadas no permite establecer una relación causa-efecto obvia y el comportamiento del sistema aparenta ser fruto del azar. Este comportamiento aleatorio puede corresponder a una variabilidad ontológica del sistema, es decir, el resultado en cada observación sigue un proceso verdaderamente espontáneo e impredecible. Sin embargo, un sistema aleatorio puede regirse por un proceso estocástico que, debido a su complejidad, no permite una descripción mediante un conjunto de leyes causales. El origen de esta aleatoriedad epistemológica es variado. Puede deberse a una sensibilidad especial a las condiciones iniciales del experimento, a un desconocimiento del conjunto completo de variables que intervienen en el proceso o a la incapacidad para observarlas [95]. En estos casos, aleatorio no es sinónimo de espontáneo ni de impredecible.

Para un sistema aleatorio que sigue un proceso estocástico, resulta difícil establecer relaciones causales entre las variable observadas. Sin embargo, la metodología estadística nos permite analizar dicha información para describir y extraer conclusiones del sistema observado. En concreto, la estadística inferencial engloba una serie métodos y técnicas que nos permite analizar un conjunto de datos y contrastarlos con información adicional con el fin estimar las variables desconocidas [7]. Dichas estimaciones permiten analizar las propiedades del proceso considerado, estudiar la evolución del sistema, así como comparar diferentes hipótesis y modelos [64].

En la Inferencia Estadística, los métodos Bayesianos son populares por su capacidad de proporcionar estimaciones en términos probabilistas. Además, la metodología Bayesiana permite manejar problemas en los que, a parte de las variables ocultas, aparece un conjunto de datos incompleto. Sin embargo, el gran potencial de este formalismo reside en la capacidad de modelar información de diferentes

fuentes, a través de distribuciones de probabilidad a priori, que completen y enriquezcan el conjunto de datos de las variables observadas.

3.1 Fundamentos de inferencia estadística

En el marco de trabajo de la estadística inferencial, la variabilidad de las propiedades que caracterizan un proceso estocástico se representan mediante un conjunto de variables aleatorias [61] [63]. Sea un conjunto de variables aleatorias observables que, por conveniencia, denotaremos como el vector columna $\mathbf{y} = [y_1, \ldots, y_G]^\top$. Además, considérese el conjunto $\mathbf{x} = [x_1, \ldots, x_G]^\top$ de variables aleatorias no observables, que interaccionan con las variables conocidas en un proceso oculto.

Un modelo estadístico del proceso considerado trata de explicar cómo se distribuyen los estados de dichas variables aleatorias. Este medida de la incertidumbre se representa mediante una función de distribución de probabilidad conjunta, que denotaremos por $p(\mathbf{y}, \mathbf{x})$. Por otro lado, el modelado de las variables aleatorias suele incluir la dependencia con ciertas constantes desconocidas $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_M]^\top$, denominadas parámetros, que caracterizan al modelo. Para resaltar esta dependencia de la probabilidad con los parámetros, dicho modelo se denomina paramétrico [64]. Dados los M parámetros de un modelo, la probabilidad conjunta de las variables consideradas se representa por una función de probabilidad condicional, que denotaremos por $p(\mathbf{y}, \mathbf{x}|\, \boldsymbol{\theta})$. Sin embargo, en un proceso con variables desconocidas, la formulación natural del modelo estadístico no permite postular la distribución conjunta de las variables ocultas y las observadas. En su lugar, resulta más intuitivo modelar la probabilidad condicional de las variables observables dadas los parámetros y las variables desconocidas, es decir $p(\mathbf{y}|\, \mathbf{x}, \boldsymbol{\theta})$.

El objetivo de la inferencia estadística es llegar a extraer conclusiones del proceso considerado más allá de una mera descripción de los datos, apoyándose en el modelo probabilista, el conjunto de observaciones y cualquier otro tipo de información disponible a priori. En particular, la inferencia inductiva prescinde de cualquier tipo de información adicional y basa todo su conocimiento en los datos. Es por ello que a las conclusiones obtenidas con estos procedimientos se denominan soluciones frecuentistas o deterministas. Por otro lado, la inferencia Bayesiana trata de modelar las cantidades desconocidas mediante distribuciones de probabilidad, que permiten incorporar información a priori y buscan un compromiso entre las configuraciones más probables a la vez que más verosímiles con los datos. La metodología Bayesiana se apoya en el teorema de Bayes para proporcionar soluciones probabilistas a posteriori.

3.1.1 Función de verosimilitud

La estadística inferencial asume un modelo observacional como hipótesis y postula la probabilidad de las variables conocidas dados los parámetros y las variables ocultas $p(\mathbf{y}|\,\mathbf{x},\boldsymbol{\theta})$. Dado un conjunto observaciones $\mathbf{y}=\mathbf{Y}$, se define la función de verosimilitud como la probabilidad condicional de las observaciones dadas el resto de cantidades desconocidas $p(\mathbf{Y}|\,\mathbf{x},\boldsymbol{\theta})$. Cabe destacar que la verosimilitud es una función de las variables ocultas y los parámetros, para un conjunto de datos concreto. En la bibliografía es frecuente encontrar la verosimilitud con la notación $\mathcal{L}(\mathbf{x},\boldsymbol{\theta})$ que incide en la dependencia exclusiva de dicha función con el segundo argumento.

Dependiendo de la naturaleza de las observaciones, en virtud de las propiedades de las distribuciones de probabilidad [64], la verosimilitud puede expresarse como un producto de funciones independientes. Diremos que un conjunto de N observaciones son estáticas cuando cada una de ellas proviene de una realización independiente de la misma experiencia. Haciendo uso de una notación matricial, representaremos por $[\mathbf{Y}]_{in}=y_i(n)$ la n-ésima observación de la i-ésima variable. Este tipo de medidas estáticas permite asumir una independencia estadística entre las observaciones de una misma variable, columnas de la matriz $\mathbf{Y}$, que constituyen una serie de datos eventuales. Por tanto, la función de verosimilitud factoriza en el producto,

$$p(\mathbf{Y}|\mathbf{x},\boldsymbol{\theta})=\prod_{n=1}^{N}p\left([y_1(n),\ldots,y_G(n)]^\top|\mathbf{x},\boldsymbol{\theta}\right). \tag{3.1}$$

Por otro lado, en un conjunto de observaciones dinámicas se realiza una única experiencia y se muestrea la evolución temporal de cada variable. En este caso, el conjunto de datos $\mathbf{Y}=[\mathbf{y}_1,\ldots,\mathbf{y}_G]$ esta formado por una serie temporal de N observaciones $\mathbf{y}_i=[y_i(1),\ldots,y_i(N)]^\top$, filas de la matriz de datos. En este tipo de medidas resulta totalmente improcedente asumir independencia entra las diferentes observaciones de la misma variable. Sin embargo, por conveniencia matemática, es frecuente asumir independencia entre las variables conocidas (filas de la matriz de datos) y las variables ocultas. Esta aproximación permite expresar la verosimilitud como un producto de probabilidades, que además de independientes se suelen considerar idénticamente distribuidas (IID). En el caso de variables IID, la verosimilitud factoriza en el producto,

$$p(\mathbf{Y}|\mathbf{x},\boldsymbol{\theta})\overset{\text{IID}}{=}\prod_{i=1}^{G}p(\mathbf{y}_i|x_i,\boldsymbol{\theta}). \tag{3.2}$$

3.1.2 Teorema de Bayes

La verosimilitud proporciona una medida de la incertidumbre, sin embargo esta función no es la distribución de probabilidad de las variables desconocidas ni de los parámetros. Esta interpretación errónea confunde la función de verosimilitud con la probabilidad de que una configuración concreta de las cantidades desconocidas modele las observaciones, es decir $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{Y})$. Esta probabilidad condicional, denominada en su formulación original por Thomas Bayes como probabilidad inversa [4], se conoce como probabilidad a posteriori [2]. El Teorema de Bayes [64] formaliza matemáticamente la relación entre la función de verosimilitud y la probabilidad a posteriori como,

$$p(\mathbf{x}, \boldsymbol{\theta}|\,\mathbf{Y}) = \frac{p(\,\mathbf{Y}|\,\mathbf{x}, \boldsymbol{\theta})\, p(\mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{Y})} \tag{3.3}$$

donde $p(\mathbf{x}, \boldsymbol{\theta})$ es la probabilidad a priori de las variables ocultas y los parámetros del modelo. La constante de normalización que aparece en le denominador de (3.3) es el resultado de marginalizar el numerador como,

$$p(\mathbf{Y}) = \int p(\,\mathbf{Y}|\,\mathbf{x}, \boldsymbol{\theta})\, p(\mathbf{x}, \boldsymbol{\theta})\, d\mathbf{x} d\boldsymbol{\theta} \tag{3.4}$$

y se conoce como verosimilitud marginal [8].

3.1.3 Divergencia de Kullback-Leibler

Un indicador adecuado para comparar distribuciones de probabilidad sobre la misma variable aleatoria es la divergencia de Kullback-Leibler [93]. Se define la divergencia de $p(x)$ respecto a $q(x)$ como la integral de la primera distribución por el logaritmo del cociente como,

$$\mathrm{D_{KL}}\left[p(x) \| q(x)\right] := \int p(x) \log \frac{p(x)}{q(x)} dx \tag{3.5}$$

Dadas dos distribuciones cualesquiera, la divergencia de Kullback-Leibler es una magnitud positiva y sólo es nula cuando dichas probabilidades sean idénticas. Sin embargo, este operador no define una métrica. Notese que la divergencia no es una función simétrica respecto a las funciones sobre las que se define [50],[93]. Estas propiedades pueden resumirse como,

$$\mathrm{D_{KL}}\left[p(x) \| q(x)\right] \geq 0 \tag{3.6}$$

$$\mathrm{D_{KL}}\left[p(x) \| q(x)\right] = 0 \Leftrightarrow p(x) \equiv q(x) \tag{3.7}$$

$$\mathrm{D_{KL}}\left[p(x) \| q(x)\right] \neq \mathrm{D_{KL}}\left[q(x) \| p(x)\right] \tag{3.8}$$

A pesar de no definir un espacio métrico, en la literatura también se puede encontrar esta función con el nombre de distancia de Kullback-Leibler o entropía

relativa. Dicha nomenclatura es propiciada por la formulación adoptada cuando la divergencia se expresa mediante una diferencia logarítmica como,

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}\left[p\left(x\right)\|\, q\left(x\right)\right] &= \int p\left(x\right)\log\frac{p\left(x\right)}{q\left(x\right)}dx \\ &= \int p\left(x\right)\log p\left(x\right)dx - \int p\left(x\right)\log q\left(x\right)dx \\ &= \mathcal{H}\left(p\left(x\right), q\left(x\right)\right) - \mathcal{H}\left(q\left(x\right)\right) \end{aligned} \tag{3.9}$$

donde $\mathcal{H}\left(q\left(x\right)\right)$ es la entropía y $\mathcal{H}\left(p\left(x\right), q\left(x\right)\right)$ la entropía cruzada [50].

3.2 Inferencia inductiva

La utilidad de la función de verosimilitud reside en su capacidad de evaluar la afinidad de una configuración paramétrica y de las variables ocultas con el conjunto de observaciones. La verosimilitud proporciona la probabilidad de observar un conjunto de datos concreto a partir de una configuración específica de los parámetros y las variables desconocidas. Esta probabilidad permite establecer un criterio de confianza basándose en a explicación más probable. Las configuraciones que poseen menor verosimilitud explican peor las observaciones, mientras que, las que proporcionan verosimilitudes mayores son más consistentes con los datos observados. Los parámetros y variables ocultas que mejor modelan las observaciones, los más verosímiles, serán los que maximizan la función verosimilitud. Se define el estimador de máxima verosimilitud (ML, acrónimo del inglés *maximum likelihood*) como el argumento que maximiza la función de verosimilitud [2] [64].

$$\begin{aligned} \left\{\hat{\mathbf{x}}_{ML}, \hat{\boldsymbol{\theta}}_{ML}\right\} &:= \arg\max_{\mathbf{x},\boldsymbol{\theta}} \left\{p\left(\mathbf{Y}\middle|\, \mathbf{x}, \boldsymbol{\theta}\right)\right\} \\ &\overset{\text{IID}}{=} \arg\max_{\mathbf{x},\boldsymbol{\theta}} \left\{\prod_{i=1}^{G} p\left(\mathbf{y}_i\middle|\, x_i, \boldsymbol{\theta}\right)\right\} \end{aligned} \tag{3.10}$$

El cálculo del máximo de la verosimilitud puede simplificarse haciendo uso de una transformación adecuada en la que desaparezcan las constantes multiplicativas. En concreto, teniendo en cuenta la monoticidad de la función logaritmo, es equivalente calcular el máximo de la verosimilitud que la de su logaritmo. Esta transformación, para variables IID, permite expresar el producto de probabilidades como una sumatoria. Sin pérdida de generalidad, se expresa el estimador ML como el argumento que maximiza el logaritmo de la función verosimilitud [64].

$$\left\{\hat{\mathbf{x}}_{ML}, \hat{\boldsymbol{\theta}}_{ML}\right\} \overset{\text{IID}}{=} \arg\max_{\mathbf{x},\boldsymbol{\theta}} \left\{\sum_{i=1}^{G} \log p\left(\mathbf{y}_i\middle|\, x_i, \boldsymbol{\theta}\right)\right\} \tag{3.11}$$

La inferencia inductiva se basa, como su nombre indica, en un razonamiento inductivo en el cual las conclusiones obtenidas para un número finito de casos observados se generalizan para la variable observada. Este método estima el valor de los parámetros y los actualiza a la luz de nuevas observaciones, añadiendo nuevos términos a la función de verosimilitud. Sin embargo, este formalismo tiende a sobreparametrizar el proceso considerado, ya que los modelos con mayor número de parámetros se ajustan mejor a las observaciones. El caso extremo, un modelo en el que hay tantos parámetros como observaciones, se ajusta a la perfección; se dice entonces que el modelo está saturado [5]. Al incluir más observaciones y más parámetros al modelo, resulta difícil distinguir una mejora real de una aportación trivial. Por este motivo se prefiere usar otros procedimientos que permiten seleccionar el modelo óptimo que explica las observaciones con el menor número de parámetros posible, siendo más sencillo, más estable y menos sesgado [73].

3.3 Inferencia Bayesiana

La inferencia Bayesiana trata de proporcionar conclusiones de las variables ocultas y los parámetros del modelo en términos probabilistas. Este enfoque exige un modelo estadístico de las cantidades de interés a través de la distribución de probabilidad a priori $p(\mathbf{x}, \boldsymbol{\theta})$. Dicha probabilidad puede elegirse subjetivamente, de manera que sólo restrinja los parámetros y las variables ocultas a unos valores físicamente admisibles. Por otro lado, si se dispone de información objetiva, puede incorporarse al modelo a través de la probabilidad a priori.

La metodología Bayesiana se apoya en la probabilidad a priori y la función de verosimilitud para calcular la probabilidad a posteriori $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{Y})$. Ésta probabilidad se puede obtener en virtud del Teorema de Bayes como,

$$p(\mathbf{x}, \boldsymbol{\theta}|\,\mathbf{Y}) = \frac{p(\mathbf{Y}|\,\mathbf{x}, \boldsymbol{\theta})\, p(\mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{Y})} \overset{\text{IID}}{=} \prod_{i=1}^{G} \frac{p(\mathbf{y}_i|\,x_i, \boldsymbol{\theta})\, p(x_i, \boldsymbol{\theta})}{p(\mathbf{y}_i)}. \tag{3.12}$$

La probabilidad a priori establece una preferencia por ciertas configuraciones del modelo. Al incorporar esta información, la probabilidad a posteriori la contrasta con los datos experimentales recogidos en la función de verosimilitud. La inferencia basada en la probabilidad a posteriori busca el compromiso entre los parámetros y las variables ocultas más verosímiles y más probables a priori. La configuración que maximize la probabilidad a posteriori, proporciona una estimación afín con las observaciones a la vez que preferida a priori. Se define el estimador de maximiza probabilidad a posteriori (MAP) como el argumento que maximiza

dicha probabilidad,

$$\begin{aligned}\left\{\hat{\mathbf{x}}_{MAP}, \hat{\boldsymbol{\theta}}_{MAP}\right\} &:= \arg\max_{\mathbf{x},\boldsymbol{\theta}} \left\{p\left(\mathbf{Y}\middle|\,\mathbf{x},\boldsymbol{\theta}\right)\right\} \\ &\overset{\text{IID}}{=} \arg\max_{\mathbf{x},\boldsymbol{\theta}} \left\{\prod_{i=1}^{G} \frac{p\left(\mathbf{y}_i\middle|\,x_i,\boldsymbol{\theta}\right) p\left(x_i,\boldsymbol{\theta}\right)}{p\left(\mathbf{y}_i\right)}\right\}.\end{aligned} \tag{3.13}$$

Al considerar una transformación logarítmica y al excluir las constantes superfluas en el cálculo del máximo, el estimador MAP se puede expresar como,

$$\left\{\hat{\mathbf{x}}_{MAP}, \hat{\boldsymbol{\theta}}_{MAP}\right\} \overset{\text{IID}}{=} \arg\max_{\mathbf{x},\boldsymbol{\theta}} \left\{\sum_{i=1}^{G} \ln p\left(\mathbf{y}_i\middle|\,x_i,\boldsymbol{\theta}\right) p\left(x_i,\boldsymbol{\theta}\right)\right\}. \tag{3.14}$$

La inferencia Bayesiana se basa en un razonamiento abductivo, en el cual se parte de un hecho experimental conocido y se valida con una hipótesis que posee cierto grado de incertidumbre. Al maximizar la probabilidad a posteriori, se obtiene un estimador en el que la verosimilitud se penaliza de acuerdo con la probabilidad a priori. La probabilidad a posteriori puede interpretarse como una regularización de la función de verosimilitud. En el caso de tener un desconocimiento absoluto del proceso, el Principio de razón insuficiente [2] establece que todas las configuraciones de los parámetros y las variables ocultas son igualmente probables. En este caso particular, la probabilidad a priori es una constante y el estimador ML coincidiría con el estimador MAP.

3.4 Metodologías de inferencia Bayesiana

La mayoría de los procesos de interés suelen contar con una gran cantidad de variables observables y ocultas. Los modelos paramétricos de mayor interés práctico y usados generalmente, se basan en estructuras jerárquicas que introducen aún más incógnitas al problema. De ésta manera, el cálculo de la probabilidad a posteriori suele a ser complicado, llegando a ser intratable analíticamente, y exige realizar aproximaciones numéricas [50].

Existe una infinidad de técnicas numéricas que tratan de calcular la probabilidad a posteriori para realizar inferencia aproximada [9]. Por un lado, las aproximaciones Bayesianas basadas en técnicas Monte-Carlo pueden llegar a obtener resultados muy precisos. Sin embargo, estos métodos demandan un gran coste computacional y resulta inviable aplicarlas en modelos con un gran número de variables. Por otro lado, existen metodologías Bayesianas que, bajo unas condiciones estrictas pero aceptables en algunos casos, permiten aproximar la probabilidad a posteriori.

3.4.1 Muestreo de Gibbs

El muestreo de Gibbs es un método de cálculo numérico que se engloba en las denominadas técnicas Monte-Carlo. Estos métodos son muy comunes en el análisis Bayesiano de datos porque, sin restriciones derivadas del coste computacional, permiten obtener aproximaciones muy exactas de la probabilidad a posteriori. En general, estas técnicas se apoyan en la integración Monte-Carlo que, dadas un conjunto de muestras $\left\{\hat{\mathbf{x}}^{(t)}\right\}_{t=0}^{T}$ simuladas a partir de la distribución de interés como,

$$\hat{\mathbf{x}}^{(t)} \sim p\left(\mathbf{x}|\,\mathbf{Y},\boldsymbol{\theta}\right), \forall t \tag{3.15}$$

permite calular el valor esperado de una variable aleatória mediante,

$$\left\langle f\left(\mathbf{x}\right)\right\rangle_{p(\,\mathbf{x}|\mathbf{Y},\boldsymbol{\theta})} = \int f\left(\mathbf{x}\right) p\left(\mathbf{x}|\,\mathbf{Y},\boldsymbol{\theta}\right) d\mathbf{x} = \frac{1}{T}\sum_{t=1}^{T} f\left(\hat{\mathbf{x}}^{(t)}\right). \tag{3.16}$$

e incluso estimar empíricamente la distribución de probabilidad como,

$$p\left(\mathbf{x}|\,\mathbf{Y},\boldsymbol{\theta}\right) \approx \frac{1}{T}\sum_{t=1}^{T} \delta\left(\mathbf{x}-\hat{\mathbf{x}}^{(t)}\right). \tag{3.17}$$

El problema que trata de resolver estos métodos es que, aunque se conozca la forma analítica de la distribución de interés, resulta difícil simular muestras a partir de su probabilidad conjunta. En particular, el muestreo de Gibbs aproxima esta probabilidad por otra proporcional a ella que, convenientemente, factoriza en distribuciones marginales e independientes que resultan más simples de manejar,

$$\begin{aligned} p\left(\mathbf{x}|\,\mathbf{Y},\boldsymbol{\theta}\right) &= p\left(x_1|\,\mathbf{Y},\boldsymbol{\theta},x_2,x_3,\ldots,x_G\right) p\left(x_2,x_3,\ldots,x_G|\,\mathbf{Y},\boldsymbol{\theta}\right) \\ &= p\left(x_1|\,\mathbf{Y},\boldsymbol{\theta},x_2,x_3,\ldots,x_G\right) p\left(x_2|\,\mathbf{Y},\boldsymbol{\theta},x_1,x_3,\ldots,x_G\right) \\ &\cdot \quad \frac{p\left(x_1,x_3,\ldots,x_G|\,\mathbf{Y},\boldsymbol{\theta}\right)}{p\left(x_1|\,\mathbf{Y},\boldsymbol{\theta}\right)} \\ &\vdots \\ &\propto \prod_{g=1}^{G} p\left(x_g|\,\mathbf{Y},\boldsymbol{\theta},\mathbf{x}-\{x_g\}\right) \end{aligned} \tag{3.18}$$

donde $\mathbf{x}-\{x_g\} = [x_1,\ldots,x_{g-1},x_{g+1},\ldots,x_G]^\top$ denota el subconjunto de variables en $\mathbf{x}$ excepto la componente x_g. Basándose en (3.18), a partir de una muestra inicial $t=0$, el muestreo de Gibbs obtiene una cadena de Markov mediante simulaciones

de las distribuciones a posteriori e independientes tal que, en la iteración t-ésima,

$$\begin{aligned}
\hat{x}_1^{(t)} &\sim p\left(x_1|\,\mathbf{Y},\boldsymbol{\theta},\hat{x}_2^{(t-1)},\hat{x}_3^{(t-1)},\ldots,\hat{x}_G^{(t-1)}\right) && (3.19)\\
\hat{x}_2^{(t)} &\sim p\left(x_2|\,\mathbf{Y},\boldsymbol{\theta},\hat{x}_1^{(t)},\hat{x}_3^{(t-1)},\ldots,\hat{x}_G^{(t-1)}\right) && (3.20)\\
&\vdots\\
\hat{x}_g^{(t)} &\sim p\left(x_g|\,\mathbf{Y},\boldsymbol{\theta},\hat{x}_1^{(t)},\ldots,\hat{x}_{g-1}^{(t)},\hat{x}_{g+1}^{(t-1)},\ldots,\hat{x}_G^{(t-1)}\right)\\
&\vdots\\
\hat{x}_G^{(t)} &\sim p\left(x_G|\,\mathbf{Y},\boldsymbol{\theta},\hat{x}_1^{(t)},\hat{x}_2^{(t)},\ldots,\hat{x}_{G-1}^{(t-1)}\right) && (3.21)\\
& && (3.22)
\end{aligned}$$

de manera que, tras un periodo $t > t'$ de estabilización de la cadena, la distribución de interés puede estimarse empíricamente, salvo constantes superfluas para el cálculo de máximos, como

$$p\left(\mathbf{x}|\,\mathbf{Y},\boldsymbol{\theta}\right) \propto \frac{1}{T-t'}\sum_{t>t'}^{T}\delta\left(\mathbf{x}-\hat{\mathbf{x}}^{(t)}\right). \tag{3.23}$$

De este modo, se puede aproximar el cálculo del estimador MAP (3.11) a través de la aproximación de la distribución de interés (3.23) mediante,

$$\hat{\mathbf{x}}_{MAP} \approx \arg\max_{\mathbf{x}}\left\{\ln\sum_{t>t'}^{T}\delta\left(\mathbf{x}-\hat{\mathbf{x}}^{(t)}\right)\right\}. \tag{3.24}$$

En el caso particular de que la distribución de interés sea una Gaussiana, el estimador MAP coincide con la media, momento de primer orden, y puede calcularse directamente a partir de (3.16) con $f\left(\mathbf{x}\right)=\mathbf{x}$ como la media muestral,

$$\begin{aligned}
\left\langle\mathbf{x}\right\rangle_{p(\mathbf{x}|\mathbf{Y},\boldsymbol{\theta})} &= \int \mathbf{x}p\left(\mathbf{x}|\,\mathbf{Y},\boldsymbol{\theta}\right)d\mathbf{x}\\
&\approx \frac{1}{T-t'}\sum_{t>t'}^{T}\int \mathbf{x}\delta\left(\mathbf{x}-\hat{\mathbf{x}}^{(t)}\right)d\mathbf{x}\\
&= \frac{1}{T-t'}\sum_{t>t'}^{T}\hat{\mathbf{x}}^{(t)}. && (3.25)
\end{aligned}$$

3.4.2 Aprendizaje variacional Bayesiano

En lugar de optimizar la probabilidad a posteriori, el método variacional Bayesiano construye un límite inferior de la verosimilitud marginal dependiente de una

distribución libre de las variables ocultas y parámetros del modelo $q(\mathbf{x}, \boldsymbol{\theta})$, que al consiserdar observaciones IID se puede expresar como,

$$\begin{aligned} p(\mathbf{Y}) &\overset{\text{IID}}{=} \prod_{i=1}^{G} p(\mathbf{y}_i) \\ &= \prod_{i=1}^{G} \int \frac{q(x_i, \boldsymbol{\theta})}{q(x_i, \boldsymbol{\theta})} p(\mathbf{y}_i | x_i, \boldsymbol{\theta}) \, p(x_i, \boldsymbol{\theta}) \, dx_i d\boldsymbol{\theta} \end{aligned} \tag{3.26}$$

Al tomar logaritmo, y haciendo uso de la desigualdad de Jensen, la verosimilitud marginal se puede expresar como un funcional dependiente de la verosimilitud, la probabilidad a priori y la distribución libre como,

$$\ln p(\mathbf{y}_i) = \ln \int \frac{q(x_i, \boldsymbol{\theta})}{q(x_i, \boldsymbol{\theta})} p(\mathbf{y}_i | x_i, \boldsymbol{\theta}) \, p(x_i, \boldsymbol{\theta}) \, dx_i d\boldsymbol{\theta} \geq \mathcal{F}[q(x_i, \boldsymbol{\theta})] \tag{3.27}$$

con

$$\mathcal{F}[q(x_i, \boldsymbol{\theta})] = \int q(x_i, \boldsymbol{\theta}) \ln \frac{p(\mathbf{y}_i | x_i, \boldsymbol{\theta}) \, p(x_i, \boldsymbol{\theta})}{q(x_i, \boldsymbol{\theta})} dx_i d\boldsymbol{\theta}. \tag{3.28}$$

El límite inferior puede expresarse como una divergencia entre la distribución libre y la probabilidad a posteriori,

$$\mathcal{F}[q(x_i, \boldsymbol{\theta})] = -\mathrm{D}_{\mathrm{KL}}[q(x_i, \boldsymbol{\theta}) \| \, p(\mathbf{y}_i | x_i, \boldsymbol{\theta}) \, p(x_i, \boldsymbol{\theta})]. \tag{3.29}$$

donde, salvo constantes de proporcionalidad, la distribución a posteriori es igual al producto de la verosimilitud y la a priori. Por extensión de las propiedades de la divergencia (3.6), el límite inferior se define como una cantidad negativa $\mathcal{F}[q(x_i, \boldsymbol{\theta})] \leq 0$ para cualquier distribución. En el caso límite,

$$\mathcal{F}[q(x_i, \boldsymbol{\theta})] = 0 \tag{3.30}$$

de acuerdo con (3.7) y (3.29), la distribución libre será proporcional a la probabilidad a posteriori,

$$q(x_i, \boldsymbol{\theta}) = p(\mathbf{y}_i | x_i, \boldsymbol{\theta}) \, p(x_i, \boldsymbol{\theta}) \propto p(x_i, \boldsymbol{\theta} | \mathbf{y}_i) \tag{3.31}$$

Por tanto, el argumento de la distribución libre (3.30) que maximiza el límite, o de manera equivalente minimiza la divergencia, será el estimador MAP,

$$\begin{aligned} \left\{ \hat{x}_{i,MAP}, \hat{\boldsymbol{\theta}}_{MAP} \right\} &= \arg\max_{x_i, \boldsymbol{\theta}} \{\ln p(x_i, \boldsymbol{\theta} | \mathbf{y}_i)\} \\ &= \arg\max_{x_i, \boldsymbol{\theta}} \{\ln q(x_i, \boldsymbol{\theta})\} \end{aligned} \tag{3.32}$$

La optimización de (3.28) es un problema que podría resolverse mediante calculo variacional. Sin embargo, obtener esta distribución libre analíticamente es difícil en la mayoría de los casos. No obstante, basándose en las simetrías de las distribuciones conjugadas, el aprendizaje variacional Bayesiano realiza una aproximación numérica que posibilita este cálculo. En un modelo conjugado, la verosimilitud y la probabilidad a priori se escogen entre un conjunto de distribuciones tales que la probabilidad a posteriori pertenezca a la misma familia que la a priori. Es decir, la probabilidad a priori $p\left(x_i, \boldsymbol{\theta} | \, \boldsymbol{\zeta}_i\right)$ y la distribución libre $q\left(x_i, \boldsymbol{\theta} | \, \boldsymbol{\xi}_i\right)$ dependen de un conjunto de hiperparámetros $\boldsymbol{\zeta}_i$ y $\boldsymbol{\xi}_i$ pertenecientes a la misma familia. Al hacer uso de distribuciones conjugadas, el cálculo de la probabilidad a posteriori se reduce a un aprendizaje de los hiperparámetros $\boldsymbol{\xi}_i$ que modelan la distribución libre. Adicionalmente, el método variacional Bayesiano escoge la probabilidad a priori en una familia de distribuciones independientes, en la que las variables ocultas y los parámetros factorizan como,

$$p\left(x_i, \boldsymbol{\theta} | \, \boldsymbol{\zeta}_i\right) \approx p\left(x_i | \, \boldsymbol{\zeta}_{i,x_i}\right) p\left(\boldsymbol{\theta} | \, \boldsymbol{\zeta}_{i,\boldsymbol{\theta}}\right) \tag{3.33}$$

$$q\left(x_i, \boldsymbol{\theta} | \, \boldsymbol{\xi}_i\right) \approx q\left(x_i | \, \boldsymbol{\xi}_{i,x_i}\right) q\left(\boldsymbol{\theta} | \, \boldsymbol{\xi}_{i,\boldsymbol{\theta}}\right). \tag{3.34}$$

Al considerar las aproximaciones (3.33) y (3.34) el limite variacional se puede reescribir,

$$\mathcal{F}\left[q\left(x_i\right), q\left(\boldsymbol{\theta}\right)\right] = \int q\left(x_i\right), q\left(\boldsymbol{\theta}\right) \ln \frac{p\left(\mathbf{y}_i | \, x_i, \boldsymbol{\theta}\right) p\left(x_i\right), p\left(\boldsymbol{\theta}\right)}{q\left(x_i\right), q\left(\boldsymbol{\theta}\right)} dx_i d\boldsymbol{\theta}. \tag{3.35}$$

Esta metodología permite desarrollar un algoritmo de Esperanza y Maximización (EM) que maximizasecuencial mente el limite inferior de la verosimilitud marginal. A partir de unos valores a priori $\hat{\boldsymbol{\xi}}_i^{(0)} = \boldsymbol{\zeta}_i$, los hiperparámetros que modelan los parámetros y las variables ocultas se van optimizando alternativamente a la vez que el limite variacional se va actualizando. Por tanto, el método variacional Bayesiano de Esperanza y Maximización (VBEM) consiste en los dos pasos siguientes, en los que una de las distribuciones libres en (3.35) es optimizada mientras la otra permanece fija:

VBE la distribución libre de las variables ocultas es optimizada, mientras que la de los parámetros permanece fija,

$$\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)} = \arg\max_{\boldsymbol{\xi}_{i,x_i}} \left\{ \mathcal{F}\left[q\left(x_i | \, \boldsymbol{\xi}_{i,x_i}\right) q\left(\boldsymbol{\theta} | \, \hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right) \right] \right\}. \tag{3.36}$$

VBM la distribución libre de los parámetros es optimizada, mientras que la de las variables ocultas permanece fija,

$$\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\xi}_{i,\boldsymbol{\theta}}} \left\{ \mathcal{F}\left[q\left(x_i | \, \hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right) q\left(\boldsymbol{\theta} | \, \boldsymbol{\xi}_{i,\boldsymbol{\theta}}\right) \right] \right\}. \tag{3.37}$$

Las reglas de actualización de las distribuciones libres se derivan de los pasos VBE y VBM definidos anteriormente (véase Apéndice A), que pueden expresarse haciendo uso de brackets para denotar valores esperados como,

$$q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{x_i}^{(t+1)}\right) \;\propto\; \mathrm{e}^{\left\langle \ln p(\mathbf{y}_i|x_i,\boldsymbol{\theta})\right\rangle_{q\left(\boldsymbol{\theta}|\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)}+\ln p(x_i)} \tag{3.38}$$

$$q\left(\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t+1)}\right) \;\propto\; \mathrm{e}^{\left\langle \ln p(\mathbf{y}_i|x_i,\boldsymbol{\theta})p(x_i)\right\rangle_{q\left(x_i|\hat{\boldsymbol{\xi}}_{x_i}^{(t+1)}\right)}+\ln p(\boldsymbol{\theta})}. \tag{3.39}$$

Posteriormente, tras cada iteración de los pasos VBE y VBM, el límite funcional se actualiza. Idealmente, se podría esperar que el límite alcanzara el valor mínimo tras un gran número de iteraciones,

$$\lim_{t\to\infty} \mathcal{F}\left[q\left(\mathbf{x}_i,\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_i^{(t)}\right)\right] = 0. \tag{3.40}$$

Sin embargo, las distribuciones a priori (3.33) fueron escogidas, por conveniencia matemática, cómo distribuciones conjugadas e independientes. Por tanto, la convergencia del límite depende de lo apropiada que sea ésta aproximación. Como alternativa, en lugar de comprobar la convergencia según (3.40), el algoritmo itera hasta que la diferencia entre dos pasos consecutivos satisfaga un criterio de convergencia con $\epsilon << 1$ tal que,

$$\left|\frac{\mathcal{F}\left[q\left(x_i,\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_i^{(t+1)}\right)\right]-\mathcal{F}\left[q\left(x_i,\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_i^{(t)}\right)\right]}{\mathcal{F}\left[q\left(x_i,\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_i^{(t+1)}\right)\right]}\right| \leq \epsilon. \tag{3.41}$$

La Figura 3.1 muestra un diagrama de flujo del método VBEM, en el que a partir de un valor inicial $t = 0$ los pasos VBE y VBM actualizan los hyperparametros de las distribuciones libres. Una vez alcanzado el criterio de convergencia para $t = T$, la distribución a posteriori se aproxima por la distribución libre

$$p\left(x_i,\boldsymbol{\theta}|\,\mathbf{y}_i\right) \approx q\left(x_i,\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_i^{(T)}\right) \tag{3.42}$$

y se calculan los correspondientes estimadores MAP como solución subóptima,

$$\left\{\hat{\mathbf{x}}_{MAP},\hat{\boldsymbol{\theta}}_{MAP}\right\} \approx \arg\max_{\mathbf{x},\boldsymbol{\theta}}\left\{q\left(x_i,\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_i^{(T)}\right)\right\}. \tag{3.43}$$

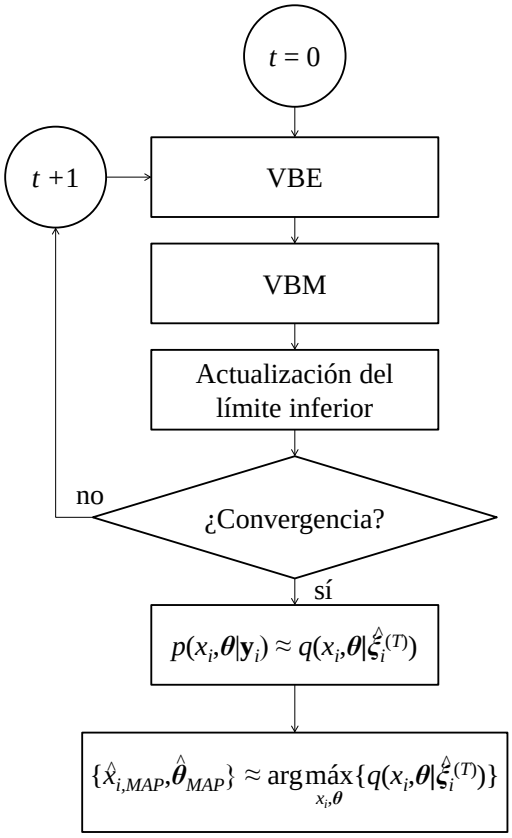


Figura 3.1: Digrama de flujo del método VBEM. A partir de un valor inicial $t = 0$ los pasos VBE y VBM actualizan los hyperparámetros de las distribuciones libres y la el límite inferior de la verosimilitud. Una vez alcanzado el criterio de convergencia para $t = T$, la probabilidad a posteriori se aproxima por la distribución subóptima y se calculan los estimadores MAP.

Chapter 4

Variational Bayesian method for microarrays time series learning and GRN reverse engineering

Gene regulatory networks (GRN) were introduced in Chapter 2 as an abstract model that tries to explain molecular interactions at a genetic level. In a GRN it is considered that the expression of a gene, referred as child, depends on the status of others, known as it parents. In this context, microarray data is a perfect candidate to objectively describing the evolution of the expression state of genes considered in the network. Therefore, we address here the problem of modeling and inferring the GRN from microarray time series.

The earlier attempts to describe GRN from a mathematical point of view is the Kauffman's model, also known as Boolean networks (BN), and its extension to probabilistic Boolean networks (PBN) [43] [84]. In this models, gene expression is quantized into binary levels: activation and inhibition. Genetic relationships are described by a predictor function that establishes logic relationships between genes. Additionally, in PBN several possible binary functions were provided with a probabilistic measure for each logic rule.

On the other hand, a GRN may be understood as a Bayesian network [38]. This model have a graphical interpretation by a belief network, with random variables as nodes and conditional dependencies as edges. Within a Bayesian framework, the complete space of the expression state of any gene is represented by a random variable. In Bayesian networks, the expression of any child is supposed to depend on some probabilistic relationships between the current state of the other variables, its parents. Although they are defined in terms of probabilities, its analysis

provides an appropriate causal interpretation of the GRN.

Bayesian methods are widely accepted for providing probabilistic solutions to inference problems. Moreover, within its formulation, the expression state of any gene may be completely described and it is capable of handling noise [32]. However, using data alone, causal inference for genetic relationships is very limited. Nevertheless, another advantage of Bayesian formalism is that any constraints on the network space may be included as a prior information.

4.1 Problem formulation and linear models

Bayesian framework provides GRN with a mathematical formalism able to cope with microarray data. In such kind of formulation, the expression profiles may be understood as observations of random variables that represents the activation status of genes. Specifically, microarray time series encode the dynamical behavior of GRN for a particular biological process. Consider a microarray data set with G genes and $N+1$ time samples denoted by

$$\mathbf{Y} = \begin{pmatrix} y_1(0) & \cdots & y_1(N) \\ \vdots & \ddots & \vdots \\ y_G(0) & \cdots & y_G(N) \end{pmatrix} \in \mathbb{R}^{G\times(N+1)} \tag{4.1}$$

with $y_i(n)$ the observed expression level of the i-th gene at the n-th time sample.

On the other hand, genetic relationships are established in a probabilistic way if the expression of two genes are significantly associated. Given a set of G genes, the structure of GRN is characterized by two important features [70]. First is the connectivity, also referred as network topology, that represents the linkage pattern of the network. This logical structure has been formally described in [91] by a set of binary latent variables, denoted as

$$\mathbf{x}_i = [x_i(1), \ldots, x_i(G)]^\top \in \{0,1\}^{G\times 1} \tag{4.2}$$

with $x_i(j) = 1$ specifies that the j-th gene is a parent of the i-th gene or $x_i(j) = 0$ otherwise. In GRN with not necessarily large sizes, where biological knowledge suggest that each gene has a limited number of parents, such kind of variable will have a high density of zero elements, i.e. the network is expected to be sparse. Second, genetic networks also specify regulatory effects between its elements, i.e. strength and type of interaction. This scheme has been mathematically represented in [91] by an additional set of weights, denoted as

$$\boldsymbol{\omega}_i = [\omega_i(1), \ldots, \omega_i(G)]^\top \in \mathbb{R}^{G\times 1} \tag{4.3}$$

with $\omega_i(j) > 0$ for gene activation and $\omega_i(j) < 0$ for gene inhibition.

4.1.1 Previous AR1 model

According to the AR1 model proposed in [89], the expression level $y_i(n)$ of the i-th child at the n-th time sample arises from a time homogeneous Markov process. Based on this assumption, the generative model fits microarray data by a linear combination of the observations at the immediately previous time step plus noise as,

$$y_i(n) = \sum_{j=1}^{G} y_j(n-1)\omega_i(j)x_i(j) + e_i(n). \tag{4.4}$$

with $e_i(n)$ the additive noisy term modeled by IID white noise.

Suitably, AR1 model considers a noise term representing any deviation on the linear assumption. However, AR1 model establishes relationships between microarray data which are supposed to be affected by experimental noise inherent to the measuring process [68]. Moreover, observations may be affected by other sources of errors proper of the stochastic nature of the biological process, that are not a part of the actual regulatory process. All these erroneous terms are part of the observations and the generative model includes it as an effect of the GRN. Therefore, AR1 model underestimates the noise and the conclusions based on this fitting may be incorrect.

4.1.2 Novel AR1MA1 model

Previous AR1 model establishes relationships between data, that are supposed to be noisy. It would be much more realistic to establish this interactions between the real expression level, the true logarithmic measure of the relative mRNA abundance. Distinguishing between the real expression level and its noisy observation, we are going to present a novel approach that fits better the nature of microarray data. We assume that gene expression follows a stochastic process, where microarray data $y_i(n)$ is a noisy observation of the real expression level, denoted by $z_i(n)$. Here, the noise term will be also denoted by $e_i(n)$ and it is supposed to cover all sources of noise described above. On the other hand, differential expression due to real regulatory effects described by the GRN are embedded in the real expression level $z_i(n)$. Therefore, the relationships between the unknown real expression level and its noisy observation may be expressed as

$$y_i(n) = z_i(n) + e_i(n). \tag{4.5}$$

Similarly to model in (4.4), we proposed a Markov process but where genetic relationships are established between the real expression levels instead of its noisy

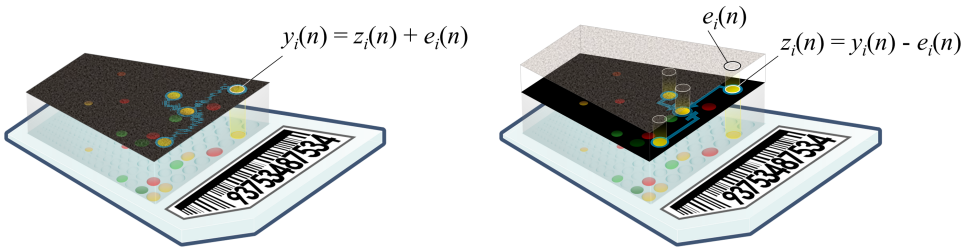
observation as

$$y_i(n) = \sum_{j=1}^{G} z_j(n-1)\omega_i(j)x_i(j) + e_i(n). \tag{4.6}$$

According to (4.5), the new model proposed in (4.6) may be expressed in terms of the observed expression level as a Box-Jenkins model, i.e. a first order autoregressive moving-average (AR1MA1)model as

$$y_i(n) = \sum_{j=1}^{G} y_j(n-1)\omega_i(j)x_i(j) - \sum_{j=1}^{G} e_j(n-1)\omega_i(j)x_i(j) + e_i(n). \tag{4.7}$$

Figure 4.1 illustrates differences between both linear models. Noise term is depicted as a distorting mask affecting the true regulatory process. On the other hand, microarray observations are a black-background image with red, green and yellow spots. Whilst AR1 model establishes relationships directly between microarray data $y_i(n)$, the AR1MA1 one distinguishes between the real expression level $z_i(n)$ and the noise $e_i(n)$.



(a) AR1 (b) AR1MA1

Figure 4.1: Illustration that characterizes microarray data as an image scanned from a DNA chip. (a) AR1 model considers the observed expression level $y_i(n)$ as the source of information to extract biological knowledge. The underlying regulatory process is distorted by the noise affecting measures. (b) AR1MA1 model distinguish between the true expression level $z_i(n)$ and its noisy observation $y_i(n)$. Noise $e_i(n)$ is depicted as a distorting mask affecting the data. The true regulatory process is revealed once this contribution is removed from data.

4.2 Variational Bayesian learning method with the AR1MA1 model

Given the set of all noise terms for each gene at every time sample as,

$$\mathbf{E} = \begin{pmatrix} e_1(0) & \cdots & e_1(N) \\ \vdots & \ddots & \vdots \\ e_G(0) & \cdots & e_G(N) \end{pmatrix} \tag{4.8}$$

AR1MA1 model in (4.7) may be conveniently expressed in a matrix formulation as,

$$\mathbf{y}_i = \mathbf{T}\mathbf{Y}^\top \mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i + \mathbf{T}\mathbf{E}^\top \mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i + \mathbf{e}_i. \tag{4.9}$$

with $\mathbf{y}_i = [y_i(1), \ldots, y_i(N)]^\top$ microarray time series and $\mathbf{e}_i = [e_i(1), \ldots, e_i(N)]^\top$ corresponding noise. Note that $n = 0$ are excluded at these multidimensional variables. Auxiliary matrix $\mathbf{T}$ defined as,

$$\mathbf{T} = \left(\begin{array}{ccc|c} 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{array}\right) \tag{4.10}$$

and $\mathbf{D}_{\boldsymbol{\omega}_i}$ a diagonal matrix with vector $\boldsymbol{\omega}_i$ as,

$$\mathbf{D}_{\boldsymbol{\omega}_i} = \begin{pmatrix} \omega_i(1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_i(G) \end{pmatrix}. \tag{4.11}$$

4.2.1 Statistical modeling: likelihood and priors

As a condition imposed by the AR1MA1 model, noise is fitted a priori by IID white noise with zero mean and unknown variance σ_i^2 as,

$$p\left(e_i(n)\right) = \mathcal{N}\left(e_i(n)|\, 0, \sigma_i^2\right), \forall n. \tag{4.12}$$

On the other hand, according to variational Bayesian framework presented in Chapter 3, variables describing the network topology $\mathbf{x}_i$ and regulatory type $\boldsymbol{\omega}_i$ must be expressed in terms of conjugate priors. Despite of $\boldsymbol{\omega}_i$ is also a hidden variable of the reverse engineering GRN problem, it will be treated as a nuisance parameter of the model, together with σ_i^2, and will be denoted for convenience as $\boldsymbol{\theta}_i = \left[\boldsymbol{\omega}_i, \sigma_i^2\right]^\top$. Therefore, given the microarray data set as in (6.1) and the priors, the main goal of the reverse engineering problem is to infer the hidden variables

$\mathbf{x}_i$ and parameters $\boldsymbol{\theta}_i$, $\forall i$. Taking into account the AR1MA1 model (4.9) and the prior of noise (4.12), likelihood may be expressed in terms of the unknowns as a multivariate Gaussian with unknown mean and variance (see Appendix B) as,

$$p\left(\mathbf{y}_i|\,\mathbf{x}_i,\boldsymbol{\theta}_i\right)=\mathcal{N}\left(\mathbf{y}_i|\,\mathbf{R}\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i,\frac{\sigma_i^2}{\gamma_i}\mathbb{1}^N\right) \tag{4.13}$$

with $\mathbf{R}=\mathbf{T}\mathbf{Y}^\top$ and

$$\gamma_i:=\frac{1}{1+\sum_{j=1}^{G}x_i^2(j)+\omega_i^2(j)}. \tag{4.14}$$

a variance scale $\gamma_i>0$ depending on the unknowns describing the GRN such as $\gamma_i=\gamma_i\left(\mathbf{x}_i,\boldsymbol{\omega}_i\right)$, that may be conveniently expressed as

$$\gamma_i=\frac{1}{1+\mathbf{x}_i^\top\mathbf{D}_{\mathbf{x}_i}\mathbf{D}_{\boldsymbol{\omega}_i}\boldsymbol{\omega}_i}. \tag{4.15}$$

According to the VBEM learning framework, the unknowns will be hyperparametrized by conjugate independent priors as in (3.33) such as the joint probability of the hidden variables and parameters factorizes as,

$$p\left(\mathbf{x}_i,\boldsymbol{\theta}_i|\,\boldsymbol{\zeta}_i\right)=p\left(\mathbf{x}_i|\,\boldsymbol{\zeta}_{\mathbf{x}_i}\right)p\left(\boldsymbol{\theta}_i|\,\boldsymbol{\zeta}_{\boldsymbol{\theta}_i}\right) \tag{4.16}$$

with $\boldsymbol{\zeta}_i=\left\{\boldsymbol{\zeta}_{\mathbf{x}_i},\boldsymbol{\zeta}_{\boldsymbol{\theta}_i}\right\}$. Additionally, the Gaussian likelihood function in (8.7) restrict these priors to the exponential family [34].

4.2.1.1 Prior of hidden variables

Despite of the hidden variables are defined in a binary space and a discrete binary prior would be more suitable, in favor to the conjugate modeling, we are going to describe the network topology by a multivariate Gaussian distribution with unknown mean and variance as,

$$p\left(\mathbf{x}_i\right)=\mathcal{N}\left(\mathbf{x}_i|\,\mathbf{m}_{\mathbf{x}_i},\sigma_0^2\mathbb{1}^G\right) \tag{4.17}$$

with hyperparameters $\boldsymbol{\zeta}_{\mathbf{x}_i}=\left\{\mathbf{m}_{\mathbf{x}_i},\sigma_0^2\right\}$ conveniently chosen for describing a priori knowledge (see Appendix).

4.2.1.2 Prior of parameters

Taking into account the gaussian distributions in (8.7) and in (4.17), the joint probability of the hidden and observable variables is also a multivariate Gaussian. Therefore, the conjugate prior for the parameters will be a Normal scaled Inverse

Gamma distribution with unknown mean, variance, scale and shape hyperparameters as

$$p\left(\boldsymbol{\omega}_i, \sigma_i^2\right) = \mathcal{NIG}\left(\boldsymbol{\omega}_i, \sigma_i^2 \middle| \mathbf{m}_{\boldsymbol{\omega}_i}, \frac{\sigma_i^2}{\bar{\gamma}_i}\mathbb{1}^G, a_i, \bar{\gamma}_i b_i\right) \tag{4.18}$$

$$= \mathcal{N}\left(\boldsymbol{\omega}_i \middle| \mathbf{m}_{\boldsymbol{\omega}_i}, \frac{\sigma_i^2}{\bar{\gamma}_i}\mathbb{1}^G\right)\mathcal{IG}\left(\sigma_i^2 \middle| a_i, \bar{\gamma}_i b_i\right) \tag{4.19}$$

with

$$p\left(\boldsymbol{\omega}_i \middle| \sigma_i^2\right) = \mathcal{N}\left(\boldsymbol{\omega}_i \middle| \mathbf{m}_{\boldsymbol{\omega}_i}, \frac{\sigma_i^2}{\bar{\gamma}_i}\mathbb{1}^G\right) \tag{4.20}$$

$$p\left(\sigma_i^2\right) = \mathcal{IG}\left(\sigma_i^2 \middle| a_i, \bar{\gamma}_i b_i\right). \tag{4.21}$$

with hyperparameters $\boldsymbol{\zeta}_{\boldsymbol{\theta}_i} = \{\mathbf{m}_{\boldsymbol{\omega}_i}, a_i, b_i\}$ conveniently chosen for describing a priori knowledge (see Appendix). Recall that the scale of variance γ_i depends on variables and parameters describing the GRN as in (4.15). Therefore, it is not an hyperparameter and it should be modeled by an hyperprior derived from (4.17) and (4.20) .

4.2.2 Free distributions

Consequently to statistical modeling in (4.17) and (8.4), the free distributions of hidden variables and parameters as in (3.34) will be in the same exponential families than it priors. Therefore, the free distribution of binary latent variables will be a multivariate Gaussian as

$$q\left(\mathbf{x}_i\right) = \mathcal{N}\left(\mathbf{x}_i \middle| \boldsymbol{\mu}_{\mathbf{x}_i}, \boldsymbol{\Sigma}_{\mathbf{x}_i}\right) \tag{4.22}$$

with hyperparameters $\boldsymbol{\xi}_{\mathbf{x}_i} = \left\{\boldsymbol{\mu}_{\mathbf{x}_i}, \boldsymbol{\Sigma}_{\mathbf{x}_i}\right\}$.

On the other hand, the free distribution of weights and variance of noise will be a Normal scaled Inverse Gamma as

$$q\left(\boldsymbol{\omega}_i, \sigma_i^2\right) = \mathcal{NIG}\left(\boldsymbol{\omega}_i, \sigma_i^2 \middle| \boldsymbol{\mu}_{\boldsymbol{\omega}_i}, \sigma_i^2\boldsymbol{\Sigma}_{\boldsymbol{\omega}_i}, \alpha_i, \beta_i\right) \tag{4.23}$$

with

$$q\left(\boldsymbol{\omega}_i \middle| \sigma_i^2\right) = \mathcal{N}\left(\boldsymbol{\omega}_i \middle| \boldsymbol{\mu}_{\boldsymbol{\omega}_i}, \sigma_i^2\boldsymbol{\Sigma}_{\boldsymbol{\omega}_i}\right) \tag{4.24}$$

$$q\left(\sigma_i^2\right) = \mathcal{IG}\left(\sigma_i^2 \middle| \alpha_i, \beta_i\right) \tag{4.25}$$

and hyperparameters $\boldsymbol{\xi}_{\boldsymbol{\omega}_i} = \left\{\boldsymbol{\mu}_{\boldsymbol{\omega}_i}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}_i}, \alpha_i, \beta_i\right\}$.

4.2.3 Suboptimal solution for AR1MA1 model

According to VBEM method based on the AR1MA1 model, namely AR1MA1-VBEM method, each free distribution is alternatively optimized and the posterior of the unknowns are approximated by the joint free ditribution. However, the likelihood as in (8.7) defined by the generative model does not satisfy the requirements of conjugate models [5]. Specifically, the dependence of γ_i on $\mathbf{x}_i$ and $\boldsymbol{\omega}_i$ breaks the symmetries of conjugate models. Moreover, the derivation of the probability of the scale variance γ_i from the priors of $\mathbf{x}_i$ is too complex to the point of being impossible analytically and the formulation of an alternative conjugate model is not possible. Hence, we are going to propose an approach that allows our statistical model to satisfy the conjugate requirements. As a suboptimal solution, we are going to remove the implicit dependency of the variance scale γ_i on $\mathbf{x}_i$ and $\boldsymbol{\omega}_i$. Specifically, we propose a fixed point approach where the scale were approximated according to the most probable value of the unknowns given the other as

$$\gamma_i\left(\mathbf{x}_i, \boldsymbol{\omega}_i\right) \approx \gamma_i\left(\bar{\mathbf{x}}_i, \bar{\boldsymbol{\omega}}_i\right) \tag{4.26}$$

with the most probable values, i.e. the modes, as

$$\bar{\mathbf{x}}_i = \arg\max_{\mathbf{x}_i} p\left(\mathbf{x}_i \middle| \mathbf{y}_i, \boldsymbol{\omega}_i, \sigma_i^2\right) \tag{4.27}$$

$$\bar{\boldsymbol{\omega}}_i = \arg\max_{\boldsymbol{\omega}_i} p\left(\boldsymbol{\omega}_i \middle| \mathbf{y}_i, \mathbf{x}_i, \sigma_i^2\right). \tag{4.28}$$

These posterior probabilities are also unknowns and should be computed at every iteration of the VBEM algorithm, extremely increasing the numerical complexity of the method. As an alternative, we propose as the best approximation to these posteriors the updated free distributions as

$$p\left(\mathbf{x}_i \middle| \mathbf{y}_i, \boldsymbol{\omega}_i, \sigma_i^2\right) \approx q\left(\mathbf{x}_i \middle| \hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t')}\right) \tag{4.29}$$

$$p\left(\boldsymbol{\omega}_i \middle| \mathbf{y}_i, \mathbf{x}_i, \sigma_i^2\right) \approx q\left(\boldsymbol{\omega}_i \middle| \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t')}, \sigma_i^2\right) \tag{4.30}$$

with t' index of the immediately last iteration of the VBEM algorithm. According to (4.22) and (4.24), the free distributions are multivariate Gaussians and its most probable values will be the means

$$\bar{\mathbf{x}}_i = \hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t')} \tag{4.31}$$

$$\bar{\boldsymbol{\omega}}_i = \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t')} \tag{4.32}$$

According to (4.31) and (4.32), scale of variance (4.26) will be

$$\bar{\gamma}_i\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t')}, \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t')}\right) = \frac{1}{1 + \sum_{j=1}^{G} \hat{\mu}_{\mathbf{x}_i}^{2(t')}(j) + \hat{\mu}_{\boldsymbol{\omega}_i}^{2(t')}(j)}. \tag{4.33}$$

Under assumption (4.33), scale $\bar{\gamma}_i$ do not depends directly on the unknowns and an approximated likelihood function, satisfying the conjugate model symmetries, may be defined as

$$p\left(\mathbf{y}_i|\,\mathbf{x}_i,\boldsymbol{\theta}_i\right) \approx \mathcal{N}\left(\mathbf{y}_i|\,\mathbf{R}\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i,\frac{\sigma_i^2}{\bar{\gamma}_i}\mathbb{1}^N\right). \tag{4.34}$$

Therefore, VBEM method requires to update the scale of the variance as an additional hyperparameter at every step as

$$\bar{\gamma}_i = \frac{1}{1+\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t')}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t')}}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t')}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t')}} \tag{4.35}$$

4.3 AR1MA1-VBEM algorithm

According to VBEM method introduced in Chapter 3, lower bound is alternatively optimized as described in (3.36) and (3.37), leading to the following hyperparameters updating rules

$$q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right) \propto \mathrm{e}^{\left\langle \ln p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)\right\rangle_{q\left(\boldsymbol{\theta}_i|\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t)}\right)}+\ln p(\mathbf{x}_i)} \tag{4.36}$$

$$q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right) \propto \mathrm{e}^{\left\langle \ln p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)p(\mathbf{x}_i)\right\rangle_{q\left(\mathbf{x}_i|\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)}+\ln p(\boldsymbol{\theta}_i)}. \tag{4.37}$$

As an initial value $t=0$, these hyperparameters may be set up same as its prior ones, such as $\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(0)} = \boldsymbol{\zeta}_{\mathbf{x}_i}$ and $\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(0)} = \boldsymbol{\zeta}_{\boldsymbol{\theta}_i}$. On the other hand, its updating rules requires to compute several expected values of likelihood and priors (see Appendix D). Details on derivation of the hyperparameters updating rules can be found in Appendixes **??**- E.3. Additionally, the AR1MA1-VBEM algorithm computes the approximated scale of variance $\bar{\gamma}_i$ as (4.35) before every VBE and VBM step as well as before updating the lower bound.

4.3.1 AR1MA1-VBE step

First, after the t-th iteration, the scale of the variance will be updated as

$$\bar{\gamma}_i = \gamma_i\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t)},\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}\right) = \frac{1}{1+\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{\top(t)}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t)}}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t)}} \tag{4.38}$$

Then, taking into account the updating rule (4.36), the hyperparameters learn-

ing rules of $\mathbf{x}_i$ that defines the VBE step may be updated as

$$\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} = \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{\top(t+1)} \left(\mathbf{y}_i^{\top} \mathbf{R} \mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}} \frac{\hat{\alpha}_i^{(t)}}{\hat{\beta}_i^{(t)}} \bar{\gamma}_i + \frac{1}{\sigma_0^2} \mathbf{m}_{\mathbf{x}_i}^{\top} \right)^{\top} \tag{4.39}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)} = \left(\bar{\gamma}_i \mathbf{B} \circ \left(\frac{\hat{\alpha}^{(t)}}{\hat{\beta}^{(t)}} \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)} \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)\top} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)} \right) + \frac{1}{\sigma_0^2} \mathbb{1}^G \right)^{-1}. \tag{4.40}$$

4.3.2 AR1MA1-VBM step

Subsequently, the scale of the variance is updated according the previous VBE step as

$$\bar{\gamma}_i = \gamma_i \left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}, \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)} \right) = \frac{1}{1 + \hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} \mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}} \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)} \mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}}} \tag{4.41}$$

Then, according to (4.37), the hyperparameters learning rules of $\boldsymbol{\omega}_i$ and σ_i^2 that defines the VBM step may be updated as

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)\top} \left(\mathbf{y}_i^{\top} \mathbf{R} \mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}} \bar{\gamma}_i + \mathbf{m}_{\boldsymbol{\omega}_i}^{\top} \mathbf{S}_{\boldsymbol{\omega}_i}^{-1} \right)^{\top} \tag{4.42}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)} = \left(\mathbf{S}_{\boldsymbol{\omega}_i}^{-1} + \mathbf{B} \circ \left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t)} \hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)\top} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)} \right) \bar{\gamma}_i \right)^{-1} \tag{4.43}$$

$$\hat{\beta}_i^{(t+1)} = b_i + \frac{1}{2} \left(\mathbf{y}_i^{\top} \mathbf{y}_i \bar{\gamma}_i + \mathbf{m}_{\boldsymbol{\omega}_i}^{\top} \mathbf{m}_{\boldsymbol{\omega}_i} + \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{\top(t+1)} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{-1(t+1)} \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} \right) \tag{4.44}$$

$$\hat{\alpha}_i^{(t+1)} = a_i + \frac{N+G}{2}. \tag{4.45}$$

4.3.3 Lower bound updating rule

After two consecutive VBE and VBM steps, the scale of the variance is updated as

$$\bar{\gamma}_i = \gamma_i \left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}, \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} \right) = \frac{1}{1 + \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{\top(t+1)} \mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}} \mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}} \hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}} \tag{4.46}$$

Finally, lower bound is updated as

$$
\begin{aligned}
& \mathcal{F}\left[q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right), q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)\right] = \\
= & -\frac{N+G}{2}\ln 2\pi + \frac{N}{2}\ln\bar{\gamma}_i - \frac{N}{2}\left(\ln\hat{\beta}_i^{(t+1)} - \Psi\left(\hat{\alpha}_i^{(t+1)}\right)\right) - \frac{G}{2}\ln\sigma_0^2 \\
- & \frac{\bar{\gamma}_i}{2}\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\mathbf{y}_i^\top\mathbf{y}_i + \bar{\gamma}_i\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\mathbf{y}_i^\top\mathbf{R}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} \\
- & \frac{\bar{\gamma}_i}{2}\mathrm{trace}\left(\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)}\right) \\
- & \frac{\bar{\gamma}_i}{2}\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{\top(t+1)}\left(\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right)\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} \\
- & \frac{1}{2\sigma_0^2}\left(\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right)^\top\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right) + \mathrm{trace}\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right) \\
- & \frac{1}{2}\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} - \mathbf{m}_{\boldsymbol{\omega}_i}\right)^\top\bar{\gamma}_i\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} - \mathbf{m}_{\boldsymbol{\omega}_i}\right) - \frac{1}{2}\mathrm{trace}\left(\bar{\gamma}_i\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)}\right) + \frac{G}{2} \\
+ & \frac{G}{2}\ln\bar{\gamma}_i + \frac{1}{2}\ln\left|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{\top(t+1)}\right| - \ln\frac{\hat{\beta}_i^{(t+1)\,\hat{\alpha}_i^{(t+1)}}\Gamma\left(a_i\right)}{\left(\bar{\gamma}_i b_i\right)^{a_i}\Gamma\left(\hat{\alpha}_i^{(t+1)}\right)} \\
+ & (\hat{\alpha}_i^{(t+1)} - a_i)\left(\ln\hat{\beta}_i^{(t+1)} - \Psi\left(\hat{\alpha}_i^{(t+1)}\right)\right) + (\hat{\beta}_i^{(t+1)} - b_i)\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}} \\
+ & \frac{G}{2}\ln 2\pi + \frac{G}{2}\ln\left|\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right| + \frac{G}{2}. \qquad (4.47)
\end{aligned}
$$

and VBEM algorithm iterates until the convergence criterion were satisfied as

$$
\left|\frac{\mathcal{F}\left[q\left(\mathbf{x}_i,\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_i^{(t+1)}\right)\right] - \mathcal{F}\left[q\left(\mathbf{x}_i,\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_i^{(t)}\right)\right]}{\mathcal{F}\left[q\left(\mathbf{x}_i,\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_i^{(t+1)}\right)\right]}\right| \leq \epsilon. \qquad (4.48)
$$

with threshold $\epsilon << 1$.

4.4 Validation by simulation

We have validated the VBEM method presented above with synthetic data by simulation. For a given number of G genes, we have set the network topology of a specific gene i' with a predefined number of parents. According to biological knowledge suggesting that each child is regulated by a limited number of parents, as a general setting, we are going to choose the network topology of this gene with

a number of parents lower than $\frac{G}{3}$. The rest of the network settings: regulatory type and topology, was generated using priors in (4.17) and (4.20) with subjective hyperparameters as discussed in section C. Once the genetic network was simulated, synthetic data were generated according to likelihood (4.34), with N time samples and different levels of noise. Instead of working directly with the variance of noise, we have used the signal-to-noise ratio (SNR) that considers the logarithm transformation of the relative variance of noise, assuming unitary variance of the signal, as

$$\mathrm{SNR} := 10 \log_{10} \frac{1}{\sigma_i^2} \tag{4.49}$$

Specifically, we have simulated data with $\mathrm{SNR} = 1, \ldots, 80$. For each given data set, we are going to infer the network topology for the gene of interest i' with the predefined connectivity pattern. Genetic relationships are established according to the following binary decision criteria

$$x_{i'}(j) = \begin{cases} 1, & q\left(x_{i'}(j) = 1\right) > q\left(x_{i'}(j) = 0\right) \\ 0, & q\left(x_{i'}(j) = 1\right) \leq q\left(x_{i'}(j) = 0\right) \end{cases} \tag{4.50}$$

For each setting, we are going to analyze the performance of AR1MA1-VBEM method. Moreover, we are going to compare its results with the ones of the VBEM method based on the AR1 model presented in [89], that will be refereed as AR1-VBEM method. Recall that such kind of settings have a highly sparse network, i.e. the binary variable $\mathbf{x}_{i'}$ describing the connectivity pattern has a large number of zeros, refereed as negatives (N). Opposite, the fraction of unitary values known as positives (P), is significantly lower. We have plotted the percentage of errors in the learned network versus the level of noise for both AR1-VBEM and AR1MA1-VBEM methods. From the point of view of Information Retrieval, at any binary decision as this kind of GRN topology learning methods, two type of errors are expected:

Type I errors , refereed as false positives (FP), when a genetic relationship between two genes is erroneously established as positive.

Type II errors , refereed as false negatives (FN), when a genetic relationship between two genes is erroneously established as negative.

On the other hand, two kind of correct outcomes are possible: true positives (TP) and true negatives (TN), when a positive or negative genetic relationships is correctly established respectively.

4.4.1 Data set with $G = 25$ and $N = 25$

As a first scenario, we have considered a data set with $G = 25$ genes, $N = 25$ samples and a gene with 8 parents. Figure 4.2 represents one of the generated data sets as a heatmap, with genes as rows and samples as columns.

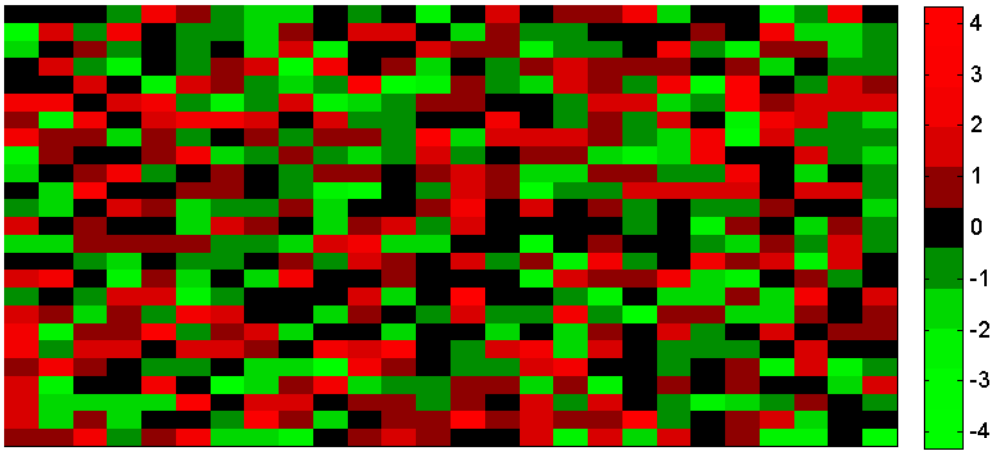


Figure 4.2: Heatmap for a synthetic data set with $G = 25$ genes, $N = 25$ samples and 8 parents. The color and intensity of each element represent the expression level of genes (rows) at every time sample (columns).

We have plotted the percentage of errors (FP and FN) in the learned network versus the level of noise for both AR1-VBEM and AR1MA1-VBEM methods. For a statistically significant result, we have performed up to 100 realizations for each value of the SNR. Figure 4.3 compares results for both VBEM methods. Additionally, we have plotted results for a uniform random assignment, considered as the most undesirable performance with a constant error level around 50%. Both VBEM methods show a dependency with SNR. For higher levels of noise, i.e. lower SNR, they produce similar results with unacceptable error rates. However, AR1MA1-VBEM outperform results obtained with AR1-VBEM method for SNR > 20. At these lower levels of noise, the behavior of AR1MA1-VBEM are satisfactory and produces an error under percentile 5%, whilst the AR1-VBEM do not.

As number of errors depends on settings of the data set, they are not good statistics for comparing results. Alternatively, the error rate between the number of actual errors and the maximum number of errors are usually given. Therefore, the false negative rate (FNR) is defined as the total number of FN over the total number of positives, whilst the false positive rate (FPR) is the fraction between

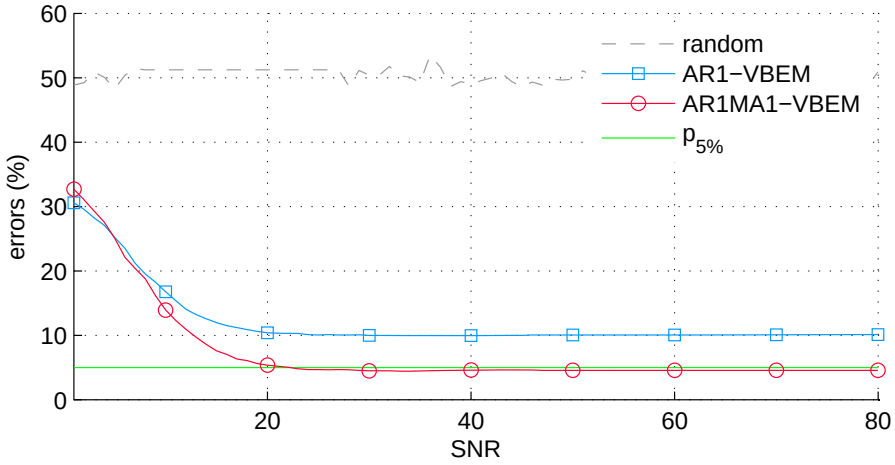


Figure 4.3: Performance of GRN inference of random assignment, AR1-VBEM method and AR1MA1-VBEM method for a synthetic data set with $G = 25$ genes, $N = 25$ samples and 8 parents. Random decision has the worst performance with a constant error level around 50%. On the other hand, VBEM methods show similar performance at higher levels of noise with unsatisfactory results. However, AR1MA1-VBEM method outperforms results of the AR1-VBEM one with SNR > 10 and it reaches a level of error under the 5% percentile, whilst AR1-VBEM method do not.

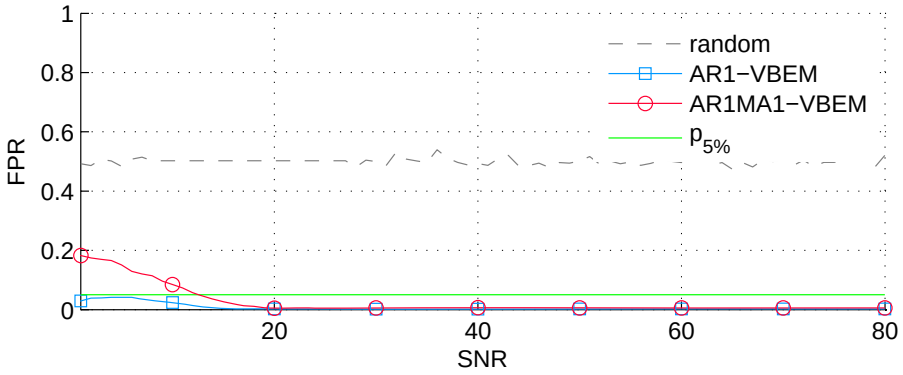
FP and the total number of negatives.

$$\text{FNR} := \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{4.51}$$

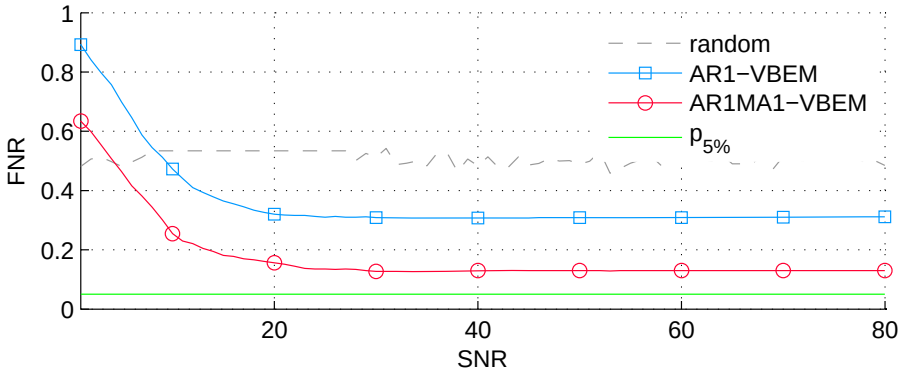
$$\text{FPR} := \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{4.52}$$

Figure 4.4 shows both FPR and FNR versus SNR with same data set as before for both VBE methods and random assignment. Specifically, in Figure 4.4(a) it is plotted the FPR, where random decision is once again the worst performance with a constant FPR about 0.5. On the other hand, both VBEM methods have a low FPR that falls to zero for SNR > 20. Such a good results for this type of error, as it was introduced above, is due to the sparsity nature of the network topology with 68% of zero elements. Additionally, Figure 4.4(b) shows the FNR with a bad performance of both VBEM methods at high levels of noise, even worst than the random assignment. However, AR1MA1-VBEM method outperforms results of the AR1-VBEM one, with a minimum FNR around 0.13 whilst the other is over 0.31 for any SNR.

Results shown above may be summarized in one plot. Figure 8.1 shows the performance of both VBEM methods for the same synthetic data set with $G = 25$ genes, $N = 25$ samples and 8 parents. Moreover, the ratio of error types is depicted



(a)



(b)

Figure 4.4: Error rates in GRN inference of random assignment, AR1-VBEM method and AR1MA1-VBEM method for a synthetic data set with $G = 25$ genes, $N = 25$ samples and 8 parents. Random assignment depicts the most undesirable scenario with constant error rate about 0.5 for any level of noise. (a) Both VBEM methods reveal a low FPR at higher levels of noise that quickly goes to zero. (b) AR1MA1-VBEM method outperforms results of the AR1-VBEM one, but no one reach a FPR under the 5% percentile.

in this plot. As shown above, both VBEM methods have similar performance at higher levels of noise with unsatisfactory results. However, in this plot it is revealed that most of errors provided by AR1-VBEM are FP. On the other hand, AR1MA1-VBEM reduces both type of errors simultaneously to the point of nearly extinguish the FP one and finally, it reaches an overall error under the 5% percentile.

Notice that relationships previously analyzed arise from a probabilistic decision and results would depend on a probability threshold. For example, binary decision in (4.50) is equivalent to a probability threshold $q\left(x_i(j)=1\right) > 0.5$. Therefore, results as shown in Figure 8.1 are susceptible to change according to the threshold considered. Higher thresholds increase our confidence in the estimations and it would outperform the precision of the learning method, also know as positive predictive value (PPV), defined as the fraction of TP over all positives retrieved. However, this also would increase the FNR.

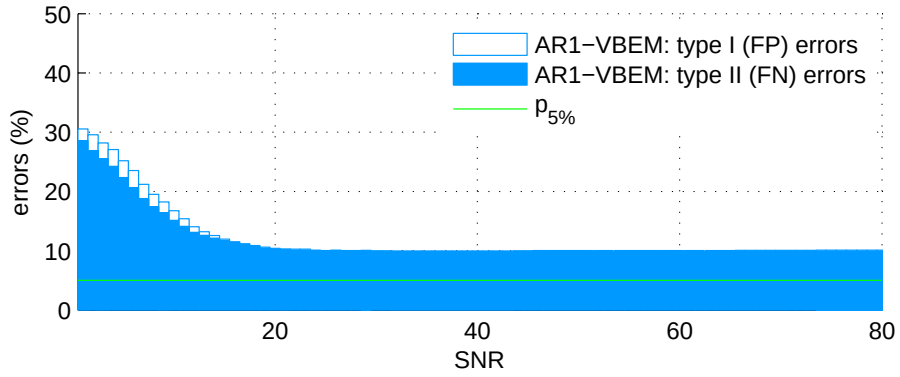
$$\mathrm{PPV} := \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{4.53}$$

Opposite, lower thresholds increase the true positive rate (TPR), also known as sensitivity or recall, defined as the fraction of positives correctly estimated. Nevertheless, it also would increase the FPR.

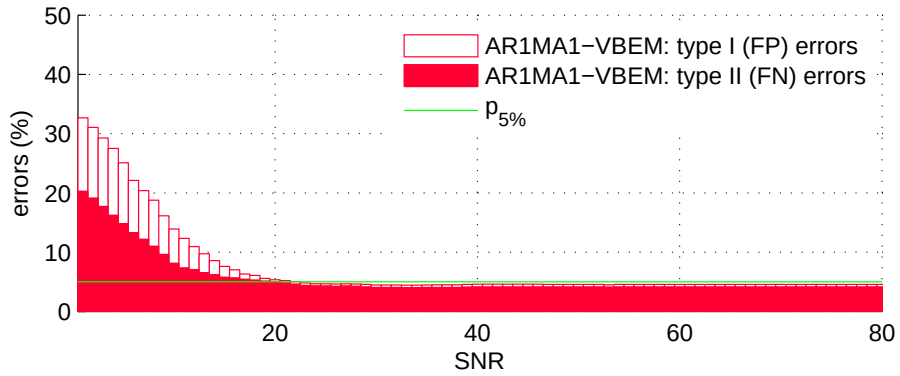
$$\mathrm{TPR} := \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1-\mathrm{FNR} \tag{4.54}$$

Ideally, one would desire to minimize both kind of errors and maximize the hits together. But, when the learning method does not allow it, one may prefer to optimize a specific one. For GRN reverse engineering problems with high levels of sparsity and a negative dominant class, it would be preferred to minimize the FNR and consequently maximizing the TPR. A common strategy for analyzing the trade-off between each error and hit rates are the receiver operating characteristic (ROC) and the Precision-Recall (PR) curves.

The ROC curve is a plot with the TPR versus the FPR as the decision threshold is changed. Intuitively, a ROC curve illustrates the confidence in any result for a desired hit rate, i.e. the specificity cost of the learning method when the sensitivity is increased. For example, in random performance it is expected to have same FPR as TPR regardless of the threshold. Therefore, the ROC curve is a line through the origin with unitary slope, known as no-discrimination (ND) line. Opposite, a perfect result would pass through point $(0,1)$. ROC analysis summarizes the performance of any method within its area under the curve (AUROCc). This statistic is 0.5 for a random performance and it would have a maximum value of one for a perfect result. In Figure 4.6 we have plotted the ND line and the ROC curve for both VBEM methods with previous settings at high level of noise with SNR = 10. Table 4.1 shows the area under both curves. It can be noticed that



(a)



(b)

Figure 4.5: Performance of GRN inference of AR1-VBEM method and AR1MA1-VBEM method for a synthetic data set with $G = 25$ genes, $N = 25$ samples and 8 parents. The overall error percentage is represented as a bar at each level of noise. The fraction FP are proportionally depicted as a filled bar and the FN fraction as an outlined bar. (a) AR1-VBEM method shows unsatisfactory results at any level of noise. (b) AR1MA1-VBEM shows a dependency with the level of noise that outperforms AR1-VBEM method and finally it reaches an error level under the 5% percentile.

AR1MA1-VBEM method outperforms the results obtained with the AR1-VBEM one with higher AUROCc.

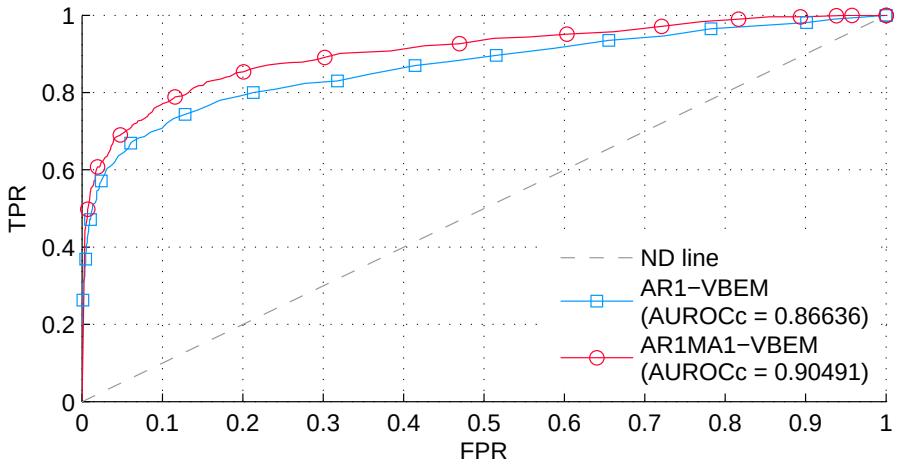


Figure 4.6: ROC analysis of AR1-VBEM method and AR1MA1-VBEM method for a synthetic data set with $G = 25$ genes, $N = 25$ samples, 8 parents and SNR $= 10$. Random assignment is illustrated by the ND line. AR1MA1-VBEM method outperforms results of the AR1-VBEM one with higher AUROCc.

In GRN reverse engineering problems, with sparse networks where negative class is much more dense than the positive one, precision-recall space is preferred for analyzing the performance of a learning method. The PR curve plots the PPV versus the TPR as the decision threshold is changed. Intuitively, PR curve reveals the ability of a method to retrieve any positive as the decision threshold is increased. If no any positive (TP or FP) is retrieved as from any probability threshold, PPV curve will diverge and PR curve will not reach point $(0, 1)$. Such case represents the limitations of the learning method to infer any genetic relationship. As in ROC analysis, the area under the PR curve (AUPRc) summarizes the overall performance of the method. In this space, the slope of the ND line will depends on the sparsity of the network. Therefore, the AUPRc for the ND line will be lower than 0.5 according to the sparsity level of the network, whilst the maximum value of one would be for a perfect performance. In Figure 4.7 we have plotted the ND line and the PR curve for both VBEM methods, for previous settings and high level of noise with SNR $= 10$. Table 4.1 shows the area under both curves. It can be noticed that AR1MA1-VBEM method outperforms the results obtained with the AR1-VBEM one with higher AUPRc. Moreover, the limitations of the AR1-VBEM to infer any genetic relationships are revealed in this representation, being impossible to reach point $(0, 1)$ whilst AR1MA1-VBEM method does.

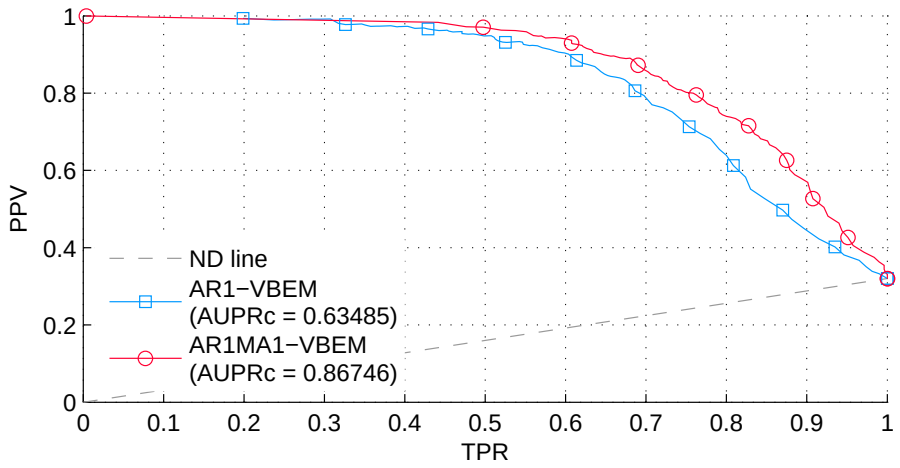


Figure 4.7: PR analysis of AR1-VBEM method and AR1MA1-VBEM method for a synthetic data set with $G = 25$ genes, $N = 25$ samples and 8 parents. Random assignment is illustrated by the ND line. AR1MA1-VBEM method outperforms results of the AR1-VBEM one with higher AUPRc.

	AR1-VBEM	AR1MA1-VBEM
AUROCc	0.8663	0.90491
AUPRc	0.63485	0.86746

Table 4.1: Area under the ROC and PR curves in the performance of AR1-VBEM and AR1MA1-VBEM methods for a data set with $G = 25$, $N = 25$, 8 parents and SNR = 25.

4.4.2 Data set with $G = 100$ and $N = 50$

We have considered a more realistic data set of bigger size with $G = 100$ genes and a reduced number of time samples $N = 50$, where studied gene has 15 parents. Figure 4.8 shows the performance of both VBEM methods for this settings. At higher levels of noise, AR1MA1-VBEM produces much more number of errors than AR1-VBEM method. Whilst, AR1-VBEM produces mainly FN errors, AR1MA1-VBEM method reduces quickly both kind of errors to the point of outperform the AR1-VBEM one for SNR ≥ 10. Moreover, AR1MA1-VBEM produces satisfactory results with errors under the 5% percentile for SNR ≥ 18 whilst AR1-VBEM method do not..

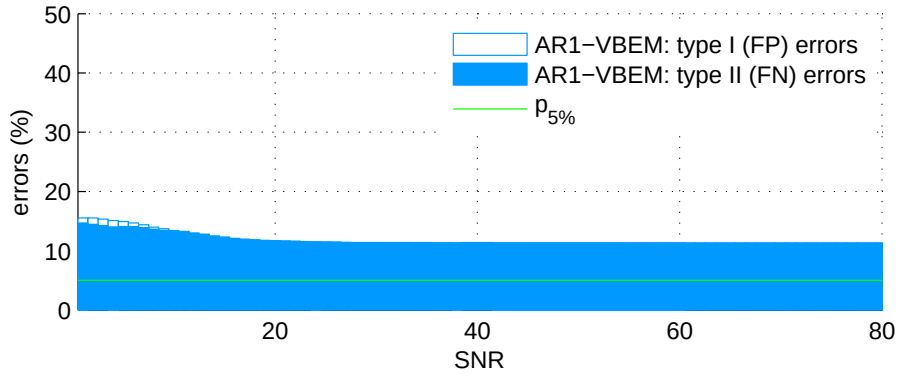
On the other hand, Figure 4.9 represents the ROC and PR curves for the AR1-VBEM, AR1MA1-VBEM method and the ND line for previous setting with $G = 100$ genes, $N = 50$ samples and 15 parents, with SNR $= 10$ where both methods produces similar error rates. As it can be notice, AR1MA1-VBEM method outperforms results obtained with the AR1-VBEM one. Table 4.2 summarizes the area under curves.

	AR1-VBEM	AR1MA1-VBEM
AUROCc	0.58009	0.87108
AUPRc	0.37187	0.67828

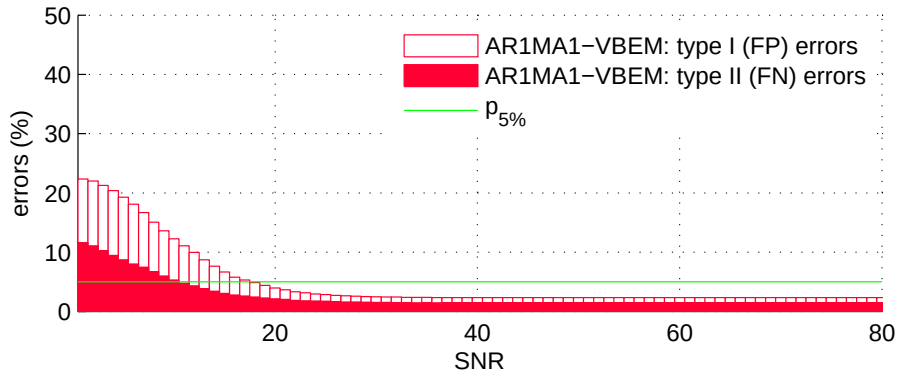
Table 4.2: Area under the ROC and PR curves in the performance of AR1-VBEM and AR1MA1-VBEM methods for a data set with $G = 100$, $N = 50$, 15 parents and SNR $= 10$.

4.5 Validation with *in-silico* data

Validation of AR1MA1-VBEM method by simulation promises an exceptional performance in reverse engineering GRN problems. However, such a good result is consequence of an idyllic scenario where data are generated by the same observational model. These results should be regarded as a ceiling away from behavior that would be expected in a real case. Therefore, apart from simulation, we have validated the AR1MA1-VBEM method with synthetic data generated by an independent observational model. Specifically, we have used GeneNetWeaver (GNW) as a benchmark generator that considers realistic networks to simulate microarray data [83]. Despite also considering synthetic data, validation of the AR1MA1-VBEM method by this tool is of special interest for at least three reasons. First, data are generated by an external model independent of the observational one that considered by the inference method. Second, instead of assuming a linear factor model as the AR1MA1 one, GNW considers an ordinary differential equa-

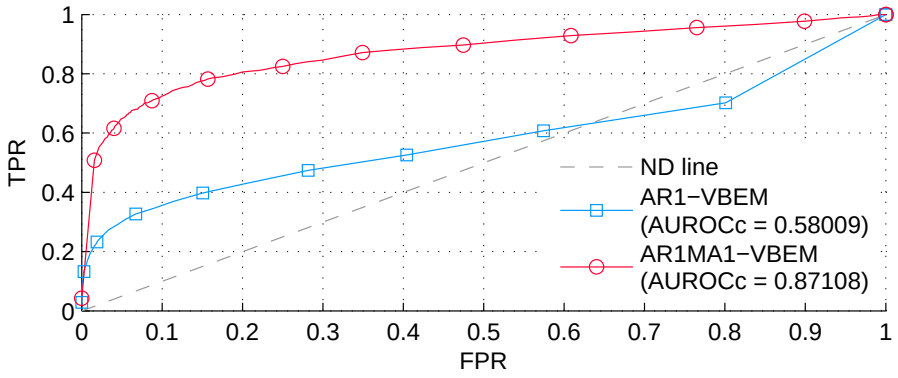


(a)

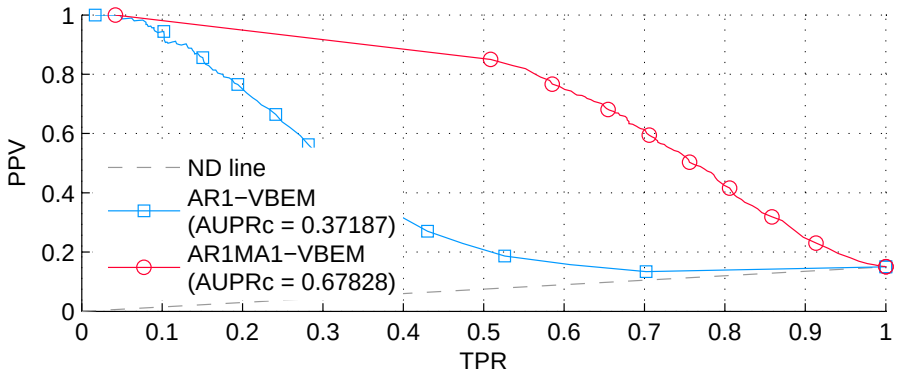


(b)

Figure 4.8: Performance of GRN inference of AR1-VBEM method and AR1MA1-VBEM method for a synthetic data set with $G = 100$ genes, $N = 50$ samples and 15 parents. The overall error percentage is represented as a bar for each level of noise. The fraction of FP are represented as a colored bar and the FN as an outlined bar.(a) AR1-VBEM method produces unsatisfactory at any level of noise. (b) AR1MA1-VBEM produces much more errors (both types) at higher levels of noise. However, AR1MA1-VBEM outperforms AR1-VBEM method providing satisfactory results under 5% percentile.



(a)



(b)

Figure 4.9: ROC and PR analysis for AR1-VBEM method and AR1MA1-VBEM method for a synthetic data set with $G = 100$ genes, $N = 50$ samples and 15 parents and SNR = 10. Random assignment is illustrated by the ND line. (a) ROC curve. (b) PR curve. AR1MA1-VBEM method outperforms the AR1-VBEM one with higher AUROCc and AUPRc.

tion model endowed with a transcriptional regulatory kinetic. Finally, GNW includes real network topologies of two biological models that simulates synthetic data within a plausible biological background. These kind of synthetic expression profiles based on real biological knowledge is known as *in-silico* data. Therefore, considering an independent observational model endowed with a biological background, results and properties derived from the performance will have a more general validity.

GNW includes the network topologies of two species: *Saccharomyces cerevisiae* and *Escherichia Coli*, the two most studied eukaryote and prokaryote organism respectively. *S.cerevisiae*, or common yeast, has been studied in biological research from years. Its genome, with more than six thousands genes, was sequenced at mid-90s and has been widely studied during its cellular and metabolic cycles. GNW encloses a GRN of the yeast with 4441 genes and 12873 documented interactions. For our validation, we are going to consider a data set of lower size. Specifically, we have extracted a subnetwork with $G = 25$ genes and we have generated a time series with $N = 51$ samples. Figure 4.10 shows the GRN (both topology and regulatory type) with 49 genetic relationships. Greedy subnetwork selection and kinetic parameters of generative model were randomly established by default settings of the software. Therefore, the sparsity level of this settings is not controlled. The extracted GRN includes genes from 10 to zero parents.

We have set up GNW with a multifactorial profile that generates data by multiple variations of the given network. Additionally, an stochastic variation of the deterministic generative model with variance 0.05 is considered. Finally, simulated data set are mapped to expression fold change by a 2-based logarithmic transformation. Figure 4.11 represents the simulated expression profile as a heatmap, with genes as rows and samples as columns.

We have considered previous data set to infer the topology of the known regulatory network with the AR1MA1-VBEM method with subjective settings as described in Appendix C. Additionally we have compared it performance with other GRN inference method. ARACNE, is an algorithm that considers an exponential probabilistic model and computes the mutual information to establishing statistical independence from gene expression time series [52]. Specifically, we have run the command line version of the ARACNE2 algorithm with default settings. Figure 8.3 represents the ROC and PR curves for both ARACNE and AR1MA1-VBEM methods and the ND line. As it can be notice, AR1MA1-VBEM method outperforms results obtained with ARACNE algorithm with higher AUROCc and AUPRc. Table 8.1 summarizes the area under curves.

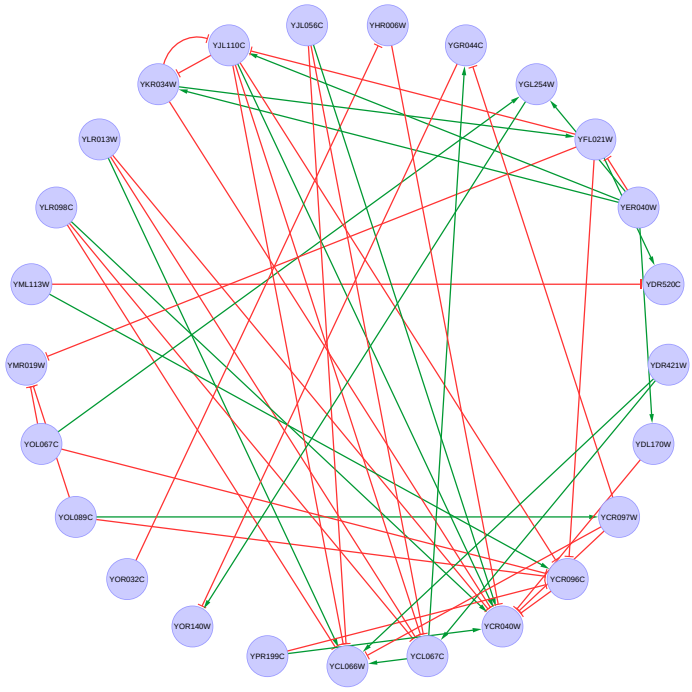


Figure 4.10: Yeast gene regulatory subnetwork with $G = 25$ extracted from the complete one from GNW. Genes are represented as nodes and regulatory interactions are depicted as edges.

	ARACNE	AR1MA1-VBEM
AUROCc	0.5263	0.6218
AUPRc	0.0931	0.2118

Table 4.3: Area under the ROC and PR curves in the performance of AR1MA1-VBEM method and ARACNE for a data set with $G = 25$ and $N = 51$.

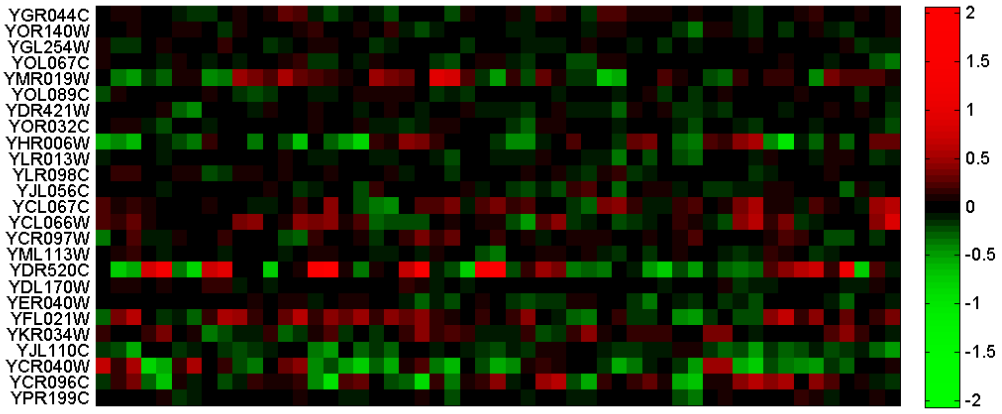
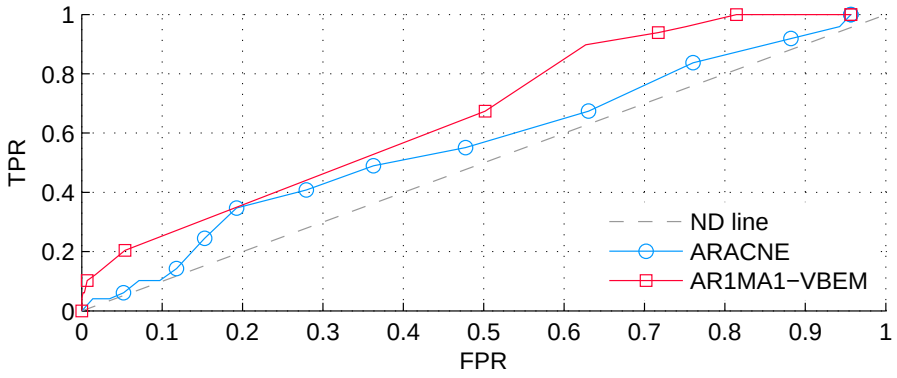
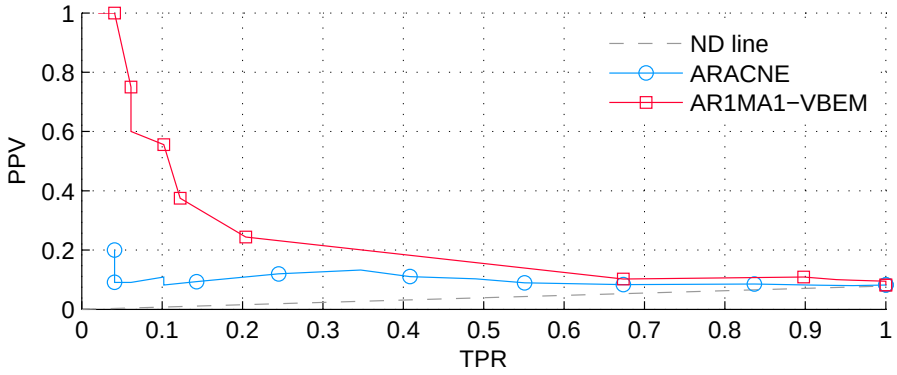


Figure 4.11: Heatmap for *in-silico* data set for yeast, the logarithmic expression fold change, simulated by multifactorial perturbations of the GRN with $G = 25$ genes and $N = 51$ time samples. The color and intensity of each element represent the expression level of genes (rows) at every time sample (columns).



(a)



(b)

Figure 4.12: ROC and PR analysis for AR1-VBEM, AR1MA1-VBEM and ARACNE methods for a synthetic data set with $G = 25$ genes and $N = 51$ time samples. Random assignment is illustrated by the ND line. (a) ROC curve. (b) PR curve. AR1MA1-VBEM method outperforms ARACNE algorithm with higher AUROCc and AUPRc.

Chapter 5

Gene regulatory network inference for yeast cell cycle

Saccharomyces cerevisiae (SCE), commom budding yeast, is the most simple eukaryotic organism. This fungi is considered by researchers as a biological model in many studies due to its easy manipulation and cycle timings. SCE was the first eukaryotic organism whose DNA was completely sequenced. Its genome is constituted by 16 chromosomes with more than 6000 genes [20]. It is believed that SCE shares about 20%-30% of its genome with human. The study of its simple but not trivial cellular mechanisms led Biologist a new insight in the global expression programs. The study of abstract models as gene regulatory networks (GRN) allows to map the metabolic pathways controlling the cell development. However, GRN are abstracts models that project different kinds of biomolecular interactionships at a genetic level. Main interest of these kind of models is its ability to describe causal relationships between genes, allowing the identification of genes sharing same ontology and figuring out the dynamical behavior of cell functions.

5.1 Yeast cell-cycle

The yeast cell-cycle is the mechanism whereby the cell grows and divides into two daughter cells. The complexity of this mechanism in SCE is neither trivial nor complex as in mammalian organisms. Molecular biologists have dissected and characterized theirs individual components and interactions to derive a consensus GRN [15]. During yeast cell cycle, DNA is duplicated and segregated to daughter cells. DNA replication and chromatin separation, process known as mitosis, occur in temporally distinct stages separated by two interphases. First, DNA is synthesized during the S-phase. At this point, during G1 interphase, if DNA damage

is detected cell-cycle is arrested. After S-phase and before mitosis, at G2 phase, DNA damage is checked again. Checkpoints are enforced by a family of molecular complexes compounded by a kninase and a protein, known as cyclin-dependent kinase (CDK). In SCE exist only one CDK, called CDC28, and two cyclins clusters compounded by nine different enzymes: CLN1-3 and CLB1-6. Genes encoding the synthesis of these molecules are usually refereed with same name than its products. The cell cycle progression requires the successive activation and inhibition of these genes, mediated by others as: (*i*) SBF, MBF and MCM1 whose products promotes the cyclins synthesis, (*ii*) CDC20, CDH1 and SCF that degrades the cyclins, (*iii*) SIC1, CDC6 and FAR1 that are CDK inhibitors or (*iv*) SWE1 that dephosphorylizes CDC28.

5.2 Yeast GRN inference from microarray time series

The molecular schemata of eukaryotic cell cycle is known in more detail for SCE than for any other organism [45]. AR1MA1-VBEM method presented in Chapter 4 has been applied in microarray time series for inferring the GRN. Specifically, the GRN of yeast cell-cycle during its G1-interphase. Spellman's expression data set has been considered [87]. This experiment measures the evolution of gene expression in two different cultures of synchronous yeast [58]. While reference one was grown in normal conditions, the experimental culture was oxygen deprived and grown in anaerobic conditions. Gene expression was profiled after every 7 minutes during two hours and a correlation analysis have identified about 800 genes involved in cell-cycle regulation.

The GRN described in [45] is considered as a ground truth of the regulatory mechanism. This network considers only 16 genes grouped in clusters, leading to a total number of $G = 11$ metagenes. Table 5.1 shows the names of considered genes and the cluster to which belongs.

cluster	genes
MBF	SWI6 and MBP1
SBF	SWI6 and SWI4
CDC14-20	CDC14 and CDC20
CLB1-2	CLB1 and CLB2
CLB5-6	CLB5 and CLB6
CLN1-2	CLN1 and CLN2

genes
CLN3
SIC1
CDH1
MCM1
SWI5

Table 5.1: Gene clusters considered in the GRN of yeast cell-cycle.

Microarray data, logarithmic expression fold change, of Spellman's experiment

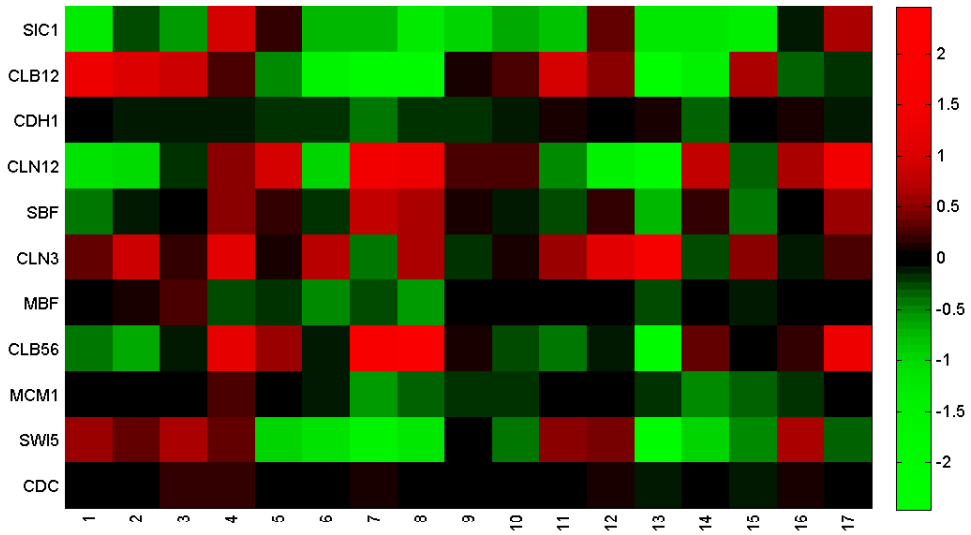


Figure 5.1: Yeast time series for $G = 11$ (meta)genes and $N = 17$ time samples.

was downloaded from GEO [3]. The expression profile of metagenes was computed as the averaged expression of genes in same cluster. Figure 5.1 shows a heatmap for considerd microarray time series with $N = 17$ time samples and $G = 11$ genes.

Inference procedure by AR1MA1-BEM method follows the scheme described in Chapter 4 with *in-silico* data, setting subjective priors but with a high detection threshold of 0.9. Therefore, genetic relationships with a higher confidence will be retrieved. Figure shows resulting GRN with a total of 65 relationships.

Main problem on validation results on real data is that, in case it were known, ground thrust corresponds to an hypothetical abstract model constructed to explain causal relationships. Whilst GRN are useful on Systems Biology studies, they are incorrect form the Molecular Biology point of view. Actual regulatory process is mediated by a variety of molecules, as transcription factors or microRNA, at different stages. However, presented application with SCE shows the potential of AR1MA1-VBEM methodology to capture this kind of functional relationships that may help Biologist to figure out the cell development machinery. Figure shows the more relevant estimations retrieved with AR1MA1-VBEM method, overlapped with the consensus GRN presented in [45].

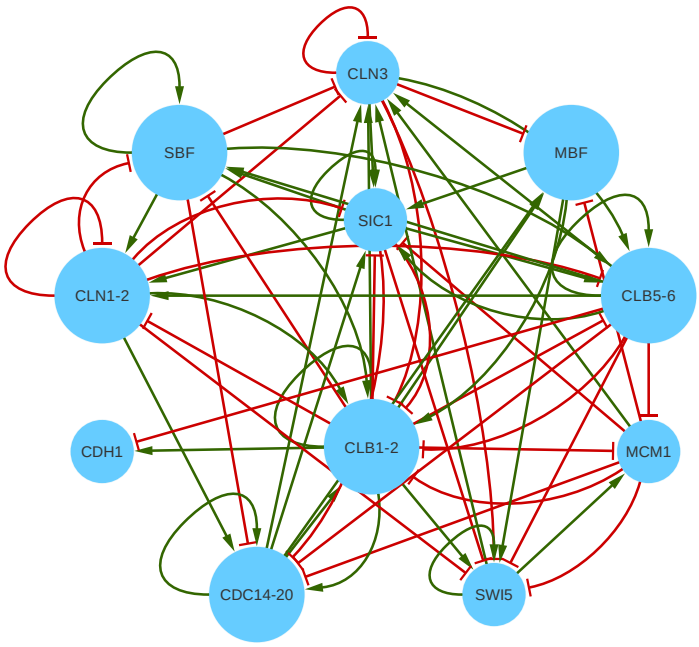


Figure 5.2: Yeast cell-cycle GRN estimated from microarray time series. Nodes are represented as genes and causal relationships as edges, form parents to children. Activation are represented with an arrowed tip while inhibition is depicted with a T-shaped tip.

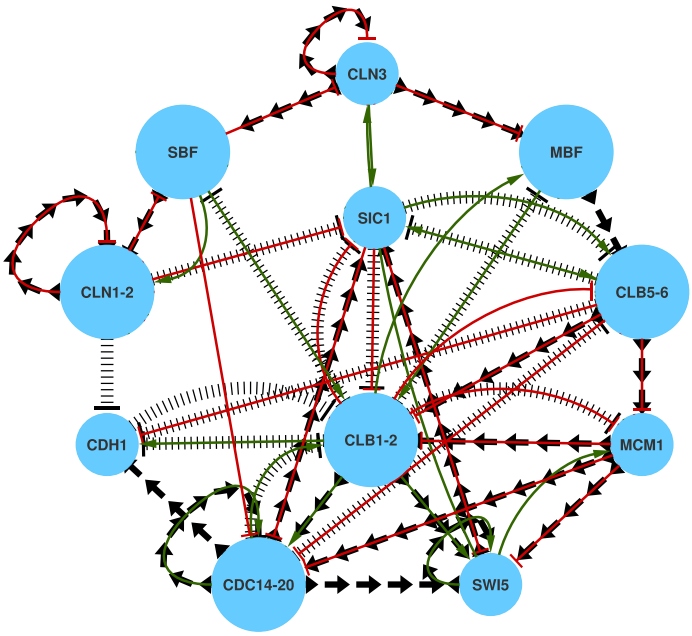


Figure 5.3: Yeast cell-cycle GRN estimated from microarray time series overlaped with the ground thrust one. Nodes are represented as genes and causal relationships as edges, form parents to children. The known activation relationships are represented by an arrow-shaped discontinuous stroke, whilst inhibition is depicted by a dashed stroke with a bold tip at target. On the other hand, estimated relationships are represented with a continuous stroke, with an arrowed tip for activation and a T-shaped tip for inhibition.

Chapter 6

Modelo Bayesiano de expansión de factores para el análisis de datos de microarrays y la estimación de perfiles de proteínas

El análisis de datos de microarrays continúa siendo un desafío de perspectivas prometedoras en la investigación Genómica, Médica y Farmacologíca [45] [94]. En concreto, el aprendizaje de modelos abstractos como las RRG, ofrecen una visión general del mecanismo de regulación genética desde un punto de vista fenomenológico. Sin embargo, el enorme desarrollo de esta, así como de otras tecnologías múltiplex, demandan un análisis más detallado que aborden el problema desde una perspectiva específica, en lugar de una simple descripción global. Este tipo de análisis, que opcionalmente puede complementar los datos de microarrays con otro tipo de información, exige un modelado más complejo a nivel organizativo y computacional.

Entre las innumerables aplicaciones del análisis de datos de microarrays, resultan de especial interés aquellas que tratan de clasificar enfermedades, de difícil diagnóstico clínico, con una base molecular [49]. El objetivo de este tipo de aproximaciones es encontrar un conjunto de genes cuyo estado de expresión constituya una firma biomolecular para una patología determinada y permita clasificarla en subtipos con diferente pronóstico y respuesta al tratamiento [86] [62]. Sin embargo, este tipo de métodos se centran en un análisis descriptivo de los perfiles de expresión, volviendo a enfocar el problema desde una perspectiva fenomenológica global. Por otro lado, otro tipo de aproximaciones proponen un análisis integra-

tivo, en el que los datos de microarrays se ven enriquecidos con otro tipo de información. Estas metodologías, buscan perfiles moleculares a mayor nivel de detalle que, además de diseccionar dicha enfermedad, puedan explicar sus orígenes [14] [55]. En este sentido, las redes reguladoras transcripcionales (RRT) son unas buenas candidatas para el análisis de patologías con base molecular por su capacidad de describir la regulación genética desde una de sus fases más tempranas.

Las RRT explican objetivamente el proceso de expresión genética desde una de sus etapas iniciales. En concreto, una RRT describe el mecanismo de regulación mediante un conjunto de proteínas funcionales, denominadas factores de transcripción (FT), encargadas de iniciar el proceso de transcripción. Los FT reconocen y se adhieren a las regiones promotoras de genes específicos, bloqueando y facilitando la adhesión del complejo transcripcional. El fenómeno de expresión incluye otros procesos de regulación a nivel postranscripcional, como la degradación del tránscrito por microARN y a niveles postranslacionales, como la transducción de señales extracelulares. Sin embargo, en una primera aproximación, se puede considerar que una RRT establece relaciones causales directas entre la abundancia de un FT y el nivel de expresión de un gen específico [36]. En este espacio transcripcional, un perfil de proteínas constituye un elemento importante para el estudio y análisis de los orígenes de la expresión genética [62]. Desgraciadamente, la obtención de perfiles proteicos es un procedimiento experimental complicado y de diseño específico, que no permite realizar estudios moleculares a grandes escalas.

En Teoría de la Información y Procesado de Señal existen numerosas técnicas para analizar un conjunto de datos en un espacio diferente al de observación, tales como PCA [92] [37], ICA [47], NCA [46] [26] [88]. El interés en estas metodologías reside en la capacidad de descomponer la información en diferentes componentes que muestren los datos en un espacio condensado, aunque menos natural en el contexto del fenómeno observado. Por otro lado, otras técnicas más recientes como *Compresing Sensing* (sondeo con compresión) tratan de reconstruir una señal a partir de un conjunto de datos reducidos, asumiendo un leve pérdida de información [13]. El muestreo con compresión se basa en la posibilidad de expresar un conjunto de datos en una base con una representación dispersa y prescindir de los términos que aportan menos información.

6.1 Formulación del problema

Los métodos Bayesianos, por su capacidad de modelar e incluir información de diferentes fuentes a través de distribuciones de probabilidad a prior, se adaptan de manera óptima al análisis de datos integrativo. En concreto, se propone un marco de trabajo que permita manejar los datos de microarray e incorpore una descripción del fenómeno de regulación a nivel transcripcional. Sea un conjunto de

datos de microarrays, concentración en escala logarítmica del mARN transcrito, con G genes y N muestras,

$$\mathbf{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1N} \\ \vdots & \ddots & \vdots \\ y_{G1} & \cdots & y_{GN} \end{pmatrix} \in \mathbb{R}^{G \times N} \tag{6.1}$$

donde $[\mathbf{Y}]_{gn} = y_{gn}$ es el nivel de expresión genético: abundancia en escala logarítmica del ARN transcrito para el gen i-ésimo en la n-ésima muestra experimental, relativa a una muestra de control. Por otro lado, considérese una RRT con F factores de transcripción que regulan la expresión de los genes de interés. Una RRT queda determinada por dos características principales: (i) su topología, es decir, el esquema de conexiones entre sus nodos y (ii) el efecto regulador de activación o inhibición en la expresión genética [71] [56]. Sea la actividad proteica, magnitud que cuantifica la abundancia de un FT, de los F factores de transcripción en cada una de las N muestras,

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{F1} & \cdots & x_{FN} \end{pmatrix} \in \mathbb{R}^{F \times N} \tag{6.2}$$

donde $[\mathbf{X}]_{fn} = x_{fn}$ es la variación en la concentración de la f-ésima proteína en la n-ésima muestra. De manera similar al nivel de expresión genético, una actividad positiva $x_{fn} > 0$ indica una mayor concentración del FT en la muestra experimental y un valor negativo $x_{fn} < 0$ su disminución respecto a la muestra de control.

6.1.1 Modelos previos

En diversas aproximaciones [74] [14] se ha considerado un modelo de factores latentes que permite conectar los datos de microarray $\mathbf{Y}$ con los perfiles de actividad proteica $\mathbf{X}$. Según este modelo, la conexión entre ambas magnitudes se establece mediante la matriz de carga

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1F} \\ \vdots & \ddots & \vdots \\ a_{G1} & \cdots & a_{GF} \end{pmatrix} \in \mathbb{R}^{G \times F} \tag{6.3}$$

donde $[\mathbf{A}]_{gf} = a_{gf}$ indica la fuerza y el tipo de regulación que el f-ésimo FT ejerce sobre el g-ésimo gen. Al considerar un término de ruido $[\mathbf{E}]_{gn} = e_{gn}$ que describa cualquier efecto no lineal, así como otro tipo de errores [68], este modelo se puede expresar como el producto y suma de matrices,

$$\mathbf{Y} = \mathbf{AX} + \mathbf{E}. \tag{6.4}$$

En una primera aproximación, la matriz de carga se identifica de manera acertada con la estructura topológica de la RRT [56], de manera que cuando una proteína no regula la expresión de un gen el coeficiente correspondiente es nulo $a_{gf} = 0$. La especificidad constatada con la que un FT regula la expresión de un gen sugiere que la matriz de carga posee una gran cantidad de elementos nulos, es decir, $\mathbf{A}$ es una variable altamente dispersa. Inicialmente, las soluciones propuestas imponían ligaduras estrictas en la matriz de carga para favorecer una solución dispersa [46] [67]. Sin embargo, aunque válidas en casos algunos particulares, estas aproximaciones son poco realistas y hacen que las variables pierdan el significado biológico en el contexto de regulación a nivel transcripcional. En lugar de estimar la actividad de las proteínas, estos métodos proyectan los datos de expresión a un espacio abstracto de metagenes con dimensiones reducidas, donde el identificar la matriz de carga con el fenómeno de regulación puede llevar a conclusiones erróneas [12].

El conocimiento y fácil acceso a las secuencias que componen el genoma de numerosas especies, incluyendo el humano, y el desarrollo de base de datos con información sobre interacciones a nivel gen-proteína, permiten estimar actualmente la estructura topológica de una RRT. En concreto, basándose en algoritmos para predecir alineamientos entre secuencias genéticas, se puede obtener una probabilidad $[\mathbf{\Pi}]_{gf} = \pi_{gf}, \forall g, f$ de que el f-ésimo factor de transcripción posea un sitio de enlace en la región promotora de g-ésimo gen [44] [53] y, por tanto, conocer a priori qué elementos de la matriz de carga son nulos, es decir,

$$p\left(a_{gf} \neq 0\right) = \pi_{gf}. \tag{6.5}$$

En un análisis previo, en el que se ha inferido dicha estructura para el genoma humano al completo, se han estimado unos niveles de dispersión de hasta el 69% (véase Apéndice G). Un modelo más acertado, que permite incorporar esta información a priori sobre la esta estructura dispersa de la RRT, propone una distribución mixta con una componente discreta en zero y contínua para el resto de valores. Concretamente, en [74] se propone una distribución Gaussiana rectificada en cero que permite expresar la probabilidad a priori de los coeficientes de carga como,

$$p\left(a_{gf}\right) = \left(1 - \pi_{gf}\right)\delta\left(a_{gf}\right) + \pi_{gf}\mathcal{N}\left(a_{gf}|\, 0, \sigma_f^2\right). \tag{6.6}$$

con $1 - \pi_{gf}$ la masa de probabilidad en cero. Por otro lado, la probabilidad de que el elemento de la matriz de carga no sea nulo se distribuye como una Gaussiana con media cero y varianza desconocida σ_f^2, que se hiperparametriza mediante una distribución inversa Gamma como,

$$p\left(\sigma_f^2\right) = \mathcal{IG}\left(\sigma_f^2 |\, \alpha_f, \beta_f\right). \tag{6.7}$$

La distribución (6.6) puede expresarse como la probabilidad marginal de una distribución Bernoulli-Gaussiana, en términos de la variable binaria $\lambda_{gf} \in \{0,1\}$ según,

$$p\left(a_{gf},\lambda_{gf}\right) = \mathcal{BN}\left(a_{gf},\lambda_{gf}|\,0,\sigma_f^2\lambda_{gf},\pi_{gf}\right) = \tag{6.8}$$

$$= \pi_{gf}^{\lambda_{gf}}\left(1-\pi_{gf}\right)^{(1-\lambda_{gf})}\mathcal{N}\left(a_{gf}|\,0,\sigma_f^2\lambda_{gf}\right) \tag{6.9}$$

que al marginalizar se puede expresar como,

$$p\left(a_{gf}\right) = \int p\left(a_{gf},\lambda_{gf}\right)d\lambda_{gf} = \int p\left(a_{gf}|\,\lambda_{gf}\right)p\left(\lambda_{gf}\right)d\lambda_{gf}$$

$$= \sum_{\lambda_{gf}=0}^{1} p\left(a_{gf}|\,\lambda_{gf}\right)p\left(\lambda_{gf}\right) \tag{6.10}$$

con

$$p\left(a_{gf}|\,\lambda_{gf}\right) = \mathcal{N}\left(a_{gf}|\,0,\sigma_f^2\lambda_{gf}\right) \tag{6.11}$$

$$p\left(\lambda_{gf}\right) = \mathcal{B}\left(\lambda_{gf}|\,\pi_{gf}\right). \tag{6.12}$$

En las metodologías basadas en este tipo de distribuciones mixtas surgen problemas de indeterminación del modelo, relacionados con la ambigüedad en signo y escala de las magnitudes a estimar [39]. Las soluciones propuestas vuelven a imponer ligaduras rígidas en las actividades de las proteica [55], aproximaciones que, además de complicar la formulación analítica del problema, son poco realistas y proporcionan estimaciones carentes de significado biológico.

6.1.2 Modelo Bayesiano de factores expandidos

Una descripción a través de un modelo de factores latentes debería permitir a la matriz de carga y a la actividad de las proteínas tomar cualquier valor del espacio real. En concreto, los coeficientes de la matriz de carga representan con valores negativos $a_{gf} < 0$ efectos de regulación negativa, es decir, inhibición de la expresión genética; mientras que, con valores positivos $a_{gf} > 0$, se indica un efecto de regulación positiva, es decir, de activación de la expresión genética. Por otro lado, los elementos nulos $a_{gf} = 0$ indican la ausencia de regulación, es decir, independencia entre un gen y dicho FT. Motivados por la capacidad de determinar a priori la estructura dispersa de la matriz de carga y por el elevado coste computacional que implica estimar todos los coeficientes para un conjunto grande de datos de expresión, proponemos una transformación de esta matriz a un espacio de dimensiones reducidas con menor nivel de dispersión.

Considérese una base ortonormal $\mathbf{H} \in \mathbb{R}^{G\times G}$ tal que,

$$\tilde{\mathbf{A}} = \mathbf{HA} \tag{6.13}$$

donde $\tilde{\mathbf{A}} \in \mathbb{R}^{G \times G}$ contiene la misma información que la matriz de carga en una representación expandida y con menor número de elementos nulos, es decir, posee menor nivel de dispersión que $\mathbf{A}$. Naturalmente, por el procedimiento y el contexto biológico en el que se extrae, la matriz de carga posee una representación altamente dispersa, donde la información sobre qué FT regula un gen determinado se localiza en un número reducido de coeficientes. Por el contrario, en su representación expandida $\tilde{\mathbf{A}}$, la misma información se codifica en una variable con mayor número de coeficientes no nulos, carentes de significado biológico. Dada la base $\mathbf{H}$, resulta inmediato recuperar la matriz de carga original a partir de su versión expandida como,

$$\mathbf{A} = \mathbf{H}^{\top} \tilde{\mathbf{A}} \tag{6.14}$$

con $\mathbf{H}^{-1} = \mathbf{H}^{\top}$. Sea $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_K] \in \mathbb{R}^{G \times K}$ un conjunto de K vectores ortonormales de manera que la matriz de carga pueda expresarse aproximadamente, en términos de pérdidas de la información, como,

$$\mathbf{A} \approx \mathbf{BC} \tag{6.15}$$

donde $\mathbf{C} \in \mathbb{R}^{K \times F}$ son los coeficientes en el subespacio de menor dimensión definido por $\mathbf{B}$, con $K < G$. De este modo, proponemos un modelo de factores latentes en una base expandida como,

$$\mathbf{Y} = \mathbf{BCX} + \mathbf{E} \tag{6.16}$$

donde se reduce el número de coeficientes a estimar $[\mathbf{C}]_{kf} = c_{kf}$ y se conserva la información estructural de la RRT a través de la relación (6.15) entre la matriz de los coeficientes expandidos y la matriz de carga original.

A pesar de conocer su distribución de elementos nulos, resulta difícil diseñar una base $\mathbf{B}$ óptima que permita aproximar la matriz de carga con el menor número de coeficientes posibles. Sin embargo, el marco de trabajo de las transformadas Wavelets se adapta adecuadamente al problema de reducción de dimensionalidad de la matriz de carga por su capacidad de segmentar la información. Las transformaciones Wavelets se caracterizan por un conjunto de vectores ortonormales que forman una base, donde los coeficientes de la transformación asociados describen propiedades estructurales de los datos con un nivel de detalle jerarquico [51]. Los vectores de la base son dilataciones y traslaciones de una función Wavelet generatriz que cumplen las condiciones de ortonormalidad. En este marco de trabajo, las bases $\mathbf{H}^{-1}$ y $\mathbf{H}$ podrían identificarse las transformadas Wavelets directa e inversa, respectivamente, mientras que $\mathbf{B}$ es la base en la que los coeficientes $\mathbf{C}$ describen el conjunto de datos al nivel de aproximación deseado. Para señales discretas, las transformadas Wavelets pueden interpretarse como un banco de filtros que iterativamente submuestrean la información hasta un nivel de detalle deseado. Por

tanto, otra ventaja de formalismo Wavelet es que permite excluir las componentes de menor información que están dominadas principalmente por ruido experimental [66].

6.2 Método Bayesiano de factores latentes expandidos

Sea el conjunto de variables que describen el ruido que afecta a los datos de microarray,

$$\mathbf{E} = \begin{pmatrix} e_{11} & \cdots & e_{1N} \\ \vdots & \ddots & \vdots \\ e_{G1} & \cdots & e_{GN} \end{pmatrix} \tag{6.17}$$

donde e_{gn} se modela como un ruido blanco, a través de una variable independiente e idénticamente distribuida (IID) por una Gaussiana de media nula y varianza desconocida σ_n^2 como,

$$p\left(e_{gn}\right) = \mathcal{N}\left(e_{gn}|\, 0, \sigma_n^2\right), \forall n. \tag{6.18}$$

Al representar las columnas de las matrices $\mathbf{A}$ y $\mathbf{C}$ por $\mathbf{a}_f = \left[a_{f1}, \ldots, a_{fG}\right]^\top$ y $\mathbf{c}_f = \left[c_{f1}, \ldots, c_{fK}\right]^\top$ respectivamente, la relación anterior puede expresarse para cada FT como,

$$\mathbf{a}_f = \mathbf{B}\mathbf{c}_f, \forall f \tag{6.19}$$

y para cada componente,

$$a_{gf} = \sum_{k=1}^{K} b_{gk}^2 c_{kf}. \tag{6.20}$$

Por otro lado, dada la base $\mathbf{B}$, los coeficientes $\mathbf{C}$ pueden expresarse como,

$$\mathbf{C} = \mathbf{B}^\top \mathbf{A} \tag{6.21}$$

con $\mathbf{B}^\top\mathbf{B} = \mathbb{1}^K$ y pseudoinversa de Moore-Penrose $\mathbf{B}^+ = \left(\mathbf{B}^\top\mathbf{B}\right)^{-1}\mathbf{B}^\top = \mathbf{B}^\top$. Del mismo modo, los coeficientes se relacionan como,

$$\mathbf{c}_f = \mathbf{B}^\top \mathbf{a}_f, \forall f = 1, \ldots, F \tag{6.22}$$

y para cada componente,

$$c_{kf} = \mathbf{b}_k^\top \mathbf{a}_f = \sum_{g=1}^{G} b_{kg}^2 a_{gf}. \tag{6.23}$$

Por tanto, al hacer uso de la relación (6.23) y (6.5), se puede expresar la probabilidad a priori de los coeficientes c_{kf} dada la base $\mathbf{B}$

$$p\left(c_{kf} \neq 0\right) = p\left(\mathbf{b}_k^\top \mathbf{a}_f \neq 0\right) = \pi_{gf}. \tag{6.24}$$

6.2.1 Modelado estadístico: probabilidades a priori

Dados los datos de microarray $\mathbf{Y}$, la base Wavelet $\mathbf{B}$ y las probabilidades a priori $\mathbf{\Pi}$; las variables desconocidas del modelo propuesto en (6.16) que se desean estimar son: $\mathbf{C}$, $\mathbf{X}$, σ_f^2 y σ_n^2. A continuación se propone un modelado estadístico de las incógnitas del problema.

6.2.1.1 Probabilidad a priori de los coeficientes

El modelado de los coeficientes c_{kf} puede ser complicado debido a la ausencia del significado biológico de esta magnitud. En lugar de modelarla directamente, proponemos partir de una una distribución de probabilidad $p\left(a_{gf}\right)$ que nos permita obtener $p\left(c_{kf}\right)$ a través de la relación (6.23). En concreto, partimos de la distribución Gaussiana rectificada en cero (6.6) usada en trabajos previos [74] [14] [55]. Al tener en cuenta la probabilidad (6.11) y la relación (6.23), la probabilidad de cada coeficiente c_{kf} dado $\boldsymbol{\lambda}_f = \left[\lambda_{1f}, \ldots, \lambda_{Gf}\right]^\top$ es otra distribución Gaussiana con media cero y varianza desconocida,

$$p\left(c_{kf} | \, \boldsymbol{\lambda}_f\right) = \mathcal{N}\left(c_{kf} | \, 0, \sigma_f^2 \sum_{g=1}^{G} b_{kg}^2 \lambda_{gf}\right) \tag{6.25}$$

donde la distribución marginal puede calcularse integrando,

$$p\left(c_{kf}\right) = \int p\left(c_{kf} | \, \boldsymbol{\lambda}_f\right) p\left(\boldsymbol{\lambda}_f\right) d\boldsymbol{\lambda}_f. \tag{6.26}$$

La marginalización en (6.26) exige la integración de las variables $\lambda_{gf}, \forall g$, cálculo complicado debido al gran número de genes que se suele considerar en este tipo de problemas. Además, una solución basada en el cálculo su probabilidad a posteriori, hace que el problema sea inviable analíticamente. Por tanto se propone como solución subóptima, de manera similar que en otro tipo de técnicas [41], una aproximación mediante una distribución Gaussiana de media y varianzas desconocidas,

$$p\left(c_{kf}\right) \approx \mathcal{N}\left(c_{kf} | \, \mu_{c_{kf}}, \tau_{kf}\right). \tag{6.27}$$

Dada la relación (6.19) y la aproximación considerada en (6.23) , la probabilidad a priori $p\left(\mathbf{a}_f\right)$ también puede expresarse como una Gaussiana,

$$p\left(\mathbf{a}_f\right) \approx \mathcal{N}\left(\mathbf{a}_f | \, \boldsymbol{\mu}_{\mathbf{a}_f}, \mathbf{\Sigma}_{\mathbf{a}_f}\right) \tag{6.28}$$

donde los hiperparámetros se relacionan como,

$$\mu_{c_{kf}} = \mathbf{b}_k^\top \boldsymbol{\mu}_{\mathbf{a}_f} = \mathbf{b}_k^\top \langle \mathbf{a}_f \rangle_{p(\mathbf{a}_f)} \tag{6.29}$$

$$\begin{aligned} \tau_{kf} &= \mathbf{b}_k^\top \boldsymbol{\Sigma}_{\mathbf{a}_f} \mathbf{b}_k \\ &= \mathbf{b}_k^\top \left(\left\langle \mathbf{a}_f \mathbf{a}_f^\top \right\rangle_{p(\mathbf{a}_f)} - \boldsymbol{\mu}_{\mathbf{a}_f} \boldsymbol{\mu}_{\mathbf{a}_f}^\top \right) \mathbf{b}_k. \end{aligned} \tag{6.30}$$

Los valores esperados en (6.29) y (6.30) poseen una solución trivial debido a la baja complejidad de (6.10) con los parámetros λ_{gf}. Considerando independientes los factores, estas esperanzas pueden expresarse como,

$$\left\langle \langle a_{gf} \rangle_{p(a_{gf}|\lambda_{gf})} \right\rangle_{p(\lambda_{gf})} = 0 \tag{6.31}$$

$$\left\langle \left\langle a_{gf}^2 \right\rangle_{p(a_{gf}|\lambda_{gf})} \right\rangle_{p(\lambda_{gf})} = \pi_{gf} \sigma_f^2. \tag{6.32}$$

de manera que la probabilidad a priori $p(a_{gf})$, dada la masa π_{gf}, se aproxima como una Gaussiana funcional (FIG, acrónimo del ingés *functional induced Gaussian*),

$$p(a_{gf}) = \mathcal{FIG}\left(a_{gf}|\, 0, \sigma_f^2, \pi_{gf}\right) \tag{6.33}$$

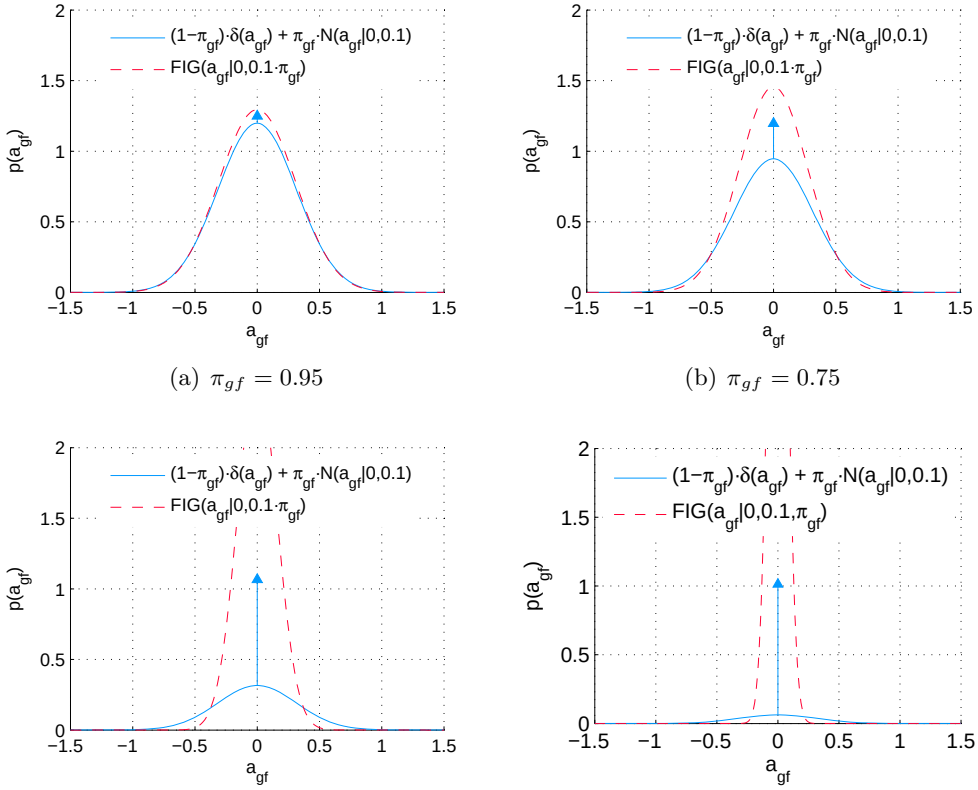
$$= \begin{cases} 0 & , \quad \pi_{gf} = 0 \\ \mathcal{N}\left(a_{bf}|\, 0, \sigma_f^2 \pi_{gf}\right) & , \quad \pi_{gf} \neq 0 \end{cases} \tag{6.34}$$

La Figura 8.6 muestra diferentes distribuciones Gaussianas rectificadas en cero, junto con su aproximación FIG correspondiente, para diferentes valores de la probabilidad a priori π_{gf}. En la Gaussiana rectificada, la masa $1 - \pi_{gf}$ proporciona la probabilidad a priori en cero, que se estima a través de bases de datos, mientras que el resto de valores se distribuyen como una Normal en torno a cero y con varianza σ_i^2. Cuando la probabilidad a priori toma un valor cercano a la unidad $\pi_{gf} \approx 1$, la aproximación FIG es una distrubición Gaussaina de varianza similar a la distribución Normal de la mezcla. Por el contrario, para valores próximos a cero $\pi_{gf} \approx 0$, domina la componente discreta de la Gaussiana rectificada y la varianza de la FIG correspondiente es muy pequeña. Para valores de probabilidad intermedios, podría plantearse la validez de la aproximación FIG. Sin embargo, las probabilidades a priori toman valores principalmente, a parte de cero, en la región $\pi_{gf} \geq 0.8$ (ver Apéndice G).

Al tener en cuenta la media (6.31) y la varianza (6.32), los hiperparámetros de la probabilidad a priori (6.27) que se calculan a través de (6.29) y (6.30) se pueden expresar como,

$$\mu_{c_{kf}} = 0 \tag{6.35}$$

$$\tau_{kf} = \sigma_f^2 \mathbf{b}_k^\top \mathbf{D}_{\boldsymbol{\pi}_f} \mathbf{b}_k \tag{6.36}$$



(a) $\pi_{gf} = 0.95$ (b) $\pi_{gf} = 0.75$

(c) $\pi_{gf} = 0.25$ (d) $\pi_{gf} = 0.05$

Figure 6.1: Distribución Gaussianas rectificada en cero junto a su aproximación FIG para diferentes valores de la probabilidad a priori y varianza $\sigma_f^2 = 0.1$.

donde $\mathbf{D}_{\boldsymbol{\pi}_f}$ es una matriz diagonal con el vector $\boldsymbol{\pi}_f = [\pi_{1f}, \ldots, \pi_{Gf}]^\top$. Por tanto, la probabilidad a priori $p(c_{kf})$ se modela mediante una distribución FIG de de media nula, varianza σ_f^2 y probabilidad a priori $\mathbf{b}_k^\top \mathbf{D}_{\boldsymbol{\pi}_f} \mathbf{b}_k$ como una distribución Gaussiana ,

$$p(c_{kf}) = \mathcal{FIG}\left(a_{gf} | \, 0, \sigma_f^2, \mathbf{b}_k^\top \mathbf{D}_{\boldsymbol{\pi}_f} \mathbf{b}_k\right) \tag{6.37}$$

Por otro lado, este resultado se puede expresar como una probabilidad a priori $p\left(\mathbf{c}_f^*\right)$, donde $\mathbf{c}_f^* \in \mathbb{R}^{K_f^* \times 1}$ es una variable que contiene los coeficientes no nulos $c_{kf}, \forall k$ tales que $\tau_{kf} \neq 0$,

$$p\left(\mathbf{c}_f^*\right) = \mathcal{N}\left(\mathbf{c}_f^* | \, \mathbf{0}^{K_f^* \times 1}, \boldsymbol{\Sigma}_{\mathbf{c}_f}\right) \tag{6.38}$$

donde la matriz de covarianzas $\boldsymbol{\Sigma}_{\mathbf{c}_f^*} \in \mathbb{R}^{K_f^* \times 1}$ es una matriz diagonal con $\tau_{kf} \neq 0, \forall k$ y K_f^* es la dimensión de la variable $\mathbf{c}_f^*$.

6.2.1.2 Probabilidad a priori de la actividad

El modelo de factores ocultos asume independientes los perfiles de actividad proteica $\mathbf{x}_n = [x_{1n}, \ldots, x_{Fn}]^\top, \forall n$. Para favorecer su cálculo analítico, se propone una distribución de la familia conjugada de su versosimilitud, que simplifique el cálculo su probabilidad a posteriori. En concreto, se propone una distribución Normal inversa Gamma que modele los perfiles de las proteínas y la varianza del ruido como,

$$\begin{aligned} p\left(\mathbf{x}_n, \sigma_n^2\right) &= \mathcal{NIG}\left(\mathbf{x}_n, \sigma_n^2 | \, \boldsymbol{\mu}_n, \kappa_n, \alpha_n, \beta_n\right) \\ &= \mathcal{N}\left(\mathbf{x}_n | \, \boldsymbol{\mu}_n, \frac{\sigma_n^2}{\kappa_n} \mathbb{1}^F\right) \mathcal{IG}\left(\sigma_n^2 | \, \alpha_n, \beta_n\right). \end{aligned} \tag{6.39}$$

donde

$$p\left(\mathbf{x}_n | \, \sigma_n^2\right) = \mathcal{N}\left(\mathbf{x}_n | \, \boldsymbol{\mu}_n, \frac{\sigma_n^2}{\kappa_n} \mathbb{1}^F\right) \tag{6.40}$$

con κ_n una escala que ajusta adecuadamente la varianza del ruido, que se hiperparametriza como una inversa-Gamma según,

$$p\left(\sigma_n^2\right) = \mathcal{IG}\left(\sigma_n^2 | \, \alpha_n, \beta_n\right). \tag{6.41}$$

La distribución a priori de la actividad proteica puede obtenerse marginalizando (6.40), obteniendose una distribución de Student como

$$\begin{aligned} p\left(\mathbf{x}_n\right) &= \int p\left(\mathbf{x}_n, \sigma_n^2\right) d\sigma_n^2 \\ &= \mathcal{St}\left(\mathbf{x}_n | \, \boldsymbol{\mu}_n, \frac{\alpha_n \kappa_n}{\beta_n}, 2\alpha_n\right) \end{aligned} \tag{6.42}$$

que, convenientemente, puede aproximarse por una Gaussiana (véase Apéndice F) según

$$p\left(\mathbf{x}_n\right) \approx \mathcal{N}\left(\mathbf{x}_n|\,\boldsymbol{\mu}_n, \frac{\beta_n}{\kappa_n\left(\alpha_n-1\right)}\mathbb{1}^F\right). \tag{6.43}$$

6.2.2 Aproximación numérica de la probabilidad a posteriori: muestreo de Gibbs

Para la estimación de las incógnitas del modelo BEFM, se propone una aproximación basada en el muestreo de Gibbs. Este método estima empíricamente las distribuciones a posterirori a partir de una cadena de Markov, obtenidas mediante simulación de distribuciones independientes de cada una de las cantidades desconocidas $\boldsymbol{\theta} = \left\{\mathbf{C}, \left\{\sigma_f^2\right\}_{f=1}^{F}, \mathbf{X}, \left\{\sigma_n^2\right\}_{n=1}^{N}\right\}$ dadas el resto. Para obtener dichas distribuciones, es necesario expresar la función verosimilitud como función de cada una de las variables a estimar. En concreto partimos del modelo presentado en (6.16) que puede expresarse como,

$$\tilde{\mathbf{Y}} = \mathbf{C}\mathbf{X} + \tilde{\mathbf{E}} \tag{6.44}$$

donde $\tilde{\mathbf{Y}} = \mathbf{B}^\top\mathbf{Y}$ y $\tilde{\mathbf{E}} = \mathbf{B}^\top\mathbf{E}$. El modelado estadístico de las distribuciones a priori en (6.7),(6.18), (6.38), (6.41) y (6.43) define un modelo exponencial conjugado con función de verosimilitud Gaussiana. Por tanto, las distribuciones a posteriori quedan definidas en la misma familia exponencial que su probabilidad a priori y el cálculo analítico se limita a unas reglas de actualización de sus hiperparámetros [50] [34].

6.2.2.1 Probabilidad a posteriori de los coeficientes

A partir del modelo BEFM en (6.44), podemos representar cada columna de la matriz $\tilde{\mathbf{Y}}$ como,

$$\tilde{\mathbf{y}}_k = \mathbf{c}_k\mathbf{X} + \tilde{\mathbf{e}}_k = \sum_{f=1}^{F} c_{kf}\mathbf{x}_f + \tilde{\mathbf{e}}_k \tag{6.45}$$

y reescribir el modelo generativo en función de cada coeficiente según,

$$\mathbf{d}_k^{(f)} := \tilde{\mathbf{y}}_k^\top - \sum_{\ell\neq f}^{F} \mathbf{x}_\ell^\top c_{k\ell} = \mathbf{x}_f^\top c_{kf} + \tilde{\mathbf{e}}_k^\top. \tag{6.46}$$

donde la distribución a priori $p\left(\tilde{\mathbf{e}}_k^\top\right)$ es una Gaussiana de media nula y varianza $\boldsymbol{\Lambda} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_N^2\right)$ como,

$$p\left(\tilde{\mathbf{e}}_k^\top\right) = \mathcal{N}\left(\tilde{\mathbf{e}}_k^\top \middle|\, \mathbf{0}^{N\times 1}, \boldsymbol{\Lambda}\right). \tag{6.47}$$

Por tanto, la función verosimilitud para cada coeficente se puede expresar como una Gaussiana,

$$p\left(\mathbf{Y}, \boldsymbol{\theta} - \left\{\mathbf{x}_f, c_{kf}, \left\{\sigma_n^2\right\}_{n=1}^N\right\} \middle| \mathbf{x}_f, c_{kf}, \left\{\sigma_n^2\right\}_{n=1}^N\right) = \mathcal{N}\left(\mathbf{d}_k^{(f)} \middle| \mathbf{x}_f^\top c_{kf}, \boldsymbol{\Lambda}\right) \quad (6.48)$$

Teniendo en cuenta la probabilidad a priori (6.37) y la función verosimilitud (6.48), la distribución a posteriori es otra FIG con hiperparámetros,

$$\begin{aligned} p\left(c_{kf} \middle| \mathbf{Y}, \boldsymbol{\theta} - \{c_{kf}\}\right) &= \mathcal{FIG}\left(c_{kf} \middle| \hat{\mu}_{kf}, \hat{\tau}_{kf}\right) && (6.49) \\ &= \begin{cases} 0 & , \quad \tau_{kf} = 0 \\ \mathcal{N}\left(c_{kf} \middle| \hat{\mu}_{kf}, \hat{\tau}_{kf}\right) & , \quad \tau_{kf} \neq 0 \end{cases} && (6.50) \end{aligned}$$

con

$$\begin{aligned} \hat{\mu}_{kf} &= \tau_{kf}\mathbf{x}_f \left(\tau_{kf}\mathbf{x}_f^\top \mathbf{x}_f + \boldsymbol{\Lambda}\right)^{-1} \mathbf{d}_k^{(f)} && (6.51) \\ \hat{\tau}_{kf} &= \tau_{kf} - \tau_{kf}\mathbf{x}_f \left(\tau_{kf}\mathbf{x}_f^\top \mathbf{x}_f + \boldsymbol{\Lambda}\right)^{-1} \mathbf{x}_f^\top \tau_{kf} && (6.52) \end{aligned}$$

donde $\tau_{kf} = \sigma_f^2 \mathbf{b}_k^\top \mathbf{D}_{\boldsymbol{\pi}_f} \mathbf{b}_k$.

6.2.2.2 Probabilidad a posteriori de la varianza de los factores

Dada la verosimilitud (6.38) y la probabilidad a priori (6.7) de la varianza de los factores, la distribución a posteriori es otra inversa Gamma de hyperparametros desconocidos,

$$p\left(\sigma_f^2 \middle| \mathbf{c}_f^*\right) = \mathcal{IG}\left(\sigma_f^2 \middle| \hat{\alpha}_f, \hat{\beta}_f\right) \quad (6.53)$$

En virtud del teorema de Bayes y excluendo los términos con $\tau_{kf} = 0$ en (6.38), la distribución a posteriori se puede expresar según,

$$\begin{aligned} p\left(\sigma_f^2 \middle| \mathbf{c}_f^*\right) &\propto p\left(\mathbf{c}_f^* \middle| \sigma_f^2\right) p\left(\sigma_f^2\right) \\ &= \mathcal{N}\left(\mathbf{c}_f^* \middle| \mathbf{0}^{K_f^*}, \boldsymbol{\Sigma}_{\mathbf{c}_f^*}\right) \mathcal{IG}\left(\sigma_f^2 \middle| \alpha_f, \beta_f\right) \\ &\propto \left(\sigma_f^2\right)^{-\alpha_f - \frac{K_f^*}{2} - 1} \mathrm{e}^{-\frac{1}{\sigma_f^2}\left(\frac{\mathbf{c}_f^{*\top} \boldsymbol{\Sigma}_{\mathbf{c}_f^*}^{-1} \mathbf{c}_f^*}{2} + \beta_n\right)} \end{aligned} \quad (6.54)$$

que permite identificar los hiperparámetros de la probabilidad a posteriori como,

$$\begin{aligned} \hat{\alpha}_f &= \alpha_f + \frac{K_f^*}{2} && (6.55) \\ \hat{\beta}_f &= \beta_f + \frac{\mathbf{c}_f^\top \left(\mathbf{B}^\top \mathbf{D}_{\boldsymbol{\pi}_f} \mathbf{B}\right)^+ \mathbf{c}_f}{2} && (6.56) \end{aligned}$$

donde $\mathbf{c}_f^{*\top}\boldsymbol{\Sigma}_{\mathbf{c}_f^*}^{-1}\mathbf{c}_f^* = \mathbf{c}_f^\top\left(\mathbf{B}^\top\mathbf{D}_{\boldsymbol{\pi}_f}\mathbf{B}\right)^+\mathbf{c}_f$ y K_f^* es el número de coeficientes no nulos en $\mathbf{c}_f^*$, es decir, el número de elementos de la diagonal principal de la matriz $\mathbf{B}^\top\mathbf{D}_{\boldsymbol{\pi}_f}\mathbf{B}$ que son diferentes de cero.

6.2.2.3 Probabilidad a posteriori de la actividad de las proteínas

A partir del modelo propuesto en (6.44), podemos representar cada fila de la matriz $\tilde{\mathbf{Y}}$ como,

$$\tilde{\mathbf{y}}_n = \mathbf{C}\mathbf{x}_n + \tilde{\mathbf{e}}_n = \sum_{f=1}^{F}\mathbf{c}_f x_{fn} + \tilde{\mathbf{e}}_n \tag{6.57}$$

y reescribir el modelo generativo en función de la actividad de cada FT para cada muestra según,

$$\mathbf{d}_k^{(f)} := \tilde{\mathbf{y}}_n - \sum_{\ell\neq f}^{F}\mathbf{c}_\ell^\top x_{\ell n} = \mathbf{c}_f x_{fn} + \tilde{\mathbf{e}}_n. \tag{6.58}$$

donde la probabilidad a priori $p\left(\tilde{\mathbf{e}}_n\right)$ continúa siendo ruido blanco modelado como,

$$p\left(\tilde{\mathbf{e}}_n\right) = \mathcal{N}\left(\tilde{\mathbf{e}}_n|\,\mathbf{0}^K, \sigma_n^2\mathbb{1}^K\right). \tag{6.59}$$

Por tanto, la función de verosimilitud para cada actividad se puede expresar como una Gaussiana,

$$p\left(\mathbf{Y}, \boldsymbol{\theta} - \left\{\mathbf{c}_f, \sigma_n^2, x_{fn}\right\}\middle|\,\mathbf{c}_f, \sigma_n^2, x_{fn}\right) = \mathcal{N}\left(\mathbf{d}_n^{(f)}\middle|\,\mathbf{c}_f^\top x_{fn}, \sigma_n^2\mathbb{1}^K\right). \tag{6.60}$$

Teniendo en cuenta la probabilidad a priori (6.40), La distribución a posteriori es una Gaussina,

$$p\left(x_{fn}|\,\mathbf{Y}, \boldsymbol{\theta} - \{x_{fn}\}\right) = \mathcal{N}\left(x_{fn}|\,\hat{\mu}_{fn}, \hat{\sigma}_n^2\right) \tag{6.61}$$

con hiperparámetros

$$\hat{\boldsymbol{\mu}}_n = \mu_{fn} + \frac{\sigma_n^2}{\kappa_n}\mathbf{c}_f^\top\left(\frac{\sigma_n^2}{\kappa_n}\mathbf{c}_f\mathbf{c}_f^\top + \sigma_n^2\mathbb{1}^K\right)^{-1}\left(\mathbf{d}_n^{(f)} - \mathbf{c}_f\mu_{fn}\right) \tag{6.62}$$

$$\hat{\sigma}_n^2 = \frac{\sigma_n^2}{\kappa_n} - \frac{\sigma_n^2}{\kappa_n}\mathbf{c}_f^\top\left(\frac{\sigma_n^2}{\kappa_n}\mathbf{c}_f\mathbf{c}_f^\top + \sigma_n^2\mathbb{1}^K\right)^{-1}\mathbf{c}_f\frac{\sigma_n^2}{\kappa_n}. \tag{6.63}$$

6.2.2.4 Probabilidad a posteriori de la varianza del ruido

A partir del modelo generativo expresado en (6.57), la verosimilitud puede expresarse como,

$$p\left(\mathbf{Y}, \boldsymbol{\theta} - \left\{\mathbf{C}, \mathbf{x}_n, \sigma_n^2\right\} \middle| \, \sigma_n^2\right) = \mathcal{N}\left(\tilde{\mathbf{y}}_n \middle| \, \mathbf{C}\mathbf{x}_n, \sigma_n^2 \mathbb{1}^K\right) \tag{6.64}$$

y al tener en cuenta la probabilidad a priori (**??**), la distribución a posteriori de la varianza de ruido es una inversa Gamma,

$$p\left(\sigma_n^2 \middle| \, \mathbf{Y}, \boldsymbol{\theta} - \left\{\sigma_n^2\right\}\right) = \mathcal{IG}\left(\sigma_n^2 \middle| \, \hat{\alpha}_n, \hat{\beta}_n\right) \tag{6.65}$$

con hiperparámetros

$$\hat{\alpha}_n = \alpha_n + \frac{K}{2} \tag{6.66}$$

$$\hat{\beta}_n = \beta_n + \frac{\left(\tilde{\mathbf{y}}_n - \mathbf{C}\mathbf{x}_n\right)^\top \left(\tilde{\mathbf{y}}_n - \mathbf{C}\mathbf{x}_n\right)}{2}. \tag{6.67}$$

6.2.3 Método BEFM para el apredizaje de microarrays

El modelo presentado en (6.16) incorpora información sobre la estructura topológica de la RRT a través de una probabilidad a priori π_{gf} de que el coeficiente correspondiente de la matriz de carga a_{gf} no sea nulo. Este tipo de conocimiento biológico se puede estimar previamente a través de métodos para predecir sitios de enlaces de los FT en el genoma de interés, apoyándose en bases de datos de interacción gen-proteína (véase Apéndice G). Además, se propone una transformación Wavelet de estas variables que reduce el número de incógnitas a estimar de $\mathbf{a}_f \in \mathbb{R}^{G\times 1}$ a $\mathbf{c}_f^* \in \mathbb{R}^{K_f^*\times 1}$, $\forall f$. Existen diversas transformadas Wavelets, basadas en diferentes tipos de funciones generatrices. Por su simplicidad, se considera una transformada Haar que permite identificar fácilmente los términos que proporcionan coeficientes de aproximación y de detalles. En concreto, se elige la base $\mathbf{B}$ como los vectores que proporcionan la coeficientes de aproximación, logrando una reducción dimensional de $K = \frac{G}{2}$.

Establecida la base $\mathbf{B}$, dados los datos de microarray $\mathbf{Y}$ y las probabilidades a priori $\mathbf{\Pi}$, el muestreo de Gibbs obtiene una cadena de Markov a partir de las probabilidades a posteriori (6.49), (6.53), (6.61) y (6.65). Este procedimiento permite estimar empíricamente la probabilidad a posteriori conjunta a patir de simulaciones de las distribuciones independientes. Tras la iteración t-ésima, el

algoritmo de Gibbs basado en el modelo BEFM obtiene la siguiente muestra como,

$$c_{kf}^{(t+1)} \sim \mathcal{FIG}\left(c_{kf} | \, \hat{\mu}_{kf}^{(t)}, \hat{\tau}_{kf}^{(t)}\right), \forall k, f \tag{6.68}$$

$$x_{fn}^{(t+1)} \sim \mathcal{N}\left(x_{fn} | \, \hat{\mu}_{fn}^{(t)}, \hat{\sigma}_{n}^{2(t)}\right), \forall f, n \tag{6.69}$$

$$\hat{\sigma}_{f}^{2(t+1)} \sim \mathcal{IG}\left(\sigma_{f}^{2} | \, \hat{\alpha}_{f}^{(t)}, \hat{\beta}_{f}^{(t)}\right), \forall f \tag{6.70}$$

$$\hat{\sigma}_{n}^{2(t+1)} \sim \mathcal{IG}\left(\sigma_{n}^{2} | \, \hat{\alpha}_{n}^{(t)}, \hat{\beta}_{n}^{(t)}\right), \forall n \tag{6.71}$$

donde los hiperparámetros $\left\{\hat{\mu}_{kf}^{(t)}, \hat{\tau}_{kf}^{(t)}, \hat{\mu}_{fn}^{(t)}, \hat{\sigma}_{n}^{2(t)}, \hat{\alpha}_{f}^{(t)}, \hat{\beta}_{f}^{(t)}, \hat{\alpha}_{n}^{(t)}, \hat{\beta}_{n}^{(t)}\right\}$ se actualizan según (6.51) - (6.67) con las muestras obtenidas tras la última iteración.

Además, para evitar problemas de indeterminación de signo en las estimaciones, se considera una corrección del signo en las muestras simuladas. En concreto, se propone aplicar un salto del signo de las actividad proteica $x_{fn}, \forall f$ (columnas de la matrix $\mathbf{X}$ y filas correspondientes de $\mathbf{C}$) en función del estado de expresión de un gen de referencia. Esta aproximación se basa en que, en algunos casos, se dispone información sobre que gen codifica una proteína concreta, que actúa como factor de transcripción. Por tanto, cabe esperar que el estado de activación genética coincida con el de actividad proteica. Cuando esta información no esté disponible, se tomará como gen de referencia el que posea mayor probabilidad a priori π_{gf}. Tras la simulación de una cadena con T muestras, suficientes para obtener una significancia estadística, la distribución a posteriori de cada incógnita se aproxima empíricamente por la distribución muestral, salvo constantes superfluas, como (3.23). Dado que los coeficientes c_{kf}^{*} y las actividades x_{fn} se distribuyen como Gaussianas, sus estimadores MAP pueden aproximarse por las medias muestrales según,

$$\hat{c}_{kf,MAP} \approx \frac{1}{T-t'} \sum_{t>t'}^{T} \hat{c}_{kf}^{(t)} \tag{6.72}$$

$$\hat{x}_{fn,MAP} \approx \frac{1}{T-t'} \sum_{t>t'}^{T} \hat{x}_{fn}^{(t)} \tag{6.73}$$

con $t \leq t'$ un periodo de estabilización de la cadena, descartado en el cálculo de las estimaciones.

6.3 Validación mediante simulación

El método basado en el modelo BEFM ha sido validado mediante simulaciones con datos sintéticos. Para un número de genes G y factores de transcripción

F, se han generado diferentes RRT con un nivel de dispersión realista, entorno al 30%. En concreto, se han generado las probabilidades a priori π_{gf}, asignando aleatóriamente una probabilidad nula al 70% de los elementos y un valor uniformemente distribuido en $\pi_{gf} \in [0.8, 1]$ al resto. Los coeficientes $\mathbf{C}$ se han generado a partir de la distribución (6.37). Por otro lado, las actividades de las proteínas se han generado para un número N de muestras con la distribución a priori (6.39), para diferentes varianzas del ruido σ_n^2 y escala $\kappa_n = \frac{1}{\sigma_n^2}$. Finalmente, los datos se han generado haciendo uso del modelo BEFM (6.16).

El muestreo se ha realizado siguiendo las reglas (6.68)-(6.71). La cadena se inicializa con valores simulados a partir de las probabilidades a priori. Para evaluar la convergencia, se ha tenido en cuenta el procedimiento descrito en [34], generando 10 cadenas paralelas para cada incógnita con $T = 10000$ muestras y un periodo de estabilización $t' = \frac{T}{2}$. Para la corrección del signo de las actividades, se han escogido como genes de referencias los que, para cada factor, poseen mayor probabilidad a priori. En el caso de haber varios genes candidatos con igual probabilidad, se ha seleccionado el que posee mayor expresión diferencial.

6.4 Datos sintéticos con $G = 50$, $N = 50$ y $F = 8$

En primer lugar, se considera un conjunto pequeño con $G = 50$ genes, $N = 50$ muestras y $F = 8$ factores de transcripción. La RRT se ha generado con una dispersión del 25%, de manera que cada FT regule la expresión de al menos un gen. La Figura 6.2 muestra un mapa de calor con la matriz de carga, que representa le efecto regulador de la red transcripcional, junto con la matriz de coeficientes y las actividades de las proteínas con las que se han generado los datos sintéticos. Tras generar la RRT y las actividades de los FT, se simulan los datos de expresión y se les añade un ruido de varianza $\sigma_n^2 = \frac{1}{10}$. La Figura 6.3 muestra un mapa de calor con el conjunto de datos sintéticos. A partir de una muestra inicial, generada con las probabilidades a priori, se han simulado 10 cadenas con $T = 10000$ muestras para cada una de las incógnitas. En la Figura 6.4 se representa una de las cadenas para la actividad de una proteína y una muestra concreta. A pesar de que la serie se estabiliza tras simular un centenar, se descartan la primera mitad de muestras. Por otro lado, en la Figura 6.5 se representan varias de las cadenas simuladas en paralelo usadas para estimar la actividad correspondiente a la proteína anterior, una vez eliminadas las muestras del periodo de estabilización. Además, se representa el histograma construido con las 10 cadenas, con un total de $T' = 50000$ muestras. El análisis de los resultados se hará en términos de la raíz del error cuadrático medio (RMSE), raíz cuadrada del promedio de la diferencia cuadrática entre cada una de las muestras $\hat{x}_{fn}^{(t)}$ y su valor verdadero x_{fn}, usado

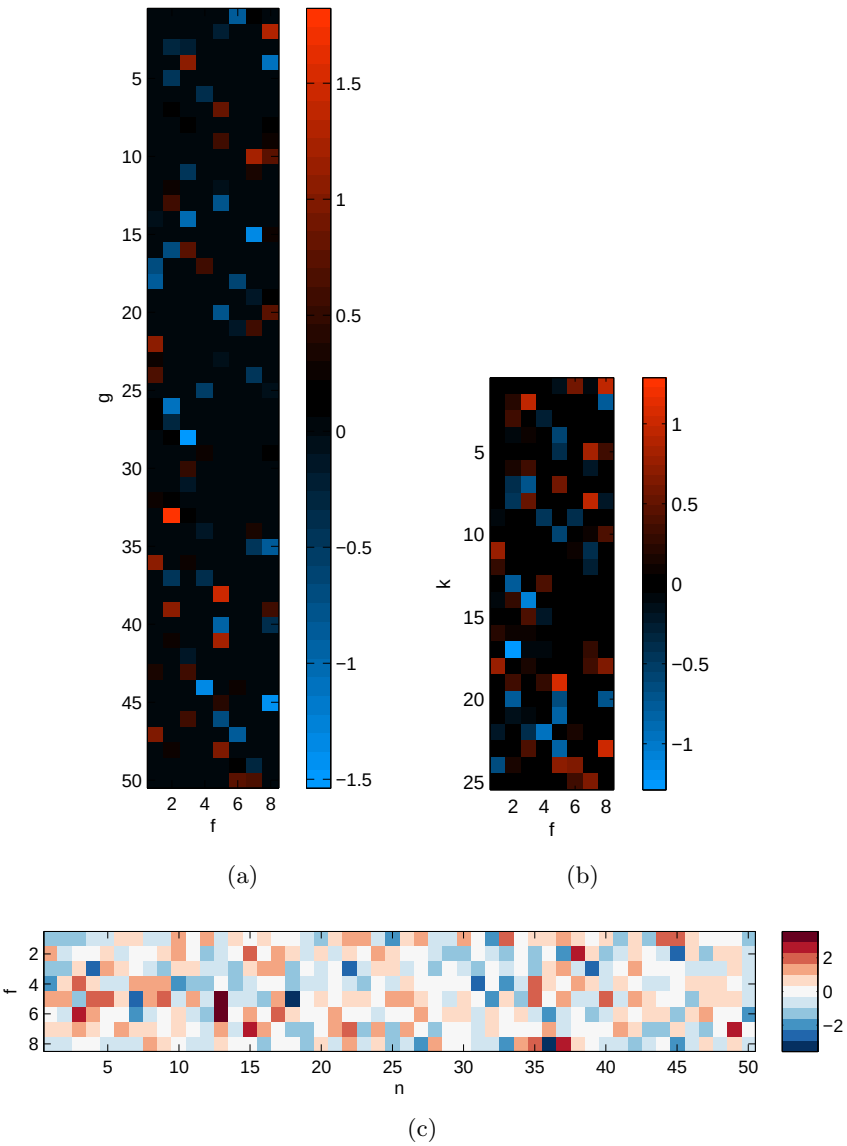


Figure 6.2: Mapas de calor con las variables descriptivas de la RRT y de los perfiles de actividad proteica simulados para generar un conjunto de datos sintético con $G = 50$ genes, $N = 50$ muestras y $F = 8$ FT. (a) Elementos de la matriz de carga a_{gf} con una dispersión del 25%. (b) Coeficientes de expansión c_{kf} con una dispersión del 45%. (c) Actividad de las proteínas x_{fn}.

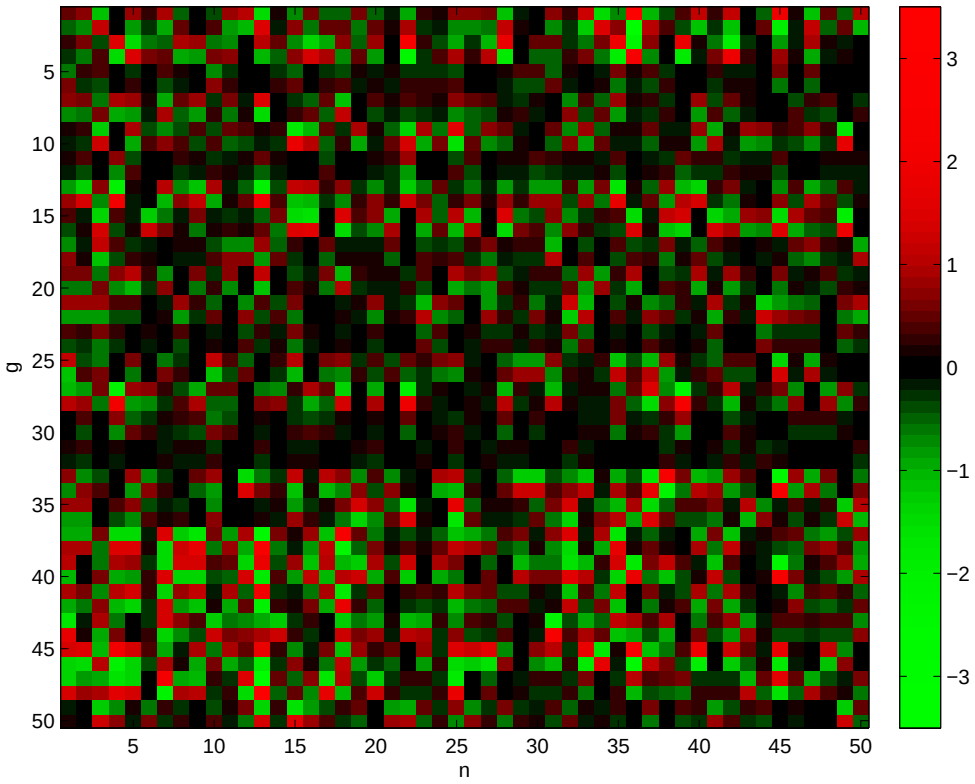


Figure 6.3: Mapa de calor con los datos sintéticos de expresión y_{gn} para $G = 50$ genes y $N = 50$ muestras con ruido de varianza $\sigma_n^2 = \frac{1}{10}$.

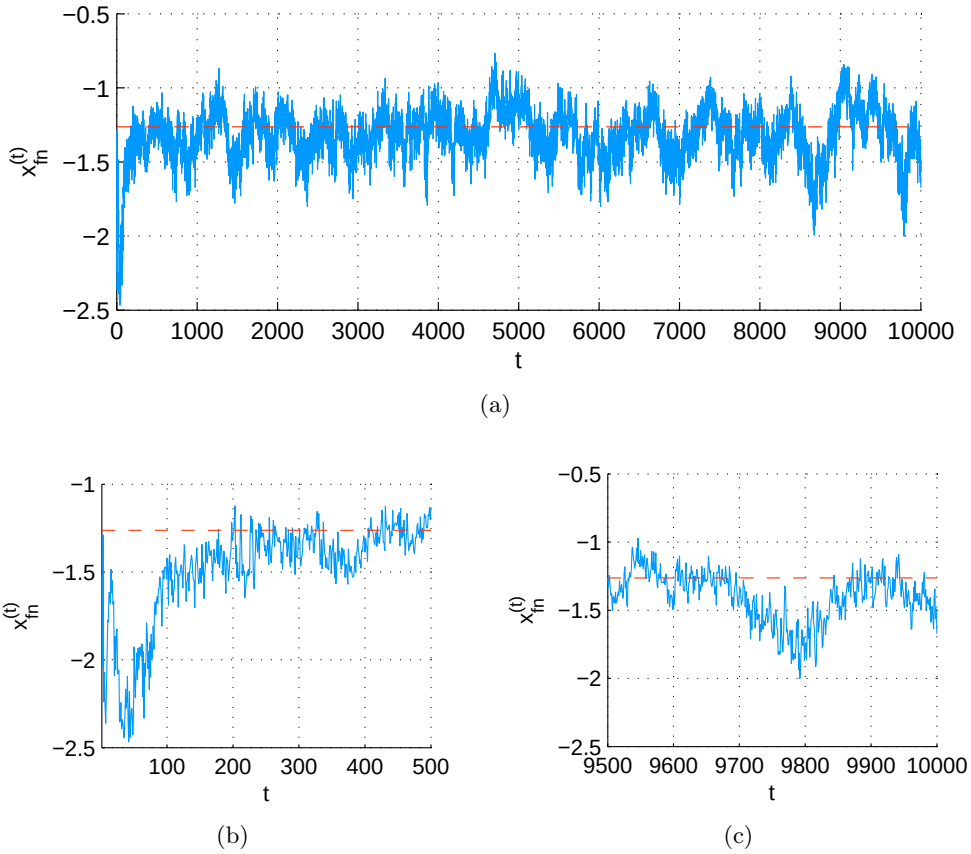


Figure 6.4: Cadena de Markov obtenida durante el muestreo de la actividad proteica x_{fn}, para el factor $f = 1$ en la muestra $n = 1$, junto con el valor verdadero usado para generar los datos (trazo discontinuo). (a) Cadena completa con $T = 10000$ muestras. (b) Detalle de las primeras 500 muestras. (c) Detalle de las últimas 500 muestras.

para generar los datos,

$$\text{RMSE} = \sqrt{\sum_{t=1}^{T'} \frac{\left(\hat{x}_{fn}^{(t)} - x_{fn}\right)^2}{T'}} \tag{6.74}$$

En la Figura 6.6 se representa en un diagrama de cajas y bigotes los resultados para las cinco mejores y las cinco peores estimaciones, con menor y mayor RMSE respectivamente. Para una visión completa de los resultados, en la Figura 6.7 se representan los valores reales de la actividad x_{fn} frente al RMSE de sus estimaciones. Del mismo modo, en la Figura 6.8 se representan los coeficientes c_{kf} usados para generar los datos frente al RMSE, excluyendo los términos nulos a priori.

Como cabía esperar, la relación entre las estimaciones y el RMSE muestran un efecto de escala multiplicativa entre los coeficientes y las actividades, problema inherente al modelo de análisis factorial. Sin embargo, esta ambigüedad resulta irrelevante para un análisis descriptivo destinado a realizar predicciones y clasificación en base a los perfiles proteicos. Por tanto, resulta igual de práctico trabajar con las estimaciones que con sus valores normalizados [74]. En concreto se propone una normalización en términos de regulón, es decir, del conjunto de genes regulados por una misma proteína. De este modo, los coeficientes de cada FT se normalizan acorde a sus valores extremos como,

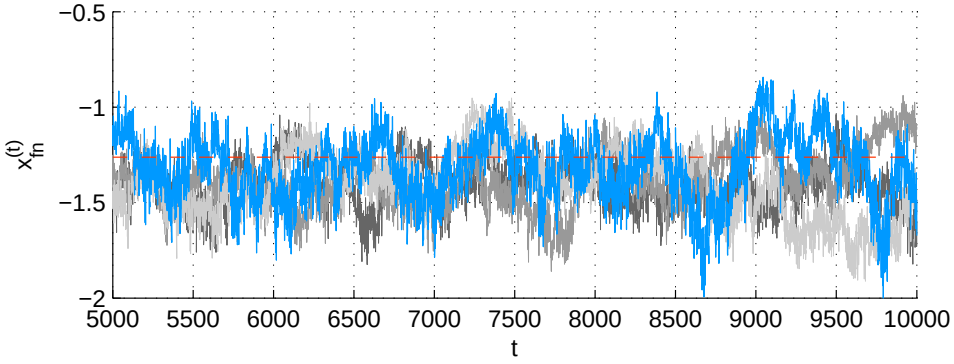
$$\tilde{\mathbf{c}}_f = \frac{\mathbf{c}_f}{\max\{\mathbf{c}_f\} - \min\{\mathbf{c}_f\}}, \forall f \tag{6.75}$$

y las actividades de las proteínas según el efecto promediado como,

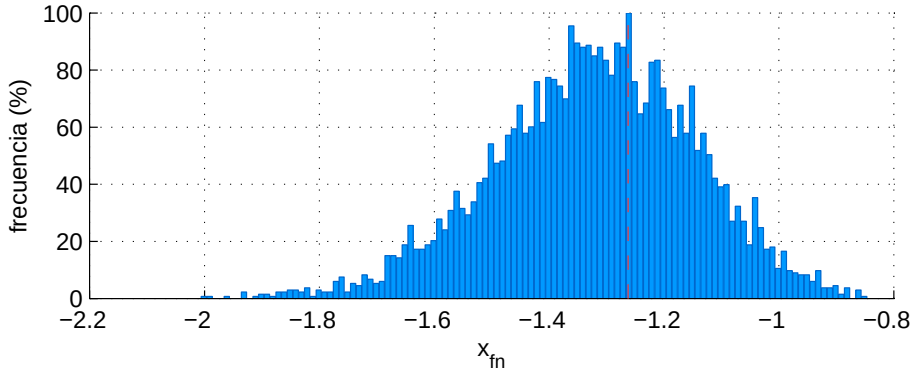
$$\tilde{x}_{fn} = \sum_{k}^{K} \frac{x_{fn}\tilde{c}_{kf}}{K_f^*}, \forall f, n. \tag{6.76}$$

La Figura 6.9 muestra las estimaciones normlizadas de la actividad proteica $\tilde{x}_{fn}$ frente a su error cuadrático (SE), diferencia respecto a su valor verdadero normalizado. Además, se representa el error cuadrático medio (MSE) promediado para el conjunto completo de perfiles. Del mismo modo, en la Figura 6.10 se representan los coeficientes estimados $\tilde{c}_{kf}$ frente a su SE, excluyendo los términos nulos a priori.

A modo descriptivo, en la Figura 6.11 se muestran los perfiles de proteínas usados para generar los datos y sus estimaciones, normalizados, a través de un mapa de calor. Por otro lado, en la Figura 6.12 se muestra el mapa de calor correspondiente a los coeficientes y su estimaciones.

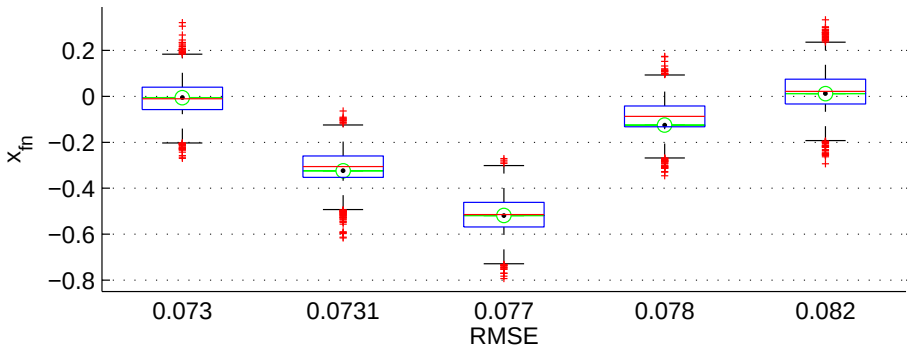


(a)

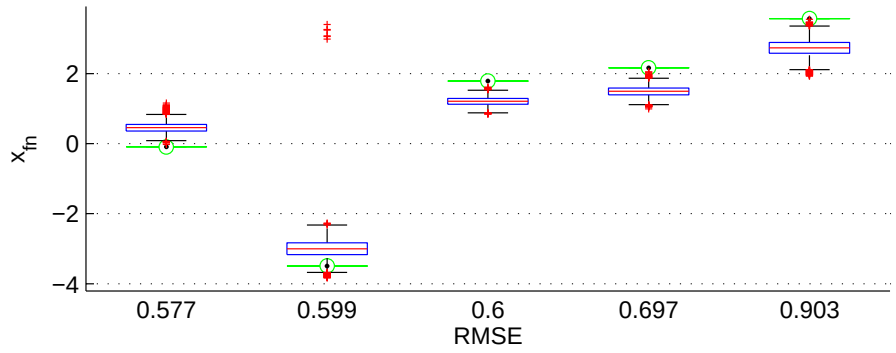


(b)

Figure 6.5: Estimación de la actividad para el factor $f = 1$ en la muestra $n = 1$ junto con su valor verdadero $x_{fn} = -1.2626$. El valor estimado es $\bar{x}_{fn} = -1.3231$ con un RMSE $= 0.1852$. (a) 5 de las 10 cadenas simuladas tras descartar las muestras del periodo de estabilización. (b) El valor estimado para este coeficiente es $\bar{x}_{fn} = -1.3231$ y su error cuadrático SE $= -0.0037$.



(a)



(b)

Figure 6.6: Diagrama de cajas y bigotes para las estimaciones de las actividades con mayor y menor RMSE. Los límites de las cajas indican los percentiles 25% y 75%, mientras que las barras de error se extienden hasta cubrir $\frac{3}{2}$ del rango intercuartil. La mediana divide la caja por la mitad, mientras que el valor real de la actividad proteica se representa mediante un segmento con una marca central. (a) Las 5 mejores estimaciones con menor RMSE. (b) Los 5 peores resultados con mayor RMSE.

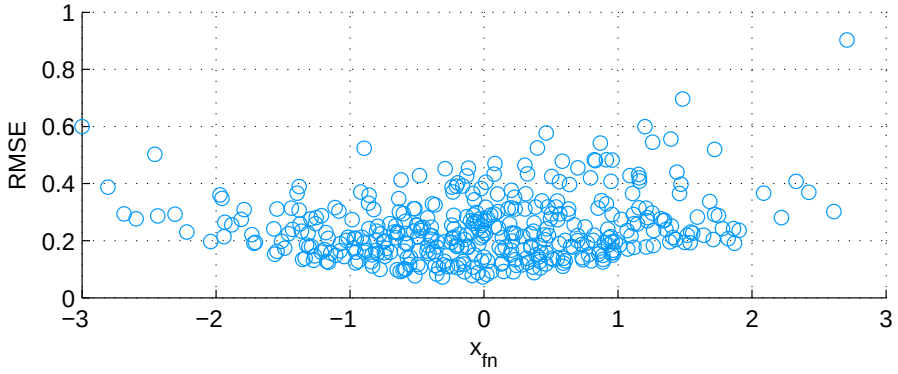


Figure 6.7: RMSE de las estimaciones frente al valor verdadero de las actividades de las proteínas x_{fn}

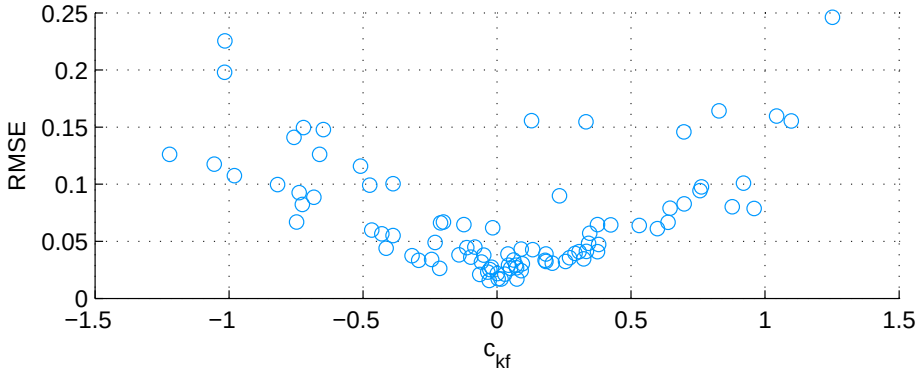


Figure 6.8: RMSE de las estimaciones frente al valor verdadero de los coeficientes c_{kf}.

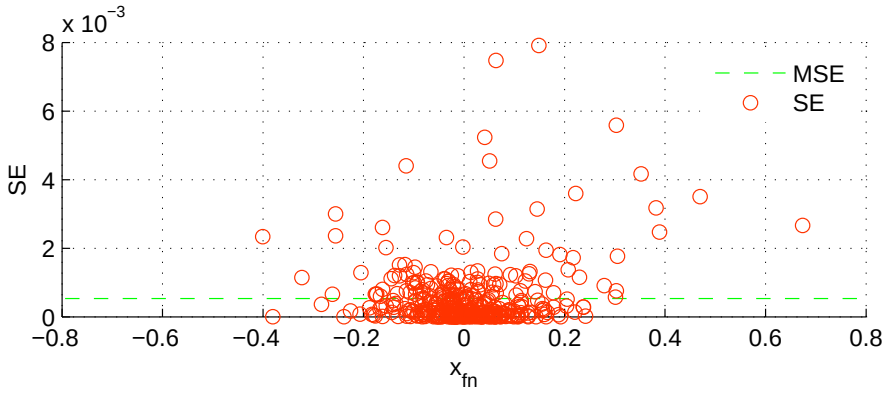


Figure 6.9: Error cuadrático (SE) frente las estimaciones normalizadas de las actividades $\tilde{x}_{fn}$, con MSE $= 5.06 \cdot 10^{-4}$.

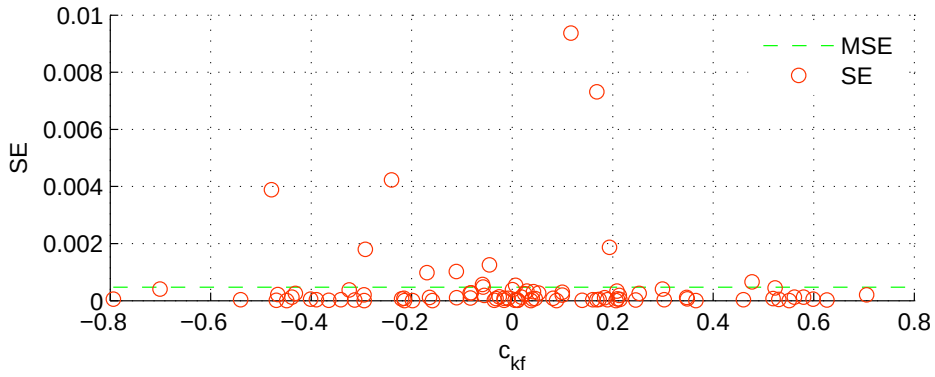


Figure 6.10: Error cuadrático (SE) frente las estimaciones normalizadas de los coeficientes $\tilde{c}_{kf}$, con MSE $= 3.12 \cdot 10^{-4}$.

6.5 Datos con $G = 100$, $N = 100$ y $F = 20$

Se ha repetido el análisis anterior para un conjunto mayor con $G = 100$ genes, $N = 100$ muestras y $F = 20$ FT. La Figura 6.13 muestra el mapa de calor correspondiente a dicho conjunto de datos de expresión.

La Figura 6.14 muestra las estimaciones de la actividad proteica $\tilde{x}_{fn}$ frente a su error cuadrático (SE), diferencia respecto a su valor verdadero normalizado y el error cuadrático medio (MSE) para el conjunto completo de perfiles. Del mismo modo, en la Figura 6.15 se representan los coeficientes estimados $\tilde{c}_{kf}$ frente a su SE, excluyendo los términos nulos a priori. Finalmente, en la Figura 6.16 se muestran los perfiles de proteínas usados para generar los datos y sus estimaciones, normalizados, a través de un mapa de calor. Por otro lado, en la Figura 6.17 se muestra el mapa de calor correspondiente a los coeficientes y su estimaciones.

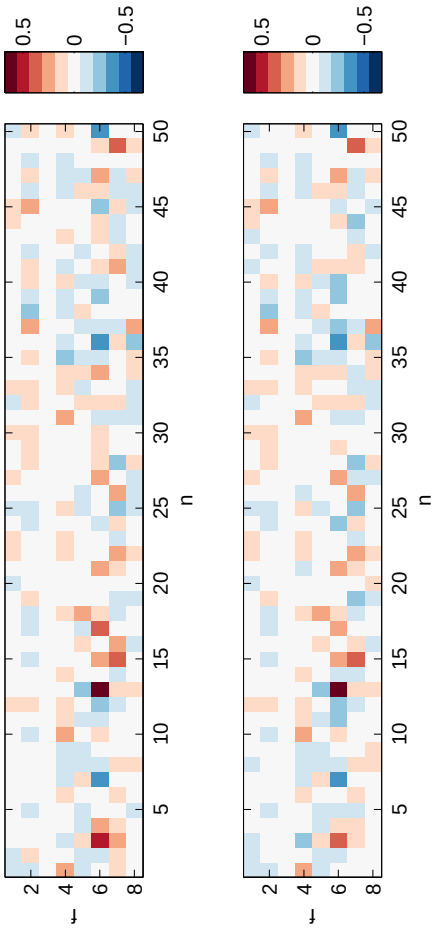


Figure 6.11: Mapa de calor con las actividades de las proteínas normalizadas $\tilde{x}_{fn}$. (a) Perfiles usados para generar los datos. (b) Estimaciones de los perfiles proteicos obtenidos mediante el método BEFM.

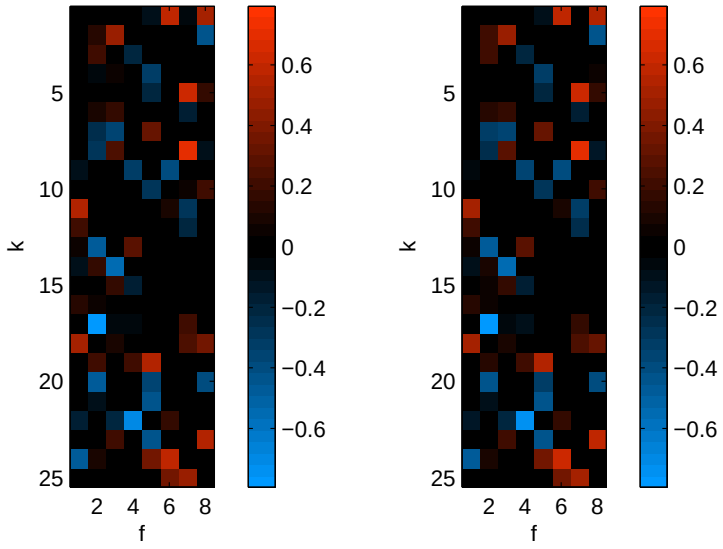


Figure 6.12: Mapa de calor con los coeficientes normalizados $\tilde{c}_{kf}$. (a) Coeficientes usados para generar los datos. (b) Coeficientes estimados mediante el método BEFM.

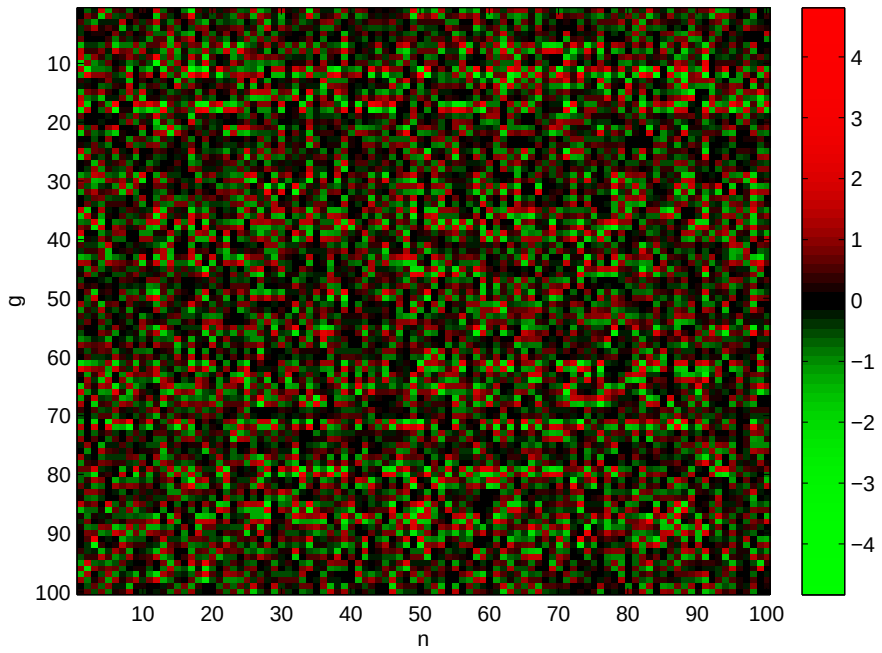


Figure 6.13: Datos sintéticos de expresión para $G = 100$ genes, $N = 100$ muestras, $F = 20$ FT y varianza del ruido $\sigma_n^2 = \frac{1}{10}$.

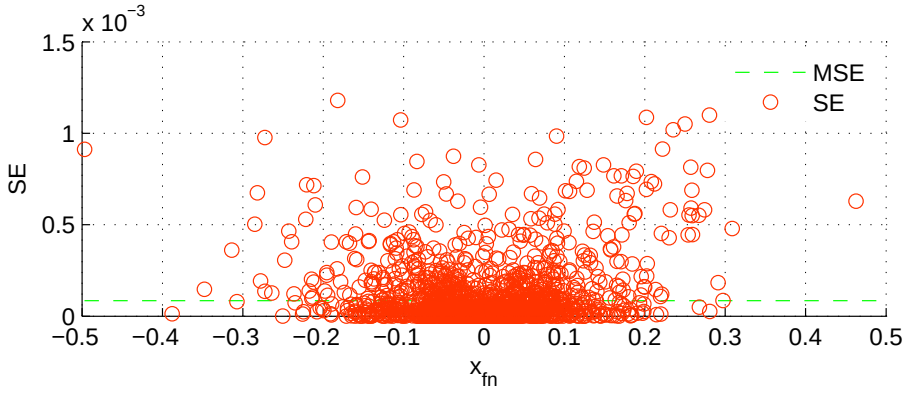


Figure 6.14: Error cuadrático (SE) frente las estimaciones normalizadas de las actividades $\tilde{x}_{fn}$, con MSE $= 5.06 \cdot 10^{-4}$.

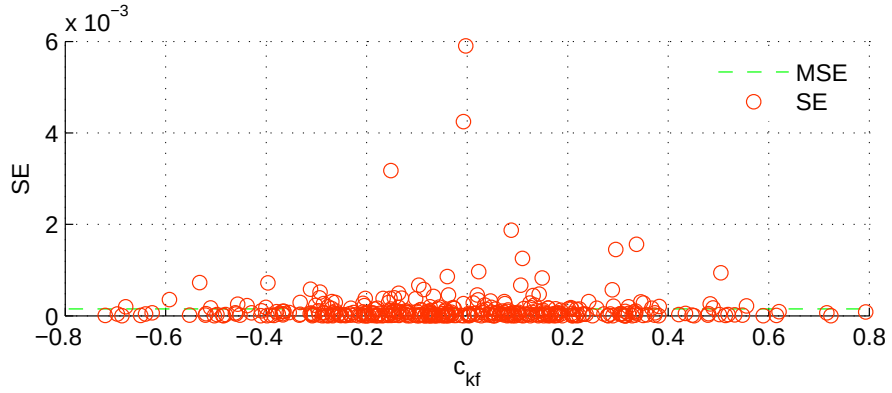


Figure 6.15: Error cuadrático (SE) frente las estimaciones normalizadas de los coeficientes $\tilde{c}_{kf}$, con MSE $= 3.12 \cdot 10^{-4}$.

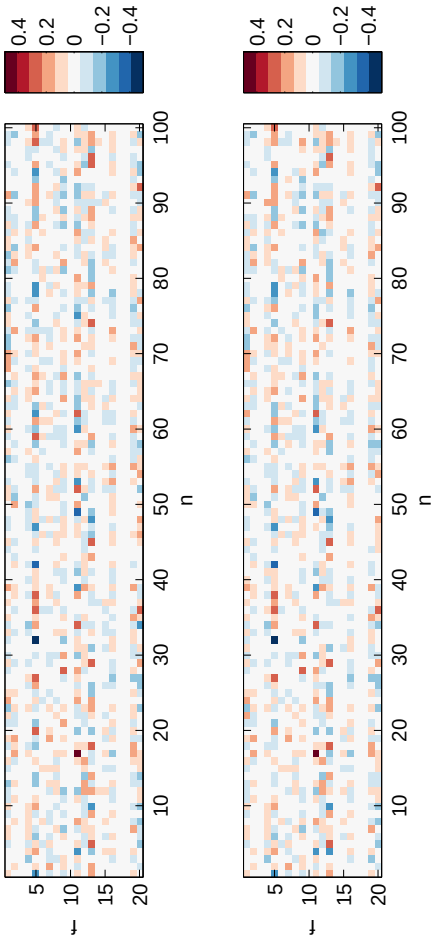


Figure 6.16: Mapa de calor con las actividades de las proteínas normalizadas $\tilde{x}_{fn}$. (a) Perfiles usados para generar los datos. (b) Estimaciones de los perfiles proteicos obtenidos mediante el método BEFM.

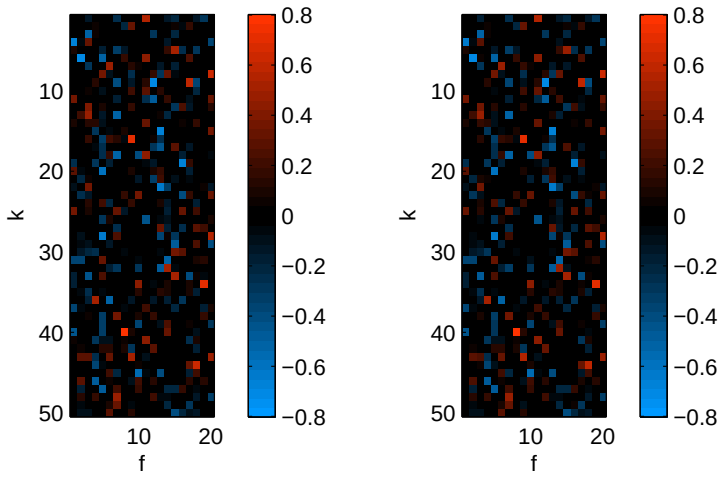


Figure 6.17: Mapa de calor con los coeficientes normalizados $\tilde{c}_{kf}$. (a) Coeficientes usados para generar los datos. (b) Coeficientes estimados mediante el método BEFM.

Chapter 7

Breast cancer subtyping method based on protein profiles classification

dedicado a Olga

Breast cancer is a malignant breast neoplasm, an abnormal proliferation of cell mass on breast tissue [65]. Besides skin cancer, breast carcinoma is the most commonly diagnosed cancer among women, about 28% of all carcinomas on women are breast cancer. Nevertheless, breast cancer is very unlikely among males, only 1% of all breast cancer cases are men. Cancer occurs as a result of mutations, or abnormal changes, in genes responsible for regulating the growth of cells. However, about 70%-80% of breast cancers occur in women who have no family history of breast cancer [85]. These occur due to genetic abnormalities that happen as a result of the aging process and life in general, rather than inherited mutations. Despite there are other significant factors (weigh or alcohol and smoking habits) the most determinant risk factors for breast cancer are gender and age, being more likely on older females.

Breast cancer usually begins in: (*i*) the cells of the lobules, which are the milk producing glands, or (*ii*) the ducts, the passages that drain milk from the lobules to the nipple. Less commonly, breast cancer can begin in the stromal tissues, which include the fatty and fibrous connective tissues of the breast. Over time, cancer cells can proliferate and spread to nearby healthy breast tissue. A non-invasive breast cancer, also called carcinoma in situ or stage zero, consist of a neoplasm in the breast tissue where the cancerous or abnormal cells have not spread beyond the breast ducts or lobules [86].

On the other hand, an invasive breast cancer, also known as infiltrating cancer, occurs when cancerous cells have spread beyond the ducts or lobules to other parts of the breast or body. Cancer cells can also invade nearby healthy breast tissue and make their way into the underarm lymph nodes. If cancer cells get into the lymph nodes, they then have a pathway into other parts of the body and can affect non-adjacent organ, process which is known as metastasis [62].

7.1 Histolopathology breast cancer classification

Breast cancers can be classified according to different schemata, each one based on different criteria and serving for different purposes related with the diagnosis, the treatment response and prognosis. The breast cancer's stage refers to a classification of the affected region and how far the cancer cells have spread beyond the original. A typical description usually considers aspects in turn: the histolopathological type, the morphology of the tumor and the metastasis nodal status [11]. According to the histopathology stage, the most common breast cancer (which represents about three-quarters of all cases) is the ductal carcinoma, an abnormal development of cell tissue on the milk ducts that comes in two forms: (*i*) ductal carcinoma in situ (DCIS), a noninvasive neoplasm confined to the mammary ducts, or (*ii*) invasive ductal carcinoma (IDC), an abnormal proliferation of cell tissue on the breast ducts. The second most common breast cancer is Lobular carcinoma also comes in two forms: (*i*) lobular carcinoma in situ (LCIS), an abnormal accumulation of cells in the breast lobules, which are connected to the milk ducts, and (*ii*) invasive lobular carcinoma (ILC) an abnormal proliferation of cells in the breast lobules.

The most common system to staging breast cancer is the TNM system that described the anatomical extent by three main characteristics [31].

Size the dimensions of affected tissue is scored for regions lower than 2 (cm), between 2-5 (cm) or higher than 5 (cm).

Nodal status the spreading to regional lymph nodes, distinguishing between different degrees according to the number of affected nodes. When lymph nodes are not affected of cancer it is usually referred as a negative nodal status. If any lymph node is involved in the cancer it is known as a positive nodal status.

Methastasis the spreading to another non-adjacent organ.

7.2 Molecular breast cancer classification

With histology description, breast cancer is a clinically heterogeneous disease where similar tumors may have different prognoses and may respond to therapy differently. An accurate classification also describes the biochemistry context of the disease [29].

7.2.1 Hormone receptor status

Receptors are proteins on the outside surfaces of cells that can attach to certain substances such as hormones [59]. Normal breast cells and some breast cancer cells have receptors that attach to estrogen and progesterone. These two hormones often fuel the growth of breast cancer cells. An important step in evaluating a breast cancer is to test if they have estrogen and progesterone receptors. Cancer cells may contain neither, one, or both of these receptors. Breast cancers that contain estrogen receptors are referred as ER-positive (ER+) cancers or ER-negative (ER-) otherwise; while those containing progesterone receptors are called PR-positive (PR+) cancers or PR-negative (PR-) if not. Women with hormone ER+ cancers tend to have a better prognosis and are much more likely to respond to hormone therapy than women with cancers without these receptors.

7.2.2 Growth receptor status

About 20% of breast cancers have too much of a growth-promoting protein called Human Epidermal growth factor Receptor 2 (HER2/neu, often just shortened to HER2). The EERB2 gene, also known as HERB2, instructs the cells to synthesize this protein[96]. Tumors with increased levels of HER2/neu are referred to as HER2-positive (HER2+). Women with HER2+ breast cancer have too many copies of the EERB2 gene, resulting in greater than normal amounts of the HER2 protein. These cancers tend to grow and spread more aggressively than other breast cancers and are treatment with drugs that target the HER2 protein, such as Trastuzumab and Lapatinib.

7.2.3 Intrinsic subtypes

A molecular classification based on microarray data, performed at earlier of year 2000, dissects breast cancer into intrinsic subtypes [65]. Results showed a correlation between a subset of genes and the clinical outcome that classified breast cancer into four subtypes, directly related with the prognosis of the disease. The list of genes that differentiates these subtypes was called the intrinsic list and is made up of several clusters of genes: (i) the luminal A and B cluster, relating to

ER expression, (*i*) genes relating to HER2 expression and (*iii*) unique set of genes called the basal cluster [86].

7.2.3.1 Luminal subtypes

The name luminal derives from similarity in expression between these tumors and the luminal epithelium of the breast. These are the most common subtypes, make up the majority of ER+ breast cancer and are characterized by expression of ER, PR, and other genes associated with ER activation. Luminal A tumors, which probably make up about 40% percent of all breast cancers, usually have high expression of ER-related genes, low expression of the HER2 cluster of genes and low expression of proliferation-related genes. Luminal A tumors carry the best prognosis of all breast cancer subtypes. On the other hand, Luminal B tumors are less common and have relatively lower, although still present, expression of ER related genes, a variable expression of the HER2 cluster, and higher expression of the proliferation cluster. Luminal B represents about 20% of breast cancers and it carry a worse prognosis than luminal A tumors.

7.2.3.2 HER2-enriched

The HER2-enriched subtype, HER2+ with ER-, makes up about 10%-15% of breast cancers and is characterized by low expression of the luminal cluster, high expression of the HER2 and proliferation gene clusters. For this reason, these tumors are typically ER-, PR- and HER2+. It is important to note that this subtype comprises only about half of clinically HER2+ breast cancer. The other half has high expression of both the HER2 and luminal gene clusters and fall in a luminal subtype. In the era before HER2-targeted therapy, this subtype carried a poor prognosis. This adverse natural history has been markedly affected by therapeutic advances in HER2-directed therapy.

7.2.3.3 Basal

The basal-like subtype, so called because of some similarity in expression to that of the basal epithelial cells, makes up about 15%-20% percent of breast cancers. It is characterized by low expression of the luminal and HER2 gene clusters. For this reason, these tumors are typically ER-, PR-, and HER2- on clinical assays, which have prompted the nickname triple negative to describe them. However, while most triple negative tumors are basal-like and most basal-like tumors are triple negative, there is significant discordance up to 30% between these two classifications methods that must be kept in mind when evaluating studies focused upon basal-like breast cancer. They also have high expression of the epidermal growth factor receptor (EGFR), as well as a unique cluster of genes called the basal cluster, which includes

basal epithelial cytokeratins 5, 14 and 17. Basal-like breast cancer has unique risk factors. Among the most intriguing is the strong association with cancers arising in women born with a mutation in the breast cancer gene 1, early onset (BRCA1) gene, in whom over 80% are basal-like. Even so, most basal-like breast cancers are sporadic and the BRCA1 gene and protein appear intact in these tumors. A commonly held, but unproven, assumption is that the BRCA1 pathway is abnormal in sporadic basal-like breast cancer, which may have therapeutic implications since this pathway is important in DNA repair.

7.3 Breast cancer classification based on protein profiles estimated by BEFM method

Breast cancer classification by traditional grading system has been enriched by more efficient methods based on gene expression. Intrinsic subtypes has suppose a breakthrough into breast cancer research, making possible the designing of clinical assays to diagnose breast cancer and for establishing therapy strategies. One of the earliest commercialized assay was the PAM50 test, that measures the expression status of 55 genes to classify breast cancer into intrinsic subtypes [62]. The experiment took into account a set of 80 gene expression profiles of different breast cancer tissues and 5 normal like samples. Other methods, widely extended for discerning which patients with breast carcinoma will benefit from hormonal or cytotoxic therapy, are Oncotype DX and MammaPrint [10]. Despite its practical application, gene expression signatures are not enough to uncover the differentiation mechanism of breast cancer and its unpredictable outcomes. The sensitivity of breast cancer to hormonal and protein changes suggests a deeper analysis at the transcriptional space. A molecular signature based on protein profiles may answer the still many doubts surrounding the diagnosis an therapy design of breast cancer. Whilst gene quantification techniques are widely available and economically accessible, protein profiling is a specific and complex procedure.

BEFM method presented in Chapter 6 has been applied in breast cancer data to infer the protein activities with classification purposes. Specifically, the set of samples with a complete clinical history in the PAM50 test have been considered, with $N = 80$ samples and $G = 55$ genes [16]. Data set includes 36 basal, 12 luminal A (luma), 7 luminal B (lumb) and 25 HER2 (her2) subtypes. Microarray data has been downloaded from GEO website [3]. Figure7.1 shows a clustered heatmap with the microarray data.

Experts have suggested $F = 17$ transcription factors potentially involved in breast cancer. Corresponding priors $\pi_{gf}, \forall g, f$ have been learned from TransFac using MATCH tool. Inference procedure have followed same scheme as in synthetic data, simulating 10 parallel chains with $T = 10000$ and a burning period of $\frac{T}{2}$. For

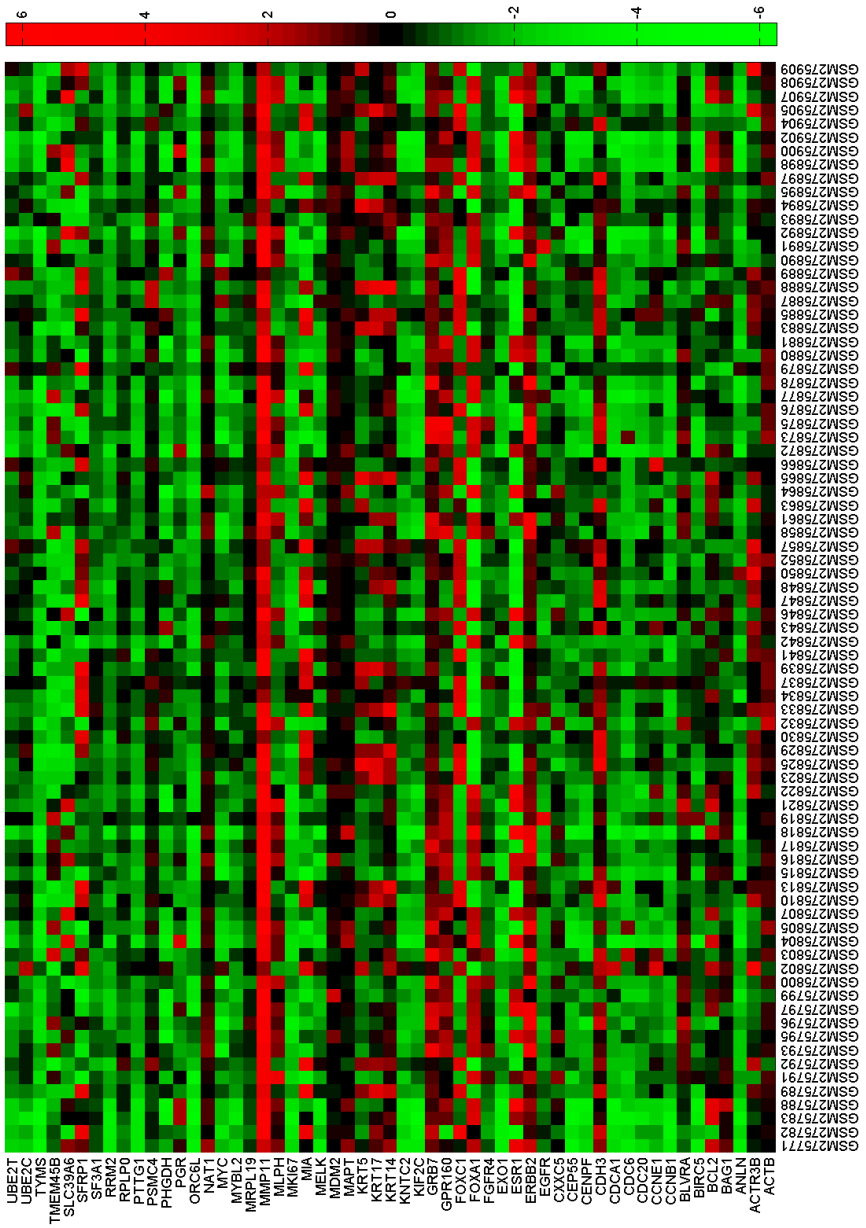


Figure 7.1: Breast cancer microarray data with the $G = 55$ genes and $N = 80$ samples considered for training the PAM50 test.

classification purposes, regulon normalizations as in (6.76) is not required and a simple normalization of each TF profiles, by columns of $\mathbf{X}$, has been applied. Figure 7.2 shows a clustergram of estimated protein activities, where samples names have been changed by its intrinsic subtype. Results shown a perfect classification at TF space, where Basal-like tumors are characterized by a down regulation of protein E2F1. This protein mediated the regulation of genes whose products are involved in DNA replication and cell proliferation [60]. On the other hand, CEBP1 protein is another marker that disects between HER2 and Luminal subtypes.

The original expression profiles has also been reconstructed with the estimated coefficients and activities, according to model BEFM (6.16) excluding the noise term. Figure 7.3 shows the clustergram of estimated expression data, where samples names have been changed by its intrinsic subtype again. Results shown again a clearly distinction between basal and the other subtypes. Moreover, known features as the high expression of ERBB2 gene on HER2 subtype or the activation on EGFR gene on Basal-like tumors are perfectly recovered.

Finally, a more intuitive representation of protein activity profiles is represented in Figure 7.4. Some relevant clinical parameters as age, nodal status, time free of relapse or time to death are represented with a colored label key. It can be seen that HER2 subtype is characterized by high activity of CEBP1 protein. Additionally, cases with lower activity of E2F1, whose overexpression is related with the apoptosis of cancerous cells [60], suffers relapse and patients finally died as cause of the disease.

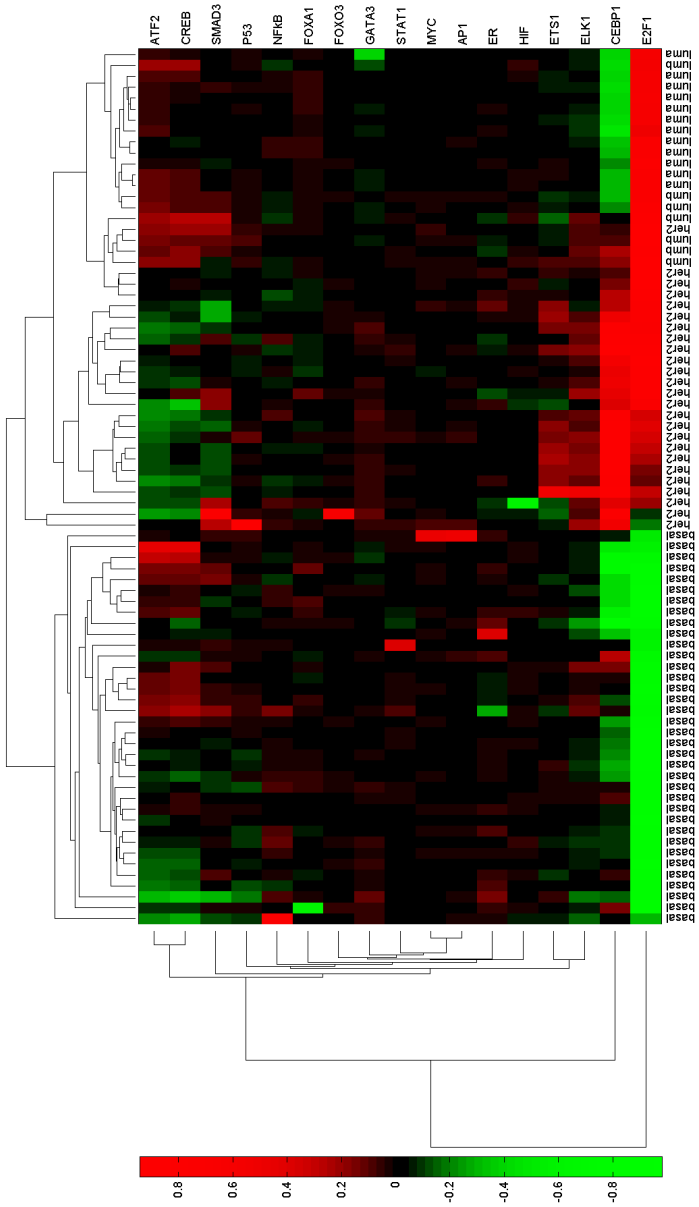


Figure 7.2: Clustergram with estimated protein activities for breast cancer data with the $F = 17$ proteins and $N = 80$ samples.

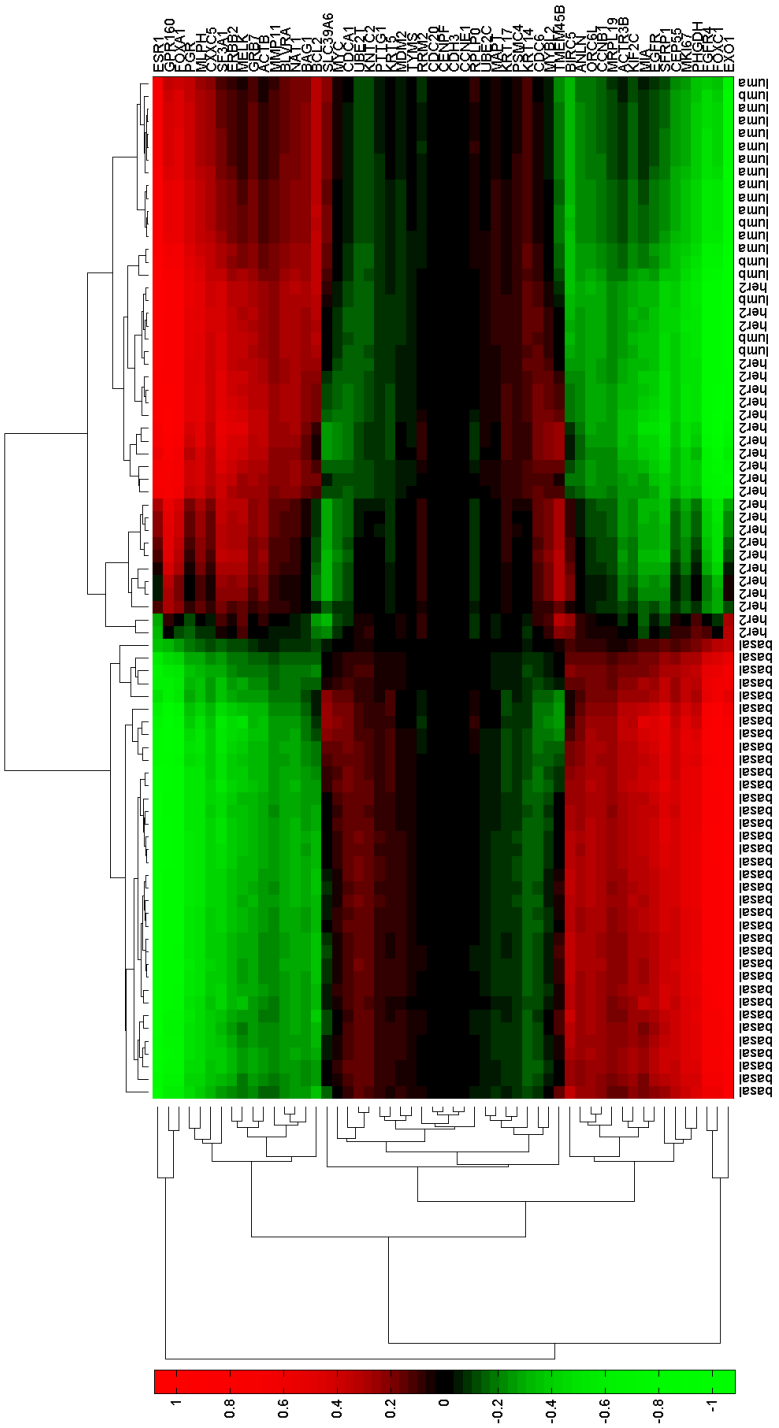


Figure 7.3: Clustergram with reconstructed microarray data by estimations with the $G = 55$ genes and $N = 100$ samples.

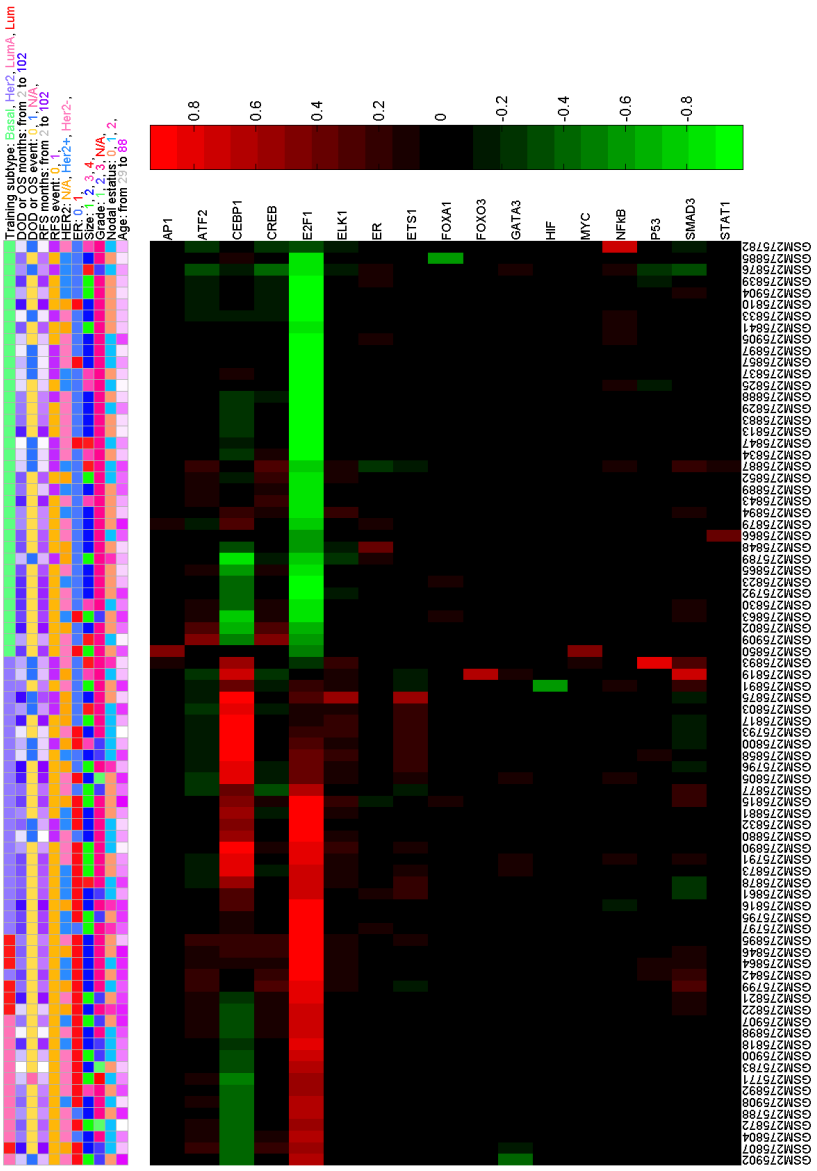


Figure 7.4: Heatmap with estimated protein activities for breast cancer data with the $F = 17$ proteins and $N = 80$ samples. Clinical data are also represented with a colored label key over heatmap. N/A: not available value. DOD (dead of disease) or OS (overall survival) event: 1 if DOD and 0 if OS. RFS (relapse free survival) event: 1 if relapse and 0 otherwise. ER (estrogen receptor): 0 if negative and 1 if postive. Size, according to TNM grading system: 1 less or equal than 2(cm), 2 higher than 2(cm) and less or equal than 5(cm), 3 higher than 5(cm) and 4 any size with direct extension to chest wall or skin. Grade, according to TNM grading system: 1 low, 2 intermediate and 3 high. Nodal Status: 0 negative, 1 positive and 2 metastasis.

Chapter 8

Conclusions and main contributions to the field

Microarray experiments are versatile tools allowing to quantify efficiently gene expression. Abstracts models learned from microarray data constitutes a valuable source of biological knowledge. Microarrays time series contains information about the regulatory mechanism conducted within the cell during its metabolic development. Uncovering the GRN form microarray data is very important in Biology Sistems to identify gene functions and how genes interact to produce diferrent responses. Main problem found on this type of analyisis is the high level of noise affecting data. Whislt more complex approaches have been proposed, linear asumptions fits conveniently micorarray data. A review of these approaches and its implications in technology has been published in [80]. Linear model proposed in this thesis assumes a Markov process that establishes relationships between data. Specifically, an AR1MA1 model that takes into account differences between the real expression level and its noisy observation is proposed. Analysis of the novel AR1MA1 model and the previosu AR1 model considered in reverse engineering problems have been published in [76]. As an inference method it is considered the VBEM approach that makes the problem tractable from a computational point of view. Analysis of the improvements of this method have been published in several conferences [78] [81]. Real data have been also considered by the AR1MA1-VBEM method. The Bayesian framework provides a powerfull formalism for modeling priors and adapting the model to a specific setting. Specifically, the results with yeast data have been published in several conferences [77] [79].

The potential of Bayesian formalism has been exploited in other approach that fits microarray data within a latent factor model. This analysis considers prior knwoledge from the regulatory process at a transcriptional level. Main difficulty in previous approaches was the way this prios knowledge was modeled and introduced

with rigid constrains. One of the main contributions of the analysis presented in this thesis is the original way in which the transcriptional network is modeled by a FIG distribution. This approach establishes a novel estrategy in the general problem of modeling sparsity and it have been published on conference [82]. Moreover a clever basis expasion model based on Wavelets is proposed for reducing the dimensionality of the problem and making it tractable from a computational point of view. BEFM is solved with a Gibbs sampling method. Results with *in-silico* and *in-vivo* data have been published on several conferences [54] [75].

SUMMARY

Summary

The genetic inheritance that a living being transmits to its offspring is stored in DNA macromolecules inside the nucleus of prokaryote cells, or in form of RNA in the cytoplasm of eukaryotic organisms. The nucleotide sequence of these nucleic acids encodes the characteristics of each individual of a species and potentially controls its cell development [1]. The part of the code that regulates a feature completely is called a gene and is the hereditary information storage unit. The central dogma of Molecular Biology describes by a unidirectional flowchart the way the genetic information is encoded, stored and transmitted from a living being to its offspring [18]. In prokaryotes, unlike in eukaryotes, DNA is not directly responsible for the cellular development. First, the information stored in DNA is transcribed into a RNA molecule. Subsequently, it is translated into a protein that is actively involved in cell metabolism. Whether the information stored within a gene is finally translated into a protein, it is said that the gene is expressed or activated.

Microarrays are experimental procedures for quantifying the expression of thousand of genes simultaneously [30]. With this technique gene expression can be massively profiled, leading to huge data sets that are widely available in public databases. Additionally, chromatin inmunoprecipitation (ChIP) [17] as well as other novel techniques such as SELEX [27], combined with gene sequencing, are allowing the construction of databases with gene-protein interactions [53]. Whilst all these experimental and computational procedures allows to measure gene expression [21] and to predict the transcriptional regulatory structure [44], other biological features of interest are difficult to quantify [19] [55]. Uncovering this kind of information is very important for many fields, such as Pharmacology and Medicine therapy, and it analysis demands help from the Computer Science community.

AR1MA1-VBEM method for GRN reverse engineering

Gene expression is a dynamical and complex process in which transcribed RNA is matured and modified at different postranciptional stages. One model that tries

to explain these interactions at a genetic level are the gene regulatory networks (GRN). GRN are abstract graphical models that simplifies the regulatory mechanisms from a phenomenological point of view [19]. In a GRN it is considered that the activation of a gene (known as child) depends on the expression status of others presented in the network (its parents). Uncovering the GRN is interesting for Systems Biology to understand how genes compete and are associated to produce complex responses and co-operative effects [24]. The problem of modeling and inferring the GRN from a microarray time series has been addressed. In a microarray, the expression levels (log-transformed relative abundance of transcribed mRNA) of thousands of genes can be evaluated simultaneously. Nevertheless, the number of temporal samples that can be obtained is usually fewer than the number of genes measured in a microarray. Therefore, the development of a mathematical tool able to estimate the GRN from a short time series is complex.

The earlier attempts to describe GRN from a mathematical point of view are the Boolean networks (BN) [43] and its extension to probabilistic Boolean networks (PBN) [84]. In this models, gene expression is quantized into binary levels: activation and inhibition. Genetic relationships are described by a predictor function that establishes logic relationships between genes. Additionally, in PBN several possible binary functions were provided with a probabilistic measure for each logic rule. On the other hand, Bayesian networks is another promising approach due to its ability to describe complex stochastic process and also handle with noise [32] [38]. Bayesian formalism describes GRN by a belief network, where the expression status of the i-th gene at the n-th time sample is represented by a random variable, denoted as $y_i(n)$. Within this framework, the expression level of any gene is not limited to a simple binary state. However, using data alone, causal inference for genetic relationships is very limited. Nevertheless, another advantage of Bayesian formalism is that any constraints on the network space may be included as a prior information.

We have revised the Bayesian linear approach introduced by Tienda-Luna and Huang in [89] [40]. In this model, GRN is mathematically described by two features [70]. First, its topology, i.e. the connectivity pattern, represented by a set of binary variables $x_i(j)$ that have a value of one if the i-th gene is a child of the j-th one and zero otherwise. Second, the regulatory type and strength, described by a set of weights $\omega_i(j)$ that have a positive value if the i-th gene activates the expression of the j-th child and negative if it inhibits. Specifically, authors propose a first order autoregressive (AR1) model where the expression level $y_i(n)$ follows a time homogeneous Markov process, where microarray data is expressed as a linear combination of the observations at the immediately previous time step plus noise

as,

$$y_i(n) = \sum_{j=1}^{G} y_j(n-1)\omega_i(j)x_i(j) + e_i(n) \tag{8.1}$$

with $e_i(n)$ the additive noisy term modeled by IID white noise with unknown variance σ_i^2. However, this model establishes relationships between the observed expression levels, $y_i(n)$, which are supposed to be noisy. It would be much more realistic to establish these relationships between the real expression level, denoted by $z_i(n) = y_i(n) - e_i(n)$. Taking into account these differences, we proposed a novel approach where genetic relationships are established between the real expression levels instead of its noisy observation, leading to a first order auto regressive moving-average (AR1MA1) model as

$$y_i(n) = \sum_{j=1}^{G} y_j(n-1)\omega_i(j)x_i(j) - \sum_{j=1}^{G} e_j(n-1)\omega_i(j)x_i(j) + e_i(n). \tag{8.2}$$

With this modification, AR1MA1 model distinguishes between data and real expression levels and constitutes a more realistic approach to the nature of the problem [78]. The inference problem have been solved using a variational Bayesian approach [5], that demands conjugate modeling of the unknowns: $x_i(j)$, $\omega_i(j)$ y σ_i^2. Specifically, binary variables are conveniently modeled by prior a Gaussian distribution as

$$p\left(\mathbf{x}_i\right) = \mathcal{N}\left(\mathbf{x}_i \middle| \, \mathbf{m}_{\mathbf{x}_i}, \sigma_0^2 \mathbb{1}^G\right) \tag{8.3}$$

with $\mathbf{x}_i = [x_i(1), \ldots, x_i(G)]^\top$ the set of variables describing the network topology of the i-th gene. On the other hand, weights and noise variance are considered as parameters and are modeled by a joint prior probability as a Normal scaled Inverse Gamma distribution as

$$p\left(\boldsymbol{\omega}_i, \sigma_i^2\right) = \mathcal{N}\left(\boldsymbol{\omega}_i \middle| \, \mathbf{m}_{\boldsymbol{\omega}_i}, \frac{\sigma_i^2}{\gamma_i} \mathbb{1}^G\right) \mathcal{IG}\left(\sigma_i^2 \middle| \, a_i, \gamma_i b_i\right) \tag{8.4}$$

with $\boldsymbol{\omega}_i = [\omega_i(1), \ldots, \omega_i(G)]^\top$ the set of weights regulating expression of the i-th gene and scale variance

$$\gamma_i = \frac{1}{1 + \sum_{j=1}^{G} x_i^2(j) + \omega_i^2(j)}. \tag{8.5}$$

Moreover, given the microarray data set

$$\mathbf{Y} = \begin{pmatrix} y_1(0) & \cdots & y_1(N) \\ \vdots & \ddots & \vdots \\ y_G(0) & \cdots & y_G(N) \end{pmatrix} \in \mathbb{R}^{G \times (N+1)} \tag{8.6}$$

with $N+1$ time samples, the likelihood function function based on the AR1MA1 model will be a multivariate Gaussian with unknown mean and variance as,

$$p\left(\mathbf{y}_i|\,\mathbf{x}_i,\boldsymbol{\theta}_i\right)=\mathcal{N}\left(\mathbf{y}_i|\,\mathbf{R}\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i,\frac{\sigma_i^2}{\gamma_i}\mathbb{1}^N\right) \tag{8.7}$$

with $\mathbf{y}_i=[y_i(1),\ldots,y_i(N)]^\top$ microarray time series, $\mathbf{R}=\mathbf{T}\mathbf{Y}^\top$, $\mathbf{T}=\left(\mathbb{1}^N\middle|\,\mathbf{0}\right)$ and $\mathbf{D}_{\boldsymbol{\omega}_i}$ a diagonal matrix with vector $\boldsymbol{\omega}_i$.

Variational Bayesian Expectation-Maximization (VBEM) method optimizes a lower bound of the marginal likelihood

$$\begin{aligned}\ln p\left(\mathbf{y}_i\right) &\geq \mathcal{F}\left[q\left(\mathbf{x}_i\right),q\left(\boldsymbol{\omega}_i,\sigma_i^2\right)\right]\\ &= \int d\boldsymbol{\omega}_i d\mathbf{x}_i q\left(\mathbf{x}_i\right)q\left(\boldsymbol{\omega}_i,\sigma_i^2\right)\ln\frac{p\left(\mathbf{y}_i|\,\mathbf{x}_i,\boldsymbol{\theta}_i\right)p\left(\mathbf{x}_i\right)q\left(\boldsymbol{\omega}_i,\sigma_i^2\right)}{q\left(\mathbf{x}_i\right),q\left(\boldsymbol{\omega}_i,\sigma_i^2\right)}\end{aligned} \tag{8.8}$$

that depends on a free distribution that factorizes in the same family than the unknowns as

$$q\left(\mathbf{x}_i\right) = \mathcal{N}\left(\mathbf{x}_i|\,\boldsymbol{\mu}_{\mathbf{x}_i},\boldsymbol{\Sigma}_{\mathbf{x}_i}\right) \tag{8.9}$$

$$q\left(\boldsymbol{\omega}_i,\sigma_i^2\right) = \mathcal{N}\left(\boldsymbol{\omega}_i|\,\boldsymbol{\mu}_{\boldsymbol{\omega}_i},\sigma_i^2\boldsymbol{\Sigma}_{\boldsymbol{\omega}_i}\right)\mathcal{IG}\left(\sigma_i^2\middle|\,\alpha_i,\beta_i\right) \tag{8.10}$$

with $\left\{\boldsymbol{\mu}_{\mathbf{x}_i},\boldsymbol{\Sigma}_{\mathbf{x}_i},\boldsymbol{\mu}_{\boldsymbol{\omega}_i},\boldsymbol{\Sigma}_{\boldsymbol{\omega}_i},\alpha_i,\beta_i\right\}$ the hyperparameters to be learned from microarray data. VBEM method demands a conjugate model that likelihood (8.7) and priors (8.3)-(8.4) do not satisfy [5] [91]. Specifically, the dependence of γ_i on $\mathbf{x}_i$ and $\boldsymbol{\omega}_i$ breaks the symmetries of conjugate models. Moreover, the derivation of the probability of the scale variance γ_i from the priors of $\mathbf{x}_i$ is too complex to the point of being impossible analytically and the formulation of an alternative conjugate model is not possible. Hence, we propose a fixed point approach where the scale of the variance is approximated according to the most probable value of the unknowns, the means of the Gaussian free distributions, as

$$\gamma_i\approx\bar{\gamma}_i=\frac{1}{1+\sum_{j=1}^{G}\mu_{\mathbf{x}_i}^2(j)+\mu_{\boldsymbol{\omega}_{\mathbf{i}i}}^2(j)} \tag{8.11}$$

Therefore, VBEM iteratively maximizes lower bound (8.8) in two steps, denoted as VBE and VBM, where one of the free distributions in (8.9) and (8.10) is optimized whilst the other is fixed. After the t-th iteration, VBE hyperparameters learning rules can be calculated as

$$\left\{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)},\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right\} = \arg\max_{q(\mathbf{x}_i)}\left\{\mathcal{F}\left[q\left(\mathbf{x}_i\right),q\left(\boldsymbol{\omega}_i,\sigma_i^2\right)\right]\right\} \tag{8.12}$$

and VBM step leads to

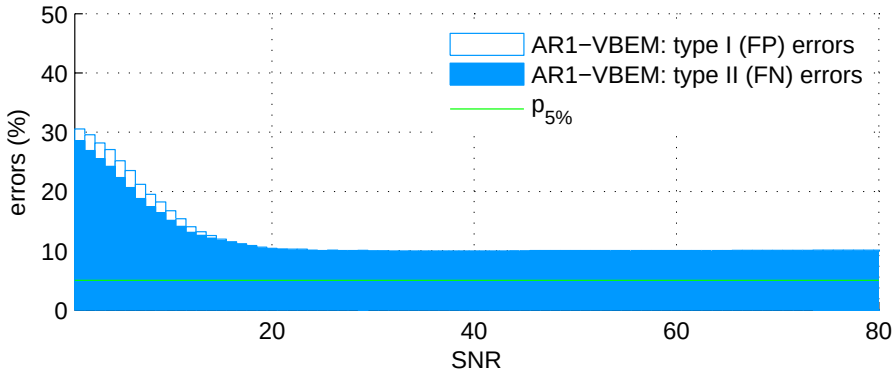
$$\left\{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)},\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)},\hat{\alpha}_i^{(t+1)},\hat{\beta}_i^{(t+1)}\right\} = \arg\max_{q\left(\boldsymbol{\omega}_i,\sigma_i^2\right)}\mathcal{F}\left[q\left(\mathbf{x}_i\right),q\left(\boldsymbol{\omega}_i,\sigma_i^2\right)\right] \tag{8.13}$$

where, before each step, the scale is updated with the mos recent updated hyper-parameters. Subsequently, after two consecutive steps, lower bound is updated and algorithm iterates until a convergence criteria were satisfied.

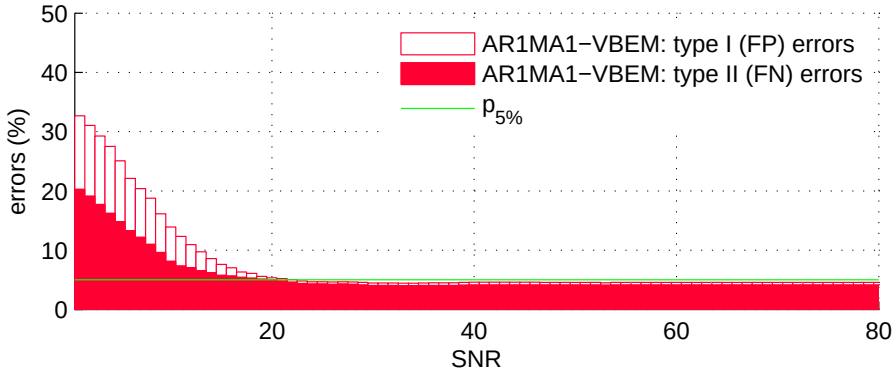
VBEM method based on AR1MA1 model, refereed as AR1MA1-VBEM, have been validated by simulation. We have simulated synthetic data sets according to priors and likelihood with different settings and levels of noise. Moreover, we have compared it performance with the VBEM method based on AR1 model, refereed as AR1-VBEM. Figure 8.1 shows the performance of both VBEM methods for same synthetic data set with $G = 25$ genes, $N = 25$ samples and a gene with 8 parents. Results has been averaged over one hundred realizations for the same gene. The percentage of errors, false negatives (FN) and false positives (FP), are represented as a bar plot versus the level of noise, expressed in terms of the signal-to-noise ratio (SNR). Moreover, the error rates is depicted in this plot as a filled or empty fraction of the bar. Both methods have a similar performance at higher levels of noise with unsatisfactory results. However, AR1MA1-VBEM produces both types of errors at lower SNR and it reduces it simultaneously as the noise level decreases. On the other hand, AR1-VBEM mainly produces FN with unsatisfactory results, over the 5% percentile, at any SNR.

Additionally, the performance of both VBEM methods have been evaluated by a receiver operating characteristic (ROC) and a Precision-Recall (PR) analysis. In ROC curve, the true positive rate (TPR) is plotted versus the false positive rate (FPR) while the probability threshold changes. On the other hand, PR curve represents the predictive positive value (PPV) versus the TPR. The overall performance is summed up into the area under ROC curve (AUROCc) and the area under PR curve (AUPRc), with a maximum value of one for the optimum performance. Figure 8.2 shows the AUROCc and AUPRc of both VBEM methods versus the noise for same synthetic data set with $G = 25$ genes, $N = 25$ samples and a gene with 8 parents. It is shown that AR1MA1-VBEM method outperforms AR1-VBEM at both ROC and PR spaces.

Moreover, AR1MA1-VBEM method have been validated with yeast *in-silico* data. GeneNetWeaver (GNW) have been used as a benchmark generator that considers realistic networks to simulate microarray data [83]. Assuming such kind of independent generative model. endowed with a biological background, results and properties derived from the performance will have a more general validity. Specifically, it has been considered a subnetwork with $G = 25$ genes and a time series with $N = 51$ samples. GRN and kinetic parameters of generative model was randomly selected by default settings. Therefore, the network topology is not controlled and it includes includes genes with 0 to 10 parents. GRN of previous data set has been inferred with AR1MA1-VBEM method and performance has been compared with other GRN inference method: ARACNE [52]. Figure 8.3 represents the ROC and PR curves for both ARACNE and AR1MA1-VBEM methods. As it can be notice,

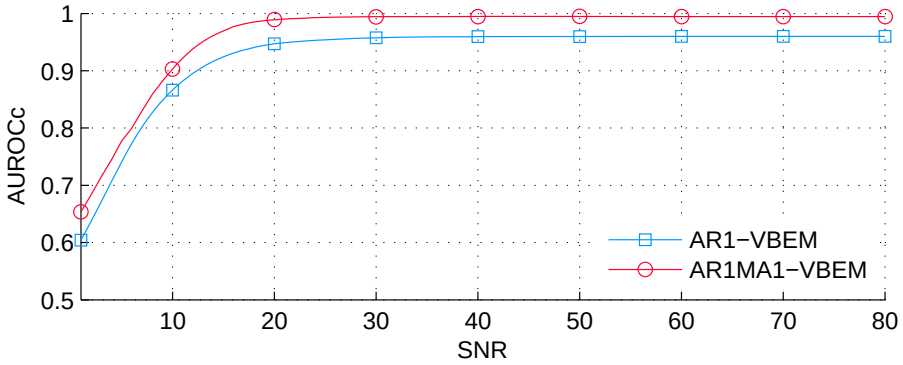


(a)

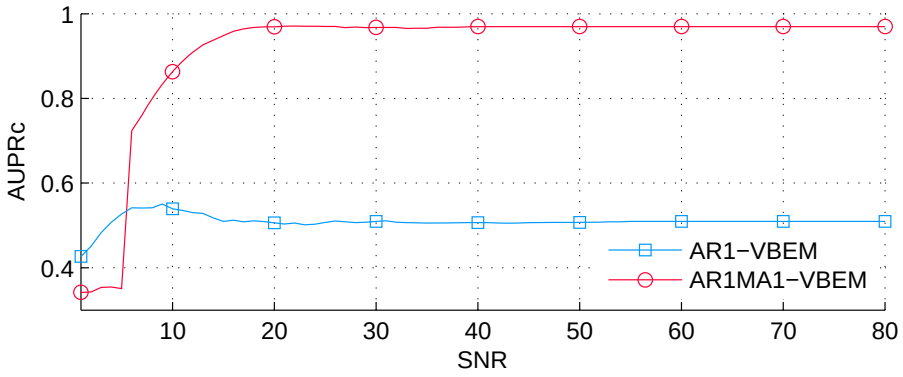


(b)

Figure 8.1: Performance of GRN inference of AR1-VBEM method and AR1MA1-VBEM method for a synthetic data set with $G = 25$ genes, $N = 25$ samples and 8 parents. The overall error percentage is represented as a bar at each level of noise. The fraction FN are proportionally depicted as a filled bar and the FP fraction as an outlined bar. (a) AR1-VBEM method shows unsatisfactory results at any level of noise. (b) AR1MA1-VBEM shows a dependency with the level of noise that outperforms AR1-VBEM method and finally it reaches an error level under the 5% percentile.



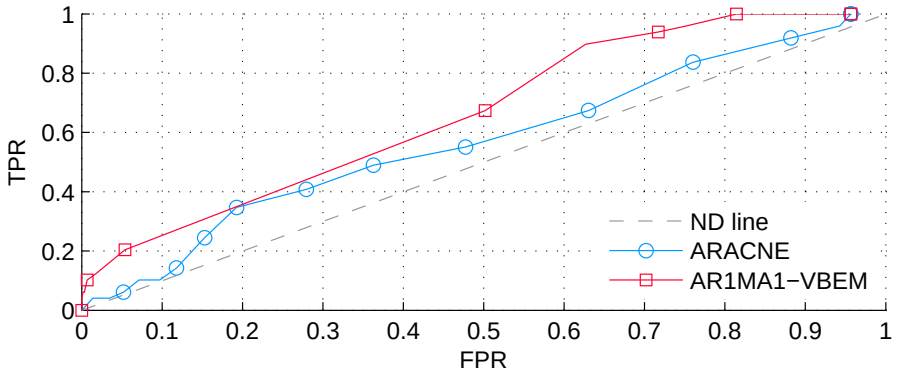
(a)



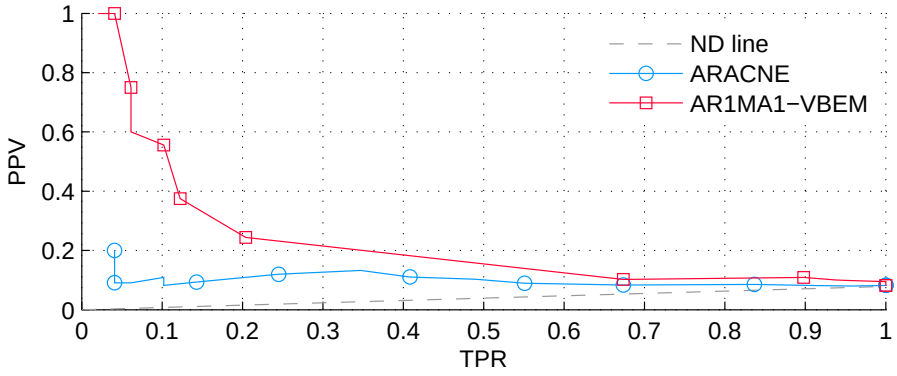
(b)

Figure 8.2: ROC and PR analysis of AR1-VBEM and AR1MA1-VBEM methods for a synthetic data set with $G = 25$ genes, $N = 25$ samples and 8 parents. The area under the curves are represented verssus the noise level. (a) AR1MA1-VBEM outperforms AR1-VBEM method providing larger AUROCc at any SNR. (b) Both VBEM methods shown unsatisfactory results at higher levels of noise, however, AR1MA1-VBEM reveal a significant improvement at a functional level of noise with SNR ≥ 10 and beyond.

AR1MA1-VBEM method outperforms results obtained with ARACNE algorithm with higher AUROCc and AUPRc. Table 8.1 summarizes the area under curves.



(a)



(b)

Figure 8.3: ROC and PR analysis for AR1-VBEM, AR1MA1-VBEM and ARACNE methods for a synthetic data set with $G = 25$ genes and $N = 51$ time samples. Random assignment is illustrated by the no discrimination (ND) line. (a) ROC curve. (b) PR curve. AR1MA1-VBEM method outperforms ARACNE algorithm with higher AUROCc and AUPRc.

Finally, AR1-VBEM method has been applied in real microarray data [79]. Specifically, we have considered Spellman's data set with $G = 11$ genes $N = 17$ temporal steps of *Saccharomyces cerevisiae* (budding yeast) cell cycle [87]. Yeast has been studied widely as a model organism, thus, most of gene functions and the metabolic pathways are well known. Specifically, its cell cycle and the GRN during this process was studied by in [45]. Figure 8.4 shows the GRN of reference and some relationships of the inferred one with AR1MA1-VBEM method.

	ARACNE	AR1MA1-VBEM
AUROCc	0.5263	0.6218
AUPRc	0.0931	0.2118

Table 8.1: Area under the ROC and PR curves in the performance of AR1MA1-VBEM method and ARACNE for a data set with $G = 25$ and $N = 51$.

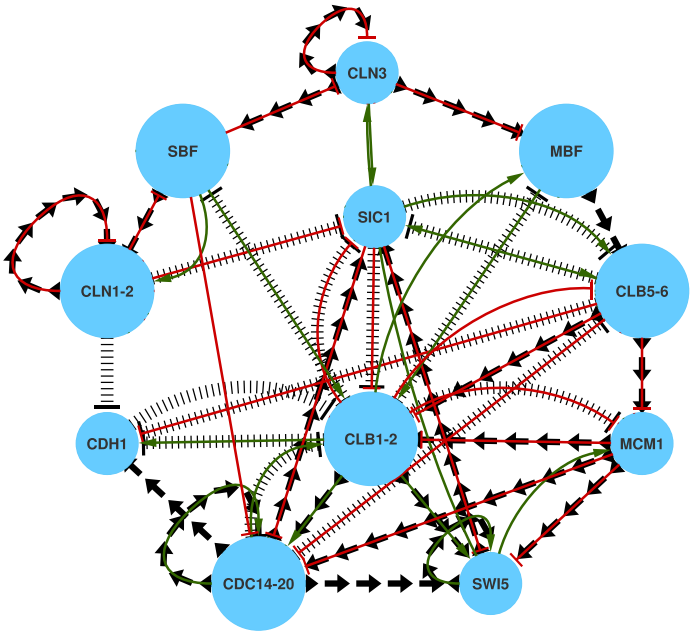


Figure 8.4: GRN of reference proposed [45] and the one estimated with AR1MA1-VBEM method for a real data set with $G = 11$ genes and $N = 17$ time samples for the Spellman's yeast experiment. GRN of reference is depicted with discontinuous arrow-shaped and dashed strokes (for gene activation or inhibition respectively). Estimated GRN is plotted as a green continuous stroke with arrowed tip for gene activation and red continuous stroke with a "'T"'-shaped tip for inhibition.

BEFM for TRN modeling and protein activities profiling

A more realistic approach to gene regulatory process are transcriptional regulatory networks (TRN). A TRN is a graphical model that explains gene expression at a transcriptional level, one of the earliest stage of the regulatory process. Gene transcription is mediated by a set of functional proteins, refereed as transcription factors (TF), that binds its promoter region and up- or down-regulates gene expression. TRN describes gene expression at a protein-gene interaction level, where TF are always source nodes and genes are targets nodes. Structure of TRN may be conveniently described by a set of coefficients $[\mathbf{A}]_{gf} = a_{gf}, \forall g, f$ that expressed the independence between the f-th TF and the g-th gene with a null value. On the other hand, this coefficient weights with positive/negative values the up/down regulatory effects. Whilst most recent experimental and computational techniques allows to infer the TRN structure, other kind of biological features of interest such as the TF abundance are difficult to measure. Protein profiles are interesting in many fields, such as disease treatment and new drug design, due to its potential to dissect clinically heterogeneous diseases with a characteristic molecular signature. The problem of modeling the TRN to infer the TF activities from microarray and prior information about the network structure, learned from databases [44], has been addressed. Specifically, the topology of the TRN is learned using TransFac database and performing TF binding site prediction [53]. This results are summarized in a set of prior probabilities $[\mathbf{\Pi}]_{gf} = \pi_{gf}, \forall g, f$, that takes a zero value whether the f-th factor do not regulates the g-th gene. A previous analysis for the entire human genome suggest that at least 69% of prior probabilities are zero. Figure 8.5 shows the distribution of estimated priors π_{gf} considering the entire genome and a small subset for an specific application on breast cancer subtyping. Therefore, the set of coefficients $\mathbf{A}$ describing the TRN is expected to be sparse, i.e. with a high density of zero elements. Hence, we propose a Bayesian approach that efficiently integrates priors for dimensionality reduction.

Previous works have intended to fit TRN with a factor model that establishes relationships between gene expression data and the TF activities. The loading matrix that links genetic and protein profiles was identified by the set of coefficients $\mathbf{A}$ describing the TRN, denoting suitably the regulatory strength and type with $a_{gf} \neq 0$ and the absence of regulatory effects with $a_{gf} = 0$. Some approaches impose rigid constrains at loading matrix [46] [67] trying to solve identifiability problems. Despite been applicable in some cases, these restrictions are not realistic assumptions and may lead to incorrect results. Actually, instead of map the expression profiles into a real TF space, these approaches projects microarray data into an abstract lower-dimension space of metagenes, where the transcriptional reg-

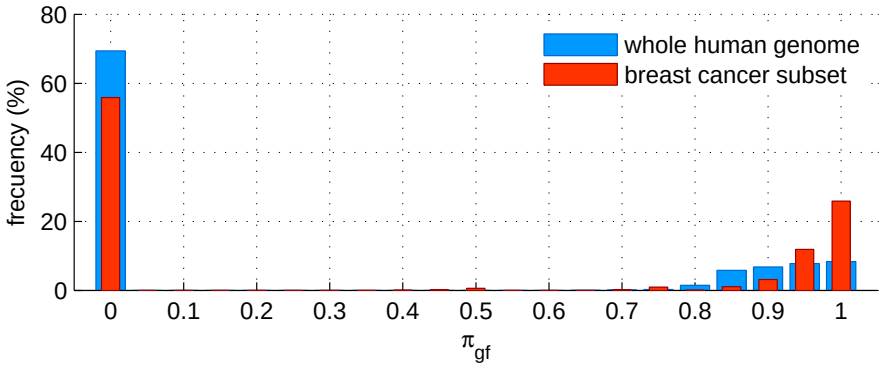


Figure 8.5: Distribution of predicted priors $\pi_{gf} \in [0,1]$ for the whole human genome with $G = 36066$ genes and $F = 850$ TFs and for a specific subset considered in breast cancer classification with $G = 55$ genes and $F = 17$ TFs. Results confirms that loading matrix may be highly sparse.

ulatory meaning has been lost [12]. An alternative wise approach, able to cope with priors, imposes probabilistic sparsity constrains based on learned structure of the TRN. Such kind of modeling considers the loading coefficients distributed by a slab-and-spiked prior [74] [14] [55] expressed in terms of a mixture Gaussian with a mass in zero as

$$p\left(a_{gf}\right) = \left(1 - \pi_{gf}\right)\delta\left(a_{gf}\right) + \pi_{gf}\mathcal{N}\left(a_{gf}|\,0, \sigma_f^2\right). \tag{8.14}$$

with $\delta\left(a_{gf}\right)$ Dirac delta function and σ_f^2 unknown variance of the f-th factor. Despite of describing conveniently the loadings, this kind of distribution complicates computation within the Bayesian framework. Therefore, a novel approach has been proposed based on a functional approximation of this distribution by a prior induced Gaussian (FIG) as

$$p\left(a_{gf}\right) = \mathcal{FIG}\left(a_{gf}|\,0, \sigma_f^2, \pi_{gf}\right) = \begin{cases} 0 & , \quad \pi_{gf} = 0 \\ \mathcal{N}\left(a_{bf}|\,0, \sigma_f^2\pi_{gf}\right) & , \quad \pi_{gf} \neq 0 \end{cases} \tag{8.15}$$

Figure 8.6 illustrates the slab-and-spiked distribution as in (8.14) and the proposed FIG approach in (8.15) with different non-zero mass probabilities. Priors closer to one $\pi_{gf} \approx 1$ describes FIG by a Gaussian with similar variance than the one in the mixture distribution. Opposite, when prior is closer to zero $\pi_{gf} \approx 0$, the discrete component of mixture distribution dominates and its corresponding FIG is a narrowed Gaussian with mean zero.

Consider $\mathbf{Y} = [\mathbf{y}_1, \ldots \mathbf{y}_N]$ microarray data set where $\mathbf{y}_n = [y_{1n}, \ldots y_{Gn}]$ is the expression profile for the n-th sample with G genes. Similarly, $\mathbf{X} = [\mathbf{x}_1, \ldots \mathbf{x}_N]$ TF activities where $\mathbf{x}_n = [x_{1n}, \ldots x_{fn}]$ is the protein profile for the n-th sample

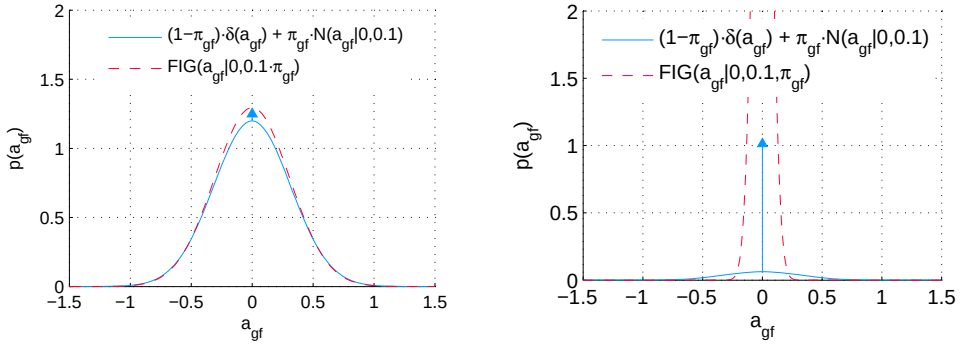


Figure 8.6: Slab and spiked distribution and its proposed FIG approximation with variance $\sigma_f^2 = 0.1$ and priors: (a) $\pi_{gf} = 0.95$ and (b) $\pi_{gf} = 0.05$.

with F transcription factors. Proposed factor model, assuming independent and identically distributed white noise $[\mathbf{E}]_{gn} = e_{gn}, \forall g, n$, may be expressed as

$$\mathbf{Y} = \mathbf{AX} + \mathbf{E} \tag{8.16}$$

Such kind of factor model, even with the sparsity constrains imposed by priors, has a large number of unknowns that complicates inference from a computational point of view. Based on the high sparsity levels of loading matrix, a novel approach that reduces the number of unknowns to be estimated has been proposed. Given the basis $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_K] \in \mathbb{R}^{G \times K}$ with K orthogonal column vectors such as loading matrix may be approximated, in terms of information losses, as

$$\mathbf{A} \approx \mathbf{BC} \tag{8.17}$$

Therefore, a Basis Expansion Factor Model (BEFM) has been proposed as

$$\mathbf{Y} = \mathbf{BCX} + \mathbf{E} \tag{8.18}$$

with $\mathbf{C} \in \mathbb{R}^{K \times F}$ expansion coefficients. Taking into account relationship (8.17) and loadings distribution (8.15), prior distribution of coefficients $[\mathbf{C}]_{kf} = c_{kf}, \forall k, f$ may be expressed as another FIG as,

$$p\left(c_{kf}\right) = \mathcal{FIG}\left(c_{gf} | \, 0, \sigma_f^2, \mathbf{b}_k^\top \mathbf{D}_{\boldsymbol{\pi}_f} \mathbf{b}_k\right) \tag{8.19}$$

with Moore-Penrose pseudoinverse $\mathbf{B}^+ = \left(\mathbf{B}^\top \mathbf{B}\right)^{-1} \mathbf{B}^\top = \mathbf{B}^\top$ and $\mathbf{D}_{\boldsymbol{\pi}_f}$ a diagonal matrix with priors $\boldsymbol{\pi}_f = [\pi_{1f}, \ldots, \pi_{Gf}]^\top$. Despite knowing the zeros distribution of

loading matrix, finding the optimum basis is complicated. As a suboptimal solution, we propose to use wavelet decomposition whereby, for a complete restoration, is represented by a set G of orthogonal vectors. To achieve dimensionality reduction, only the $K = \frac{G}{2}$ vectors corresponding to the approximation coefficients of the Haar wavelet have been considered.

The rest of unknowns are expressed in terms of conjugate modeling in favor to a sampling solution. Specifically, TF activities and noise variance was modeled by a Normal scaled Inverse prior distribution with unknowns mean, variance, scale and shape hyperparameters as,

$$\begin{aligned} p\left(\mathbf{x}_n, \sigma_n^2\right) &= \mathcal{NIG}\left(\mathbf{x}_n, \sigma_n^2 \middle| \boldsymbol{\mu}_n, \kappa_n, \alpha_n, \beta_n\right) \\ &= \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_n, \frac{\sigma_n^2}{\kappa_n} \mathbb{1}^F\right) \mathcal{IG}\left(\sigma_n^2 \middle| \alpha_n, \beta_n\right) \end{aligned} \tag{8.20}$$

and the variance of each factor is hyperparametrized by an Inverse Gamma distribution as,

$$p\left(\sigma_f^2\right) = \mathcal{IG}\left(\sigma_f^2 \middle| \alpha_f, \beta_f\right). \tag{8.21}$$

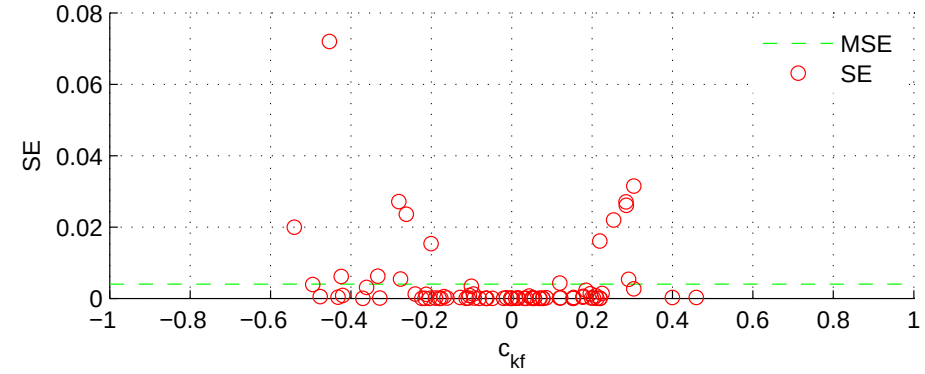
Unknowns $\{\mathbf{C}, \mathbf{X}\}$ have been inferred by a Gibbs sampler approach. Given a microarray data set $\mathbf{Y}$, priors $\mathbf{\Pi}$, basis $\mathbf{B}$ and samples at the immediately previous step $\left\{\hat{\mathbf{C}}^{(t)}, \hat{\mathbf{X}}^{(t)}\right\}$, new samples were taken from posteriors as

$$\hat{\mathbf{c}}_f^{(t+1)} \sim p\left(\mathbf{c}_f \middle| \mathbf{Y}, \left[\hat{\mathbf{c}}_1^{(t+1)}, \ldots, \hat{\mathbf{c}}_{f-1}^{(t+1)}, \hat{\mathbf{c}}_{f+1}^{(t)}, \ldots, \hat{\mathbf{f}}_F^{(t)}\right], \hat{\mathbf{X}}^{(t)}\right), \forall f \tag{8.22}$$

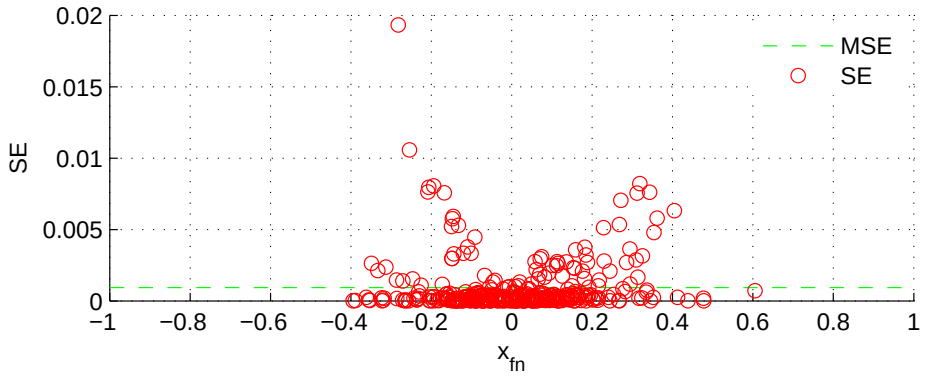
$$\hat{\mathbf{x}}_n^{(t+1)} \sim p\left(\mathbf{x}_n \middle| \mathbf{Y}, \left[\hat{\mathbf{x}}_1^{(t+1)}, \ldots, \hat{\mathbf{x}}_{n-1}^{(t+1)}, \hat{\mathbf{x}}_{n+1}^{(t)}, \ldots, \hat{\mathbf{x}}_N^{(t)}\right], \hat{\mathbf{C}}^{(t+1)}\right), \forall n \tag{8.23}$$

Additionally, to avoid sign ambiguity on estimations, a sign flipping is proposed. If it is known whether the f-th TF is a product of the g-th gene, the sign of protein activities x_{fn} will be flipped according to the sign of gene expression y_{gn}. If not, gene with the highest prior π_{gf} will be taken as reference.

BEFM has been validated by simulation. Synthetic data sets have been generated according to priors (8.19), (8.20) and (8.21) with different settings and levels of noise. Sampling and convergence test has been performed as in [34], simulating up to $T = 10000$ samples with $\frac{T}{2}$ burning period. Figure 8.7 shows the performance of BEFM for a synthetic data set with $G = 42$ genes, $N = 50$ samples, $F = 7$ TF and $\sigma_n^2 = \frac{1}{10}$. TRN structure has been generated randomly with a 30% of non-zero elements. Normalized results has been plotted for each unknown in terms of squared errors (SE). Additionally, it has been plotted the averaged mean squared error (MSE) for the whole estimation. Results shown an excellent performance of BEFM, able to capture the regulatory process and estimating the protein profiles.



(a)



(b)

Figure 8.7: Performance of BEFM method for a synthetic data set with with $G = 42$ genes, $N = 50$ samples, $F = 7$ TF, 30% of sparsity level and noise with $\sigma_n^2 = \frac{1}{10}$. Squared errors (SE) are represented versus the true value for each coefficient. The overall performance is represented by the mean squared error (MSE). (a) Results for (non-zero) coefficients $\mathbf{C}$ with MSE = 0.00403. (b) Results for TF activities $\mathbf{X}$ with MSE = 0.0093.

Additionally, BEFM has been analyzed using real data set. Specifically, it has been considered breast cancer data. Breast cancer is an histopathology heterogeneous disease with different prognosis between patients under same treatment and with apparently same clinical history. Previous microarray-based classification dissects breast cancer into different intrinsic subtypes with relevant implications in diagnosis and treatment design [65] [86]. Specifically, it has been considered microarray data with $G = 55$ genes, the ones considered for training the PAM50 classification test, and $N = 61$ samples of two different subtypes [62]. Moreover, it has been taken into account $F = 17$ factors, suggested by experts as potential proteins involved in breast cancer. Priors were estimated using TransFact and MATCH tool to predict binding sites along the 2kbp promoter region of each gene included in the data set. Figure 8.8 shows a clustergram with the inferred protein profiles, with TF as rows and samples as columns, labeled with the known intrinsic subtype. It is shown how estimated activities dissects breast cancer between Basal, the most aggressive subtype with worst prognosis, and HER2. Moreover, protein E2F1 is suggested as a potential TF with a differential behavior in Basal and HER2 subtypes.

Future work

Microarray experiments are versatile tools allowing to quantify eficiently gene expression. Abstracts models learned from microarray data constitutes a valuable source of biological knowledge. AR1MA1-VBEM method fits better the nature of microarray time series, handling properly experimental and inherent noise of biological process, being able to estimate the GRN. Moreover this method is based on a variational Bayesian approach, reducing the computational complexity of inference process. Results on in-silico data shows an acceptable performance, better than other well accepted methods. Moreover, it application to real data reveal the ability of the method to estimating well known genetic relationships. On the other hand, BEFM deals with a more realistic model fitting microarray data in a transcriptional space. This approach eficiently models sparsity at same time it reduces computational complexity. Results on real data shows a promising application of this method to disease classification. However, new applications of microarray techniques are generating new kind of data, quantifying different stages of the regulatory process. A more general analysis is needed for integrating different sources of biological knowledge leading to more precise results.

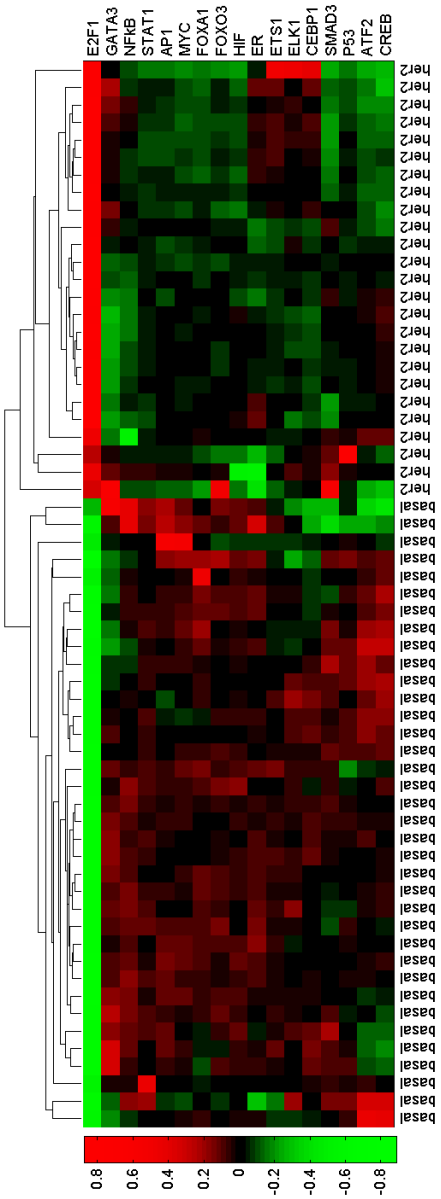


Figure 8.8: Clustergram with inferred protein activities of $F = 17$ TF and breast cancer data with $G = 55$ genes and $N = 61$ time samples. Protein E2F1 is suggested as a potential TF with a clear differential behavior in Basal and HER2 subtypes.

Appendix A

Derivation of the VBEM updating rules

Variational Bayesian method considers a lower bound of the marginal likelihood, as in (3.35), depending on a free distribution that factorizes into hidden variables and parameters. Optimization of this lower bound is performed by the Expectation Maximization method, in which one of the free distributions is optimized while the other is fixed. At the VBE step, the hidden variables free distribution is optimized whilst the parameters one is fixed. Therefore, after the t-th iteration, the lower bound to be optimized will be

$$\begin{aligned} & \mathcal{F}\left[q\left(x_i\right), q\left(\boldsymbol{\theta}\middle|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\right] \\ = & \int q\left(x_i\right)\left(\int q\left(\boldsymbol{\theta}\middle|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\ln\frac{p\left(\mathbf{y}_i\middle|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)}{q\left(x_i\right)}d\boldsymbol{\theta}\right)dx_i \\ + & \int q\left(x_i\right)\left(\int q\left(\boldsymbol{\theta}\middle|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\ln\frac{p\left(\boldsymbol{\theta}\right)}{q\left(\boldsymbol{\theta}\middle|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)}d\boldsymbol{\theta}\right)dx_i \\ = & \int F_{x_i}\left[q\left(x_i\right)\right]dx_i + \varphi\left[q\left(\boldsymbol{\theta}\middle|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\right] \end{aligned} \tag{A.1}$$

with $\varphi\left[q\left(\boldsymbol{\theta}\middle|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\right]$ constant and

$$F_{x_i}\left[q\left(x_i\right)\right] = q\left(x_i\right)\int q\left(\boldsymbol{\theta}\middle|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\ln\frac{p\left(\mathbf{y}_i\middle|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)}{q\left(x_i\right)}d\boldsymbol{\theta}. \tag{A.2}$$

Functional in (A.1) just depends on (A.2) and not on any of its derivatives. Therefore, extremal $q\left(x_i\middle|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)$ that maximizes this functional may be obtained,

in virtue of Euler-Lagrange theorem, equating the first derivative to zero as

$$\begin{aligned}
& \frac{\partial}{\partial q\left(x_i\right)}\left(F_{x_i}\left[q\left(x_i\right)\right]\right) \\
= & \frac{\partial}{\partial q\left(x_i\right)}\left(q\left(x_i\right)\int q\left(\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\ln\frac{p\left(\mathbf{y}_i|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)}{q\left(x_i\right)}d\boldsymbol{\theta}\right) \\
= & \int q\left(\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\ln\frac{p\left(\mathbf{y}_i|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)}{q\left(x_i\right)}d\boldsymbol{\theta} \\
= & \int q\left(\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\left(\ln p\left(\mathbf{y}_i|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)-\ln q\left(x_i\right)\right)d\boldsymbol{\theta} \\
+ & q\left(x_i\right)\int q\left(\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\frac{\partial}{\partial q\left(x_i\right)}\left(\ln\frac{p\left(\mathbf{y}_i|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)}{q\left(x_i\right)}\right)d\boldsymbol{\theta} \\
= & \int q\left(\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\ln p\left(\mathbf{y}_i|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)d\boldsymbol{\theta}-\ln q\left(x_i\right) \\
- & q\left(x_i\right)\int\frac{q\left(\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)}{q\left(x_i\right)}d\boldsymbol{\theta} \\
= & \int q\left(\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\ln p\left(\mathbf{y}_i|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)d\boldsymbol{\theta}-\ln q\left(x_i\right)-1=0. \qquad \text{(A.3)}
\end{aligned}$$

Solving and introducing λ_{x_i} as normalization constant results in

$$q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)=\lambda_{x_i}\mathrm{e}^{\int q\left(\boldsymbol{\theta}|\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)\ln p(\mathbf{y}_i|x_i,\boldsymbol{\theta})p(x_i)d\boldsymbol{\theta}-1} \qquad \text{(A.4)}$$

with

$$\ln\lambda_{i,x_i}=1-\int q\left(x_i|\,\hat{\boldsymbol{\xi}}_{x_i}^{(t+1)}\right)dx_i. \qquad \text{(A.5)}$$

Finally, using brackets to denote expected values and living out multiplicative constants, the learning rule of VBE step may be expressed as

$$q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)\propto\mathrm{e}^{\left\langle\ln p(\mathbf{y}_i|x_i,\boldsymbol{\theta})\right\rangle_{q\left(\boldsymbol{\theta}|\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t)}\right)}+\ln p(x_i)}. \qquad \text{(A.6)}$$

Similarly, the parameter's free distribution is optimized whilst the hidden variables are fixed at the VBM step. In this case, the lower bound to be optimized

will be

$$\begin{aligned}
& \mathcal{F}\left[q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right), q\left(\boldsymbol{\theta}\right)\right] \\
= & \int q\left(\boldsymbol{\theta}\right)\left(\int q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\ln\frac{p\left(\mathbf{y}_i|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)p\left(\boldsymbol{\theta}\right)}{q\left(\boldsymbol{\theta}\right)}dx_i\right)d\boldsymbol{\theta} \\
- & \int q\left(\boldsymbol{\theta}\right)\left(\int q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\ln q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)dx_i\right)d\boldsymbol{\theta} \\
= & \int F_{\boldsymbol{\theta}}\left[q\left(\boldsymbol{\theta}\right)\right]d\boldsymbol{\theta} + \phi\left[q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\right] \qquad \text{(A.7)}
\end{aligned}$$

with $\phi\left[q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\right]$ constant and

$$F_{\boldsymbol{\theta}}\left[q\left(\boldsymbol{\theta}\right)\right] = q\left(\boldsymbol{\theta}\right)\int q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\ln\frac{p\left(\mathbf{y}_i|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)p\left(\boldsymbol{\theta}\right)}{q\left(\boldsymbol{\theta}\right)}dx_i \qquad \text{(A.8)}$$

Functional in (A.7) just depends on (A.8) and not on any of its derivatives. Therefore, extremal $q\left(\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t+1)}\right)$ that maximizes this functional may be obtained, in virtue of Euler-Lagrange theorem, equating the first derivative to zero as

$$\begin{aligned}
& \frac{\partial}{\partial q\left(\boldsymbol{\theta}\right)}\left(F_{\boldsymbol{\theta}}\left[q\left(\boldsymbol{\theta}\right)\right]\right) \\
= & \frac{\partial}{\partial q\left(\boldsymbol{\theta}\right)}\left(q\left(\boldsymbol{\theta}\right)\int q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\ln p\left(\mathbf{y}_i|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)dx_i\right) \\
+ & \frac{\partial}{\partial q\left(\boldsymbol{\theta}\right)}\left(q\left(\boldsymbol{\theta}\right)\int q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\ln\frac{p\left(\boldsymbol{\theta}\right)}{q\left(\boldsymbol{\theta}\right)}dx_i\right) \\
= & \int q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\ln p\left(\mathbf{y}_i|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right)dx_i \\
+ & \int q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\ln\frac{p\left(\boldsymbol{\theta}\right)}{q\left(\boldsymbol{\theta}\right)}dx_i \\
+ & q\left(\boldsymbol{\theta}\right)\int q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\frac{\partial}{\partial q\left(\boldsymbol{\theta}\right)}\left(\ln\frac{p\left(\boldsymbol{\theta}\right)}{q\left(\boldsymbol{\theta}\right)}\right)dx_i \\
= & \int q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\left(\ln p\left(\mathbf{y}_i|\,x_i,\boldsymbol{\theta}\right)p\left(x_i\right) + \ln p\left(\boldsymbol{\theta}\right)\right)dx_i \\
- & \ln q\left(\boldsymbol{\theta}\right) - q\left(\boldsymbol{\theta}\right)\int\frac{q\left(x_i|\,\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)}{q\left(\boldsymbol{\theta}\right)}dx_i = 0. \qquad \text{(A.9)}
\end{aligned}$$

Solving and introducing $\lambda_{\boldsymbol{\theta}}$ as normalization constant results in

$$q\left(\boldsymbol{\theta}|\,\hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t+1)}\right) = \lambda_{\boldsymbol{\theta}}\mathrm{e}^{\int q\left(x_i|\hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)\left(\ln p\left(\mathbf{y}_i|x_i,\boldsymbol{\theta}\right)p\left(x_i\right)+\ln p\left(\boldsymbol{\theta}\right)\right)dx_i-1} \qquad \text{(A.10)}$$

with

$$\ln \lambda_{\boldsymbol{\theta}} = 1 - \int q\left(\boldsymbol{\theta} | \, \hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t+1)}\right) d\boldsymbol{\theta}. \tag{A.11}$$

Once again, living out superfluous terms, the VBM updating rule may be expressed as

$$q\left(\boldsymbol{\theta} | \, \hat{\boldsymbol{\xi}}_{i,\boldsymbol{\theta}}^{(t+1)}\right) \quad \propto \quad \mathrm{e}^{\left\langle \ln p(\,\mathbf{y}_i | x_i, \boldsymbol{\theta}) p(x_i) \right\rangle_{q\left(\,x_i | \hat{\boldsymbol{\xi}}_{i,x_i}^{(t+1)}\right)} + \ln p(\boldsymbol{\theta})}. \tag{A.12}$$

Appendix B

Derivation of Gaussian likelihood for the AR1MA1 model

Linear models as the AR one introduced in (4.4) or the AR1MA1 presented in (4.9) includes an IID white noise term. By definition, white noise is modeled as a Gaussian distribution with zero mean and unknown variance as in (4.12). Therefore, likelihood will be a multivariate Gaussian distribution as

$$p\left(\mathbf{y}_i|\,\mathbf{x}_i,\boldsymbol{\theta}_i\right)=\mathcal{N}\left(\mathbf{y}_i|\,\boldsymbol{\mu}_{\mathbf{y}_i},\boldsymbol{\Sigma}_{\mathbf{y}_i}\right) \tag{B.1}$$

with mean and variance depending on the unknowns.

For the likelihood function of the AR1MA1 model, the mean may be computed as

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{y}_i} &= \langle\mathbf{y}_i\rangle_{p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)}=\left\langle\mathbf{T}\mathbf{Y}^\top\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i\right\rangle_{p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)}\\ &+ \cancelto{0}{\left\langle\mathbf{T}\mathbf{E}^\top\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i\right\rangle_{p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)}} + \cancelto{0}{\langle\mathbf{e}_i\rangle_{p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)}}\\ &= \mathbf{R}\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i\end{aligned} \tag{B.2}$$

with $\mathbf{R}=\mathbf{T}\mathbf{Y}^\top$.

On the other hand, covariance matrix may be computed as

$$
\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{y}_i} &= \left\langle (\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i})(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i})^\top \right\rangle_{p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)} \\
&= \left\langle \left(\mathbf{T}\mathbf{E}^\top \mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i + \mathbf{e}_i\right)\left(\mathbf{T}\mathbf{E}^\top \mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i + \mathbf{e}_i\right)^\top \right\rangle_{p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)} \\
&= \left\langle \mathbf{T}\mathbf{E}^\top \mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i\mathbf{x}_i^\top \mathbf{D}_{\boldsymbol{\omega}_i}^\top \mathbf{E}\mathbf{T}^\top \right\rangle_{p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)} \\
&+ \cancelto{0}{\left\langle \mathbf{T}\mathbf{E}^\top \mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i\mathbf{e}_i^\top \right\rangle_{p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)}} + \left\langle \mathbf{e}_i\mathbf{e}_i^\top \right\rangle_{p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)} \\
&= \left\langle \mathbf{T}\mathbf{E}^\top \left(\mathbf{T}\mathbf{E}^\top\right)^\top \right\rangle_{p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)} \left(\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i\mathbf{x}_i^\top \mathbf{D}_{\boldsymbol{\omega}_i}^\top\right)^\top + \sigma_i^2 \mathbb{1}^N \\
&= \sigma_i^2 \sum_{j=1}^{G} x_i^2(j)\omega_i^2(j)\mathbb{1}^N + \sigma_i^2 \mathbb{1}^N \\
&= \sigma_i^2 \left(1 + \sum_{j=1}^{G} x_i^2(j)\omega_i^2(j)\right)\mathbb{1}^N \\
&= \sigma_i^2 \left(1 + \mathbf{x}_i^\top \mathbf{D}_{\mathbf{x}_i}\mathbf{D}_{\boldsymbol{\omega}_i}\boldsymbol{\omega}_i\right)\mathbb{1}^N
\end{aligned}
\tag{B.3}
$$

Appendix C

Subjective hyperparameters

Priors of hidden variables and parameters have been chosen to define a conjugate model as (4.17) and (8.4). Despite of hidden variables are defined as discrete binary variables, in favor to the conjugate modeling, a multivariate Gaussian distribution with unknown mean and variance was considered as

$$p\left(\mathbf{x}_i\right) = \mathcal{N}\left(\mathbf{x}_i \middle| \, \mathbf{m}_{\mathbf{x}_i}, \sigma_0^2 \mathbb{1}^G\right) \tag{C.1}$$

When no prior information of the GRN is available, subjective hyperparameters may be chosen. Assuming that both states have same probability a priori, i.e. $p\left(x_i(j)=0\right) = p\left(x_i(j)=1\right) = \frac{1}{2}$, the mean may be computed as

$$\begin{aligned} m_{\mathbf{x}_i}(j) &= \left\langle x_i(j) \right\rangle_{p(x_i(j))} \\ &= \sum_{x_i(j)}^{\{0,1\}} x_i(j) p\left(x_i(j)\right) = \frac{1}{2}, \forall j \end{aligned} \tag{C.2}$$

Additionally, variance may be computed as

$$\begin{aligned} \sigma_0^2 &= \left\langle x_i^2(j) - m_{\mathbf{x}_i}^2(j) \right\rangle \\ &= \sum_{x_i(j)}^{\{0,1\}} x_i^2(j) p\left(x_i(j)\right) - \frac{1}{4} = \frac{1}{4} \end{aligned} \tag{C.3}$$

On the other hand, given the variance of noise σ_i^2, weights describing regulatory type are also modeled by a multivariate Gaussian as in (4.20) with unknown mean and scale of variance as

$$p\left(\boldsymbol{\omega}_i \middle| \, \sigma_i^2\right) = \mathcal{N}\left(\boldsymbol{\omega}_i \middle| \, \mathbf{m}_{\boldsymbol{\omega}_i}, \frac{\sigma_i^2}{\bar{\gamma}_i} \mathbb{1}^G\right) \tag{C.4}$$

with $\bar{\gamma}_i$ scale of variance that is approximated as (4.33). Opposite to $x_i(j)$, $\omega_i(j)$ is a continuous parameter describing the regulatory strength. Moreover, weights specifies the regulatory type, with negative or positive values, for inhibition or activation respectively. Assuming that any state is possible a priori $p\left(\omega_i(j) \in \mathbb{R}^+\right) = p\left(\omega_i(j) \in \mathbb{R}^-\right)$, a subjective mean may be set up as

$$\begin{aligned} m_{\boldsymbol{\omega}_i}(j) &= \langle \omega_i(j) \rangle_{p(\omega_i(j))} \\ &= \int_{-\infty}^{0} \omega_i(j) p\left(\omega_i(j)\right) + \int_{0}^{+\infty} \omega_i(j) p\left(\omega_i(j)\right) = 0, \forall j \end{aligned} \tag{C.5}$$

According to priors hyperparameters in (C.2) and (C.5), a subjective approach of the scale of variance $\bar{\gamma}_i$ as in (4.33) should be

$$\bar{\gamma}_i \approx \gamma_i \left(\mathbf{m}_{\mathbf{x}_i}, \mathbf{m}_{\boldsymbol{\omega}_i}\right) = 1. \tag{C.6}$$

Finally, variance of noise is modeled by an Inverse Gamma distribution as in (4.21) with unknown shape and scale as

$$p\left(\sigma_i^2\right) = \mathcal{IG}\left(\sigma_i^2 \middle| a_i, \bar{\gamma}_i b_i\right). \tag{C.7}$$

By definition, the variance $\sigma_i^2 \in (0, +\infty)$ is a continuous parameter with positive values. A priori, all states should be intended to be equally-probable. How to set these hyperparameters subjectively is discussed in the bibliography with insufficient success [35]. The most extended set up considers both hyperperameters or just the scale one lower than one $b << 1$ [34]. However, this settings concentrates the probability in a narrow range of σ_i^2, leaving out other possible values. Alternatively, we are going to consider a setting with

$$a = 2 \tag{C.8}$$

$$b = \frac{1}{a} \tag{C.9}$$

Figure C.1 compares different settings with the one porposed above. By fixing $a = 2$, the second moment of σ_i^2 is not defined and the tail of the distribution will be as flat as possible. Moreover, with $b = \frac{1}{2}$ the scale parameters is not large or small enough to discriminate variances of noise closer or higher than zero.

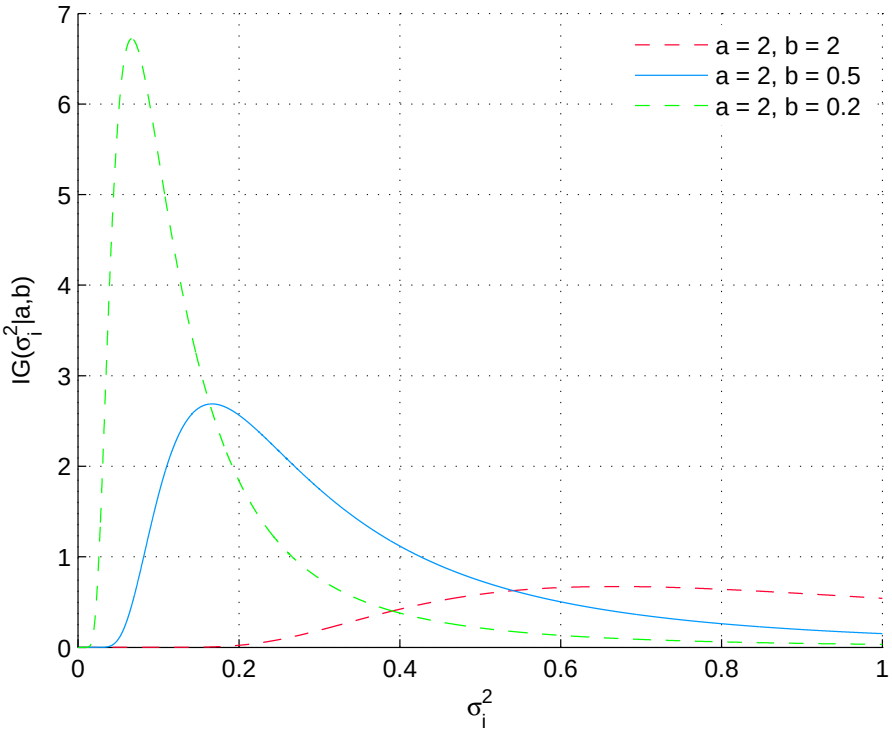


Figure C.1: By fixing $a = 2$, the second moment of σ_i^2 is not defined and the tail of the distribution will be as flat as possible. Moreover, with $b = \frac{1}{2}$ the scale parameters is not large or small enough to discriminate variances of noise closer or higher than zero.

Appendix D

Expected values

The VBE updating rule in (4.36) requires to compute the expected value of the log-likelihood as

$$
\begin{aligned}
& \left\langle \ln p\left(\mathbf{y}_i|\,\mathbf{x}_i,\boldsymbol{\theta}_i\right)\right\rangle_{q\left(\boldsymbol{\theta}_i|\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t)}\right)} \\
= & -\frac{N}{2}\ln 2\pi+\frac{N}{2}\ln\bar{\gamma}_i+\frac{N}{2}\left\langle \ln\sigma_i^2\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t)}\right)} \\
- & \frac{\bar{\gamma}_i}{2}\left\langle \frac{1}{\sigma_i^2}\left\langle \left(\mathbf{y}_i-\mathbf{R}\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i\right)^\top\left(\mathbf{y}_i-\mathbf{R}\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i\right)\right\rangle_{q\left(\boldsymbol{\omega}_i|\sigma_i^2,\hat{\boldsymbol{\xi}}_{\boldsymbol{\omega}_i}^{(t)}\right)}\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t)}\right)} \\
= & -\frac{N}{2}\ln 2\pi+\frac{N}{2}\ln\bar{\gamma}_i-\frac{N}{2}\left\langle \ln\sigma_i^2\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t)}\right)}-\frac{\bar{\gamma}_i}{2}\mathbf{y}_i^\top\mathbf{y}_i\left\langle \frac{1}{\sigma_i^2}\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t)}\right)} \\
+ & \bar{\gamma}_i\mathbf{y}_i^\top\mathbf{R}\left\langle \frac{1}{\sigma_i^2}\left\langle \mathbf{D}_{\boldsymbol{\omega}_i}\right\rangle_{q\left(\boldsymbol{\omega}_i|\sigma_i^2,\hat{\boldsymbol{\xi}}_{\boldsymbol{\omega}_i}^{(t)}\right)}\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t)}\right)}\mathbf{x}_i \\
- & \frac{\bar{\gamma}_i}{2}\mathbf{x}_i^\top\left\langle \frac{1}{\sigma_i^2}\left\langle \mathbf{D}_{\boldsymbol{\omega}_i}^\top\mathbf{B}\mathbf{D}_{\boldsymbol{\omega}_i}\right\rangle_{q\left(\boldsymbol{\omega}_i|\sigma_i^2,\hat{\boldsymbol{\xi}}_{\boldsymbol{\omega}_i}^{(t)}\right)}\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t)}\right)}\mathbf{x}_i \\
= & -\frac{N}{2}\ln 2\pi+\frac{N}{2}\ln\bar{\gamma}_i-\frac{N}{2}\left(\ln\hat{\beta}_i^{(t)}-\Psi\left(\hat{\alpha}_i^{(t)}\right)\right)-\frac{\bar{\gamma}_i}{2}\mathbf{y}_i^\top\mathbf{y}_i\frac{\hat{\alpha}^{(t)}}{\hat{\beta}^{(t)}} \\
+ & \bar{\gamma}_i\mathbf{y}_i^\top\mathbf{R}\left\langle \frac{1}{\sigma_i^2}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)\top}}\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t)}\right)}\mathbf{x}_i \\
- & \frac{\bar{\gamma}_i}{2}\mathbf{x}_i^\top\left\langle \frac{1}{\sigma_i^2}\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)\top}+\sigma_i^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t)}\right)\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t)}\right)}\mathbf{x}_i
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\bar{\gamma}_i - \frac{N}{2}\left(\ln\hat{\beta}_i^{(t)} - \Psi\left(\hat{\alpha}_i^{(t)}\right)\right) - \frac{\bar{\gamma}_i}{2}\mathbf{y}_i^\top\mathbf{y}_i\frac{\hat{\alpha}^{(t)}}{\hat{\beta}^{(t)}} \\
&+ \bar{\gamma}_i\mathbf{y}_i^\top\mathbf{R}\frac{\hat{\alpha}^{(t)}}{\hat{\beta}^{(t)}}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}}\mathbf{x}_i - \frac{1}{2}\mathbf{x}_i^\top\bar{\gamma}_i\mathbf{B}\circ\left(\frac{\hat{\alpha}^{(t)}}{\hat{\beta}^{(t)}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)\top} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t)}\right)\mathbf{x}_i \qquad \text{(D.1)}
\end{aligned}
$$

On the other hand, VBM updating rule in (4.37) requires to compute the expected value of the logarithm of the likelihood times the hidden variables prior as

$$
\begin{aligned}
&\left\langle \ln p\left(\mathbf{y}_i|\,\mathbf{x}_i,\boldsymbol{\theta}_i\right)p\left(\mathbf{x}_i\right)\right\rangle_{q\left(\mathbf{x}_i|\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)} \\
&= -\frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\bar{\gamma}_i - \frac{N}{2}\ln\sigma_i^2 - \frac{G}{2}\ln 2\pi - \frac{G}{2}\ln\sigma_0^2 \\
&- \frac{\bar{\gamma}_i}{2\sigma_i^2}\left\langle\left(\mathbf{y}_i - \mathbf{R}\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i\right)^\top\left(\mathbf{y}_i - \mathbf{R}\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i\right)\right\rangle_{q\left(\mathbf{x}_i|\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)} \\
&- \frac{1}{2\sigma_0^2}\left\langle\left(\mathbf{x}_i - \mathbf{m}_{\mathbf{x}_i}\right)^\top\left(\mathbf{x}_i - \mathbf{m}_{\mathbf{x}_i}\right)\right\rangle_{q\left(\mathbf{x}_i|\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)} \\
&= -\frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\bar{\gamma}_i - \frac{N}{2}\ln\sigma_i^2 - \frac{G}{2}\ln 2\pi - \frac{G}{2}\ln\sigma_0^2 \\
&- \frac{\bar{\gamma}_i}{2\sigma_i^2}\mathbf{y}_i^\top\mathbf{y}_i + \frac{\bar{\gamma}_i}{\sigma_i^2}\mathbf{y}_i^\top\mathbf{R}\left\langle\mathbf{D}_{\mathbf{x}_i}\right\rangle_{q\left(\mathbf{x}_i|\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)}\boldsymbol{\omega}_i \\
&- \frac{\bar{\gamma}_i}{2\sigma_i^2}\boldsymbol{\omega}_i^\top\left\langle\mathbf{D}_{\mathbf{x}_i}^\top\mathbf{B}\mathbf{D}_{\mathbf{x}_i}\right\rangle_{q\left(\mathbf{x}_i|\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)}\boldsymbol{\omega}_i \\
&- \frac{1}{2\sigma_0^2}\left(\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right)^\top\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right) + \operatorname{trace}\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right) \\
&= -\frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\bar{\gamma}_i - \frac{N}{2}\ln\sigma_i^2 - \frac{G}{2}\ln 2\pi - \frac{G}{2}\ln\sigma_0^2 \\
&- \frac{\bar{\gamma}_i}{2\sigma_i^2}\mathbf{y}_i^\top\mathbf{y}_i + \frac{\bar{\gamma}_i}{\sigma_i^2}\mathbf{y}_i^\top\mathbf{R}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}}\boldsymbol{\omega}_i \\
&- \frac{\bar{\gamma}_i}{2\sigma_i^2}\boldsymbol{\omega}_i^\top\left(\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right)\boldsymbol{\omega}_i \\
&- \frac{1}{2\sigma_0^2}\left(\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right)^\top\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right) + \operatorname{trace}\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right) \qquad \text{(D.2)}
\end{aligned}
$$

and following for the expectation over all the unknowns

$$
\begin{aligned}
& \left\langle \left\langle \ln p\left(\mathbf{y}_i|\,\mathbf{x}_i,\boldsymbol{\theta}_i\right) p\left(\mathbf{x}_i\right)\right\rangle_{q\left(\mathbf{x}_i|\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)} \right\rangle_{q\left(\boldsymbol{\theta}_i|\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)} \\
= & -\frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\bar{\gamma}_i - \frac{N}{2}\left\langle \ln\sigma_i^2\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)} - \frac{G}{2}\ln 2\pi - \frac{G}{2}\ln\sigma_0^2 \\
- & \left\langle \frac{\bar{\gamma}_i}{2\sigma_i^2}\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)} \mathbf{y}_i^\top\mathbf{y}_i + \left\langle \frac{\bar{\gamma}_i}{\sigma_i^2}\mathbf{y}_i^\top\mathbf{R}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}}\left\langle\boldsymbol{\omega}_i\right\rangle_{q\left(\boldsymbol{\omega}_i|\hat{\boldsymbol{\xi}}_{\boldsymbol{\omega}_i}^{(t+1)}\right)}\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)} \\
- & \left\langle \frac{\bar{\gamma}_i}{2\sigma_i^2}\left\langle \boldsymbol{\omega}_i^\top\left(\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right)\boldsymbol{\omega}_i\right\rangle_{q\left(\boldsymbol{\omega}_i|\hat{\boldsymbol{\xi}}_{\boldsymbol{\omega}_i}^{(t+1)}\right)}\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)} \\
- & \frac{1}{2\sigma_0^2}\left(\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right)^\top\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right) + \operatorname{trace}\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right) \\
= & -\frac{N+G}{2}\ln 2\pi + \frac{N}{2}\ln\bar{\gamma}_i - \frac{N}{2}\left(\ln\hat{\beta}_i^{(t+1)} - \Psi\left(\hat{\alpha}_i^{(t+1)}\right)\right) - \frac{G}{2}\ln\sigma_0^2 \\
- & \frac{\bar{\gamma}_i}{2}\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\mathbf{y}_i^\top\mathbf{y}_i + \bar{\gamma}_i\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\mathbf{y}_i^\top\mathbf{R}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} \\
- & \left\langle \frac{\bar{\gamma}_i}{2\cancel{\sigma_i^2}}\operatorname{trace}\left(\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\cancel{\sigma_i^2}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)}\right)\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)} \\
- & \left\langle \frac{\bar{\gamma}_i}{2\sigma_i^2}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{\top(t+1)}\left(\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right)\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)} \\
- & \frac{1}{2\sigma_0^2}\left(\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right)^\top\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right) + \operatorname{trace}\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right) \\
= & -\frac{N+G}{2}\ln 2\pi + \frac{N}{2}\ln\bar{\gamma}_i - \frac{N}{2}\left(\ln\hat{\beta}_i^{(t+1)} - \Psi\left(\hat{\alpha}_i^{(t+1)}\right)\right) - \frac{G}{2}\ln\sigma_0^2 \\
- & \frac{\bar{\gamma}_i}{2}\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\mathbf{y}_i^\top\mathbf{y}_i + \bar{\gamma}_i\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\mathbf{y}_i^\top\mathbf{R}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} \\
- & \frac{\bar{\gamma}_i}{2}\operatorname{trace}\left(\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)}\right) \\
- & \frac{\bar{\gamma}_i}{2}\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{\top(t+1)}\left(\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right)\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} \\
- & \frac{1}{2\sigma_0^2}\left(\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right)^\top\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right) + \operatorname{trace}\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right). \qquad \text{(D.3)}
\end{aligned}
$$

Additionally, computation of the lower bound requires the expected values of

following square forms

$$\begin{aligned}
&\left\langle \left(\boldsymbol{\omega}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}\right)^\top \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{-1(t+1)}}{\sigma_i^2} \left(\boldsymbol{\omega}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}\right) \right\rangle_{q\left(\boldsymbol{\theta}_i | \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)} \\
= &\left\langle \left(\cancel{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}} - \cancel{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}}\right)^\top \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{-1(t+1)}}{\sigma_i^2} \left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}\right) \right\rangle_{q\left(\sigma_i^2 | \hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)} \\
+ &\operatorname{trace}\left(\left(\cancel{\sigma_i^2 \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)}}\right)^{-1} \cancel{\sigma_i^2 \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)}} \right) = \operatorname{trace}\left(\mathbb{1}^G\right) = G
\end{aligned} \tag{D.4}$$

and

$$\begin{aligned}
&\left\langle \frac{\bar{\gamma}_i}{\sigma_i^2} \left(\boldsymbol{\omega}_i - \mathbf{m}_{\boldsymbol{\omega}_i}\right)^\top \left(\boldsymbol{\omega}_i - \mathbf{m}_{\boldsymbol{\omega}_i}\right) \right\rangle_{q\left(\boldsymbol{\theta}_i | \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)} \\
= &\left\langle \left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} - \mathbf{m}_{\boldsymbol{\omega}_i}\right)^\top \frac{\bar{\gamma}_i}{\sigma_i^2} \left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} - \mathbf{m}_{\boldsymbol{\omega}_i}\right) \right\rangle_{q\left(\sigma_i^2 | \hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)} + \operatorname{trace}\left(\frac{\bar{\gamma}_i}{\cancel{\sigma_i^2}} \cancel{\sigma_i^2} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)} \right) \\
= &\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} - \mathbf{m}_{\boldsymbol{\omega}_i}\right)^\top \bar{\gamma}_i \frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}} \left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} - \mathbf{m}_{\boldsymbol{\omega}_i}\right) + \operatorname{trace}\left(\bar{\gamma}_i \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)}\right)
\end{aligned} \tag{D.5}$$

Appendix E

VBEM updating rules

E.1 VBE learning rules

According to conjugate model properties, the optimized free distribution of $\mathbf{x}_i$ as in (4.22) at the $(t+1)$-th iteration results in a multivariate Gaussian as

$$\begin{aligned} q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right) &= \mathcal{N}\left(\mathbf{x}_i|\,\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}, \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right) \\ &\propto \mathrm{e}^{-\frac{1}{2}\left(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\right)^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{-1(t+1)}\left(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\right)} \\ &\propto \mathrm{e}^{-\frac{1}{2}\mathbf{x}_i^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{-1(t+1)}\mathbf{x}_i+\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{-1(t+1)}\mathbf{x}_i}. \end{aligned} \tag{E.1}$$

On the other hand, from expectation (D.1), the VBE updating rule in (4.36) may be conveniently expressed in terms of the hidden variables as

$$\begin{aligned} q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right) &\propto \mathrm{e}^{\left(\mathbf{y}_i^\top \mathbf{R}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}}\frac{\hat{\alpha}^{(t)}}{\hat{\beta}^{(t)}}\bar{\gamma}_i+\frac{1}{\sigma_0^2}\mathbf{m}_{\mathbf{x}_i}^\top\right)\mathbf{x}_i} \\ &\cdot \mathrm{e}^{-\frac{1}{2}\mathbf{x}_i^\top\left(\bar{\gamma}_i\mathbf{B}\circ\left(\frac{\hat{\alpha}^{(t)}}{\hat{\beta}^{(t)}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)\top}+\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)}\right)+\frac{1}{\sigma_0^2}\mathbb{1}^G\right)\mathbf{x}_i} \end{aligned} \tag{E.2}$$

with $\bar{\gamma}_i$ approximated scale of variance as in (4.33) given the means of $\mathbf{x}_i$ and $\boldsymbol{\theta}_i$ at the immediately previous step as

$$\bar{\gamma}_i = \gamma_i\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t)}, \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}\right) = \frac{1}{1+\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{\top(t)}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t)}}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t)}}. \tag{E.3}$$

Comparing free distribution in (E.2) with the VBE updating rule in (E.2), the

hyperparameters learning rules may be identified as

$$\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} = \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{\top(t+1)} \left(\mathbf{y}_i^\top \mathbf{R} \mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}} \frac{\hat{\alpha}_i^{(t)}}{\hat{\beta}_i^{(t)}} \bar{\gamma}_i + \frac{1}{\sigma_0^2} \mathbf{m}_{\mathbf{x}_i}^\top \right)^\top \tag{E.4}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)} = \left(\bar{\gamma}_i \mathbf{B} \circ \left(\frac{\hat{\alpha}^{(t)}}{\hat{\beta}^{(t)}} \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)} \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)\top} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)} \right) + \frac{1}{\sigma_0^2} \mathbb{1}^G \right)^{-1}. \tag{E.5}$$

E.2 VBM learning rules

After each VBE step, at the $(t+1)$-th iteration, the approximated scale of variance $\bar{\gamma}_i$ is updated as

$$\bar{\gamma}_i = \gamma_i \left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}, \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)} \right) = \frac{1}{1 + \hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} \mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}} \mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}} \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}}. \tag{E.6}$$

On the other hand, the optimized free distribution of $\boldsymbol{\theta}_i$ as in (4.37) results in a Normal scaled Inverse Gamma as

$$\begin{aligned} q\left(\boldsymbol{\theta}_i \middle| \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)} \right) &= \mathcal{N}\left(\boldsymbol{\omega}_i \middle| \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}, \sigma_i^2 \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)} \right) \mathcal{IG}\left(\sigma_i^2 \middle| \hat{\alpha}_i^{(t+1)}, \hat{\beta}_i^{(t+1)} \right) \\ &\propto \mathrm{e}^{-\frac{G}{2}\ln\sigma_i^2 - \frac{1}{2\sigma_i^2}\left(\boldsymbol{\omega}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}\right)^\top \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{-1(t+1)} \left(\boldsymbol{\omega}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}\right)} \\ &\cdot \mathrm{e}^{-\frac{\hat{\beta}_i^{(t+1)}}{\sigma_i^2} - (\hat{\alpha}_i^{(t+1)}+1)\ln\sigma_i^2} \\ &\propto \mathrm{e}^{-\frac{1}{2\sigma_i^2}\left(2\hat{\beta}_i^{(t+1)} + \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{\top(t+1)} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{-1(t+1)} \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}\right) - (\hat{\alpha}_i^{(t+1)}+1+\frac{G}{2})\ln\sigma_i^2} \\ &\cdot \mathrm{e}^{-\frac{1}{2\sigma_i^2}\boldsymbol{\omega}_i^\top \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{-1(t+1)} \boldsymbol{\omega}_i + \frac{1}{2\sigma_i^2} \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{\top(t+1)} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{-1(t+1)} \boldsymbol{\omega}_i}. \end{aligned} \tag{E.7}$$

Taking into account the parameters hyperprior (8.4), that may be expressed as

$$\begin{aligned} p\left(\boldsymbol{\theta}_i\right) &= \mathcal{N}\left(\boldsymbol{\omega}_i \middle| \mathbf{m}_{\boldsymbol{\omega}_i}, \frac{\sigma_i^2}{\bar{\gamma}_i} \mathbb{1}^G \right) \mathcal{IG}\left(\sigma_i^2 \middle| a_i, \bar{\gamma}_i b_i \right) \\ &\propto \mathrm{e}^{-\frac{G}{2}\ln\sigma_i^2 - \frac{\bar{\gamma}_i}{2\sigma_i^2}\left(\boldsymbol{\omega}_i - \mathbf{m}_{\boldsymbol{\omega}_i}\right)^\top \left(\boldsymbol{\omega}_i - \mathbf{m}_{\boldsymbol{\omega}_i}\right) - \frac{\bar{\gamma}_i b_i}{\sigma_i^2} - (a_i+1)\ln\sigma_i^2} \\ &\propto \mathrm{e}^{-\frac{\bar{\gamma}_i}{2\sigma_i^2}\boldsymbol{\omega}_i^\top \boldsymbol{\omega}_i + \frac{\bar{\gamma}_i}{2\sigma_i^2}\mathbf{m}_{\boldsymbol{\omega}_i}^\top \boldsymbol{\omega}_i - \frac{\bar{\gamma}_i}{2\sigma_i^2}\left(\mathbf{m}_{\boldsymbol{\omega}_i}^\top \mathbf{m}_{\boldsymbol{\omega}_i} + 2b_i\right) - (a_i+1+\frac{G}{2})\ln\sigma_i^2}. \end{aligned} \tag{E.8}$$

and from expectation in (D.2), the VBM updating rule in (4.37) may be conveniently expressed as

$$\begin{aligned} q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right) \;\propto\; & \mathrm{e}^{(a_i+1+\frac{N+G}{2})\ln\sigma_i^2} \\ & \cdot\; \mathrm{e}^{-\frac{1}{2\sigma_i^2}\left(\bar{\gamma}_i\mathbf{y}_i^\top\mathbf{y}_i+2\bar{\gamma}_i b_i+\bar{\gamma}_i\mathbf{m}_{\mathbf{x}_i}^\top\mathbf{m}_{\mathbf{x}_i}\right)} \\ & \cdot\; \mathrm{e}^{\frac{1}{\sigma_i^2}\left(\bar{\gamma}_i\mathbf{y}_i^\top\mathbf{R}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}}+\bar{\gamma}_i\mathbf{m}_{\mathbf{x}_i}^\top\right)\boldsymbol{\omega}_i} \\ & \cdot\; \mathrm{e}^{-\frac{1}{2\sigma_i^2}\boldsymbol{\omega}_i^\top\left(\bar{\gamma}_i\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)}+\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)+\bar{\gamma}_i\mathbb{1}_G\right)\boldsymbol{\omega}_i} \end{aligned} \tag{E.9}$$

Comparing free distribution in (E.7) with the VBM updating rule in (E.9), the hyperparameters learning rules may be identified as

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)\top}\left(\mathbf{y}_i^\top\mathbf{R}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}}\bar{\gamma}_i+\mathbf{m}_{\boldsymbol{\omega}_i}^\top\mathbf{S}_{\boldsymbol{\omega}_i}^{-1}\right)^\top \tag{E.10}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)} = \left(\mathbf{S}_{\boldsymbol{\omega}_i}^{-1}+\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)\top}+\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\bar{\gamma}_i\right)^{-1} \tag{E.11}$$

$$\hat{\beta}_i^{(t+1)} = b_i+\frac{1}{2}\left(\mathbf{y}_i^\top\mathbf{y}_i\bar{\gamma}_i+\mathbf{m}_{\boldsymbol{\omega}_i}^\top\mathbf{m}_{\boldsymbol{\omega}_i}+\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{\top(t+1)}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{-1(t+1)}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}\right) \tag{E.12}$$

$$\hat{\alpha}_i^{(t+1)} = a_i+\frac{N+G}{2}. \tag{E.13}$$

E.3 Lower bound updating rules

After each VBM step, at the $(t+1)$-th iteration, the approximated scale of variance $\bar{\gamma}_i$ is updated as

$$\bar{\gamma}_i = \gamma_i\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)},\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t)}\right) = \frac{1}{1+\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}}. \tag{E.14}$$

After to consecutive VBE and VBM steps, VBEM method requires the lower bound updating as

$$
\begin{aligned}
&\mathcal{F}\left[q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right), q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)\right] = \\
&\int q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right) q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right) \ln \frac{p\left(\mathbf{y}_i|\,\mathbf{x}_i,\boldsymbol{\theta}_i\right) p\left(\mathbf{x}_i\right) p\left(\boldsymbol{\theta}_i\right)}{q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right) q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)} d\mathbf{x}_i d\boldsymbol{\theta}_i \\
= & \int q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right) q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right) \ln p\left(\mathbf{y}_i|\,\mathbf{x}_i,\boldsymbol{\theta}_i\right) p\left(\mathbf{x}_i\right) d\mathbf{x}_i d\boldsymbol{\theta}_i \\
+ & \int q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right) q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right) \ln \frac{p\left(\boldsymbol{\theta}_i\right)}{q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)} d\mathbf{x}_i d\boldsymbol{\theta}_i \\
- & \int q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right) q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right) \ln q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right) d\mathbf{x}_i d\boldsymbol{\theta}_i \\
= & \left\langle \left\langle \ln p\left(\mathbf{y}_i|\,\mathbf{x}_i,\boldsymbol{\theta}_i\right) p\left(\mathbf{x}_i\right)\right\rangle_{q\left(\mathbf{x}_i|\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)} + \ln \frac{p\left(\boldsymbol{\theta}_i\right)}{q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)} \right\rangle_{q\left(\boldsymbol{\theta}_i|\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)} \\
- & \left\langle \ln q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)\right\rangle_{q\left(\mathbf{x}_i|\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)} \qquad \text{(E.15)}
\end{aligned}
$$

with

$$
\begin{aligned}
&\left\langle \ln \frac{p\left(\boldsymbol{\theta}_i\right)}{q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)} \right\rangle_{q\left(\boldsymbol{\theta}_i|\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)} = \\
= & -\frac{G}{2}\cancel{\left\langle \ln 2\pi\sigma_i^2\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)}} + \frac{G}{2}\ln\bar{\gamma}_i \\
- & \left\langle \frac{\bar{\gamma}_i}{2\sigma_i^2}\left(\boldsymbol{\omega}_i - \mathbf{m}_{\boldsymbol{\omega}_i}\right)^\top\left(\boldsymbol{\omega}_i - \mathbf{m}_{\boldsymbol{\omega}_i}\right)\right\rangle_{q\left(\boldsymbol{\theta}_i|\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)} \\
+ & \ln\frac{\left(\bar{\gamma}_i b_i\right)^{a_i}}{\Gamma\left(a_i\right)} - \left(a_i+1\right)\left\langle \ln\sigma_i^2\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)} - \left\langle \frac{b_i}{\sigma_i^2}\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)} \\
+ & \cancel{\frac{G}{2}\left\langle 2\pi\ln\sigma_i^2\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)}} + \frac{1}{2}\ln\left|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{\top(t+1)}\right| \\
+ & \left\langle \frac{1}{2\sigma_i^2}\left(\boldsymbol{\omega}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}\right)^\top \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{-1(t+1)}\left(\boldsymbol{\omega}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)}\right)\right\rangle_{q\left(\boldsymbol{\theta}_i|\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)} \\
- & \ln\frac{\hat{\beta}_i^{(t+1)\,\hat{\alpha}_i^{(t+1)}}}{\Gamma\left(\hat{\alpha}_i^{(t+1)}\right)} + \left(\hat{\alpha}_i^{(t+1)}+1\right)\left\langle \ln\sigma_i^2\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)} + \left\langle \frac{\hat{\beta}_i^{(t+1)}}{\sigma_i^2}\right\rangle_{q\left(\sigma_i^2|\hat{\boldsymbol{\xi}}_{\sigma_i^2}^{(t+1)}\right)}
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{2}\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} - \mathbf{m}_{\boldsymbol{\omega}_i}\right)^\top \bar{\gamma}_i \frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}} \left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} - \mathbf{m}_{\boldsymbol{\omega}_i}\right) - \frac{1}{2}\text{trace}\left(\bar{\gamma}_i \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)}\right) + \frac{G}{2} \\
&+ \frac{G}{2}\ln\bar{\gamma}_i + \frac{1}{2}\ln\left|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{\top(t+1)}\right| - \ln\frac{\hat{\beta}_i^{(t+1)\,\hat{\alpha}_i^{(t+1)}}\,\Gamma(a_i)}{(\bar{\gamma}_i b_i)^{a_i}\,\Gamma\left(\hat{\alpha}_i^{(t+1)}\right)} \\
&+ (\hat{\alpha}_i^{(t+1)} - a_i)\left(\ln\hat{\beta}_i^{(t+1)} - \Psi\left(\hat{\alpha}_i^{(t+1)}\right)\right) + (\hat{\beta}_i^{(t+1)} - b_i)\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}
\end{aligned} \tag{E.16}
$$

and

$$
\begin{aligned}
&\left\langle \ln q\left(\mathbf{x}_i | \hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)\right\rangle_{q\left(\mathbf{x}_i|\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)} \\
&= -\frac{G}{2}\ln 2\pi - \frac{G}{2}\ln\left|\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right| \\
&- \frac{1}{2}\left\langle \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\right)^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{-1(t+1)} \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\right)\right\rangle_{q\left(\mathbf{x}_i|\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right)} \\
&= -\frac{G}{2}\ln 2\pi - \frac{G}{2}\ln\left|\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right| - \frac{1}{2}\text{trace}\left(\cancel{\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}}\right) \\
&- \frac{1}{2}\left(\cancel{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}} - \cancel{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}}\right)^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{-1(t+1)} \left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\right) \\
&= -\frac{G}{2}\ln 2\pi - \frac{G}{2}\ln\left|\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right| - \frac{1}{2}\text{trace}\left(\mathbb{1}^G\right) \\
&= -\frac{G}{2}\ln 2\pi - \frac{G}{2}\ln\left|\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right| - \frac{G}{2}
\end{aligned} \tag{E.17}
$$

From expected values in (D.3), (E.16) and (E.17), the updating rules of the lower bound may be expressed as

$$
\begin{aligned}
& \mathcal{F}\left[q\left(\mathbf{x}_i|\,\hat{\boldsymbol{\xi}}_{\mathbf{x}_i}^{(t+1)}\right), q\left(\boldsymbol{\theta}_i|\,\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_i}^{(t+1)}\right)\right] = \\
= & -\frac{N+G}{2}\ln 2\pi + \frac{N}{2}\ln\bar{\gamma}_i - \frac{N}{2}\left(\ln\hat{\beta}_i^{(t+1)} - \Psi\left(\hat{\alpha}_i^{(t+1)}\right)\right) - \frac{G}{2}\ln\sigma_0^2 \\
- & \frac{\bar{\gamma}_i}{2}\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\mathbf{y}_i^\top\mathbf{y}_i + \bar{\gamma}_i\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\mathbf{y}_i^\top\mathbf{R}\mathbf{D}_{\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} \\
- & \frac{\bar{\gamma}_i}{2}\mathrm{trace}\left(\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)}\right) \\
- & \frac{\bar{\gamma}_i}{2}\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{\top(t+1)}\left(\mathbf{B}\circ\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)}\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{\top(t+1)} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right)\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} \\
- & \frac{1}{2\sigma_0^2}\left(\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right)^\top\left(\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}^{(t+1)} - \mathbf{m}_{\mathbf{x}_i}\right) + \mathrm{trace}\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right)\right) \\
- & \frac{1}{2}\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} - \mathbf{m}_{\boldsymbol{\omega}_i}\right)^\top\bar{\gamma}_i\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}}\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}_i}^{(t+1)} - \mathbf{m}_{\boldsymbol{\omega}_i}\right) - \frac{1}{2}\mathrm{trace}\left(\bar{\gamma}_i\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{(t+1)}\right) + \frac{G}{2} \\
+ & \frac{G}{2}\ln\bar{\gamma}_i + \frac{1}{2}\ln\left|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}_i}^{\top(t+1)}\right| - \ln\frac{\hat{\beta}_i^{(t+1)\,\hat{\alpha}_i^{(t+1)}}\,\Gamma\left(a_i\right)}{\left(\bar{\gamma}_i b_i\right)^{a_i}\Gamma\left(\hat{\alpha}_i^{(t+1)}\right)} \\
+ & \left(\hat{\alpha}_i^{(t+1)} - a_i\right)\left(\ln\hat{\beta}_i^{(t+1)} - \Psi\left(\hat{\alpha}_i^{(t+1)}\right)\right) + \left(\hat{\beta}_i^{(t+1)} - b_i\right)\frac{\hat{\alpha}_i^{(t+1)}}{\hat{\beta}_i^{(t+1)}} \\
+ & \frac{G}{2}\ln 2\pi + \frac{G}{2}\ln\left|\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{(t+1)}\right| + \frac{G}{2}. \qquad \text{(E.18)}
\end{aligned}
$$

Appendix F

Gaussian prior approximation for the Student's t-distribution

Given the joint prior of the activities and noise variance by a Norma Scaled Inverse Gamma distribution as in (6.43), the marginalization over the σ_n^2 leads to a three parameters multivariate Student's t-distribution of $\mathbf{x}_n$ as,

$$
\begin{aligned}
p\left(\mathbf{x}_n\right) &= \int p\left(\mathbf{x}_n, \sigma_n^2\right) d\sigma_n^2 \\
&= \int \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_n, \frac{\sigma_n^2}{\kappa_n}\mathbb{1}^F\right) \mathcal{IG}\left(\sigma_n^2 \middle| \alpha_n, \beta_n\right) d\sigma_n^2 \\
&= \int \left(2\pi\frac{\sigma_n^2}{\kappa_n}\right)^{-\frac{F}{2}} \mathrm{e}^{-\frac{\kappa_n}{2\sigma_n^2}\left(\mathbf{x}_n-\boldsymbol{\mu}_n\right)^\top\left(\mathbf{x}_n-\boldsymbol{\mu}_n\right)} \frac{\beta_n^{\alpha_n}}{\mathcal{G}\left(\alpha_n\right)} \left(\sigma_n^2\right)^{-\alpha_n-1} \mathrm{e}^{-\frac{\beta_n}{\sigma_n^2}} d\sigma_n^2 \\
&= \frac{\beta_n^{\alpha_n}}{\mathcal{G}\left(\alpha_n\right)} \left(\frac{\kappa_n}{2\pi}\right)^{\frac{F}{2}} \int \mathrm{e}^{-\frac{1}{\sigma_n^2}\left[\frac{\kappa_n}{2}\left(\mathbf{x}_n-\boldsymbol{\mu}_n\right)^\top\left(\mathbf{x}_n-\boldsymbol{\mu}_n\right)+\beta_n\right]} \left(\sigma_n^2\right)^{-\alpha_n-1-\frac{F}{2}} d\sigma_n^2 \\
&\overset{v.c.}{=} \left\{t = \frac{1}{\sigma_n^2}\frac{\kappa_n\left(\mathbf{x}_n-\boldsymbol{\mu}_n\right)^\top\left(\mathbf{x}_n-\boldsymbol{\mu}_n\right)+2\beta_n}{2}\right\} \\
&= \frac{\beta_n^{\alpha_n}}{\mathcal{G}\left(\alpha_n\right)} \left(\frac{\kappa_n}{2\pi}\right)^{\frac{F}{2}} \left(\frac{\kappa_n\left(\mathbf{x}_n-\boldsymbol{\mu}_n\right)^\top\left(\mathbf{x}_n-\boldsymbol{\mu}_n\right)+2\beta_n}{2}\right)^{-\alpha_n-\frac{F}{2}} \int t^{\alpha_n+\frac{F}{2}-1}\mathrm{e}^{-t}dt \\
&= \frac{\beta_n^{\alpha_n}}{\beta_n^{\alpha_n+\frac{F}{2}}} \left(\frac{\kappa_n}{2\pi}\right)^{\frac{F}{2}} \left(\frac{\kappa_n\left(\mathbf{x}_n-\boldsymbol{\mu}_n\right)^\top\left(\mathbf{x}_n-\boldsymbol{\mu}_n\right)}{2\beta_n}+1\right)^{-\alpha_n-\frac{F}{2}} \frac{\mathcal{G}\left(\alpha_n+\frac{F}{2}\right)}{\mathcal{G}\left(\alpha_n\right)}
\end{aligned}
$$

$$= \left(\frac{\kappa_n}{2\beta_n\pi}\right)^{\frac{F}{2}} \left(\frac{\kappa_n \left(\mathbf{x}_n - \boldsymbol{\mu}_n\right)^\top \left(\mathbf{x}_n - \boldsymbol{\mu}_n\right)}{2\beta_n} + 1\right)^{-\frac{2\alpha_n+F}{2}} \frac{\mathcal{G}\left(\frac{2\alpha_n+F}{2}\right)}{\mathcal{G}\left(\alpha_n\right)}$$

$$= \mathcal{S}\mathrm{t}\left(\mathbf{x}_n | \, \boldsymbol{\mu}_n, \frac{\alpha_n\kappa_n}{\beta_n}\mathbb{1}^F, 2\alpha_n\right) \tag{F.1}$$

Following the results described in [9], the mean and variance of a Student's t-distribution as in (F.1) will be,

$$\left\langle \mathbf{x}_n \right\rangle_{p(\mathbf{x}_n)} = \boldsymbol{\mu}_n \tag{F.2}$$

$$\left\langle \mathbf{x}_n\mathbf{x}_n^\top - \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top \right\rangle_{p(\mathbf{x}_n)} = \frac{\beta_n}{\kappa_n\left(\alpha_n - 1\right)}\mathbb{1}^F \tag{F.3}$$

As discussed in [9], for higher degrees of freedom the Student's t-distribution can be conveniently approximated by a Gaussian. Figure F.1 shows different settings for a Student's t-distributions as in (6.43) and the its corresponding Gaussian approximation with (F.2) and (F.3).

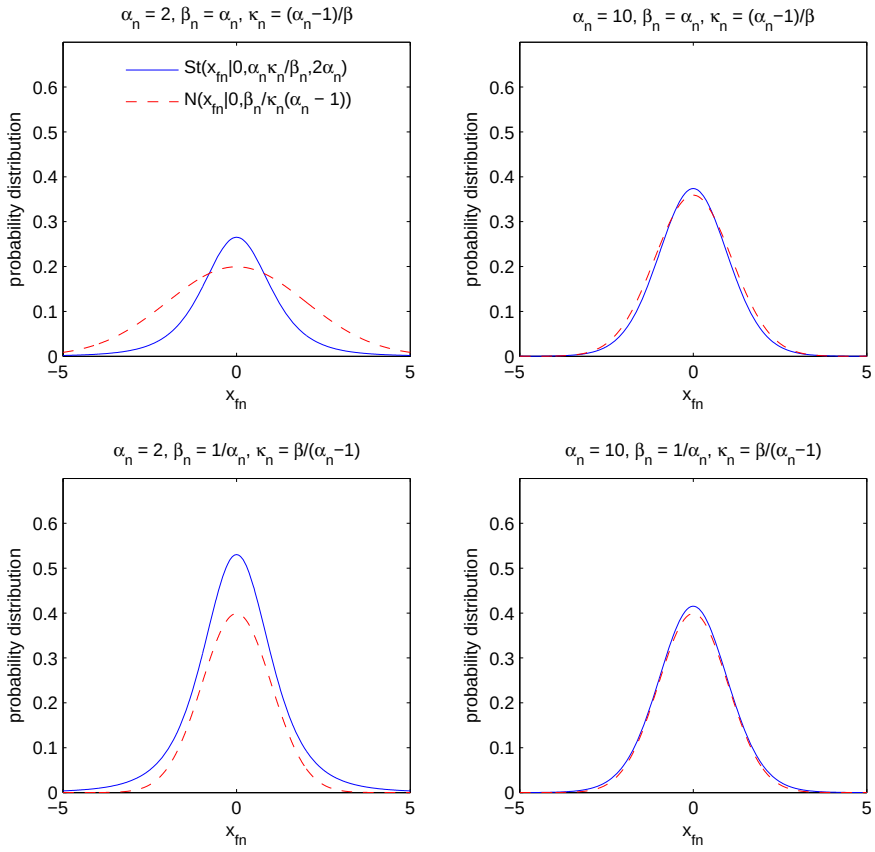


Figure F.1: Sutendt's t-distribution as in (F.1) with $\mu_{fn} = 0$ and different settings for α_n, β_n and κ_n and its Gaussian approximation with (F.2) and (F.3).

Apéndice G

Predicción de sitios de enlace de factores de transcripción

La técnica ChIP-on-chip, así como otros procedimientos experimentales más actuales [27], permiten documentar las interacciones entre genes y proteínas. Además, a partir de métodos computacionales y otros procedimientos experimentales se puede obtener la secuencia de nucleidos, del orden 500 pares de bases (bp), que compone la región activa del gen responsable de la interacción [69]. Esta información se procesa para caracterizar los sitos de enlaces mediante estructuras menores, denominadas *motifs*, constituidas por un patrón de pocos nucleótidos muy conservados, que se conservan en diferentes experimentos [53] [69]. Este tipo de información se almacena como secuencias cortas de cuatro nucleótidos, donde en algunas posiciones de la cadena se contempla la posibilidad de observar alguna de las otras tres bases nitrogenadas. La Tabla G.1 muestra un patrón de ejemplo con 10 secuencias diferentes, cada una compuesta por una secuencia de nucleótidos diferente pero con cierto parecido estructural, donde hay constancia de la interacción con una proteína hipotética.

Otra forma de almacenar este tipo de datos es mediante matrices de frecuencias de posiciones (PFM, acrónimo del inglés *position frecuency matrix*), en las que se recuenta el número de veces que uno de las cuatro bases aparece en una posición concreta. La Tabla G.2 muestra la PFM del ejemplo anterior, donde las frecuencias más altas revela las similitudes entre las secuencias del patrones observado.

Alternativamente, las PFM se suelen representar como una transformación logarítmica en base dos de la frecuencia observada respecto a la probabilidad de un hipotético proceso aleatório. Esta transformación describe la PFM como una matriz de pesos (PWM, acrónimo del inglés *position weight matrix*) en términos de entropía relativa. Esta información se puede resumir en el vector de información del *motif*, que se calcula como el valor esperado de la PWM. Gráficamente, el

Nº	Secuencia
1	GAGGTAAAC
2	TCCGTAAGT
3	CAGGTTGGA
4	ACAGTCAGT
5	TAGGTCATT
6	TAGGTACTG
7	ATGGTAACT
8	CAGGTATAC
9	TGTGTGAGT
10	AAGGTAAGT

Tabla G.1: Ejemplo de un patrón con 10 secuencias, cada una compuesta por 9 bases, donde se observa que una proteína interacciona con un gen.

Nucleótido	1	2	3	4	5	6	7	8	9
A	0.3	0.6	0.1	0.0	0.0	0.6	0.7	0.2	0.1
C	0.2	0.2	0.1	0.0	0.0	0.2	0.1	0.1	0.2
G	0.1	0.1	0.7	1.0	0.0	0.1	0.1	0.5	0.1
T	0.4	0.1	0.1	0.0	1.0	0.1	0.1	0.2	0.6

Tabla G.2: Matriz de frecuencias de posiciones (PFM) en la que se contabiliza la frecuencia de observación de un nucleótido concreto en un conjunto de secuencias. Por ejemplo, en la 4º y en la 5º posición las bases G y T se conservan para todas las secuencias del patrón.

vector de información se suele representar como un logotipo en el que, en cada posición, se representan las cuatro bases con un tamaño proporcional al contenido en información. La Figura (G.1) muestra el logotipo del conjunto de secuencias del ejemplo anterior.

Figura G.1: Logotipo del *motif* de ejemplo. La región mejor conservada corresponde a las posiciones 4 y 5 del patrón.

El potencial de este tipo de base de datos, donde se almacenan las PFM, reside en la capacidad de predecir sitios de enlaces de proteínas funcionales que actúan como factores de transcripción (FT). Actualmente, las técnicas de secuenciación permiten obtener la composición nucleótida de genomas tan complejos como el humano. Este tipo de información también está disponible en bases de datos [33] y permiten conocer la región promotora que precede a cada gen. Por tanto, dada una secuencia de interés, se desea conocer qué proteínas poseen un potencialmente sitio de enlace y pueden actuar como factores de la transcripción. Los métodos usados para realizar este tipo de estimaciones se basan en medidas de similitud entre el alineamiento de la secuencia de interés y el *motif* de una proteína. Resultan de especial interés aquellas aproximaciones que proporcionan un estadístico en términos de probabilidad.

Actualmente, existen diferentes bases de datos y herramientas para realizar este tipo de tareas [33]. En concreto, la base de datos TransFac y su paquete específico para la predicción de sitios de enlaces de FT, MATCH, es uno de los recursos más extendidos [44] [53]. TransFac integra diferente tipo de información relevante para caracterizar la interacción gen-proteína, incluyendo las PFM de factores de transcripción para diferentes especies, entre las que se encuentra la humana. Por otro lado, MATCH proporciona estadísticos basados en la similitud entre la secuencia de interés y regiones especialmente conservadas en la PFM, denominadas *core*. En particular, este método permite procesar secuencias largas de varios kilo-bp y estimar la probabilidad de enlace entre dicha secuencia y un conjunto de FT. Además, éste algoritmo selecciona las PFM de manera que minimice el número de falsos positivos (FP) o falsos negativos (FN) en sus predicciones.

La información extraída según este procedimiento permiten estimar la estruc-

tura topológica de la red de regulación transcripcional (RRT). En una RRT se describe el efecto de regulación que una proteína ejerce sobre un número de genes. Formalmente, esta estructura se suele describir mediante unos coeficientes a_{gf} que cuantifica el efecto regulador que la f-ésima proteína ejerce sobre el g-ésimo gen. En concreto, esta magnitud toma un valor nulo $a_{gf} = 0$ cuando el FT no regula dicho gen. Mientras que la predicción de sitios de enlaces de FT permite establecer este tipo de relaciones causales entre genes y proteínas, el efecto de activación o inhibición de la expresión queda indeterminado. Se sabe que el número de genes regulados por una proteína concreta, denominado regulón, es muy limitado. Esta especificad con la que un FT controla la expresión de un gen sugiere que la variable que describe el regulón de un FT para un genoma con G genes, $\mathbf{a}_f = [a_{1f}, \ldots, a_{Gf}]^\top$, posea una gran cantidad de elementos nulos, es decir, $\mathbf{a}_f$ es una variable dispersa. Los niveles de dispersión estructural de la RRT dependen de la especia considerada, pero se estima que puede llegar a niveles del hasta el 80 %, con sólo el 20 % de los coeficientes diferentes de cero.

Para confirmar esta propiedad estructural de las RRT, se ha estimado la distribución de elementos nulos en $\mathbf{a}_f$ para un conjunto de FT. Se ha hecho uso de TransFac (2009.4) y MATCH para predecir los sitios de enlaces de $F = 850$ FT humanos incluidos en la base de datos, con un perfil que asegura minimizar el número de FN. Para ello, se han predicho los sitios de enlaces de cada FT a lo largo de la región promotora de cada gen, con una longitud de 2 kilo-bp. El resultado de este análisis es un conjunto de probabilidades π_{gf} de que el f-ésimo FT regule la expresión del g-ésimo gen, es decir,

$$p\left(a_{gf} \neq 0\right) = \pi_{gf}. \tag{G.1}$$

En primer lugar se ha tenido en cuenta el genoma humano completo, con $G = 36066$ genes y los $F = 850$ FT incluidos en TransFac, para el que se ha estimado un nivel de dispersión mínimo del 69 %. Por otro lado, en un subconjunto más específico en el que se han considerado $G = 55$ genes y $F = 17$ FT con funciones específicas en el cáncer de mama, se han estimado niveles mínimos del 57 %. La Figura muestra un histograma con la distribución de las probabilidades π_{gf} para cada subconjunto G.2.

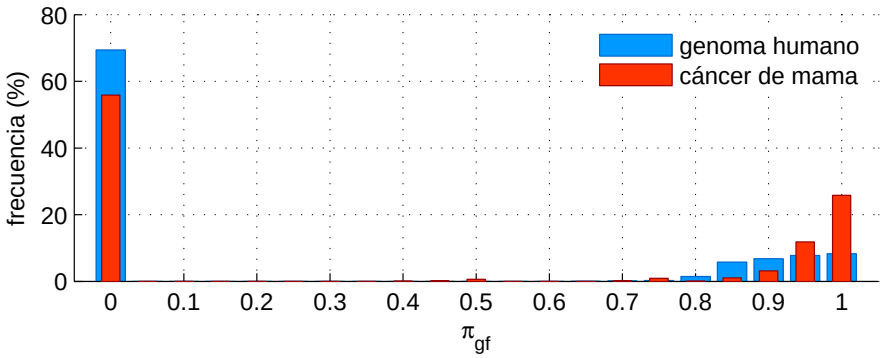


Figura G.2: Distribución de $\pi_{gf} \in [0, 1]$, probabilidad que el f-ésimo factor regula la expresión del g-ésimo gen. Para el genoma humano al completo se han considerado $G = 36066$ genes y los $F = 850$ FT disponibles en TransFac (2009.4). En el subconjunto del cáncer de mama se han tenido en cuenta $G = 55$ genes y $F = 17$ FT con funciones específicas para esta enfermedad.

Bibliografía

[1] P. Baldi and G. W. Hatfield. *DNA microarray and gene expression*. Cambridge University Press, 2002. 9, 115

[2] V. Barnett. *Comparative Statistical Inference*. John Wiley & Sons, 1973. 26, 27, 29

[3] T. Barret. Ncbi geo: archive for functional genomics data sets. *Nucleic acids research*, 39:1005–1010, 2011. 65, 105

[4] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. 26

[5] M. J. Beal and Z. Ghahramani. The variational bayesian EM algorithm for incomplete data with application to scoring graphical model structures. *Bayesian Statistics*, 7, 2003. 28, 44, 117, 118

[6] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7:559–583, 2004. 4

[7] J. M. Bernardo. *Bioestadística*. Barcelona: Vicens-Vives, 1981. 23

[8] J. M. Bernardo and A. F. Smith. *Bayesian Theory*. John Wiley & Sons, 1994. 26

[9] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006. 29, 152

[10] J. U. Blohmer, M. Rezai, S. Kummel, and W. Eiermann. Using the 21-gene assay to guide adjuvant chemotherapy decision-making in early-stage breast cancer: a cost-effectiveness evaluation in the german setting. *Journal on Medical Economics*, 2012. 105

[11] BreastCancer.Org. Understanding breast cancer. 102

[12] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *PNAS*, 101:4164–4169, 2004. 6, 72, 125

[13] E. J. Candes and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25:21–30, 2008. 70

[14] C. M. Carvalho, J. Chang, J. E Lucas, J. R. Nevins, Q. Wang, , and M. West. High-dimensional sparse factor modelling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103:1438–1456, 2008. 70, 71, 76, 125

[15] K. C. Chen, L. Calzone, A. Csikasz-Nagyan, R. Cross, B. Novak, and J. J. Tyson. Integrative analysis of cell cycle control in budding yeast. *Moelcular Biology of the Cell*, 15:3841–3862, 2004. 63

[16] S. K. Chia, V. H. Bramwell, D. Tu, and T. O. Nielsen. A 50 gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clinical Cancer Research*, 18:4465–4472, 2012. 105

[17] P. Collas. The current state of chromatin immunoprecipitation. *Nucleic Acids Research*, 45:87–100, 2010. 18, 115

[18] F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970. 115

[19] S. Das, D. Caragea, W. H. Hsu, and S. M. Welch, editors. *Computational Methodologies in Gene Regulatory Networks*. IGI Global, 2008. 19, 115, 116

[20] Saccharomyces Genome Database. Saccharomyces genome database (sgd). 63

[21] A. Datta and E. R. Dougherty. *Introducction to genomic signal processing with control*. CRC Press, 2007. 9, 115

[22] M. de Hoon, S. Imoto, and S. Miyano. Inferring gene regulatory networks from time-ordered gene expression data using differential equations. *Pacific Symposium on Biocomputing*, 9:17–28, 2003. 5

[23] H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9:67–103, 2002. 5

[24] A. de la Fuente, P. Brazhnik, and P. Mendes. Linking the genes: inferring quantitative gene networks from microarray data. *Trends in Genetics*, 18:395–398, 2002. 19, 116

[25] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997. 18

[26] K. Devarajan. Non-negative matrix factorization: An analytical and interpretive tool in computational biology. *PLOS Computational Biology*, 4:1–12, 2008. 6, 70

[27] M. Djordjevic. SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. *Biomolecular Engineering*, 24:179–189, 2007. 21, 115, 155

[28] Balleza E, E. R. Alvarez-Buylla, A. Chaos, S. Kauffman, I. Shmulevich, and M. Aldana. Critical dynamics in genetic regulatory networks: Examples from four kingdoms. *PLOS one*, 3:1–10, 2008. 5

[29] P. Eroles, A. Bosch, J. A. Pérez-Fidalgo, and A. Lluch. Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. *Journal on Medical Economics*, 38:698–707, 2012. 103

[30] S. Fodor, J. Leighton Read, M. Pirrung, L. Stryer, A T Lu, and D Solas. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251:767–773, 1990. 115

[31] Union for International Cancer Control. *TNM clasiffication of Malignant tumors*. John Wiley & Sons, 2009. 102

[32] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000. 5, 38, 116

[33] E. Gadaleta and N. R. Lemoine C. Chealala. Online resources of cancer data: barriers, benefits and lessons. *Briefs in Bioinformatics*, 12:52–63, 2010. 157

[34] A. Gelman. *Bayesian Data Analysis*. Chapman & Hall, 2003. 42, 80, 85, 127, 138

[35] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:1–19, 2006. 138

[36] T. Gong, J. Xuan, L. Chen, and R. B. Riggins. Motif-guided sparse decomposition of gene expression data for regulatory module identification. *BMC Bioinformatics*, 12:1–16, 2011. 70

[37] Y. Guan and J. G. Dy. Sparse probabilistic principal component analysis. *Journal of Machine Learning Research*, 5:185–192, 2009. 6, 70

[38] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models

of genetic regulatory networks. *Pacific Symposium on Biocomputing*, 6:422–433, 2001. 5, 37, 116

[39] R. Henao and O. Winther. Sparse linear identifiable multivariate modeling. *Journal of Machine Learning Research*, 12:863–905, 2011. 73

[40] Y. Huang, I. M. Tienda-Luna, and Y. Wang. A survey of statistical models for reverse engineering gene regulatory networks. *IEEE Signal Process Magazine*, 26:76–97, 2009. 4, 5, 19, 116

[41] Y. Huang and J. Zhang. A generalized probabilistic data association multiuser detector. In *Information Theory, (ISIT 2004). International Symposium on*, pages 529–531, 2004. 76

[42] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33:245–254, 2003. 19

[43] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22:437–467, 1969. 5, 37, 116

[44] A. E. Kel, E. Gobling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. MATCH : a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, 31:3576–3579, 2003. 6, 21, 72, 115, 124, 157

[45] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. The yeast cell-cycle network is robustly designed. *PNAS*, 101:4781–4786, 2004. 20, 64, 65, 69, 122, 123, 174

[46] J. C. Liao, R. Boscolo, Y. L. Yang, L. M. Tran, C. Sabatti, and V.P. Roychowdhury. Network component analysis: Reconstruction of regulatory signals in biological systems. *PNAS*, 100:15522–15527, 2003. 6, 70, 72, 124

[47] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18:51–60, 2002. 6, 70

[48] M. López, P.Mallorquín, and M. Vega. Microarrays y biochips de ADN. Technical report, Fundación Genoma España, 2002. 9, 15

[49] J. Loscalzo, I. Kohane, and A. L. Barabasi. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. *Molecular Systems Biology*, 3:1–11, 2007. 69

[50] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2002. 26, 27, 29, 80

[51] S. Mallat. *A Wavelet Tour of Signal Processing*. ELSEVIER, 2009. 74

[52] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla-Favera, and A. Califano. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 27:2263–2270, 2011. 59, 119

[53] V. Matys and O. V. Kel-Margoulis. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34:108–110, 2006. 6, 19, 72, 115, 124, 155, 157

[54] J. Meng, M. Sanchez-Castillo, I.M. Tienda-Luna, and Y. Huang. Prediction of cancer subtypes using bayesian factor network model. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 995–996, 2011. 4, 112

[55] J. Meng, J. Zhang, Y. Qi, Y. Chen, and Y. Huang. Uncovering transcriptional regulatory networks by sparse bayesian factor model. *EURASIP Journal on Advances in Signal Processing*, 2010:1–18, 2010. 4, 70, 73, 76, 115, 125

[56] S. Narasimhan, R. Rengaswamy, and R. Vadigepalli. Structural properties of gene regulatory networks: Definitions and connections. *BMC Bioinformatics*, 6:158–170, 2009. 71, 72

[57] KEGG: Kyoto Encyclopedia of Genes and Genomes. Kegg pathway database. 20

[58] Yeast Cell Cycle Analysis Project of Stanford University. Download data. 64

[59] D. S. Oh, M. A. Troester, J. Usary, and C. M. Perou. Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. *JOURNAL OF CLINICAL ONCOLOGY*, 24:1656–1664, 2006. 103

[60] M. Palaiologou, J. Koskinas, M. Karanikolas, and E. Fatourou D. G. Tiniakos. E2f-1 is overexpressed and pro-apoptotic in human hepatocellular carcinoma. *VIRCHOWS ARCHIV*, 460:439–446, 2012. 107

[61] A. Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2006. 24

[62] J. S. Parker. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinicical Oncology*, 27:1160–1167, 2009. 69, 70, 102, 105, 129

[63] P. Z. Peebles. *Principios de probabilidad, variables aleatorias y señales aleatorias*. Mc Graw Hilll, 2008. 24

[64] D. Peña. *Fundamentos de estadística*. Alianza Editorial, 2008. 23, 24, 25, 26, 27

[65] C. M. Perou and T. Sorlie. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000. 101, 103, 129

[66] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *Biostatistics*, 12:1338–1351, 2005. 75

[67] Q. Qi, Y. Zhao, M. Li, and R. Simon. Non-negative matrix factorization of gene expression profiles: a plug-in for BRB-arraytools. *Bioinformatics*, 25:545–547, 2009. 72, 124

[68] J. M. Raser. Noise in gene expression origins, consequences and control. *Science*, 309:2010–2013, 2005. 39, 71

[69] J. E Reid, K. J. Evans, N. Dyer, L. Wernisch, and S. Ott. Variable structure motifs for transcription factor binding sites. *BMC Genomics*, 11:1–18, 2010. 19, 21, 155

[70] A. Ribeiro, R. Zhu, and S. A. Kauffman. A general modeling strategy for gene regulatory networks with stochastic dynamics. *Journal of Computational Biology*, 9:1603–1609, 2006. 19, 38, 116

[71] M. Ronen, R. Rosenberg, and B. I. Shraimanand U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *PNAS*, 99:10555–10560, 2002. 71

[72] R. Rouzier and W. F. Symmans C. M. Perou. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clinical Cancer Research*, 11:5678–5685, 2005. 5

[73] J. J. K. Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996. 28

[74] C. Sabatti and G. M. James. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22:739–746, 2006. 71, 72, 76, 89, 125

[75] M. Sanchez-Castillo, J. Meng, I.M. Tienda-Luna, and Y. Huang. Basis-expansion factor models for uncovering transcription factor regulatory networks. In *Statistical Signal Processing Workshop (SSP), 2011 IEEE*, Ann Arbor, Michigan, 2012. 112

[76] M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Modificación del método em variacional bayesiano para el estudio de microarrays temporales. In *XXIV Simposio nacional de la Unión Científica Internacional de Radio (URSI 2009)*, pages 136–137, Santander, Spain, 2009. 111

[77] M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Análisis bayesiano de microarrays temporales para la estimación de redes de reguladoras de genes. In *XXV Simposio nacional de la Unión Científica Internacional de Radio (URSI 2010)*, Bilbao, Spain, 2010. 111

[78] M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Modified variational method for genes regulatory network learning. In *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*, pages 1781–1784, Beijing, China, 2010. 111, 117

[79] M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Revision of the variational bayesian method for uncovering genes regulatory network. In *Genomic Signal Processing and Statistics (GENSIPS), 2011 IEEE International Workshop on*, pages 206–209, San ANtonio, Texas, 2011. 111, 122

[80] M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Methods and recent patents for modeling and uncovering gene regulatory networks. *Recent Patents on Signal Processing*, 2:88–95, 2012. 5, 111

[81] M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Microarray time series modeling and variational bayesian method for reverse engineering gene regulatory networks. In *International Conference on Advances in Signal and Image Processing (SPITASP 2012)*, Dubai, United Arab Emirates, 2012. 111

[82] M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Modelo factorial bayesiano para el aprendizaje de la red de regulación transcripcional. In *XXVII Simposio nacional de la Unión Científica Internacional de Radio (URSI 2012)*, Elche, Spain, 2012. 112

[83] T. Schaffter, D. Marbach, and D. Floreano. Genenetweaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 7:1–17, 2006. 56, 119

[84] I. Shmulevich and E. R. Dougherty. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18:261–274, 2002. 5, 37, 116

[85] American Cancer Society. American cancer society annual report 2011. 101

[86] T. Sorlie and C. M. Perou. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS*, 98:10869–10874, 2001. 69, 101, 104, 129

[87] P. T. Spellman. Comprehensive identification of cell cycle-regulated genes of the yeast SCE by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998. 20, 64, 122

[88] D. M. Witten R. Tibshirani. A penalized matrix decomposition with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10:515–534, 2009. 70

[89] I. M. Tienda-Luna, Y. Huang, and Y. Yin. Uncovering gene regulatory networks from time-series microarray data with variational bayesian structural expectation maximization. *EURASIP Journal on Bioinformatic and and Systems Biology*, 1, 2007. 4, 5, 39, 48, 116

[90] I. M. Tienda-Luna, Y. Yin, M. C. Carrion-Perez, Y. Huang, M. Sanchez H. Cai, and Yufeng Wang. Inferring the skeleton cell cycle regulatory network of malaria parasite using comparative genomic and variational bayesian approaches. *Genetica*, 132:131–142, 2008. 5

[91] I. M. Tienda-Luna, Y. Yin, and Y. Huang. Constructing gene networks using variational bayesian variable selection. *Artificial life*, 14:65–79, 2008. 4, 38, 118

[92] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11:443–482, 1999. 70

[93] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004. 26

[94] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS*, 98:11462–11467, 2001. 5, 69

[95] S. Wolfram. *A New Kind of Science*. Wolfram Media, 2002. 23

[96] J. Xiang, X. Pan, J. Xu, and Q. Wei. Human epidermal growth factor receptor 2 protein expression between primary breast cancer and paired asynchronous local-regional recurrences. *Experimental and Therapeutic Medicine*, 2:1187–1191, 2011. 103

[97] J. Xie and P. S. Crooke. A computational model of quantitative chromatin immunoprecipitation (ChIP) analysis. *Cancer Informatics*, 6:138–176, 2008. 18

Índice de figuras

Índice de cuadros

Índice alfabético