FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA QUÍMICA Y BIOTECNOLOGÍA
DEPARTAMENTO DE INGENIERIA CIVIL INDUSTRIAL

INVESTIGACIÓN, MODELACIÓN Y RECONSTRUCCIÓN DE REDES DE REGULACIÓN TRANSCRIPCIONALES UTILIZANDO UN ENFOQUE DE PROBLEMAS INVERSOS.

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN BIOTECNOLOGÍA E INGENIERO CIVIL INDUSTRIAL

DANIEL HERMES LILLO EGAÑA

PROFESOR GUÍA:

JUAN ASENJO DE LEUZE DE LANCIZOLLE

MIEMBROS DE LA COMISIÓN:

IVAN RAPAPORT ZIMERMANN

AXEL OSSES ALVARADO

NICOLAS FIGUEROA GONZÁLEZ

SANTIAGO DE CHILE

AGOSTO 2010

A todos los que me acompañaron y apoyaron en estos largos años de formación.

A mis amigos, compañeros, hermanos y a mi familia, por estar siempre ahí.

A ti Javi, por apoyarme, quererme y entenderme siempre.

Y en especial para mis padres… gracias por todo.
Fofo, Moma, esto es para Uds.

Resumen

Investigación, modelación y reconstrucción de redes de regulación transcripcionales utilizando un enfoque de problemas inversos.

El nivel de complejidad y organización de la materia en los seres vivos es sorprendente y a la vez fascinante. Aun cuando no se sepa por qué existe o cómo se genera, se cree firmemente en el paradigma que dicta que gran parte de la información para el surgimiento de tal complejidad está codificada, de alguna forma, en los genes. Al respecto las redes de regulación transcripcionales cumplen un rol fundamental, por cuanto son capaces de responder a estímulos externos y controlar de forma precisa los genes, y por ende, las proteínas que son expresadas en un momento particular.

Debido a la incapacidad para directamente observar y comprender el funcionamiento de estos sistemas de regulación, surge la necesidad de utilizar métodos indirectos basados en el análisis de experimentos de expresión genética. Lamentablemente la mayor parte de estos procedimientos son de naturaleza estadística, por lo que ignoran el trasfondo netamente biológico del sistema analizado. Por ende los resultados carecen del sentido biológico necesario para hacerlos útiles e interpretables a dicho nivel. En el último tiempo nuevos enfoque de análisis basados en la naturaleza estructural del problema han sido introducidos. Al respecto, NCA (Network Component Analysis) ha demostrado un gran potencial y ventaja sobre otro tipo de aplicaciones, permitiendo al investigador reconstruir los parámetros desconocidos de redes de regulación transcripcionales.

En el presente estudio se pretende reproducir las bases de dicha técnica, con el objetivo de comprender sus ventajas e identificar sus limitaciones. Luego se busca modificar y extender la técnica base, buscando incluir en el método nuevas funcionalidades, que en conjunto con reconstruir la red permitan la inclusión de información adicional con el fin de obtener resultados más precisos. El enfoque propuesto permite incluir la varianza de los datos de microarrays utilizado en el análisis, así como suposiciones en los parámetros a estimar del modelo, obteniendo así reconstrucciones que bajo ciertas condiciones se muestran más precisas que con el método original. Los métodos son probados y validados extensamente en redes sintéticas obteniendo resultados que ilustran la gran capacidad y robustez frente a errores de la técnica desarrollada.

Por otra parte, un nuevo enfoque basado en la técnica heurística de recocido simulado es desarrollado. Con éste se espera encontrar redes alternativas a la propuesta, que expliquen de manera alternativa los resultados obtenidos y reduzcan así la gran cantidad de información respecto a la estructura de la red que NCA requiere para trabajar.

Agradecimientos

En estas líneas quisiera agradecer a todas las personas e instituciones que de alguna forma me ayudaron y apoyaron en esta etapa. En primer lugar al ICDB que como instituto se encargaron de patrocinar y financiar este proyecto. Y por supuesto a todas las personas que en él me apoyaron. A mi profesor guía, el Dr. Juan Asenjo, por su incondicional apoyo y guía durante el desarrollo de mi memoria, por presentarme un tema que la verdad me dejo en extremo conforme, y en especial por su ayuda en lo que se refirió a mis problemas con la doble titulación. Además, por supuesto, por darme toda la libertad y confianza a la hora de trabajar. Gracias también a mis profesores co-guías y miembros de la comisión. Al profesor Ivan R. y Axel O. por su gran disposición y tiempo invertido en el desarrollo de esta memoria. Agradezco mucho las reuniones semanales que tuvimos, en las que siempre me ayudaron a aclarar mis dudas y a enfocar de buena forma el trabajo. Su ayuda en el ámbito de las matemáticas fue fundamental en el desarrollo del trabajo, y sus sugerencias me permitieron llegar a buen puerto. También muchas gracias al profesor Nicolas F., por su apoyo en representación del Departamento de Industrias, y por su buena disposición en todo momento. Espero que el nuevo proyecto que estoy desarrollando en mi Magister con tu ayuda termine de tan buena forma como éste. Gracias también por los consejos y críticas de los demás miembros del instituto.

No puedo dejar de agradecer al Departamento de Ingeniería Química y Biotecnología y al de Ingeniería Civil Industrial, así como a todos sus profesores y funcionarios, por su orientación y formación en estos años de estudio. Además agradezco a todos los cursos y profesores que me recibieron como auxiliar y ayudante, y a todo lo que esto me permitió aprender. Finalmente me gustaría agradecer a la Escuela de Ingeniería, en especial a Alfredo Lucas, quien me dio la confianza y me permitió trabajar en la universidad desde un lugar poco común para un alumno. Fue sin duda una experiencia enriquecedora de la cual aprendí mucho.

Fuera del ámbito académico me gustaría dar las gracias, en primer lugar, a mis padres y familia. Nunca dejé de sentir su apoyo, el cual fue fundamental en momentos en que vi difícil las cosas. Gracias a mi padre por educarme siempre para ser el mejor, y por enseñarme a tomar las cosas con calma cuando se ven difíciles pero permaneciendo siempre fuerte y de pie. Y por supuesto a mi madre por darme más de un consejo, y entregarme todo su cariño y apoyo. Siempre lo sentí así. Gracias a mi hermano y hermanas. En especial a la menor por mantener vivo mi espíritu de niño.

No puedo olvidar tampoco a mis amigos y compañeros, sin los cuales todo hubiese sido mucho más difícil. A mis compañeros de universidad, en especial a Carlitos, Chicho, Sergio y Cami, y a tantos otros que aprendí a conocer y que me entregaron los mejores años en la universidad. Y por supuesto a mis amigos de la vida. Pese a que el tiempo y las circunstancias nos ha separado físicamente, me llevo de cada uno de Uds. lo mejor. Gracias Isa, Gori, Javi, Kokure, David, Martín, Diego, Ale, Cata, y en general a todos los que se consideran parte de mi vida. En especial gracias a mi gran amiga y compañera, Javi, por aguantarme todos estos años, aceptarme, apoyarme, entenderme y quererme. Nada hubiera sido igual sin tu compañía.

Gracias a todos.

Índice de Contenidos

Resumen ..... **ii**
Agradecimientos ..... **iii**
Índice de Contenidos ..... **iv**
Índice de figuras ..... **viii**
Índice de tablas ..... **x**
Símbolos y nomenclatura ..... **xii**
1. Introducción **1**
1.1 Introducción a la problemática 1
1.2 Marco teórico 3
1.2.1 Antecedentes generales: Control de expresión génica 3
1.2.2 Regulación transcripcional de la expresión de genes 5
1.2.3 Perfiles de expresión génicos 7
1.2.4 Enfoque de Biología de Sistemas 8
1.2.5 Problemas Inversos 10
1.2.6 Aplicación a redes de regulación genéticas 11
1.3 Descripción del proyecto y justificación 11
1.3.1 Problemática 11
1.3.2 Descripción del proyecto 12
1.3.3 Fundamentación y relevancia 13
1.4 Objetivos 14
1.4.1 Objetivo general 14
1.4.2 Objetivos específicos 14
1.5 Alcances y limitaciones del estudio 14
1.6 Indicaciones sobre la información presentada 15
2. Metodología y métodos **16**
2.1 Modelación de sistemas biológicos 16
2.2 Modelo de regulación transcripcional 16
2.2.1 Red y Señales de regulación (Actividad de los Factores de Transcripción) 16
2.2.2 Modelo matemático 17
2.3 Modelo y red de regulación transcripcional 21
2.4 Problema inverso de reconstrucción 23

2.5 Enfoque de reconstrucción NCA: Network Component Analysis 23
2.5.1 Enfoque de reconstrucción 26
2.5.2 Enfoque de estimación de parámetros 26
2.6 Implementación numérica básica 27
2.6.1 Función de minimización 27
2.6.2 Normalización de resultados 27
2.6.3 Algoritmo de optimización bi-lineal alternada 28
2.7 Metodología de pruebas sintéticas 31
2.7.1 Algunos detalles sobre el enfoque de pruebas sintéticas 33
2.7.2 Medición de errores en pruebas sintéticas 34
2.8 Programación en MatLab 35
3. Resultados y Discusiones 36
3.1 Problemas mal condicionados y otros problemas 36
3.2 Enfoque de resolución basado en el gradiente funcional. 37
3.2.1 Sobre la utilización del gradiente funcional 39
3.3 Criterios NCA y su implicancia 40
3.4 Sobre la normalización de resultados 43
3.5 Programación en MatLab 45
3.6 Reprogramación métodos originales 45
3.6.1 NCA básico (NCAbasic) 45
3.6.2 NCA básico general (gNCAbasic) 50
3.6.3 Sobre métodos de regularización 55
3.6.4 NCA básico general con regularización (gNCAreg) 55
3.6.5 Métodos programados 60
3.7 Modificación a métodos originales 60
3.7.1 Sobre errores de convergencia 61
3.7.2 Sobre mínimos locales 62
3.8 Nuevos métodos NCA 63
3.9 Errores en los datos y NCA considerando confiabilidad 63
3.9.1 Utilización de matrices de confiabilidad 65
3.10 Regularización general. Suposiciones a priori en $\boldsymbol{A}$ y $\boldsymbol{P}$ 73
3.10.2 NCA general con suposiciones (agNCAr) 75

3.11 Grado de certeza en las suposiciones de $\boldsymbol{A}$ y $\boldsymbol{P}$ 78
3.11.1 Definición y cálculo de confiabilidad de suposiciones 80
3.11.2 NCA general con confiabilidad en suposiciones (acgNCAreg)....83
3.12 NCA General 87
3.13 Un nuevo enfoque: Interpolación optimal....90
3.13.1 Confiabilidad de las reconstrucciones....92
3.14 Resumen de métodos NCA....93
3.15 Pruebas sintéticas 93
3.16 Pruebas sintéticas: Grupo 1 94
3.16.1 Nivel 1....94
3.16.2 Nivel 2....106
3.16.3 Nivel 3....119
3.16.4 Nivel 4....128
3.16.5 Nivel 5....145
3.17 Pruebas sintéticas: Grupo 2 148
3.17.1 Nivel 1: Redes gigantes 148
3.17.2 Nivel 2: Análisis por partes de redes que no cumplen NCA 150
3.17.2 Nivel 3: Otras experiencias....155
3.18 Redes alternativas: Un nuevo enfoque 158
3.18.1 Experimentos para detectar unión TF-gen....159
3.18.2 Criterio de redes equivalentes 160
3.18.2 Simulated Annealing (Recocido simulado)....161
3.18.3 Definición de la técnica de recocido simulado....161
3.18.4 Aplicación a NCA: Implementación y comentarios. 163
3.18.5 Medición de errores entre estructuras 174
3.18.6 Recocido Simulado NCA 175
3.19 Pruebas sintéticas: Grupo 3 175
3.19.1 Prueba simple con una red pequeña. 176
Conclusiones 183
Anexos 188
Anexo 1: Demostración de teoremas....188
Teorema 1: Solución esencialmente única en gNCA....188

Extraído de Liao J.C. (2005) 188
Anexo 2: Particularidades matemáticas 193
Diferenciación de matrices 193
Anexo 3: Funciones MatLab 195
Programas y funciones creadas 195
Anexo 4: Programas en MatLab 197

Índice de figuras

Figura 1: Células epiteliales y neuronas de células de mamíferos. 3
Figura 2: Niveles de control de expresión génica. 4
Figura 3: Resumen de mecanismos de regulación transcripcional. 6
Figura 4: Procedimiento experimento microarray. 8
Figura 5: Disciplinas que interactúan en la Biología de Sistemas. 9
Figura 6: Proceso de integración de la información realizado por la Biología de Sistemas. 10
Figura 7: Red de regulación transcripcional. 17
Figura 8: Modelación efecto $\boldsymbol{TF's}$ sobre un gen.. 18
Figura 9: Representación esquemática en forma de red del proceso de regulación transcripcional.. 22
Figura 10: Representación gráfica curvas de iso-error.. 28
Figura 11: Método de optimización bi-lineal alternada. 29
Figura 12: Red de regulación transcripcional generada sintéticamente. 32
Figura 13: Matrices reducidas de A.. 41
Figura 14: Efecto gráfico de la normalización. 43
Figura 15: Efecto no deseado de la normalización.. 44
Figura 16: Existencia de mínimos locales 48
Figura 17: Distribución del residuo de ajuste obtenido con NCAbasic para 200 adivinaciones iniciales 49
Figura 18: Red sintética experimentos knock-out 53
Figura 19: Efecto parámetro de regularización en el ajuste a los datos. 59
Figura 20: Convergencia de error en iteraciones. 61
Figura 21: Frecuencia del residuo utilizando método modificado 62
Figura 22: Distribución de suposiciones. 81
Figura 23: Esquema construcción matrices precisión de suposiciones. 83
Figura 24: Ajuste reconstrucción. Pruebas G1N1E1 95
Figura 25: Ajuste erróneo en P producto de un mínimo local. Pruebas G1N1E1. 97
Figura 26: Frecuencia errores reconstrucción NCA. Red 1. Pruebas G1N1E1 98
Figura 27: Frecuencia errores reconstrucción NCA. Red 2. Pruebas G1N1E1 99
Figura 28: Ajuste gráfico reconstrucción. Pruebas G1N1E2 100
Figura 29: Distribucón de los errores de ajuste frente a diferentes adivinaciones. Pruebas G1N1E2. 101
Figura 30: Representación esquematica del numero de iteraciones. Pruebas G1N1E3 102
Figura 31: Distribución del error. Pruebas G1N1E5. 106
Figura 32: Distribución del error de los datos. Pruebas G1N2. 107
Figura 33: Ajuste gráfico de la reconstrucción utilizando nca_n. Pruebas G1N2E1. 110
Figura 34: Distribución del error. Pruebas G1N2E2. 113
Figura 35: Ajuste gráfico $\boldsymbol{TFA's}$ reconstruidos. Pruebas G1N2E3. 115
Figura 36: Ajuste gráfico $\boldsymbol{TFA's}$ reconstruidos con un error del 100%. Pruebas G1N2E3. 117
Figura 37: Esquema gráfico de convergencia utilizando suposiciones para los datos. Pruebas G1N4E1. 130
Figura 38: Red hipotética a particionar. Pruebas G2N2. 151
Figura 39: Particiones de la red propuesta. Pruebas G2N2. 151

Figura 40: Red no particionable de manera perfecta. Pruebas G2N2. 152
Figura 41: Posibles particiones de la red propuesta. Pruebas G2N2. 152
Figura 42: Representación gráfica del algoritmo de recocido simulado.....163
Figura 43: Representación esquemática de los cambios puntuales en la estructura de conexión. 166
Figura 44: Distribución de los errores a cambios puntuales en la estructura. Agregar conexiones.....167
Figura 45: Distribución de los errores a cambios puntuales en la estructura. Eliminar conexiones. 168
Figura 46: Distribución de los errores a cambios puntuales en la estructura. Permutar conexiones. 169
Figura 47: Distribución de los errores a cambios puntuales en la estructura con error en los datos. Agregar conexiones. 170
Figura 48: Distribución de los errores a cambios puntuales en la estructura con error en los datos. Eliminar conexiones. 170
Figura 49: Distribución de los errores a cambios puntuales en la estructura con error en los datos. Permutar conexiones. 171
Figura 50: Distribución de los errores a cambios puntuales en la estructura con otra estructura base. Agregar conexiones. 172
Figura 51: Distribución de los errores a cambios puntuales en la estructura con otra estructura base. Eliminar conexiones. 172
Figura 52: Distribución de los errores a cambios puntuales en la estructura con otra estructura base. Permutar conexiones 173
Figura 53: Visualización gráfica del algoritmo de recocido simulado 1. Pruebas G3N1E1. 177
Figura 54: Visualización gráfica del algoritmo de recocido simulado 2. Pruebas G3N1E1. 178
Figura 55: Visualización gráfica del algoritmo de recocido simulado 1. Pruebas G3N1E2. 179
Figura 56: Visualización gráfica del algoritmo de recocido simulado 2. Pruebas G3N1E2.. 180
Figura 57: Visualización gráfica del algoritmo de recocido simulado 2. Pruebas G3N1E2. 181

Índice de tablas

Tabla 1: Resumen métodos NCA. 93
Tabla 2: Resumen errores utilizando nca_n. Pruebas G1N1E1. 96
Tabla 3: Mínimo local Red 1 utilizando nca_n. Pruebas G1N1E1. 96
Tabla 4: Resumen errores utilizando nca_n. Pruebas G1N1E2. 100
Tabla 5: Resumen de errores utilizando nca_n. Pruebas G1N1E3. 102
Tabla 6: Distribución del error. Pruebas G1N1E3.. 103
Tabla 7: Resumen reconstrucción red 1. Pruebas G1N1E4. 104
Tabla 8: Resumen de errores, red pequeña. Pruebas G1N1E5. 105
Tabla 9: Resumen de errores de las matrices de datos utilizadas. Pruebas G1N2E1. 108
Tabla 10: Resumen de reconstrucción utilizando nca_n. Pruebas G1N2E1. 109
Tabla 11: Resumen reconstrucción utilizando nca_n. Red diferente. Pruebas G1N2E1.. 111
Tabla 12: Resumen reconstrucción utilizando gnca_n. Pruebas G1N2E1. 112
Tabla 13: Resumen reconstrucción utilizando gnca_reg_n y diferentes errores en los datos. Pruebas G1N2E2. 112
Tabla 14: Resumen de la reconstrucción utilizando 3 métodos. Pruebas G1N2E3.. 113
Tabla 15: Resumen de la reconstrucción frente a errores extremos. . Pruebas G1N2E3. 116
Tabla 16: Resumen de la reconstrucción frente a diferentes concentraciones de errores. Pruebas G1N2E4. 118
Tabla 17: Resumen de la reconstrucción utilizando $\boldsymbol{Ep}$. Pruebas G1N3E1.. 121
Tabla 18: Resumen de la reconstrucción utilizando $\boldsymbol{Ep2}$. Pruebas G1N3E1. 122
Tabla 19: Resumen de la reconstrucción en red de tamaño medio. Pruebas G1N3E2. 123
Tabla 20: Resumen de la reconstrucción en red de tamaño grande. Pruebas G1N3E2. 123
Tabla 21: Resumen de la reconstrucción. Red 2.Pruebas G1N3E3.. 124
Tabla 22: Resumen de la reconstrucción. Red 2 modificada. Pruebas G1N3E3. 125
Tabla 23: Resumen de la reconstrucción. Red 5. Pruebas G1N3E3. 125
Tabla 24: Resumen de la reconstrucción. Red 6. Pruebas G1N3E3. 126
Tabla 25: Resumen de la reconstrucción. Red 18. Pruebas G1N3E3. 126
Tabla 26: Resumen de la reconstrucción utilizando $\boldsymbol{Ep1}$. Pruebas G1N3E4. 127
Tabla 27: Resumen de la reconstrucción utilizando $\boldsymbol{Ep2}$. Pruebas G1N3E4. 127
Tabla 28: Resumen de la reconstrucción utilizando $\boldsymbol{Abr}$. Pruebas G1N4E1. 129
Tabla 29: Resumen de la reconstrucción utilizando $\boldsymbol{Ab}$. Pruebas G1N4E1. 130
Tabla 30: Resumen de la reconstrucción utilizando $\boldsymbol{Abn}$. Pruebas G1N4E1. 131
Tabla 31: Resumen de la reconstrucción utilizando $\boldsymbol{Abr}$ y $\boldsymbol{Em1}$. Pruebas G1N4E2. 132
Tabla 32: Resumen de la reconstrucción utilizando $\boldsymbol{Abr}$ y $\boldsymbol{Em2}$. Pruebas G1N4E2. 132
Tabla 33: Resumen de la reconstrucción utilizando $\boldsymbol{Ab}$ y $\boldsymbol{Em1}$. Pruebas G1N4E2.. 133
Tabla 34: Resumen de la reconstrucción utilizando $\boldsymbol{Ab}$ y $\boldsymbol{Em2}$. Pruebas G1N4E2.. 133
Tabla 35: Resumen de la reconstrucción utilizando $\boldsymbol{Abn}$ y $\boldsymbol{Em1}$. Pruebas G1N4E2 133
Tabla 36: Resumen de la reconstrucción utilizando $\boldsymbol{Abn}$ y $\boldsymbol{m}$. Pruebas G1N4E2 133
Tabla 37: Resumen de la reconstrucción utilizando $\boldsymbol{Pb1}$ y $\boldsymbol{Em1}$. Pruebas G1N4E3 134
Tabla 38: Resumen de la reconstrucción utilizando $\boldsymbol{Pb1}$ y $\boldsymbol{Em2}$. Pruebas G1N4E3 134

Tabla 39: Resumen de la reconstrucción utilizando $\boldsymbol{Pb2}$ y $\boldsymbol{Em1}$. Pruebas G1N4E3.....135
Tabla 40: Resumen de la reconstrucción utilizando $\boldsymbol{Pb2}$ y $\boldsymbol{Em2}$. Pruebas G1N4E3.....135
Tabla 41: Resumen de la reconstrucción utilizando $\boldsymbol{Pbr}$ y $\boldsymbol{Em1}$. Pruebas G1N4E3.....135
Tabla 42: Resumen de la reconstrucción utilizando $\boldsymbol{P_{br}}$ y $\boldsymbol{E}_{m2}$. Pruebas G1N4E3. 135
Tabla 43: Resumen de la reconstrucción utilizando $\boldsymbol{Ab}$ y $\boldsymbol{Em1}$. Pruebas G1N4E4.. 137
Tabla 44: Resumen de la reconstrucción utilizando $\boldsymbol{Ab}$ y $\boldsymbol{Em2}$. Pruebas G1N4E4. 137
Tabla 45: Resumen de la reconstrucción utilizando $\boldsymbol{Ab}$, $\boldsymbol{Em1}$y $\boldsymbol{D}$. Pruebas G1N4E4. 138
Tabla 46: Resumen de la reconstrucción utilizando $\boldsymbol{Ab}$, $\boldsymbol{Em2}$y $\boldsymbol{D}$. Pruebas G1N4E4. 138
Tabla 47: Resumen de la reconstrucción utilizando $\boldsymbol{Ab}$, $\boldsymbol{Em1}$y $\boldsymbol{D}$ modificada. Pruebas G1N4E4. 139
Tabla 48: Resumen de la reconstrucción utilizando $\boldsymbol{Ab}$, $\boldsymbol{Em2}$y $\boldsymbol{D}$ modificada. Pruebas G1N4E4. 139
Tabla 49: Resumen de la reconstrucción utilizando $\boldsymbol{Em1}$. Pruebas G1N4E5.. 141
Tabla 50: Resumen de la reconstrucción utilizando $\boldsymbol{Em2}$. Pruebas G1N4E5. 141
Tabla 51: Resumen de la reconstrucción utilizando $\boldsymbol{Em1}$, flexibilidad exagerada. Pruebas G1N4E5.....142
Tabla 52: Resumen de la reconstrucción. Flexibilidad 1.000. Pruebas G1N4E6. 143
Tabla 53: Resumen de la reconstrucción. Flexibilidad 10.000. Pruebas G1N4E6. 143
Tabla 54: Resumen de la reconstrucción. Flexibilidad 1.000.000. Pruebas G1N4E6. 143
Tabla 55: Resumen de la reconstrucción. Flexibilidad 1.000. Pruebas G1N4E7. 144
Tabla 56: Resumen de la reconstrucción. Flexibilidad 10.000. Pruebas G1N4E7. 144
Tabla 57: Resumen de la reconstrucción. Flexibilidad 1.000.000. Pruebas G1N4E7. 144
Tabla 58: Resumen de la reconstrucción utilizando GgNCAreg. Pruebas G1N5E1. 145
Tabla 59: Resumen de la reconstrucción utilizando GgNCAreg y otra matriz de datos. Pruebas G1N5E1. 146
Tabla 60: Resumen de la reconstrucción utilizando GgNCAreg. Aumento de flexibilidad. Pruebas G1N5E1. 147
Tabla 61: Resumen de la reconstrucción. Pruebas G1N5E2. 147
Tabla 62: Resumen de la reconstrucción de una red grande utilizando NCAbasic. Pruebas G2N1E1.....148
Tabla 63: Resumen de la reconstrucción de una red grande utilizando NCAbasic. Datos con error. Pruebas G2N1E1.....148
Tabla 64: Resumen de la reconstrucción de una red grande utilizando gNCAbasic. Pruebas G2N1E1. ...149
Tabla 65: Resumen de la reconstrucción de una red de tamaño real. Pruebas G2N1E2. 149
Tabla 66: Resumen reconstrucción red particionable. Pruebas G2N2E1 153
Tabla 67: Resumen reconstrucción red particionable y error en los datos. Pruebas G2N2E1. 153
Tabla 68: Resumen reconstrucción utilizando particiones de la red. Pruebas G2N2E1. 153
Tabla 69: Detalle parámetros utilizados. Pruebas G3N1E1. 176
Tabla 70: Detalle parámetros utilizados. Pruebas G3N1E2 179

Símbolos y nomenclatura

Símbolo	Significado
$\boldsymbol{B^{-1}}$	Matriz de precisión de las suposiciones de P.
$\boldsymbol{B}$	Matriz de varianza de las suposiciones de P.
$\boldsymbol{D^{-1}}$	Matriz de precisión de las suposiciones de A.
$\boldsymbol{D}$	Matriz de varianza de las suposiciones de A.
$\boldsymbol{F_A}$	Flexibilidad de las suposiciones de A.
$\boldsymbol{F_P}$	Flexibilidad de las suposiciones de P.
$\boldsymbol{Z_A}$	Sub espacio vectorial que resume la estructura de la matriz A.
$\boldsymbol{Z_P}$	Sub espacio vectorial que resume la estructura de la matriz P.
$[\boldsymbol{mRNA}]$	Concentración de $mRNA$.
$\mathbb{P}$	Símbolo de *"Problema a resolver"*.
$\boldsymbol{A}$	Matriz de conexiones CS.
$\boldsymbol{CS - CS's}$	Control Strenght - (Grado - Grados de influencia de la conexión gen-TF)
$\boldsymbol{DNA}$	Ácido desoxirribonucleico.
$\boldsymbol{E}$	Matriz de datos. Expresión de los genes en diferentes experimentos.
$\boldsymbol{P}$	Matriz de actividad de los $TF's$ en diferentes experimentos.
$\boldsymbol{RNA}$	Ácido ribonucleico.
$\boldsymbol{TF - TF's}$	Transcriptional Factor – Factors (Factor de Transcripción – Factores de Transcripción).
$\boldsymbol{TFA}$	Actividad de los $TF's$.
$\boldsymbol{b}$	Vector de datos para ejemplos genéricos.
$\boldsymbol{i = \{1, \dots, N\}}$	Subíndice para el número de genes.
$\boldsymbol{j = \{1, \dots, L\}}$	Subíndice para el número de $TF's$.
$\boldsymbol{k = \{1, \dots, M\}}$	Subíndice para el número de experimentos.
$\boldsymbol{mRNA}$	RNA mensajero.
$\boldsymbol{x}$	Vector de incógnitas para ejemplos genéricos.

1. Introducción

El presente trabajo de tesis, titulado ***"Investigación, modelación y reconstrucción de redes de regulación transcripcionales utilizando un enfoque de problemas inversos"*** se inserta en el contexto de uno de los tantos proyectos de investigación desarrollados por el ICDB[1] destinados a conocer más los mecanismos de dinámica celular y de los sistemas biológicos en general. En este instituto convergen profesionales y expertos en diferentes ciencias (matemáticas, física, química, biología y computación) por lo que el intercambio de conocimiento e interacción producida privilegia y estimula el desarrollo de este tipo de actividades. Las áreas y temáticas de investigación son variadas, y el enfoque va dirigido a generar conocimiento que pueda ser luego llevado a aplicaciones provechosas en un contexto nacional e internacional.

En particular, el proyecto descrito en este documento forma parte de la unidad de investigación denominada **"Expresión Genética y Modelación Matemática y Bioinformática"**, donde se llevan a cabo esfuerzos para modelar los mecanismos y vías que intervienen en la expresión genética a nivel celular. La Matemática y el enfoque de Biología de Sistemas es básico a la hora de reproducir los sistemas biológicos, por cuanto permiten una mayor comprensión al analizarlos a una escala más entendible y abordable. La convergencia de distintas disciplinas es lo que permite este tipo de desarrollos. En particular, el aporte de la Ingeniería Civil Industrial en sus ámbitos de operaciones, análisis de redes, técnicas y heurísticas de optimización numéricas y estimación estadística de parámetros se torna relevante. A continuación se pretende presentar algunos antecedentes que justifican la elección y desarrollo del tema.

1.1 Introducción a la problemática

"Los organismos vivos son compatibles con las leyes químicas y físicas de interacción de la materia, pero no son consecuencias de ellas". (Jaques Monod. El Azar y la necesidad, 1970).

El nivel de complejidad y organización de la materia en los seres vivos es sorprendente y a la vez fascinante. Aun cuando no se sepa por qué existe o cómo se genera, se cree firmemente en el paradigma que dice que gran parte de la información para el surgimiento de tal complejidad esta codificada, de alguna forma, en los genes. Con el descubrimiento del código genético en la década de los sesenta, se pudo comprobar que la información contenida en los genes determina la estructura (y por ende la funcionalidad) de las proteínas involucradas en el metabolismo celular. Pronto quedó de manifiesto que los genes sólo poseen la información para construir las piezas del rompecabezas, pero no la necesaria para armarlo, al menos no de manera directa. De esta manera, aún después de tener secuenciado el genoma completo de varios organismos, se continúa sin poder armar el rompecabezas de la vida, es decir, sin poder explicar cómo la información contenida en los genes determina la estructura, funcionamiento y la gran complejidad de los seres vivos.

[1] Institute for Cell Dynamics and Biotechnology

Aparentemente no hay suficiente espacio en las secuencias codificadores del genoma para albergar la información necesaria que determine las características fenotípicas de los seres vivos. Las funciones biológicas básicas que lleva a cabo un recién nacido, tales como respirar, mantener su corazón latiendo, comer, llorar, etc., no son aprendidas, si no que forman parte del programa genético con el que nace todo bebé. ¿Cuántas neuronas y conexiones sinápticas entre ellas se necesitan para que un recién nacido lleve a cabo dichas funciones biológicas? No se sabe con certeza, pero tan sólo en el sentido del olfato están involucradas más de 40 millones de neuronas, cada una con 10 mil conexiones sinápticas en promedio [1][2] y más de 1000 genes [2]. Pero lo que es aún más desconcertante es que no se sabe en qué lugar está guardada la información de la arquitectura neuronal de un bebé, es decir, donde se encuentra el respaldo que determina la forma en que se deben establecer las millones y millones de conexiones neuronales para que un ser vivo sobreviva. Es claro que en los casi 26 mil genes del ser humano [3] no hay espacio para almacenar de manera explícita esta información, más la información de los diferentes tipos celulares, más la información de la secuencia temporal de las miles de reacciones químicas que ocurren dentro de la célula. Y sin embargo, la evidencia experimental parece confirmar una y otra vez el paradigma genético: la información de gran parte de la estructura y funcionamiento de los seres vivos está contenida en los genes [4].

Pero el paradigma, al menos en parte, ha ido modificándose. Durante décadas se estudiaron genes individuales, intentando asociar un gen, o un grupo reducido de éstos con alguna característica fenotípica específica. Este enfoque reduccionista ha permitido comprender ciertos fenómenos, pero no es suficiente a la hora de estudiar los organismos como un todo [5]. El punto relevante en esta discusión es que el estudio experimental y teórico de los genomas en los últimos años ha evidenciado que la complejidad de los organismos vivos no es el resultado directo del número de genes, sino de los intrincados mecanismos de regulación de la expresión genética en todo el genoma [6] [7]. Si bien la célula es la expresión más simple de vida, la manera en que ésta se organiza, se desarrolla y responde a estímulos externos está regulada por una compleja combinación de biomoléculas, que en conjunto se encargan de hacer de la célula la unidad funcional, estructural y de vida básica por excelencia.

Entender de qué manera funciona esta regulación y de qué forma diversas funcionalidades de las células son expresadas frente a estímulos específicos, es básico en aplicaciones biotecnológicas, con el fin de predecir perturbaciones causadas por manipulación genética, condiciones ambientales, compuestos tóxicos o agentes farmacéuticos. El genoma de un organismo es básico al respecto. Es una entidad dinámica, que involucra una serie de proteínas y otras moléculas que se unen a sitios específicos, regulando de esta forma la expresión de proteínas. Cada célula, su forma, funcionalidad y capacidad, es el resultado del set de proteínas que su genoma expresa, y por ende, de las complejas relaciones de regulación existentes entre los miles de genes y los factores regulatorios.

[2] Las referencias entre [ ] son detalladas en la sección Bibliografía al final del documento.

1.2 Marco teórico

1.2.1 Antecedentes generales: Control de expresión génica

El DNA de un organismo contiene toda la información necesaria para codificar las proteínas y las moléculas RNA necesarias para la construcción de sus células. Sin embargo, en un momento dado o para un tipo celular determinado, sólo un subconjunto de proteínas es utilizado, acorde a las necesidades o funcionalidades de la célula [8] [9]. En el caso de los microorganismos, por ejemplo, como las bacterias o levaduras, las condiciones ambientales tales como el medio de cultivo, temperatura o el pH, pueden hacer variar la expresión de diferentes proteínas, con el fin de adaptar su metabolismo a estas nuevas condiciones. En organismos pluricelulares sucede lo mismo. En la Figura 1 es posible observar 2 células de mamíferos: Células del epitelio y neuronas.

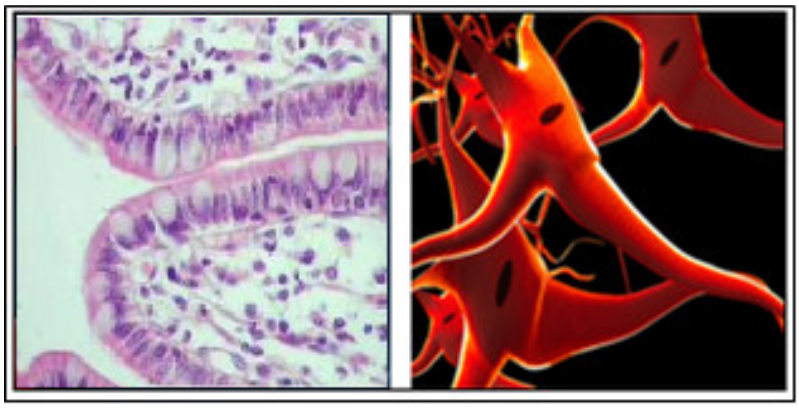


Figura 1: Células epiteliales y neuronas de células de mamíferos.
Fuente: Elaboración propia.

Observando las diferencias morfológicas de ambas, es difícil pensar que pertenecen al mismo organismo, y que por ende poseen el mismo material genético[3]. El punto clave al respecto son los sistemas de controles génicos, que se aseguran que tan sólo algunos genes se expresen, y por ende que el conjunto de proteínas sintetizadas correspondan a aquellas que la célula necesita dado su estado o funcionalidad. Luego, cada célula es el producto de un patrón de expresión genética específico, involucrando diferentes tipos de regulación y la interacción de miles de genes [10].

Es posible regular la expresión génica en muchas etapas, coincidentes con la vía que conduce del DNA al RNA y finalmente a una proteína funcional. En la *Figura 2* se muestran las etapas principales asociadas a la expresión de una proteína. De esta forma, es posible identificar 6 etapas en las que se puede controlar la expresión genética [11].

[3] El material genético se encuentra en el núcleo de ambas células, y es común a un organismo sin importar de que célula se trate.

1. ***Control transcripcional:*** *Correspondiente al control del proceso de transcripción, que copia en una molécula de mRNA la información del gen contenida en el genoma.*
2. ***Control del procesamiento de RNA****: Control correspondiente a las etapas de procesamiento del RNA en el núcleo, como son el corte y empalme y modificaciones CAP y poliA.*
3. ***Control del transporte del RNA****: En relación a la exportación del RNA maduro al citoplasma.*
4. ***Control traduccional****: Correspondiente al control en la etapa de transcripción del mRNA a una proteína.*
5. ***Control de la degradación:*** *Desestabilización selectiva de moléculas de mRNA.*
6. ***Control de actividad proteica:*** *Activando o inactivando selectivamente las proteínas ya sintetizadas.*

En la mayor parte de los organismos procariontes los controles del tipo transcripcional son los más importantes, en el sentido que aseguran que no se generen productos intermedios que luego no serán utilizados [12] [13] [11].

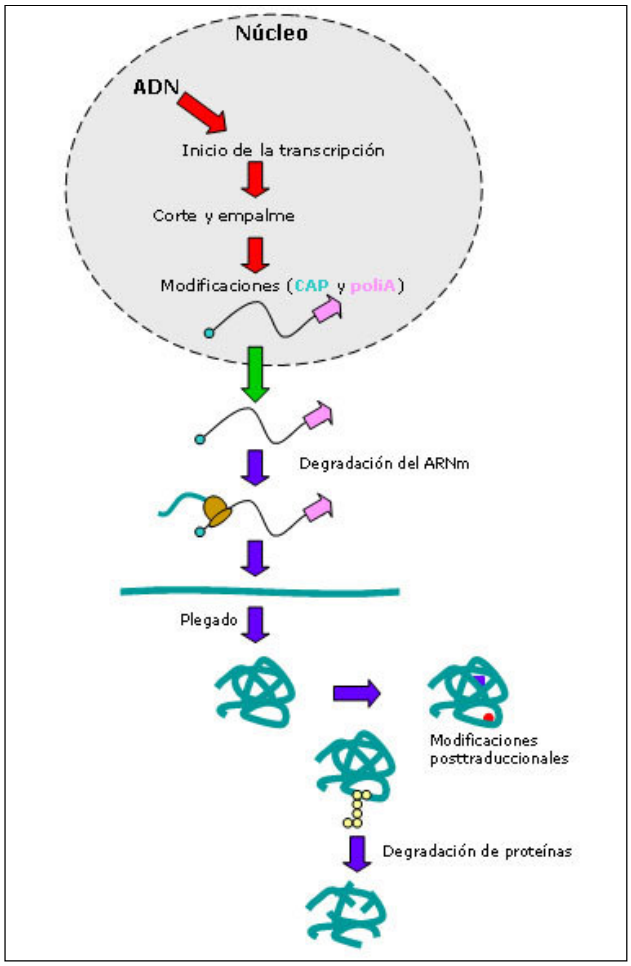


Figura 2: *Niveles de control de expresión génica.*
En la figura se resumen los diferentes niveles de control de expresión génica. Al inicio del proceso, la transcripción.
Fuente: Albert, Biología Molecular de la Célula.

1.2.2 Regulación transcripcional de la expresión de genes

Como se ha mencionado, el control de la transcripción es el sistema de regulación más importante de la expresión genética en procariontes, y no menos importante en eucariontes. La transcripción es la primera de las etapas por la que pasa la información codificada en el DNA para transformarse en una proteína funcional, y consiste en la copia de dicha información en un transcrito de mRNA. La molécula encargada de dicha copia es la RNA polimerasa, que se sitúa al inicio del gen respectivo, reclutando nucleótidos a medida que va leyendo la molécula de DNA, con el fin de construir el mRNA asociado. El control de dicho proceso es lo que se conoce como regulación transcripcional.

La regulación de este tipo es mediada por moléculas especializadas conocidas con el nombre de factores de transcripción ($TF's$ [4]), que se unen selectivamente a la secuencia de los genes, activando o reprimiendo su transcripción al mediar de forma importante la interacción del DNA con la RNA polimerasa. El proceso, tanto de activación como de inhibición, comienza con una señal externa que modifica la actividad de los $TF's$. En algunos casos dicho fenómeno actúa por medio de ligandos, que interactúan con la proteína modificando su conformación y estado de actividad, y por ende afectando la expresión de los genes. Dichos ligandos pueden corresponder a moléculas o a ciertas modificaciones estructurales como son fosforilaciones o glucocilaciones [4].

En el caso de los $TF's$ activadores, la regulación recibe el nombre regulación positiva, y el ligando puede actuar activando (provocando la expresión del gen) o desactivando la actividad del factor de transcripción (deteniendo la expresión del gen). De la misma manera, en el caso de los $TF's$ inhibidores, la regulación se dice negativa. De esta manera, el ligando puede unirse al TF activándolo y provocando su acción represora, o desactivando su acción inhibidora, provocando la transcripción del gen [14]. Esto se resume en la Figura 3 siguiente.

[4] Por su nombre en inglés, T*ranscriptional Factors*,

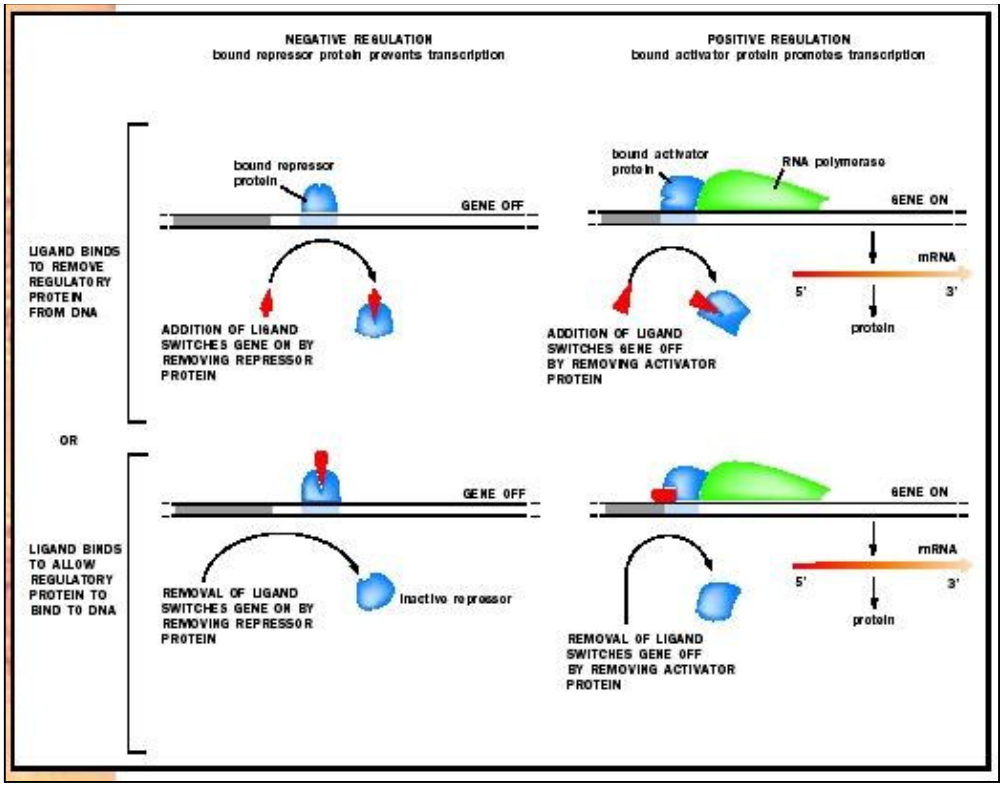


Figura 3: Resumen de mecanismos de regulación transcripcional.
Fuente: Albert, Biología Molecular de la Célula.

Por lo general existe una secuencia promotora, ubicada antes del gen en particular, a la cual los $TF's$ activadores se unirán. Una vez provocada dicha unión, el resto de la maquinaria transcripcional tomará su lugar y la transcripción se iniciará. Los $TF's$ inhibidores en cambio, bloquean el sitio de unión de la RNA polimerasa, impidiendo el proceso.

Es necesario, sin embargo, aclarar que el proceso es algo más complicado que lo mencionado, principalmente en las células eucariontes. Las fuentes que aumentan la complejidad de este sistema de regulación son de 2 tipos y son las que otorgan la versatilidad y capacidad a este medio de control [11]. En primer lugar, la maquinaria de transcripción que se arma en torno a una secuencia promotora es extremadamente compleja y consiste de numerosos $TF's$ que actúan en conjunto sobre el gen. Algunos de estos corresponden a factores generales (aquellos necesarios para que la RNA polimerasa cumpla su función), mientras que otros son específicos al gen en particular que regulan. Además de esto, los $TF's$ no necesitan unirse directamente a la secuencia promotora, algunos tienen la capacidad de influir en el promotor aun estando a miles de nucleótidos de distancia (es el caso de los *enhancers*), por lo que el potencial regulatorio es casi ilimitado [4]. En segundo lugar, la actividad de los $TF's$ no posee solo 2 estados (de encendido y apagado). De acuerdo a los estímulos que reciban, estas proteínas pueden pasar por diferentes estados de actividad, afectando por ende de diferente manera la expresión de los genes, en el sentido de velocidad y cantidad [4]. En resumen, será la actividad combinada de activadores o inhibidores, actuando selectivamente sobre las secuencias promotoras de los genes la que determinará sus expresiones.

1.2.3 Perfiles de expresión génicos

Un perfil de expresión génico corresponde a los niveles de expresión relativos (en relación a un estado de referencia) de los $mRNA$[5] de todos los genes de un organismo. Hace algunas décadas, obtener este tipo de información era una tarea de proporciones (utilizando técnicas de hibridación como Northern y Southern Blot [15]), pero actualmente las técnicas de alta resolución como los microarrays la convierten en una tarea relativamente sencilla [16]. Un microarray es una versión masiva de los experimentos antes mencionados, en donde fácil y rápidamente es posible medir el nivel de expresión de los mRNA de todos los genes para un organismo particular, y para un experimento o condición determinada.

La base de la técnica es sencilla y consiste en 3 etapas principales [17]:

1. Un oligonucleótido, con una secuencia complementaria al mRNA de interés es inmovilizada en una superficie. El proceso se realiza de forma automatizada, lo que permite fijar nucleótidos muy cerca uno de otros y con una gran precisión. Repitiendo el procedimiento, con secuencias complementarias a todos los genes del organismo en estudio se logra obtener un chip de microarray que resume parte de la información del genoma en pocos centímetros cuadrados. Actualmente se encuentran disponibles en el mercado kits comerciales para distintos tipos de organismos.

2. Posteriormente se extrae el RNA de una célula blanco y se convierte a cDNA utilizando nucleótidos marcados en forma fluorescente. El resultado se hibridiza con los oligonucleótidos del microarray, con el fin de detectar la presencia de dicho mRNA en la célula en estudio. Es común la utilización de una muestra blanco y una de referencia, marcadas con colores distintos, con el fin de observar en un mismo experimento la expresión relativa a dicho estado.

3. Finalmente el microarray es escaneado y analizado.

En la Figura 4 se observa un esquema del proceso.

[5] RNA mensajero.

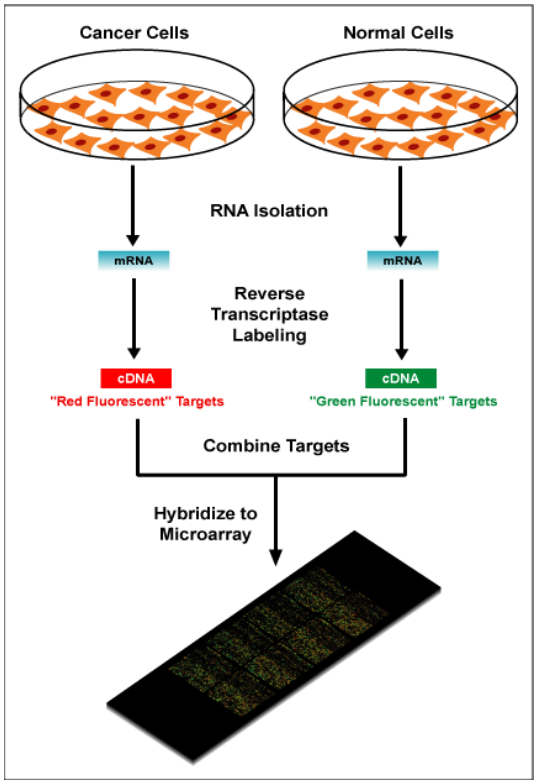


Figura 4: Procedimiento experimento microarray.
Fuente: **www.molecularstation.com**

1.2.4 Enfoque de Biología de Sistemas

La biología de sistemas es una nueva rama de la biología, una suerte de revolución en el campo de las ciencias biológicas, pero con repercusiones en otras disciplinas como la informática, las matemáticas y la física y en particular la física de los sistemas complejos [18]. Sin el afán de ser extensos, y como una forma de explicar esta disciplina, es posible hacer un paralelo con la Biología Molecular, que durante años asumió un camino directo entre los genes, proteínas y funciones biológicas; si bien este enfoque más bien reduccionista ha generado importantes conocimientos, no permite entender el comportamiento de las células como un todo [19]. Durante el último tiempo se ha tenido acceso a un tipo de información mucho más compleja y de mucho mayor dimensión, por lo que la aplicación de técnicas del área de la bioinformática, la física y la ingeniería se han hecho indispensables con el fin de analizar y poder interpretar estos datos [20].

El objetivo de la Biología de Sistemas consiste en integrar esta información a fin de generar un mayor entendimiento de las interacciones entre los componentes de los sistemas vivos, y por ende, de sus procesos biológicos. La herramienta fundamental para llevar a cabo esta misión es el desarrollo de modelos matemáticos, técnicas de procesamiento de datos y simulaciones, como una manera de simplificar y obtener información respecto a un mundo que es bastante desconocido. Algunas características que definen la Biología de Sistema son [21]:

- Estudio de los sistemas biológicos a un nivel global. No se enfoca en entender las partes, sino la forma en que éstas interaccionan.
- Contrasta con la aproximación clásica: un gen, una proteína.
- Integra el conocimiento de diferentes disciplinas (genómica, transcriptómica, proteómica, metabolómica, fisiología, patología, informática, física, ingeniería, análisis de redes, etc.)
- Hace uso de una gran cantidad de datos procedentes de estudios experimentales de alta resolución.
- Propone y valida modelos matemáticos para explicar los fenómenos biológicos estudiados.
- Utiliza simulaciones matemáticas con afanes predictivos.
- Valida modelo mediante comparación de datos y simulaciones numéricas.

En la siguiente figura se representa esquemáticamente el gran número de disciplinas que interactúan en la Biología de Sistemas, aportando con herramientas que permitan analizar y comprender de mejor manera los fenómenos biológicos.

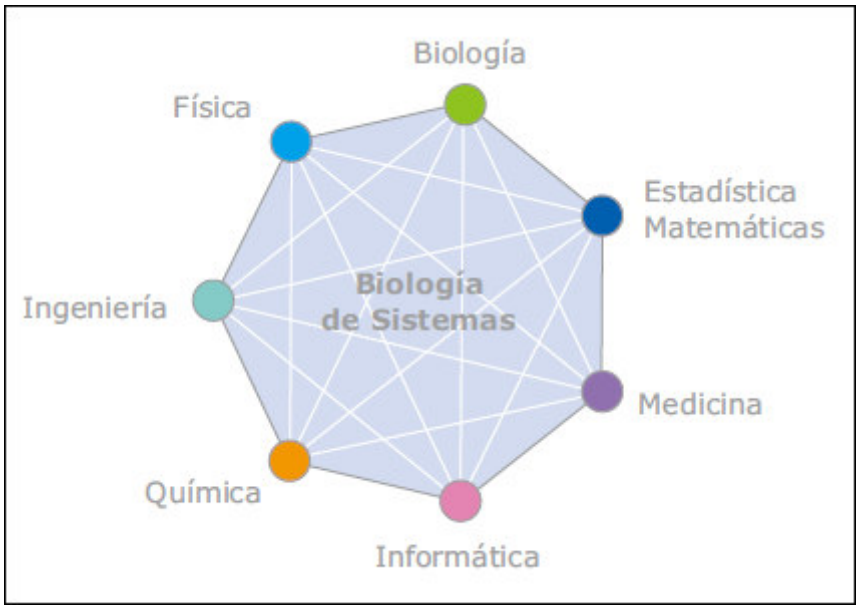


Figura 5: Disciplinas que interactúan en la Biología de Sistemas.
Fuente: Elaboración propia.

En la figura siguiente se resumen los pasos seguidos por la Biología de Sistemas en el proceso de análisis e investigación de los sistemas biológicos.

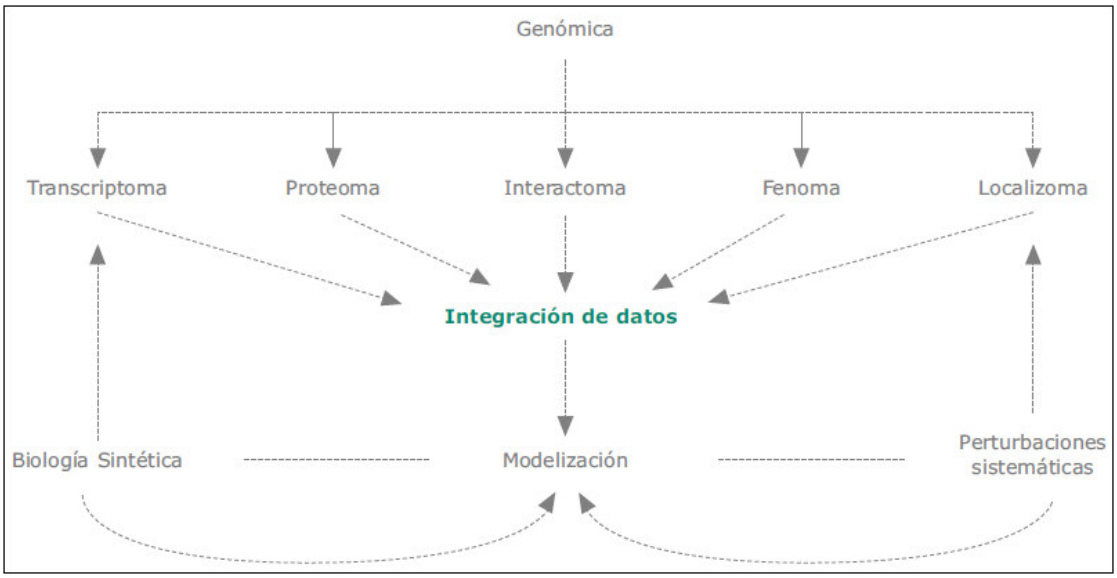


Figura 6: Proceso de integración de la información realizado por la Biología de Sistemas.
Fuente: Elaboración propia.

1.2.5 Problemas Inversos

Los problemas inversos surgen de forma natural en la Biología de Sistemas, y también en otras ramas de las ciencias que hacen uso intensivo de modelos y datos con el fin de integrar información y entregar una mayor compresión del fenómeno en estudio [22]. La Biología de Sistemas busca explicar algún fenómeno, que para efectos prácticos es modelado matemáticamente mediante un conjunto de ecuaciones que lo definen. Por supuesto, dichos modelos poseerán una serie de parámetros que es necesario suponer, estimar o medir, con el fin de validar dicha teoría y poder hacer extensivo su uso como herramienta de predicción [23]. Un problema directo sería aquél en que conociendo con exactitud la definición del modelo y sus leyes, y el valor de sus parámetros y la interacción entre sus partes es posible predecir el resultado (output) de dicha interacción[6].

El problema inverso sería por lo tanto conocer el output[7], e intentar definir qué está pasando en el proceso, descifrando las interacciones que se dan, y los valores actuales de los parámetros [24]. Es precisamente eso lo que se realiza en la Biología de Sistemas. La era post-genómica ha desarrollado técnicas de inmensa capacidad, que permiten obtener experimentalmente una gran cantidad de datos que es necesario interpretar y emplear para completar y definir de la forma más acertada posible los modelos utilizados para dichos fenómenos.

[6] Se puede ver como una máquina bien conocida y definida, a la que si se ingresa un insumo sabemos perfectamente qué se obtendrá.
[7] Por ejemplo la medición de un resultado bajo ciertas condiciones.

1.2.6 Aplicación a redes de regulación genéticas

La interacción y la dinámica de las redes de regulación genéticas puede ser modelada mediante diferentes enfoques [25], pero en todos ellos se tendrá un modelo que combinará una red[8] que relacione de alguna manera a los diferentes genes, y alguna regla que definirá la manera en que estos interactuarán al ser regulada su expresión. Esto implica necesariamente la existencia de ciertos parámetros, algunos de los cuales definirán el modelo, mientras que otros tendrán que ver con su estado o nivel de regulación[9] particular en algún instante.

Conociendo la total definición del modelo, y de los parámetros para un momento determinado, sería posible predecir con exactitud los resultados del mismo, por ejemplo, el nivel de expresión de cada gen de un organismo. La idea, por supuesto, es mediante el enfoque de problema inverso realizar lo contrario, y mediante los datos dilucidar los parámetros más relevantes del modelo. Una vez realizado esto, y validados los resultados[10], es posible utilizar el modelo con afanes predictivos.

En la regulación transcripcional, es posible modelar el proceso mediante una red compuesta por 2 tipos de elementos [26]: Nodos de regulación, correspondientes a los TF's involucrados, y nodos regulados, correspondientes a los genes de interés.

1.3 Descripción del proyecto y justificación

1.3.1 Problemática

Como ya se ha mencionado, los mecanismos de regulación genética son los responsables de la complejidad de los organismos y de asegurar que éste responda de manera certera frente a estímulos externos. Son estos sistemas de regulación, esta topología inducida por la redes que relacionan reguladores transcripcionales con secuencias promotoras, las que determinan los genes qué se expresen en un momento determinado, y por lo tanto, la expresión fenotípica de un ser vivo. Lamentablemente, es poco lo que se conoce respecto a estos sistemas regulatorios. Si bien es posible medir el *output* del sistema (o sea, medir el nivel de expresión de los genes) mediante las técnicas de alto rendimiento como los microarray, los intrincados mecanismos regulatorios, y la forma precisa en que los factores de transcripción responden frente a estímulos e interactúan con los genes es todavía un misterio. Pese a que se continúa acumulando información respecto al estado celular, no se obtiene real información sobre el mecanismo que controla dicha expresión.

[8] Representada por cierta topología.
[9] O estado de expresión génica.
[10] Comparando, por ejemplo, una predicción del modelo con otro set de datos.

Es ya conocido que no es posible estudiar los genes por separado para obtener una visión realista de la dinámica celular, y dado que el seguimiento y medición de los factores que intervienen en la regulación de los genes *in vivo* es realmente complicada, se hace necesaria la utilización de métodos indirectos que permitan desentrañar, al menos en parte, el misterio que continua siendo la dinámica y regulación celular.

Los organismos vivos, bacterias, levaduras y células animales han sido utilizados extensamente en los últimos años para diversos fines. Con la tecnología actual, es posible intervenir el genoma de un organismo, y al menos en parte, redirigir sus funciones biológicas hacia ciertos objetivos. De esta manera, y haciendo uso extenso de la ingeniería genética, ha sido posible la producción industrial en bacterias y levaduras de diferentes proteínas recombinantes y otros productos químicos. El problema radica en que se ha estado utilizando una *maquinaria biológica* de la que en realidad se conoce poco respecto de su funcionamiento. Y esto puede traducirse, por ejemplo, en un bajo rendimiento de las aplicaciones industriales y un total desconocimiento de la manera óptima de trabajar con un organismo en particular.

La problemática general se puede resumir en el poco conocimiento del funcionamiento y regulación de los sistemas biológicos, y consecuentemente la pérdida del gran potencial que significaría el develar, al menos parcialmente, la forma y los mecanismos mediante los cuales las células regulan y controlan sus funciones. En esta misma línea, es necesario destacar la gran ventaja que significaría comprender mejor el funcionamiento y la regulación celular. Los puntos de acción que se abrirían son numerosos: Aplicaciones que aprovechen mejor el metabolismo celular, traducidas en producciones industriales más eficientes y de mayor rendimiento, así como el gran potencial en medicina y desarrollo de terapias, al poder conocer con mayor exactitud la forma en que virus y bacterias se desarrollan en los organismos.

1.3.2 Descripción del proyecto

El proyecto propuesto se enfoca en el estudio e investigación de una novedosa técnica de análisis de redes de regulación, conocida como NCA: Network Component Analysis, que se presenta como una herramienta para ayudar a comprender y reconstruir las redes de regulación genéticas de diferentes organismos [26]. Dicha técnica hace uso de herramientas matemáticas para, en base a información de la expresión genética, reconstruir la red que genera dichos resultados. El enfoque de problemas inversos ha sido ampliamente utilizado en el último tiempo, mediante aplicaciones estadísticas como el análisis de componentes principales o el análisis de componentes independientes que, a fin de cuentas, imponen restricciones con el fin de disminuir la dimensionalidad de los datos que carecen de una intuición o explicación biológica [27] [28]. Por el contrario, NCA utiliza fundamentos de carácter biológicos (principalmente referentes a información *a priori* respecto a la topología de la red) para restringir la reconstrucción de las redes, por lo que es presumible la obtención de resultados representativos de la realidad.

En base a esto el proyecto se centra, en primer lugar, en el estudio extenso de la técnica NCA, tanto a un nivel biológico como matemático, y la consecuente reproducción computacional de las funciones en

MatLab. Junto a esto, se espera generar propuestas de mejoras, tanto a nivel de método como funcional, extendiendo la capacidad y alcances de la técnica, que puedan ser implementadas prácticamente al ser programadas en la misma plataforma.

Posteriormente se pretende testear extensamente el funcionamiento de la técnica y de las nuevas funcionalidades propuestas en experimentos sintéticos, que permitan apreciar el alcance y las limitaciones de las mismas, permitiendo su reformulación y/o corrección hasta que cumpla con las expectativas propuestas. Se espera obtener además un alto grado de conocimiento y experiencia en este tipo de aplicaciones, y en el análisis de redes en general, a fin de obtener resultados útiles para otras áreas de investigación y aplicaciones.

1.3.3 Fundamentación y relevancia

La fundamentación y utilidad del proyecto se justifica principalmente por el gran beneficio y potencial que significa conocer la forma en que las células regulan sus funciones. Entre otras ventajas, permitiría predecir el comportamiento de los organismos a modificaciones en su genoma vía genética, optimizando de esa manera la redirección de su metabolismo a la producción de compuestos de interés. Por otra parte se podría predecir también el efecto de medicamentos y drogas en las células, de manera de mejorar las técnicas de terapias en el ámbito de la medicina. Desentrañar los misterios de la regulación celular, y conocer de modo preciso de qué manera y mediante qué mecanismos la expresión de los genes es regulada posee un gran poder predictivo, y permitiría la generación de nuevas técnicas y enfoques en ingeniería genética y medicina.

Como un ejemplo relevante se podría mencionar el caso del cáncer: No se conoce con exactitud la manera en que funcionan las células cancerígenas, y por ende, el desarrollo de terapias efectivas es altamente complicado.

En otro aspecto, es necesario aclarar que el enfoque NCA y de problemas inversos utilizados para abordar el problema, no está solo restringido a la temática mencionada; existen muchos problemas, tanto en biología como en otras áreas de las ciencias, que pueden ser analizados con las técnicas y conocimientos que se pretende generar. Por lo tanto los beneficios pueden extenderse a otras temáticas.

1.4 Objetivos

1.4.1 Objetivo general

El objetivo general del proyecto se resume en lo siguiente:

- Modelar matemáticamente redes transcripcionales de regulación genética de sistemas biológicos sintéticos y reales a fin de aplicar técnicas de optimización y de problemas inversos, que permitan reconstruir y analizar los parámetros de dichos sistemas en base a información experimental, mejorando la técnica base NCA tanto en su aplicación como en su función.

1.4.2 Objetivos específicos

- Generación de una herramienta de análisis provechosa, que resulte beneficiosa a otras áreas de estudio.

- Reproducción computacional de las herramientas NCA (Network Component Analysis) en MatLab.

- Testeo extenso de la funcionalidad y limitaciones de la técnica en experimentos sintéticos.

- Proponer mejoras y/o nuevas funcionalidades a las técnicas, en base a los puntos débiles identificados.

- Generación de conocimiento y experiencia en beneficio para otras aéreas de investigación.

1.5 Alcances y limitaciones del estudio

El presente estudio se enfoca principalmente en el desarrollo de una técnica que pueda ser usada de forma exhaustiva en el análisis de redes. De la misma manera, y junto con desarrollar una serie de funcionalidades orientadas en su uso y utilidad al análisis en un plano biológico, se espera entender el estado del arte en este tipo de temáticas, que pueda quedar plasmado en las experiencias realizadas, sirviendo como guía en futuras investigaciones. Se desea por sobre todo plasmar el sentido de intuición e interpretación de los diferentes análisis que es posible desarrollar, a fin de generar la experiencia que permita discernir entre el mejor enfoque o combinación de técnicas a aplicar a un problema particular.

El estudio se centra exclusivamente en el análisis de redes sintéticas, como una forma de validar los métodos propuestos. No se trabaja utilizando redes de organismos reales, ni datos de microarrays, debido principalmente al tiempo limitado que el presente estudio presenta.

Se trabaja también introduciendo un nuevo enfoque de análisis, basado en el método heurístico de recocido simulado. En relación a dicho punto se pretende principalmente presentar la motivación, y

desarrollar los fundamentos del método adaptado a NCA, y presentar algunas pruebas interesantes que demuestran su potencial utilización y capacidad. Además se dejan preguntas abiertas respecto a la técnica e interpretación en el sentido biológico y matemático de los resultados, a la vez que se sugieren algunas modificaciones a la técnica que puedan mejorar los resultados obtenidos al incorporar un análisis más sofisticado.

1.6 Indicaciones sobre la información presentada

En los siguientes capítulos se describe detalladamente los pasos seguidos para el desarrollo de este proyecto, y los principales resultados que se obtuvieron.

El Capítulo 2 está destinado a describir la metodología y procedimientos utilizados. Se comienza describiendo el modelo de regulación transcripcional con el que se trabajará, así como la base de la técnica con enfoque de problemas inversos NCA. Se describe también el algoritmo de optimización base utilizado en el problema descrito. Posteriormente se menciona la metodología de experiencias sintéticas con la cual se testearán y validarán los resultados.

En el Capítulo 3 se describen y discuten cabalmente los resultados obtenidos, los aportes realizados y las diferentes pruebas efectuadas. Se comienza describiendo cada uno de los métodos basados en NCA reproducidos, para luego fundamentar las nuevas funcionalidades propuestas para el método. Se revisa y explica también en extenso la modelación y las técnicas matemáticas utilizadas, a fin de desarrollar la intuición que permita interpretar de mejor manera los resultados. Posteriormente se da paso a la descripción de diferentes experiencias sintética que permiten testear el funcionamiento de cada método e identificar posibles fuentes de error en los mismos. Los métodos son además discutidos y analizados, con el fin de concluir respecto a su utilización. Finalmente se menciona un nuevo enfoque propuesto, que se propone como el inicio de nuevas temáticas de investigación que permite relajar la gran cantidad de información requerida por los métodos NCA.

En el Capítulo 4 se describen las principales conclusiones y nuevas líneas de análisis e investigación propuestas.

Finalmente, y en los capítulos restantes, se presenta toda la información anexa de relevancia para quienes deseen ahondar más en el tema.

2. Metodología y métodos

2.1 Modelación de sistemas biológicos

Como ya se ha mencionado, la necesidad de entender la estructura y dinámica de los sistemas biológicos y la incapacidad para observar el funcionamiento celular directamente, ha llevado a que las matemáticas y otras ciencias se hallan unido a la biología, con el fin de modelar dichos fenómenos y mecanismos. El fin principal de este enfoque y en general el de la biología de sistemas, es ayudar a comprender de mejor manera el funcionamiento celular, mediante un proceso de modelamiento y experimentación. Es así como dichos modelos pueden ser validados y utilizados para predecir el comportamiento celular frente a otras condiciones e interpretar las observaciones [21].

Para dicho procedimiento se hace extenso uso de las herramientas matemáticas y de diferentes experiencias, que permiten representar la realidad en ecuaciones de diversos tipos y obtener una aproximación razonable de los parámetros reales de dichos modelos. Si bien el proceso de modelación implica simplificaciones, y en último término una abstracción de la realidad, ha demostrado ser una valiosa herramienta de análisis e interpretación en diversas áreas de la ciencia [29].

2.2 Modelo de regulación transcripcional

2.2.1 Red y Señales de regulación (Actividad de los Factores de Transcripción)

En la regulación transcripcional de la expresión genética intervienen principalmente 2 actores: Los factores de transcripción ($TF's$) y los genes. Es necesario, por supuesto, entender el proceso para lograr su abstracción en la forma de un modelo matemático. Como ya se ha mencionado, los $TF's$ se unen selectivamente a las secuencias promotoras de algunos genes (o interactúan con ella a distancia) promoviendo o inhibiendo en diferentes grados la expresión del respectivo gen en forma de su transcrito de $mRNA$. Dicha grado de expresión (o en otras palabras la concentración de $mRNA$ que se obtendrá) dependerá de varios factores, entre los cuales los más relevantes son el grado de actividad del factor de transcripción (o señal de regulación) y el grado de influencia o control intrínseco que cada TF tiene sobre determinados genes (asumiendo que algunos $TF's$ pueden ser más afines a ciertos genes, independiente de su actividad). La actividad de un factor de transcripción (TFA) más que el nivel de expresión del TF es lo que determina la regulación transcripcional. Debido a modificaciones pos-traducionales y pos-transcripcionales, no existe una correlación significante entre TFA y el nivel de expresión del TF, por lo que un enfoque que determine funciones fisiológicas e interacciones de $TF's$ basado en el análisis de perfiles de TFA tiene una mayor fundamentación que uno realizado únicamente en los $TF's$ por si solos [30]. TFA será definido como la fracción unida a DNA de los factores de transcripción.

Finalmente, será la influencia conjunta de varios $TF's$ sobre un gen, activándolo algunos, o inhibiéndolo otros, lo que producirá un transcrito o copia de dicho gen en una cierta concentración. Dicho fenómeno se puede representar fácilmente por una red de regulación transcripcional, en la que de forma esquemática se especifica qué $TF's$ intervienen, qué genes son los regulados y qué parejas interactúan. Aún más, es posible visualizar por el color de las líneas y su grosor, el tipo de regulación (azul para activación y rojo para inhibición) y el grado de influencia. En la Figura 7 se representa dicha red para el caso de 4 $TF's$ y 7 genes.

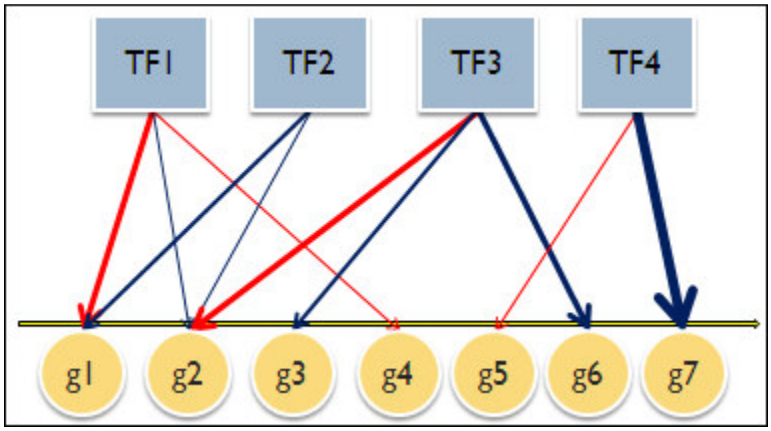


Figura 7: Red de regulación transcripcional.
Fuente: Elaboración propia.

Es necesario destacar que dicha representación es sólo un aspecto del problema a modelar, a saber, la estructura de la red de regulación, dado que no toma en cuenta las señales de regulación (la actividad particular de cada factor de regulación).

2.2.2 Modelo matemático

En un momento determinado, sobre un gen pueden actuar diferentes $TF's$ que van a provocar diversos efectos en la transcripción del mismo, de acuerdo a su actividad y al tipo y grado de influencia que tengan sobre el gen respectivo. Esto se puede representar esquemáticamente en la siguiente figura.

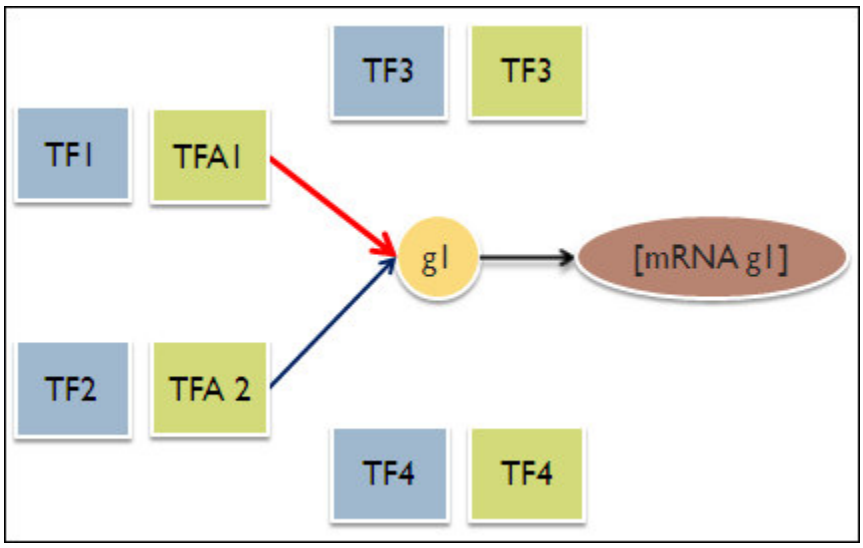


Figura 8: Modelación efecto $\boldsymbol{TF's}$ sobre un gen.
Fuente: Elaboración propia.

Como se puede observar, el estado activo (o señal de regulación) de los $TF's$ 1 y 2 en un momento determinado es lo que provoca la expresión en cierta concentración de $mRNA$. El TF 1 se encuentra inhibiendo mientras que el TF 2 activando, ambos con distinto grado de influencia. El resto de los factores de transcripción, si bien no actúan sobre el gen ejemplificado, sí pueden hacerlo sobre el resto de los genes.

De forma generalizada consideremos un organismo con N genes (especificados con el sub-índice $i = 1, \dots, N$) y L factores de transcripción ($j = 1, \dots, L$).

Además sean:

- $\boldsymbol{TFA_j}$: Actividad o señal de regulación del factor de transcripción j, con $j = 1, \dots, L$.
- $\boldsymbol{CS_{ij}}$: Grado de influencia (Control Strength) que ejerce el TF j sobre el gen i, con $i = 1, \dots, N$ y $j = 1, \dots, L$.
- $[\boldsymbol{mRNA}]_{\boldsymbol{i}}$: Concentración del mRNA correspondiente al gen i, con $i = 1, \dots, N$.

Se hace notar que el parámetro CS_{ij} resume toda la información estructural de la red antes especificada. El número o magnitud de dicho parámetro especificará el grado de influencia, mientras que el signo, positivo o negativo, indicará si se trata de una influencia del tipo activadora o inhibidora. Finalmente los pares j, i de dicho parámetro especificarán la dirección de las conexiones de la red (indicando los genes que controlan cada $TF's$). Si el parámetro es diferente de 0, habrá una conexión en la red, mientras que un valor 0 indicará que no existe relación entre un TF y un gen determinado.

El cambio en la concentración del $mRNA$ de un gen particular que ejercen los diferentes $TF's$ puede ser relacionado mediante distintos modelos, algunos de características no lineales. Sin embargo, el más

versátil quizás, es el modelo log-lineal, similar al utilizado en la teoría de sistemas bioquímicos [31], cinética química y muchos análisis de ingeniería, como el caso de los fenómenos de transporte [32]. De esta manera, para el gen i se tiene:

$$\frac{d[mRNA]_i}{dt} = k_1 \prod_{j=1}^{L} TFA_j(t)^{CS_{ij}}$$

Ecuación 1

En la Ecuación 1, el cambio en la concentración de $mRNA$ en el tiempo (t) es proporcional a la actividad de los $TF's$ elevados al coeficiente de influencia CS respectivo.

Por otra parte, el $mRNA$ en el tiempo experimenta un efecto de degradación, que es proporcional a su concentración. Esto es:

$$\frac{d[mRNA]_i}{dt} = -k_2 \cdot [mRNA]_i(t)$$

Ecuación 2

El efecto combinado de ambos efectos entrega la siguiente ecuación diferencial para la variación de la $[mRNA]_i$ en el tiempo. Esto es:

$$\frac{d[mRNA]_i}{dt} = k_1 \prod_{j=1}^{L} TFA_j(t)^{CS_{ij}} - k_2 \cdot [mRNA]_i(t)$$

Ecuación 3

Existirá una ecuación de este tipo para cada gen $i = 1, \dots, N$.

Una primera aproximación consiste en considerar una escala de tiempo en que un estado cuasi-estacionario para la $[mRNA]_i$ sea válido [33]. Así se tiene de la Ecuación 3:

$$\frac{d[mRNA]_i}{dt} = 0$$

$$k_1 \prod_{j=1}^{L} TFA_j(t)^{CS_{ij}} = k_2 \cdot [mRNA]_i(t)$$

$$[mRNA]_i = \frac{k_1}{k_2} \prod_{j=1}^{L} TFA_j{}^{CS_{ij}}$$

Ecuación 4

Por lo que la $[mRNA]_i$ en estado cuasi-estacionario estará dada por la expresión de la Ecuación 4. Se destaca con fines explicativos, que la expresión anterior entrega la concentración de $mRNA$ del gen i en un estado cuasi-estacionario, correspondiente a un experimento o condición en particular; por ejemplo, algún tipo de nutriente, temperatura o condición de pH. Así, TFA_j se puede considerar como la actividad del factor de regulación respectivo en el estado cuasi-estacionario. Se considerará de ahora en adelante que se tienen M experimentos, simbolizados con el índice $k = 1, \dots, M$, en donde para cada uno de estos y para cada gen, se puede escribir una ecuación como la anterior, donde la dependencia k de la concentración de $mRNA$ y de la señal de regulación, corresponde al estado cuasi-estacionario de ambas variables para el experimento k.

$$[mRNA]_i(k) = \frac{k_1}{k_2} \prod_{j=1}^{L} TFA_j{}^{CS_{ij}}\,(k)$$

Ecuación 5

Los experimentos $k = 1, \dots, M$, pueden considerarse indistintamente diferentes condiciones experimentales independientes, o series de tiempo (instantes en los que se toman mediciones), en los que puede considerarse válida la aproximación de estado cuasi-estacionario.

Considerando el experimento $k = k_0$ como el experimento de referencia (el correspondiente a condiciones estándar por ejemplo), es posible dividir la expresión genérica de la Ecuación 5 por su correspondiente del experimento de referencia.

$$\frac{[mRNA]_i(k)}{[mRNA]_i(k_0)} = \prod_{j=1}^{L} \frac{TFA_j{}^{CS_{ij}}(k)}{TFA_j{}^{CS_{ij}}(k_0)}$$

Ecuación 6

Tomando logaritmo a la expresión anterior.

$$log\left(\frac{[mRNA]_i(k)}{[mRNA]_i(k_0)}\right) = \sum_{j=1}^{L} CS_{ij} \cdot log\left(\frac{TFA_j(k)}{TFA_j(k_0)}\right)$$

Ecuación 7

Las expresiones $[mRNA]_i(k)/[mRNA]_i(k_0)$ y a $TFA_j(k)/TFA_j(k_0)$ corresponden a concentraciones y actividades relativas para el $mRNA$ y los $TFA's$, por lo que son unidades adimensionales. Por otra parte, al considerar el logaritmo de ambas expresiones se obtendrá un resultado de interés:

- Si $log\left(\frac{[mRNA]_i(k)}{[mRNA]_i(k_0)}\right) > 0$, se tendrá una sobreexpresión respecto al experimento de referencia.
- Si $log\left(\frac{[mRNA]_i(k)}{[mRNA]_i(k_0)}\right) < 0$, se tendrá una sub-expresión respecto al experimento de referencia.

- Si $log\left(\frac{[mRNA]_i(k)}{[mRNA]_i(k_0)}\right) \approx 0$, la expresión será similar al experimento de referencia.

Lo mismo es válido por supuesto para la expresión de logaritmo de la señal de regulación de cada TF, obteniendo una mayor, menor o similar actividad respecto al punto de referencia.

Como se mencionó, se posee una expresión como la de la Ecuación 7 para cada combinación gen-experimento, un total de NxM ecuaciones. De esta forma, es posible escribir el set de ecuaciones log-lineales dada por dicha expresión en forma matricial de la forma siguiente:

$$E = A \cdot P$$

Ecuación 8

Donde E es una matriz de dimensión NxM, y la entrada (i,k) corresponde a $E_{ik} = log\left(\frac{[mRNA]_i(k)}{[mRNA]_i(k_0)}\right)$, A posee dimensión NxL y la entrada (i,j) corresponde a $A_{ij} = CS_{ij}$ y P posee dimensión LxM y la entrada (j,k) corresponde a $P_{jk} = log\left(\frac{TFA_j(k)}{TFA_j(k_0)}\right)$.

2.3 Modelo y red de regulación transcripcional

Dado el modelo anterior, es fácil relacionarlo con el esquema en red antes descrito para el mismo proceso. Como se observa en la Figura 7, los $TF's$ interaccionan con los genes, activándolos o inhibiéndolos según sea el caso. Siendo más específicos, es la actividad de los $TF's$ o señales de regulación la que va a provocar el efecto en los transcritos de $mRNA$, por lo que un esquema como el de la figura siguiente sería más exacto.

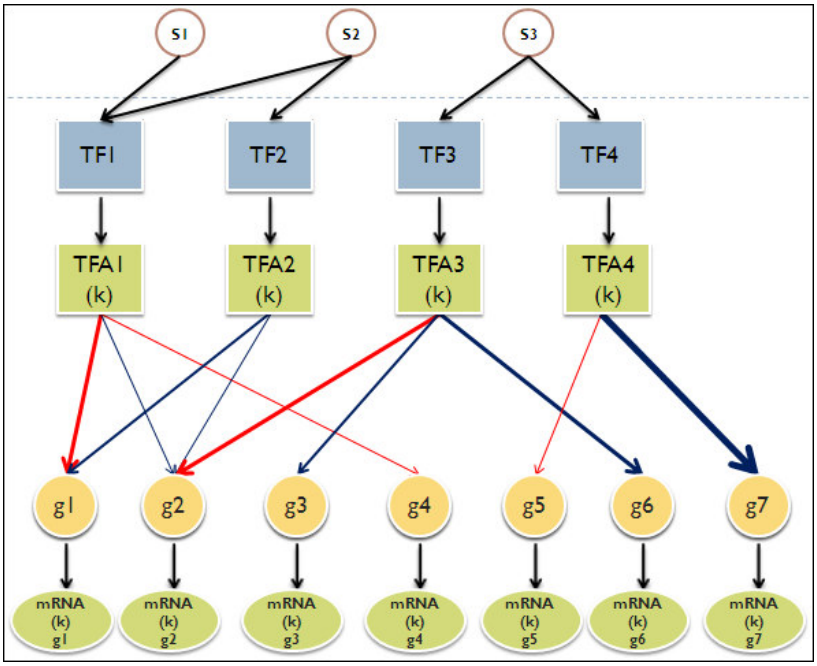


Figura 9: Representación esquemática en forma de red del proceso de regulación transcripcional.
Fuente: Elaboración propia.

En dicho esquema, las señales ambientales en un experimento k interactúan con los $TF's$ modificando su actividad. A la vez estos actúan con los genes como se ha explicado antes, produciendo sus transcritos en una cierta concentración. De esta forma tenemos un sistema en forma de red en el que se pueden identificar las siguientes partes principales y de interés:

- Una salida (output) correspondiente a la concentración del $mRNA$ de los genes para un experimento en particular.
- Un proceso, que indica las conexiones entre $TF's$ y genes, el tipo de influencia (activación o inhibición) y el grado o fuerza de esta, independiente de que experimento se considere[11].
- Una señal de regulación representada por la actividad de los $TF's$ para el experimento estudiado, que traduce las señales ambientales. Un estado o señal del proceso.[12]

Observando la expresión correspondiente a la Ecuación 8 es posible distinguir varias cosas. En primer lugar, el lado izquierdo de la ecuación correspondiente a la matriz E indica la concentración relativa de mRNA para cada combinación gen-experimento. Esto corresponde a la salida de la red antes descrita. La matriz A del lado derecho guarda toda la información correspondiente a las conexiones (proceso) que se dan entre los TF's y genes. De acuerdo a como se ha definido el parámetro CS_{ij}, la entrada A_{ij} indicará si el gen i está conectado con el TF j; si dicho valor es cero no hay conexión, mientras que cualquier valor

[11] Este punto es importante. Si bien la actividad de los $TF's$ variará de un experimento a otro, los parámetros de la matriz A que indican las conexiones, son constantes a lo largo de los mismos.
[12] Se destaca una vez más que dichas actividades variarán de un experimento a otro, indicando de alguna forma el estado de sistema.

diferente indicará el tipo y fuerza de la misma. Finalmente la matriz P, que resume la actividad relativa de cada TF en cada experimento corresponde al estado o señal de regulación del proceso.

De esta manera en A y P se resume toda la información respecto a la red de regulación transcripcional que genera el perfil de expresión genética observado en los diferentes experimentos. Así es posible identificar un problema inverso como el descrito en la introducción, en el que dado un output de un sistema, se busca reconstruir el sistema que genera dichos datos, estimando los parámetros representativos de éste; las señales de regulación de los $TF's$ en los diferentes experimentos, y la información de la conexión de la red o proceso. En este caso particular, se busca descomponer la matriz E de expresión genética en las matrices A y P, que estimen de forma adecuada los parámetros reales.

2.4 Problema inverso de reconstrucción

El objetivo es resolver el problema inverso descrito con anterioridad. El problema directo es trivial: conocido el proceso (las conexiones, y el tipo y grado de estas) y las señales de regulación que se desencadenan en cada experimento consecuencia de las señales ambientales, es posible obtener mediante las ecuaciones que rigen el proceso la matriz de expresión genética E, que indica la concentración de $mRNA$ de cada gen en cada experimento. El problema inverso consiste en reconstruir los parámetros que rigen el proceso, dada la matriz medida E. Una forma de enfocarlo es resolviendo el siguiente problema, en que se busca minimizar el ajuste de las estimaciones de A y P a los datos.

$$\mathbb{P}: \quad \min_{A,P} J(E - A \cdot P) \; = \min_{A,P} J(\Gamma)$$

Ecuación 9

, donde $\Gamma = E - A \cdot P$ es el residuo o grado de ajuste, y $J(\cdot)$ es alguna función o norma adecuada al problema[13].

2.5 Enfoque de reconstrucción NCA: Network Component Analysis

NCA[14] es un enfoque de reconstrucción, utilizado con el fin de entregar una estimación de los parámetros de una red como la anterior. Descomponer E en las 2 matrices que contengan dicha información de acuerdo al problema $\mathbb{P}$ descrito en la Ecuación 9 no es un asunto trivial, ya que existirán infinitas posibilidades sin restricciones adicionales. En efecto, sean A y P tal que:

$$\{A, P\} = argmin(\mathbb{P})$$

$$E - A \cdot P = \Gamma$$

Se puede considerar entonces cualquier matriz invertible X, de forma tal que:

[13] Si se quiere, es equivalente a descomponer E en las matrices A y P, minimizando su diferencia.
[14] Siglas de Network Component Analysis.

$$\bar{A} = A \cdot X$$

$$\bar{P} = X^{-1} \cdot P$$

Y dado que:

$$\bar{A} \cdot \bar{P} = A \cdot X \cdot X^{-1} \cdot P = A \cdot P$$

, $\bar{A}$ y $\bar{P}$ descomponen a E de la misma manera (con el mismo residuo), y las posibilidades son por ende infinitas. En el trabajo de J.C. Liao [26] [34] se derivan algunas condiciones de suficiencia de información, que mediante restricciones del tipo estructural en A y P aseguran una descomposición de E esencialmente única (en el sentido que se verá en lo siguiente).

Las restricciones en A y P se basan en información a priori (parcial) que se puede obtener de la literatura o de otras fuentes, y que restringen en espacio de soluciones de la reconstrucción. Una vez se tiene claro cuáles son los $TF's$ y genes relevantes, así como un set de experimentos validos (y por ende se conoce la dimensión de las matrices) se imponen una serie de restricciones a la estructura de la red resumida en A, indicando de manera parcial que genes no se relacionan con un TF particular; al CS_{ij} correspondiente se le asigna el valor de 0. De la misma manera, es posible restringir alguna de las entradas de P a 0, asociándolas a experimentos knock-out[15] en las que se tiene certeza que la actividad del TF correspondiente no varía de un experimento a otro. El objetivo detrás de este procedimiento es que los 0 impuestos en las matrices A y P, basados en información a priori, no necesitan ser estimados, y por ende, disminuyen los grados de libertad del problema. En base a lo anterior, y con el fin de definir con claridad el problema, se definirían los siguientes conceptos. Sean los sub-espacios vectoriales siguientes, que resumen de cierta manera los ceros impuestos a las estructuras de las matrices buscadas.

$$Z_A = \left\{A \in \mathbb{M}^{NxL}(\mathbb{R}) \;/\; A_{ij} = 0 \; para \; unos \; ciertos \; (i,j) \; dados\right\}$$

$$Z_P = \left\{P \in \mathbb{M}^{LxM}(\mathbb{R}) \;/\; P_{jk} = 0 \; para \; unos \; ciertos \; (j,k) \; dados\right\}$$

Ecuación 10

Luego, el problema de reconstrucción modificado es encontrar A y P que pertenezcan a dichos conjuntos, y que minimicen algún tipo de función del residuo $(E - A \cdot P)$. Esto es:

$$\mathbb{P} = \min_{A,P} \; J(E - A \cdot P)$$

$$s.t. \;\; A \in Z_A \;\; y \;\; P \in Z_P$$

Ecuación 11

Así:

$$\{A, P\} = argmin(\mathbb{P})$$

[15] Un experimento en el que se ha alterado genéticamente al organismo y por lo tanto se tiene certeza del nivel de actividad relativo nulo de algún TF.

En [26] y [34] se deriva el siguiente teorema, que como ya se ha mencionado impone condiciones necesarias y suficientes en A y P de forma que la solución al problema anterior sea esencialmente única.

Teorema 1 (Solución esencialmente única de $\mathbb{P}$).

Dada una matriz E, y ciertas restricciones a la estructura de la red dadas por los conjuntos Z_A y Z_P, las condiciones necesarias y suficientes para una descomposición esencialmente única de acuerdo a $\mathbb{P}$ son:

1. $A \in Z_A$ tiene rango columna completo.
2. Cada matriz reducida de G, G_{rj} $(j = 1, \dots, L)$ posee rango $L - 1$.
3. $P \in Z_P$ tiene rango fila completo.

La matriz G y sus matrices reducidas son definidas como sigue:

$$G = \begin{pmatrix} A_{c1} & A & 0 & \dots & 0 \\ A_{c1} & 0 & A & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ A_{cL} & 0 & 0 & 0 & A \\ P_{c1}^t & & & & \\ \vdots & Q(P_{c1}^t) & Q(P_{c2}^t) & \dots & Q(P_{cL}^t) \\ P_{cL}^t & & & & \end{pmatrix}$$

Ecuación 12

, donde M_{ci} es la i-esima columna de la matriz M y M_{ci}^t es la i-esima columna de la matriz M^t(la traspuesta de M). La matriz $Q(v)$ es una matriz diagonal en bloque, en que cada bloque es el vector v. En la definición de G lo importante es respetar el patrón de ceros de A y P. En las otras entradas (las libres a ser estimadas) puede colocarse 1's o valores aleatorios. Las matrices $G_{rj}(j = 1, \dots, L)$ se obtiene eliminando en G todas las filas que tengan un cero en la primera columna. Así:

$$G_r = \begin{pmatrix} 1 & L & & & \\ & & & & \\ \vdots & G_{r1} & G_{r2} & \dots & G_{rL} \\ 0 & & & & \end{pmatrix}$$

Ecuación 13

En los resultados se discutirán las implicancias matemáticas y biológicas de dichas condiciones.

La descomposición de E en A y P de acuerdo a $\mathbb{P}$ poseerá solución esencialmente única, en el sentido de que dos soluciones diferirán solo en una matriz diagonal. Así, si $\{A_1, P_1\} = argmin(\mathbb{P})$ y encontramos otra solución $\{A_2, P_2\}$ tal que $E - A_1 \cdot P_1 = \Gamma = E - A_2 \cdot P_2$, entonces se tendrá necesariamente que:

$$A_2 = A_1 \cdot X^{-1}$$

$$P_2 = X \cdot P_1$$

, con X una matriz diagonal.

Es posible resumir el problema anterior y sus implicancias en 2 problemas equivalentes, uno de reconstrucción y otro de estimación, que pasarán a definirse en lo siguiente. Ambos enfoques son 2 caras de una misma moneda, y serán utilizados indistintamente a lo largo del documento.

2.5.1 Enfoque de reconstrucción

El enfoque de reconstrucción consiste en reconstruir los parámetros de la red de regulación[16] mediante un enfoque de problema inverso, en el cual los datos entregan la información que permite descomponer la matriz de datos E en las matrices A y P, de forma que estas últimas se ajusten de la mejor forma posible a los datos. Se aprecia que el hecho de descomponer la matriz de datos en estas 2 matrices, es sólo consecuencia del modelo linealizado que se ha utilizado, y no de los problemas inversos en general.

2.5.2 Enfoque de estimación de parámetros

El problema propuesto puede ser visualizado también como un problema de estimación de parámetros, que será precisamente el enfoque propuesto para desarrollar la estrategia de resolución general. No se debe confundir conceptos; el problema tratado sigue siendo uno del tipo inverso. Lo comentado en esta sección simplemente busca aclarar que la problemática, a un nivel matemático, puede ser descompuesta en distintos problemas análogos a los que se resuelven con técnicas de estimación de parámetros. MCO[17] por ejemplo.

Analizando el problema general dado por la Ecuación 9, se hace notar que el mismo puede ser escrito de la siguiente manera, utilizando como norma de ajuste una genérica.

$$\mathbb{P}: \quad \min_{A,P} \|E - A \cdot P\|$$

Lo anterior puede ser descompuesto en M problemas independientes, cada uno de los cuales corresponde a uno de las columnas de E. Esto es:

$$\mathbb{P}_k: \quad \min_{A,P} \|E_{ck} - A \cdot P_{ck}\| \qquad k = \{1, \dots, M\}$$

Ecuación 14

, donde E_{ck} y P_{ck} corresponden a la columna k de E y P respectivamente. Los problema anteriores, escritos de dicha manera, son perfectamente homologables a uno de estimación de parámetro, en el que es posible utilizar MCO para obtener estimadores de buenas características de P_{ck} dado la variable dependiente E_{ck} [35]. Por supuesto en este caso A no es una matriz de regresores fijos, y tampoco

[16] A saber, los pesos de la red que relaciona genes y $TF's$, y la actividad de los mismos en los diferentes experimentos considerados.
[17] MCO (Mínimos Cuadrados Ordinarios)

estocástica, sino más bien incógnitas a ser estimadas. Luego, el problema es algo más complicado que lo anterior, pero aun posible de resolver mediante dicha estrategia, como se verá a continuación:

2.6 Implementación numérica básica

2.6.1 Función de minimización

Si bien el teorema anterior asegura una descomposición esencialmente única, no provee una estrategia para resolver el problema directamente. Más aun, no es posible encontrar una solución analítica, y una solución numérica, como MCO no se puede utilizar directamente debido a que ambas matrices A y P son desconocidas. Aún así es posible utilizar un algoritmo que hace uso de la teoría base de MCO, en un proceso de optimización alternada.

La función de minimización es básica al respecto, por lo que es necesaria definirla correctamente. Para este caso y en adelante, se utilizará la norma Frobenius del residuo[18], correspondiente a la siguiente expresión:

$$\|E - A \cdot P\|_F = \|\Gamma\|_F = \sum_i \sum_k \sqrt{(E - A \cdot P)_{ik}{}^2}$$

En el caso anterior se está trabajando con matrices. Si v es un vector, entonces la norma Frobenius es equivalente a:

$$\|v\|_F{}^2 = \sum_i v_i{}^2 = v^t \cdot v$$

2.6.2 Normalización de resultados

Al ser la solución al problema $\mathbb{P}$ esencialmente única[19], existirán diferentes soluciones que se diferenciarán solo en una matriz diagonal X. Dicha salvedad se puede solucionar normalizando los resultados, a fin de obtener una solución única. El criterio de normalización es arbitrario, y el escogido será normalizar la matriz A por una matriz tal que el promedio absoluto de las entradas no nulas de cada columna sea igual a la unidad. En base a lo propuesto por J.C. Liao en su documento original, X será definida como:

$$X_{jj} = \frac{1}{n} \sum_{i=1}^{N} |A_{ij}|$$

Ecuación 15

[18] La raíz de la suma de las entradas al cuadrado.

[19] Bajo el supuesto que se cumplan los criterios NCA por supuesto en relación a la estructura impuesta a las matrices buscadas.

, donde n es el número de entradas no nulas en la columna respectiva.

2.6.3 Algoritmo de optimización bi-lineal alternada

El método propuesto es similar al utilizado en el documento original NCA [26], que hace uso de la convexidad de la función de optimización. Sin embargo, a fin de implementar un método general para las modificaciones propuestas, se hará uso de la noción de gradiente de la función objetivo. En la Figura 10 se representa esquemáticamente una curva de iso-errores, conformada por todas las combinaciones de matrices A y P que generan un residuo mínimo para una matriz de datos determinada (obviamente todas difieren en una matriz diagonal, y al normalizar se llegará a la solución $\underline{A}$ y $\underline{P}$ mostrada).

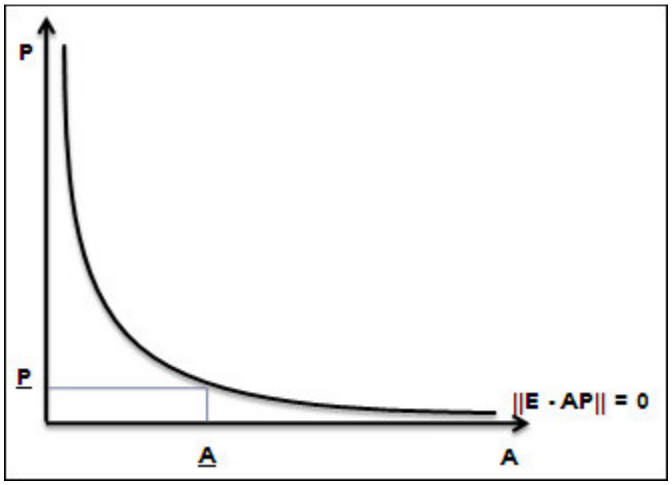


Figura 10: Representación gráfica curvas de iso-error.
Fuente: Elaboración propia.

Partiendo desde un punto cualquiera en este espacio, el objetivo es converger de forma alternada al mínimo del problema. La estrategia general consiste en partir desde un punto cualquiera del espacio (Z_A, Z_P) y moverse de forma alternada, en dirección del gradiente hacia el mínimo del funcional[20]. El movimiento en forma alternada es necesario, dado que tanto A como P son incógnitas. Este procedimiento se refiere a dejar fija alguna de dichas matrices (A por ejemplo) y encontrar un P que minimice:

$$\min_{P} \|E - A \cdot P\|_F^{\,2}$$

$$s.t. \;\; P \in Z_P$$

Ecuación 16

[20] Local o global como se verá luego.

, donde E y A se consideran constantes. Luego, tomando el P que minimiza el problema anterior fijo, se encuentra un A que resuelve el mismo problema anterior. El procedimiento se repite alternadamente para ambas variables, generando una sucesión P^n y A^n que se espera converjan a las matrices soluciones de $\mathbb{P}$. Esto se resume esquemáticamente en la Figura 11.

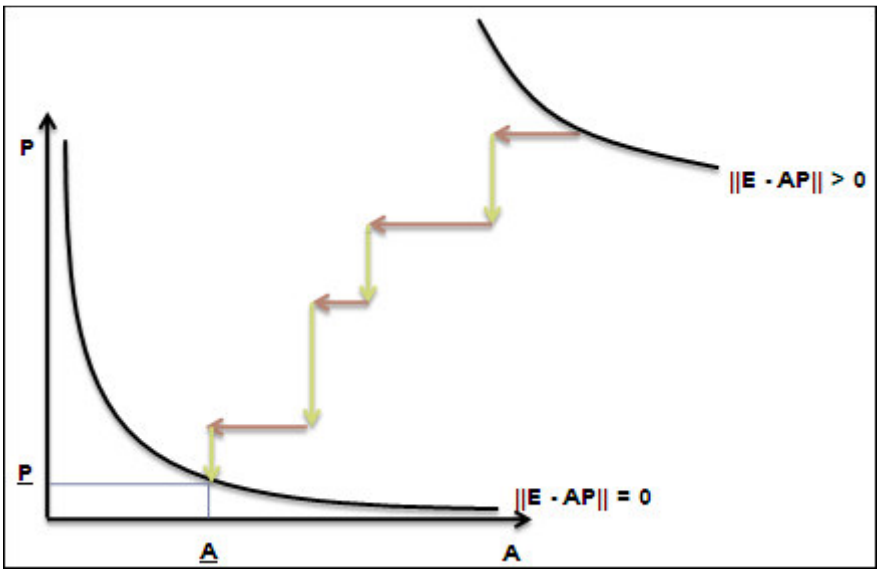


Figura 11: Método de optimización bi-lineal alternada.
Fuente: Elaboración propia.

Un punto importante a considerar son las dimensiones de las matrices en el problema descrito por la Ecuación 16. E y P no son matrices columnas, por lo que el problema es algo más complejo que una estimación de parámetros vía MCO. De todas maneras, dicha salvedad es fácilmente solucionable si se nota que cada columna de E se obtiene multiplicando la matriz A por la misma columna de la matriz P, como se vio antes. Luego se tendrán M problemas autónomos (uno para cada columna de E) que pueden ser resueltos independientemente. Cada uno de estos problemas será del tipo:

$$\min_{x} \|b - A \cdot x\|_F^{\,2}$$

Ecuación 17

, donde b y x son vectores columnas, y A una matriz de constantes de las dimensiones adecuadas[21]. Con el fin de ejemplificar, y siguiendo la notación que se usará, A será una matriz de dimensión (NxL). Lo que se busca en el problema anterior es ajustar $A \cdot x$ a la matriz b. Si A fuera invertible, la solución sería trivial, y correspondería a la formula analítica dada por:

[21] No confundir con la matriz A que define la red en estudio. En este caso, A simplemente es la matriz de constante del problema genérico $A \cdot x = b$.

$$x^* = A^{-1} \cdot b$$

Pero normalmente A no es invertible; en estos sistemas por lo general hay más ecuaciones que incógnitas ($N \gg L$), por lo que se busca un x, tal que $A \cdot x$ se ajuste en la mejor forma posible a los datos, en el sentido de minimizar la función objetivo (mínimos cuadrados en el caso de la Ecuación 17[22]). Para resolver este problema es posible utilizar variados métodos, como la factorización QR utilizada en el documento original. Notar además que no todas las entradas de A y P necesitan ser estimadas, por cuanto existen restricciones de ceros en algunas de ellas dadas las características de la red.

Lo anterior aplicado al problema original se resume en el siguiente algoritmo de resolución.

1. Se genera una adivinación inicial para los parámetros a estimar de A y P, de forma tal que estas pertenezcan a Z_A y a Z_P respectivamente: $A^{(0)}$ y $P^{(0)}$.
2. Dado $A^{(0)}$, encontrar $P^{(1)}$ que resuelva:

$$\min_{P^{(1)}} \left\| E - A^{(0)} \cdot P^{(1)} \right\|_F^2$$
$$s.t.\ P^{(1)} \in Z_P$$

   Esto se logra dividiendo el problema anterior en M problemas (uno para cada columna en E) de la forma ($k = 1, \dots, M$):

$$\min_{P_{Rck}{}^{(1)}} \left\| E_{ck} - A_{Rk}{}^{(0)} \cdot P_{Rck}{}^{(1)} \right\|_F^2$$
$$s.t.\ P_{Rcj}{}^{(1)} \in Z_P$$

   , donde E_{ck} es la columna k de E y P_{Rck} es la columna k de P reducida; esto es la columna k de P original, a la que se han eliminado las entradas restringidas a cero en Z_P (a fin de no estimarlas de nuevo). A_{Rk} corresponde a la matriz A reducida, a la cual se han eliminado las columnas correspondientes a las filas eliminadas en P_{Rck}.

   De esta manera, y dadas las condiciones NCA especificadas con anterioridad, los problemas anteriores tienen una solución $\boldsymbol{P_{Rck}}^{(1)}$ que puede ser obtenida vía factorización QR u otro método..

3. Dado $P^{(1)}$, encontrar $A^{(1)}$ que resuelva:

[22] Minimizar la norma F es equivalente a minimizar la norma 2, que sería equivalente a la suma de los cuadrados para la cual está definida MCO. La equivalencia viene de notar que la función raíz cuadrada es una función creciente.

$$\min_{A^{(1)}} \left\| E - A^{(1)} \cdot P^{(1)} \right\|_F^2$$
$$s.t.\ A^{(1)} \in Z_A$$

Esto es equivalente a resolver el problema siguiente, el que convenientemente está escrito en la misma forma que la Ecuación 17.

$$\min_{A^{(1)}} \left\| E^t - {P^{(1)}}^t \cdot {A^{(1)}}^t \right\|_F^2$$
$$s.t.\ A^{(1)} \in Z_A$$

La estrategia de resolución es similar, dividiendo el problema anterior en N problemas (uno para cada columna en E^t) de la forma ($i = 1, \dots, N$):

$$\min_{{A_{Rri}}^{(1)}} \left\| {E_{ri}}^t - {P_{Ri}}^{(1)^t} \cdot {A_{Rri}}^{(1)^t} \right\|_F^2$$
$$s.t.\ {A_{Rri}}^{(1)} \in Z_P$$

, donde ${E_{ri}}^t$ es la traspuesta de la fila i de E y ${A_{Rri}}^{(1)^t}$ es la traspuesta de fila i de A reducida; esto es la traspuesta de la fila i de A original, a la que se han eliminado las entradas restringidas a cero en Z_A. ${P_{Ri}}^{(1)^t}$ corresponde a la matriz P traspuesta reducida, a la cual se han eliminado las columnas correspondientes a las entradas eliminadas en ${A_{Rri}}^{(1)^t}$.

De esta manera, y dadas las condiciones NCA especificadas con anterioridad, los problemas anteriores tienen una solución $\boldsymbol{A_{Rri}}^{(1)}$ que puede ser obtenida vía factorización QR.

4. Se repite el paso 2. y 3. Hasta que el funcional evaluado en $A^{(n)}$ y $P^{(n)}$, $\left\| E - A^{(n)} \cdot P^{(n)} \right\|_F^2$ sea menor que un cierto grado de tolerancia.
5. Finalmente las matrices A y P son normalizadas por una matriz X definida como en la Ecuación 15.

2.7 Metodología de pruebas sintéticas

La metodología de pruebas sintéticas es la manera que se utilizará para testear y validar los diversos métodos creados y modificados. Dado que con la técnica planteada se busca reconstruir los parámetros de un modelo desconocido, este tipo de prueba permite corroborar aspectos del funcionamiento, a fin de identificar fortalezas y debilidades, así como solucionar diversos problemas que se planteen. De la misma forma, dichas pruebas permiten mejorar el "arte" y la experiencia e intuición necesarias en este tipo de aplicaciones, de forma tal que conociendo la realidad y características de un problema determinado, aplicar la mejor combinación de técnicas que permita resolverlo de la manera más precisa posible.

La estructura general de este enfoque, que se utilizará a lo largo del resto de este documento, consiste en generar aleatoriamente los parámetros de las matrices P y A, esto es la actividad o señales de regulación del set de $TF's$ en estudio en diversos experimentos, y los parámetros que definen la estructura de la red de regulación. Una vez escogida la dimensión del proceso (número de genes, $TF's$ y experimentos), se hace uso de una función programada en Matlab que genera los parámetros procurando que las redes construidas sean compatibles con el criterio NCA y los valores de la matriz P tengan fundamento en el sentido biológico.[23] A modo de ejemplo, las siguientes matrices han sido generadas aleatoriamente, representando la red que se muestra en la Figura 12.

A =

$$\begin{bmatrix} 0 & 0 & 1.1080 \\ 0.7600 & 4.0111 & 1.0538 \\ 0 & -2.6656 & 0 \\ 0.1686 & 3.0089 & -0.0267 \\ 0 & 0 & 1.8649 \\ 2.7439 & 0 & 0 \\ 0.3276 & -0.3545 & 0 \end{bmatrix}$$

P =

$$\begin{bmatrix} -0.0227 & 0.1324 & -0.0558 & -0.0074 \\ -0.0773 & 0.0338 & -0.0017 & -0.0698 \\ 0.1263 & 0.0838 & 0.0346 & 0.0942 \end{bmatrix}$$

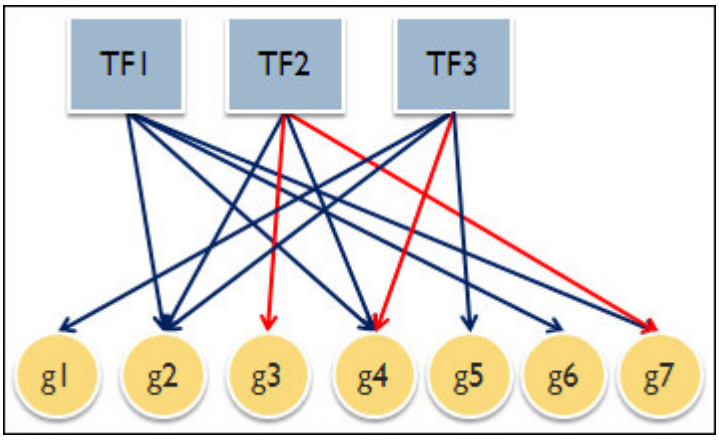


Figura 12: Red de regulación transcripcional generada sintéticamente.
Fuente: Elaboración propia.

[23] Debido a que los cambios en actividad de un experimento a otro no son a gran escala, el logaritmo de la actividad relativa estará por lo general en valores cercanos a cero.

Por supuesto en dicho diagrama no se representa las señales de regulación de cada TF en los experimentos considerados y resumidos en la matriz P, y tampoco el grado de influencia representado por el grosor de la línea de conexión.

El siguiente paso del procedimiento, teniendo el proceso completo especificado, es generar la matriz de expresión genética (concentración relativa de los $mRNA$ respectivos en cada experimento) para los valores considerados; esto es, multiplicar la matriz A y P de acuerdo al modelo propuesto. Dicha matriz se representa en lo siguiente.

$$E = \begin{bmatrix} 0.1400 & 0.0929 & 0.0384 & 0.1043 \\ -0.1942 & 0.3245 & -0.0129 & -0.1864 \\ 0.2060 & -0.0901 & 0.0047 & 0.1861 \\ -0.2398 & 0.1218 & -0.0156 & -0.2138 \\ 0.2356 & 0.1563 & 0.0646 & 0.1756 \\ -0.0624 & 0.3633 & -0.1530 & -0.0203 \\ 0.0200 & 0.0314 & -0.0176 & 0.0223 \end{bmatrix}$$

Cabe destacar que dicha matriz se obtendría vía experimentación[24], y por ende los valores exactos que se obtienen por el modelo son los que se obtendrían en un experimento 100% exacto. Como antes, cada columna de E representa una condición experimental distinta, en donde se resumen los niveles de expresión génicas dados por la Ecuación 7 para cada gen. La matriz A es independiente del experimento (es lo que resume el proceso en sí), y por último la matriz P entrega en cada columna las señales individuales de los $TF's$ en cada experimento.

El enfoque de pruebas sintéticas consiste en partir de la matriz E (ya sea la versión exacta mostrada con anterioridad o perturbaciones que simulen errores de medición) e intentar reconstruir las matrices A y P. Los resultados son luego contrastados con los valores reales a fin de corroborar la exactitud del método, errores obtenidos y visualizar el comportamiento de la reconstrucción frente a la alteración de diversas variables.

2.7.1 Algunos detalles sobre el enfoque de pruebas sintéticas

El enfoque de validación y testeo antes propuesto tiene una serie de detalles que es necesario considerar al momento de crear las redes sintéticas, estas se generan considerando una cierta densidad de ceros dentro de la matriz A, esto es, la densidad de conexiones existentes entre $TF's$ y genes. Como se puede corroborar en la literatura, la densidad de ceros en redes de regulación es de un 75-80%, por lo que una vez creadas redes completas (donde no existan $TF's$ sin parejas), los 2 primeros criterios NCA se cumplen fácilmente [36, 37]. El 3° criterio requiere una condición crítica, en donde el número de

[24] Microarrays por ejemplo.

experimentos M debe ser mayor o igual que el número L de $TF's$. En redes grandes esta condición es en extremo difícil de conseguir (luego se discuten en más detalle las razones físicas y prácticas), por lo tanto, con un afán netamente funcional, en las redes generadas se considerará un experimento más que $TF's$. De esta forma se asegura la unicidad de la solución vía NCA. Por otra parte la red original vendrá normalizada, a fin de poder compararla satisfactoriamente con los datos reconstruidos (que por defecto vienen normalizados). Un punto importante respecto al enfoque descrito es el tipo de error utilizado para comparar los valores reales con la reconstrucción. Esto se detalla a continuación:

2.7.2 Medición de errores en pruebas sintéticas[25]

Sean A_{on} (A real normalizada) ∈ M^{NxL} y $P_{on} \in M^{LxM}$ las matrices reales y normalizadas que se obtienen al crear una red sintética. Sea además E_r la matriz de expresión genética real (la obtenida teóricamente bajo condiciones experimentales 100% precisas en la red anterior) y E_m[26] la versión medida (generalmente con error agregado a las coordenadas simulando el error experimental de medición). Se llamará A y P a las versiones reconstruidas utilizando NCA y la información de E_m (esta puede ser E_r o una versión con error). Dado que al considerar diferentes redes, estas pueden tener tamaños bastante dispares, no es clara la forma en que se puede comparar el error entre las matrices reales y las reconstruidas. De forma general sea $X_r \in M^{RxS}$ la versión real de una matriz genérica, y X la versión reconstruida o medida. Se definen las siguientes medidas de error de X en relación a su correspondiente real.

1. **Error cuadrático medio (MSE[27]).** Sea n el número total de entradas no nulas, o más exactamente, el número total de entrada de la cual no se tiene certeza en X. Este error está definido como:

$$e(Xr, X) = \frac{1}{n} \cdot \sum_{r=1}^{R} \sum_{s=1}^{S} (X_{r_{rs}} - X_{rs})^2$$

Ecuación 18

2. **Matriz de errores porcentuales.** Corresponde a los errores porcentuales por cada entrada de X. La coordenada (r, s) de la matriz de errores porcentuales se define como:

$$E(X_r, X)_{rs} = 100 \cdot \left| \frac{X_{r_{rs}} - X_{rs}}{X_{r_{rs}}} \right|$$

Ecuación 19

[25] Error relativo referido a la comparación de los valores reconstruidos y reales para redes sintéticas.
[26] O indistintamente E_p cuando se habla de una matriz de datos medida promedio, como se verá más adelante.
[27] MSE: Mean Square Error.

3. **Error porcentual medio.** Corresponde al promedio de las entradas de $E(Xr, X)$. Sea como antes n el número total de entradas no nulas. El error porcentual medio es definido como:

$$p(X_r, X) = \frac{1}{n} \cdot \sum_{r=1}^{R} \sum_{s=1}^{S} E(X_r, X)_{rs}$$

Ecuación 20

Por otra parte, las expresiones anteriores se aplicarán en el cálculo de los siguientes errores de interés:

I. Error de reconstrucción de A respecto a A_{on}.
II. Error de reconstrucción de P respecto a P_{on}.
III. Error de medición de E_m respecto a E_r.
IV. Error de ajuste a los datos medidos de E_m respecto a $A \cdot P$ (o matriz E reconstruida).
V. Error de ajuste a los datos reales de E_r respecto a $A \cdot P$ (o matriz E reconstruida).

Los 2 últimos errores se refieren simplemente a un *data fitting*, o sea, la norma F de la diferencia de ambas expresiones.

Oro punto importante respecto a la medición de errores es la concentración relativa de las diferentes entradas en la matriz de error porcentual. La concentración del error de una entrada se definirá como el porcentaje del total del error que es atribuido a dicha coordenada.

Sea ϵ el error total de una matriz X respecto a X_r, esto es, la suma de las entradas de $E(X_r, X)$. La matriz de concentración de errores se define por la siguiente expresión.

$$C(X_r, X)_{rs} = 100 \cdot \frac{E(X_r, X)_{rs}}{\epsilon}$$

Ecuación 21

2.8 Programación en MatLab

MatLab fue utilizado extensamente como la plataforma que permitió programar los métodos de reconstrucción y realizar todas las pruebas y simulaciones detalladas en este documento. En los anexos se resume una lista de las funciones de mayor relevancia creadas, su uso y función.

3. Resultados y Discusiones

En el siguiente capítulo, el más extenso del presente documento se describe y discuten todos los aspectos relevantes a la temática, los aportes desarrollados y las diferentes pruebas que se utilizaron con el fin de testear los resultados. El desarrollo será más bien cronológico, y en los resultados y discusiones de algunas pruebas o aspectos del problema, se motivará el desarrollo de otros puntos de acción comentados de la misma manera más adelante. Quizás esto pueda resultar un poco confuso, pero se piensa que es realmente necesario destacar el resultado de algunas pruebas como el *gatillante* de otras distintas, o del desarrollo de nuevos enfoques en base a la necesidad de resolver problemas identificados.

3.1 Problemas mal condicionados y otros problemas

Un problema que surge inherente a los problemas inversos, es la existencia de problemas mal condicionados, y no solo eso, el problema genérico propuesto en la Ecuación 9 y Ecuación 11 no es un problema lineal[28] por lo que la existencia de mínimos locales es otro asunto con el cual será necesario lidiar. En esta sección se dedicarán algunas palabras al que se piensa es el mayor de los 2 problemas: La existencia de matrices mal condicionadas en los algoritmos iterativos, que pueden desestabilizar la solución.

En cualquier tipo de problema inverso, se dice que este está bien condicionado en el sentido de Hadamard si cumple las siguientes características [38]:

1. Hay existencia de una solución.
2. Hay unicidad de la solución.
3. Hay una dependencia continua de la solución con respecto a los datos.

En el caso del problema tratado, los 2 primeros puntos se cumplen si se consideran correctas las suposiciones del criterio NCA, a pesar de existir solamente una solución esencialmente única. El 3° punto en cambio no es algo trivial, y es lo que a lo largo del documento se referirá al mal condicionamiento. No existe una dependencia continua de los datos cuando en un problema inverso del tipo ya descrito, pequeñas variaciones en los datos medidos[29] suponen grandes variaciones en las matrices reconstruidas. Si este es el caso, errores en la medición de los datos acarrearán grandes variaciones en la estimación de los parámetros, y por lo tanto el método será poco robusto. Existen por cierto algunas herramientas para evitar o disminuir este problema, que serán discutidas en otra sección.

[28] Pertenece a la familia de problemas log-lineales.
[29] Específicamente en la matriz E.

3.2 Enfoque de resolución basado en el gradiente funcional.

Como se vio en el punto 2.6.3 Algoritmo de optimización bi-lineal alternada, el problema básico de NCA que involucra descomponer E en las matrices A y P, se resuelve minimizando el funcional dado por la Ecuación 11 y optimizando alternadamente respecto a las matrices incógnitas. Cada iteración, como se vio, puede ser asociada a un problema del tipo descrito en la Ecuación 17, donde la solución (que no es trivial dado que generalmente el sistema resulta ser sobreestimado) puede ser obtenida numéricamente utilizando factorización QR u otro método. En lo siguiente se resumirá un método de optimización basado en el gradiente funcional que permite obtener una expresión para el mínimo del problema que es fácilmente programable en este problema básico, y en otros más complicados detallados luego.

Nótese que la Ecuación 17 es equivalente a la siguiente expresión:

$$\min_{x} \ (b - A \cdot x)^t \cdot (b - A \cdot x)$$

Ecuación 22

El funcional se puede descomponer como sigue:

$$(b - A \cdot x)^t \cdot (b - A \cdot x) = (b^t - x^t \cdot A^t) \cdot (b - A \cdot x) = b^t b - b^t Ax - x^t A^t b + x^t Ax$$

Pero como $b^t Ax$ es una constante, $(b^t Ax)^t = x^t A^t b$. Así se tiene:

$$f(x) = (b - A \cdot x)^t \cdot (b - A \cdot x) = b^t b - 2b^t Ax + x^t Ax$$

Ecuación 23

Finalmente es posible obtener el gradiente de la función a minimizar siguiendo las reglas de derivación de vectores, formas lineales y formas cuadráticas detalladas en los anexos.

$$\frac{df(x)}{dx} = \nabla f(x) = 0 - 2A^t b + 2A^t Ax$$

Ecuación 24

El mínimo estará dado en donde el gradiente vale cero, esto es:

$$\nabla f(x^*) = 0$$

$$A^t Ax^* = A^t b$$

$$\boldsymbol{x^* = (A^t \cdot A)^{-1} A^t \cdot b}$$

Ecuación 25

Notemos que aquí la expresión $A^t A$ es una matriz cuadrada, por lo que al cumplirse las condiciones NCA, la inversa existirá sin problemas.

Calculando la segunda derivada del funcional, se ve que efectivamente se trata de un mínimo.

$$\frac{d^2 f(x)}{d^2 x} = \mathcal{H} f(x) = 2A^t A$$

Ecuación 26

Que bajo ciertas condiciones generales es semi-definida positiva. Se hace notar además, que dado la forma del funcional, el resultado coincide con la expresión para los estimadores MCO [39]. Además, dadas las características del problema, en cada iteración será necesario resolver M de estos problemas para P y N para A, respetando por supuesto aquellas entradas en las que se impone la presencia de un cero (o equivalentemente, que la solución en cada iteración pertenezca a Z_A y a Z_P).

Lo anterior aplicado al problema original se resume en el siguiente algoritmo modificado.

1. Se genera una adivinación inicial para los parámetros a estimar de A y P, de forma tal que estas pertenezcan a Z_A y a Z_P respectivamente: $A^{(0)}$ y $P^{(0)}$.
2. Dado $A^{(0)}$, encontrar $P^{(1)}$ que resuelva:

$$\min_{P^{(1)}} \left\| E - A^{(0)} \cdot P^{(1)} \right\|_F^2$$
$$s.t.\ P^{(1)} \in Z_P$$

   Esto se logra dividiendo el problema anterior en M problemas (uno para cada columna en E) de la forma ($k = 1, \dots, M$):

$$\min_{P_{Rck}{}^{(1)}} \left\| E_{ck} - A_{Rk}{}^{(0)} \cdot P_{Rck}{}^{(1)} \right\|_F^2$$

$$s.t.\ P_{Rck}{}^{(1)} \in Z_P$$

   , donde E_{ck} es la columna k de E y P_{Rck} es la columna k de P reducida; esto es la columna k de P original, a la que se le han eliminado las entradas restringidas a cero en Z_P (a fin de no estimarlas de nuevo). A_{Rk} corresponde a la matriz A reducida, a la cual se le han eliminado las columnas correspondientes a las filas eliminadas en P_{Rck}.

   De esta manera, y dado las condiciones NCA especificadas con anterioridad, los problemas anteriores tienen una solución dada por la Ecuación 25:

$$\boldsymbol{P_{Rck}{}^{(1)} = \left(A_{Rk}{}^{(0)^t} \cdot A_{Rk}{}^{(0)} \right)^{-1} A_{Rk}{}^{(0)^t} \cdot E_{ck}}$$

3. Dado $P^{(1)}$, encontrar $A^{(1)}$ que resuelva:

$$\min_{A^{(1)}} \left\| E - A^{(1)} \cdot P^{(1)} \right\|_F^2$$
$$s.t.\ A^{(1)} \in Z_A$$

Esto es equivalente a resolver el problema siguiente, que conveniente, el que convenientemente está escrito en la misma forma que la Ecuación 17.

$$\min_{A^{(1)}} \left\| E^t - {P^{(1)}}^t \cdot {A^{(1)}}^t \right\|_F^2$$
$$s.t.\ A^{(1)} \in Z_A$$

La estrategia de resolución es similar, dividiendo el problema anterior en N problemas (uno para cada columna en E^t) de la forma ($i = 1, \dots, N$):

$$\min_{{A_{Rri}}^{(1)}} \left\| {E_{ri}}^t - {P_{Ri}}^{(1)^t} \cdot {A_{Rri}}^{(1)^t} \right\|_F^2$$
$$s.t.\ {A_{Rri}}^{(1)} \in Z_P$$

, donde ${E_{ri}}^t$ es la traspuesta de la fila i de E y ${A_{Rri}}^{(1)^t}$ es la traspuesta de la fila i de A reducida; esto es la traspuesta de la fila i de A original, a la que se le han eliminado las entradas restringidas a cero en Z_A. ${P_{Ri}}^{(1)^t}$ corresponde a la matriz P traspuesta reducida, a la cual se le han eliminado las columnas correspondientes a las entradas eliminadas en ${A_{Rri}}^{(1)^t}$.

De esta manera, y dado las condiciones NCA especificadas con anterioridad, los problemas anteriores tienen una solución dada por la Ecuación 17:

$$\boldsymbol{{A_{Rri}}^{(1)} = \left({P_{Ri}}^{(1)} \cdot {P_{Ri}}^{(1)^t} \right)^{-1} {P_{Ri}}^{(1)} \cdot {E_{ri}}^t}$$

4. Se repite el paso 2. y 3. Hasta que el funcional evaluado en $A^{(n)}$ y $P^{(n)}$, $\left\| E - A^{(n)} \cdot P^{(n)} \right\|_F^2$ sea menor que un cierto grado de tolerancia.
5. Finalmente las matrices A y P son normalizadas por una matriz X definida como en la Ecuación 15.

3.2.1 Sobre la utilización del gradiente funcional

Como comentario, es necesario destacar que el procedimiento puede ser utilizado con funcionales más complicados que el dado por la Ecuación 11, a los que en un principio no pueden aplicarse otros métodos como la factorización QR. Se verá a continuación que de acuerdo al tipo de reconstrucción que se quiera realizar (mediante qué procedimiento) o tipo de información a utilizar, la función de optimización puede adquirir formas realmente complicadas (nada parecido al sistema lineal anterior), en

las que en un principio no existe la claridad de cómo obtener el mínimo. Este método permite superar dicho inconveniente pudiendo adaptarse a casi cualquier función que se utilice.

3.3 Criterios NCA y su implicancia

En 2.5 Enfoque de reconstrucción NCA, se describió un teorema del documento original [33] que bajo ciertas condiciones que deben cumplir las estructuras de A y P asegura una descomposición de los datos esencialmente única en las matrices buscadas. La estructura de las matrices (no los datos que se quiere reconstruir) resume información a priori que se tiene del sistema en estudio, por lo que las condiciones NCA impuestas en el Teorema 1 son una suerte de suficiencia de información para la reconstrucción. Se analizará la implicancia física-biológica de cada una de las condiciones.

Rango completo columna de A

Esta primera condición es bastante básica e implica que el rango de A sea igual a L. Para corroborar el criterio se generan al azar datos para las entradas de la matriz, por lo que la condición se traduce en que las columnas de A deben ser linealmente independientes, y por ende que no se den algunos de los siguientes casos: En primer lugar, ninguna columna puede tener sólo ceros: esto se traduciría en que algún TF no controla genes, y por ende podría ser eliminado del análisis.

Además es necesario que no existan 2 TF que controlen a los mismos genes y con la misma magnitud. Si bien biológicamente es posible que 2 TF controlen a los mismos genes (por ejemplo un activador e inhibidor) es poco probable que lo controlen con la misma magnitud; en dicho caso, sería el mismo TF y se podría eliminar alguno de ellos en el análisis, y por ende dicha información repetida. Un caso particular es el de aquellos TF que controlan un solo gen, lo cual biológicamente es posible en una primera instancia. Esto implicaría que sin importar los valores, ambos TF serían linealmente dependientes y se perdería la primera condición NCA. En un principio dicho problema podría resolverse considerando ambos TF como uno solo (tan solo en este caso particular) que englobe el efecto conjunto de ambos.

Finalmente, es necesario que ninguna columna sea una combinación lineal de otras. Si bien no existen razones físicas o biológicas para que dicha condición se cumpla, es poco probable que se dé dada la gran densidad de ceros que tiene la matriz A.

Matrices reducidas de G, G_{rj} $(j = 1, \dots, L)$ deben tener rango $L - 1$

Esta es quizás la condición más extraña que impone NCA y que es necesario analizar en detalle para entender su significado y las instancias en que es violada. En primer lugar, es necesario destacar que la condición general que se discute en este punto es menos restrictiva que la siguiente[30]:

- Al eliminar un nodo de regulación en A (una columna) junto a todos los nodos de salida (eliminar las filas correspondientes a los genes que controla el TF que se está eliminando) el

[30] Si se cumple esta, se cumple la anterior.

sistema resultante debe tener rango $(L-1)$. Se llamarán a dichos sistemas matrices reducidas de A, A_{rj} $(j=1,\dots,L)$. Pueden verse como un conjunto de redes reducidas, en donde se ha eliminado un TF junto a todos los genes que éste controlaba. Esto se visualiza para la red ficticia mostrada en la Figura 13, en donde se representa la red inducida por la matriz A, y las correspondientes matrices reducidas para cada columna (TF) de esta.

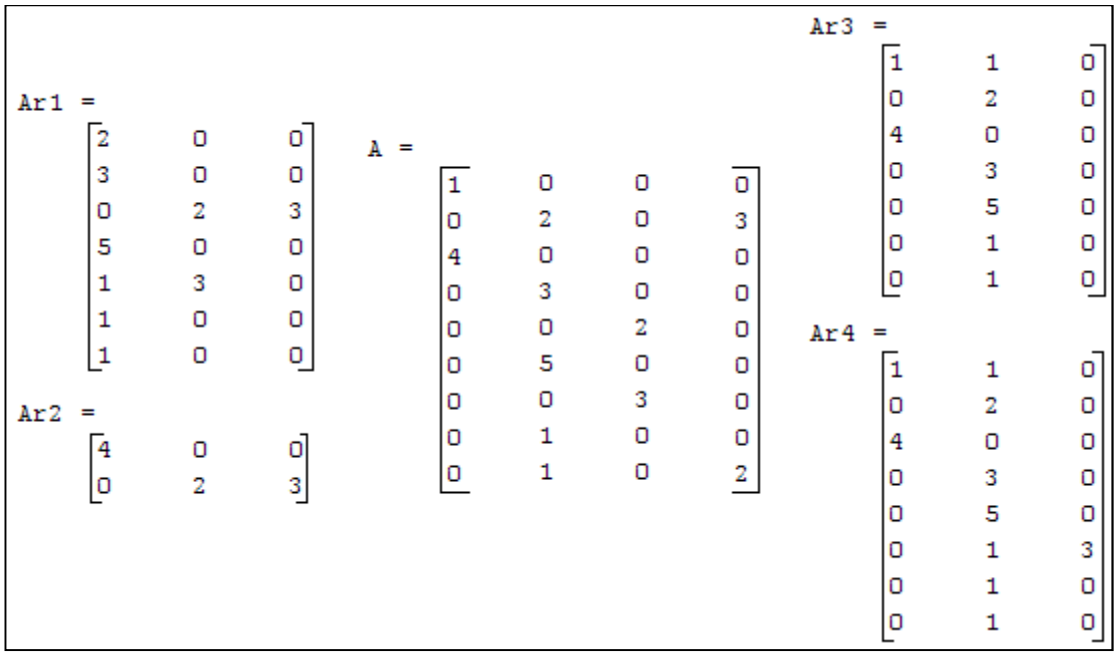


Figura 13: Matrices reducidas de A.
Fuente: Elaboración propia.

Los problemas surgen cuando el rango de cualquiera de las matrices reducidas es diferente a $(L-1)$. Esto puede deberse a 2 problemas con la red:

I. En una matriz reducida de A el número de genes puede ser menor que el de reguladores, y por lo tanto no podrá haber rango completo de columna. Es el caso observado en A_{r2} en donde existen 3 reguladores y 2 genes. Por supuesto, en el mejor de los casos en esta matriz el rango podría ser 2, y no 3 como sería necesaria para cumplir con el criterio NCA. Dicho problema se produce cuando un regulador controla una cantidad extensa de genes, por lo que al eliminar las filas correspondientes, la cantidad de genes en la matriz reducida es menor que la de reguladores. Más exactamente, si en un TF se tiene que el número de genes regulados (H) y el número de genes totales (N) es tal que: $N-H<L-1$ existirá un conflicto como el comentado. Biológicamente esto no debiese ser un problema, ya que en general $N \gg L$ y por ende es difícil dada la densidad de ceros identificada en A que la condición anterior se cumpla. No será normal que un TF controle demasiados genes.

II. Es posible que exista una columna de ceros en alguna matriz reducida. Esto se produce cuando los genes regulados por algún TF corresponden a un subconjunto de los

controlados por otro. Si esto es así, existirá una columna de ceros en la columna correspondiente a la que controla menos genes y en la matriz reducida del TF que controla a genes que engloban los del anterior. Es el caso de la matriz reducida A_{r3} de la figura anterior. Biológicamente no existe razón alguna para tal condición, ya que perfectamente es posible existan TF que actúen sobre los mismos genes. La condición tiene una necesidad matemática, por lo que será necesario cumplirla de una u otra forma para utilizar el método. Una posibilidad es eliminar el TF con conflicto (junto a los genes que controla por supuesto), y realizar el análisis con una nueva red que no tenga problemas. Si bien es una aproximación menos precisa que la obtenida al utilizar la red completa, permitiría obtener información de los demás TF's.

Otra opción consiste en hacer uso de experimentos knock-out, que como se ha comentado, imponen ceros en la matriz P[31]. En esta caso, no es posible utilizar las matrices reducidas de A, pero si la matriz G y su correspondiente reducida G_r. Si se observa la estructura de esta matriz, a partir de la segunda columna se encuentran las matrices G_{rj}, $j = 1, \dots, L$. Cada una de estas matrices por construcción, corresponden a la matriz reducida de A respectiva y debajo de esta, distintas columnas de P. Si P no posee restricciones de ceros, las matrices G_{rj} corresponderán a las mismas A_{rj} ya que solo habrá ceros bajo ellas, y estudiar el rango de cada sub-matriz de G_r es equivalente al desarrollo comentado con anterioridad. En el caso de que en P hallan restricciones de ceros debido a experimentos knock-out habrá un aporte a las matrices G_{rj}, que se expresarán como otras matrices bajo las matrices A_{rj}. El punto de interés es que al estudiar estas matrices G_{rj} sin experimentos knock-out, será lo mismo que estudiar el rango A_{rj}, y por ende se seguirá violando el método[32]. Si es que existen estas restricciones en P habrá un aporte extra, y puede darse el caso de que las G_{rj} que antes violaban el criterio, ahora lo cumplan. El objetivo es imponer condiciones en las filas de P que correspondan a los mismos $TF's$ que presentaban problemas (la columna que mostraba tan solo ceros al obtener alguna reducida de A).

III. P tiene rango completo de fila.

Principalmente se traduce en la condición necesaria[33] de que $M \geq L$. Si se cumple es en extremo probable, *a priori*, que se cumpla la condición. En estricto rigor la condición debiese ser comprobada *a posteriori*, una vez realizada la reconstrucción.

[31] Tal cual se hace con A.

[32] Estudiar ambas matrices entrega el mismo resultado si no existen restricciones en P.

[33] No suficiente, ya que a pesar de dicha condición, podría darse el caso de la existencia de filas linealmente dependientes.

3.4 Sobre la normalización de resultados

Como se comentó en la sección 2.6.2 Normalización de resultados, la normalización de los resultados de la reconstrucción es indispensable para obtener un resultado único y comparable con los valores reales de los parámetros a reconstruir (los que tienen que estar normalizados mediante el mismo criterio lógicamente). De la misma forma, el criterio de normalización es necesario a la hora de comparar 2 reconstrucciones, que de otra forma y pese a ajustarse igualmente bien a los datos, hubiesen sido numéricamente diferentes. Esto puede representarse esquemáticamente en la siguiente figura.

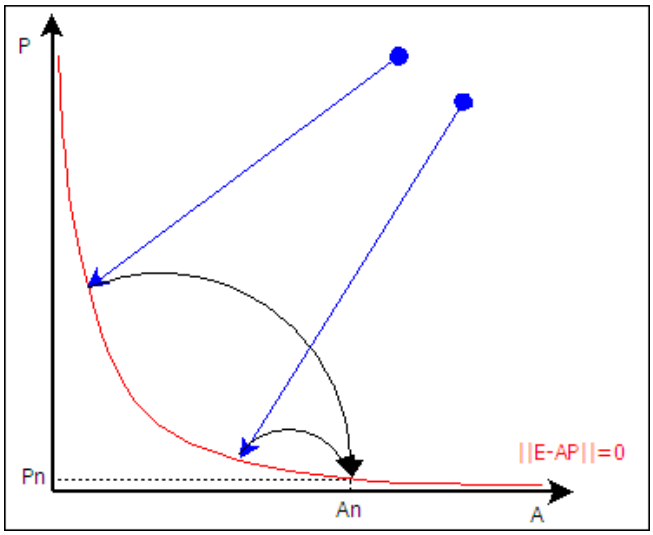


Figura 14: Efecto gráfico de la normalización.
Fuente: Elaboración propia.

Se observa que 2 adivinaciones iniciales pueden converger a matrices (A, P) diferentes (pese a generar ambas el mismo ajuste a los datos), combinaciones que quedan representadas por la curva de iso-error representada en rojo en el esquema. La normalización de los resultados rompe dicha ambigüedad, haciendo "saltar" los datos desde el punto de convergencia, a un punto de la curva de iso-error, donde los resultados se encuentran normalizados de acuerdo a alguna regla: (A_n, P_n).

La elección de la regla de normalización depende en parte de las características del problema. J.C.Liao en su documento original [26] propone normalizar los resultados multiplicando las matrices por una matriz X definida por la Ecuación 15. Esto es, una matriz diagonal de la forma:

$$X_{jj} = \frac{1}{n}\sum_{i=1}^{N}\left|A_{ij}\right|$$

Sin embargo, mediante las pruebas realizadas se llegó a la conclusión de que dicha elección no era la mejor funcionalmente hablando. Aun cuando se comporta bien en la mayoría de los casos, presenta un efecto no deseado que alteraba los resultados obtenidos. A modo de ejemplo imaginar que 2 puntos cualquiera (combinación de matrices (A, P)) a lo largo de la curva de iso-error, difieren sólo en que los

datos de una de sus columna tiene signos contrarios. Si esto es así, y dado que la matriz de normalización X definida como en la Ecuación 15 presentará solo entradas positivas, ambas matrices, pese a tener el mismo ajuste no podrán, al momento de normalizarse corresponder a las mismas matrices. Ambas matrices convergerán a matrices normalizadas similares, pero que tendrán signos contrarios en la columna descrita. En el caso de las entradas negativas, por ejemplo, los valores positivos de la matriz de normalización no podrán cambiar su signo para hacerlas coincidir con los datos normalizados de la matriz con entradas positivas. De esta forma, el espacio de soluciones se restringe, y dicha normalización solo conducirá a resultados únicos cuando el signo de sus entradas coincida con el de la matriz normalizada. A modo esquemático, y siguiendo el ejemplo de la Figura 14, en la siguiente figura se representa dicho esquema.

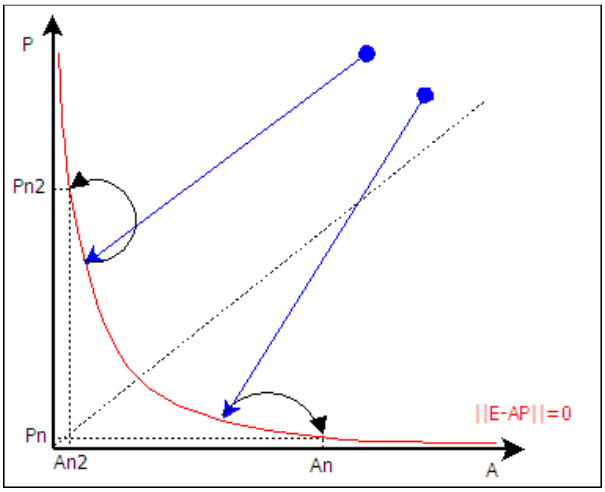


Figura 15: Efecto no deseado de la normalización.
Fuente: Elaboración propia.

El problema anterior puede solucionarse definiendo otro tipo de normalización que no presente dicho inconveniente. Para esto se propone la siguiente expresión, donde la matriz diagonal de normalización está definida por:

$$X_{jj} = \frac{1}{n}\sum_{i=1}^{N} A_{ij}$$

Ecuación 27

Si bien dicha matriz soluciona lo comentado con anterioridad, su efecto en los datos presenta menos cualidades interpretativas. Utilizando el valor absoluto en la ecuación anterior, el promedio absoluto de cada columna de A correspondería a 1, por lo que las magnitudes eran fácilmente comparables al ser su valor una medida relativa del grado de influencia del TF respectivo, en relación a un gen particular. Pese a esto, y a lo largo del resto del documento, se utilizará esta última expresión para normalizar los resultados.

3.5 Programación en MatLab

Con el objetivo de generar una herramienta que pueda ser testeada y utilizada en el análisis de redes de diferentes tamaños, se hace imprescindible el uso de plataformas computaciones de cálculo que automaticen los métodos mediante una programación de los mismos. El programa escogido para el desarrollo es MatLab [40], dada su gran capacidad de cálculo numérico y la facilidad de programación del mismo. Si bien se consideraron otras plataformas como Mathematica [41] y Octave, se piensa que MatLab ofrece una plataforma más dinámica de utilización, y de la misma manera, su extensa utilización en el ámbito académico y científico hacen más fácil la distribución de los resultados y métodos creados. Con el fin de aclarar dudas y averiguar la mejor manera de responder a las diferentes necesidades de los métodos programados se utilizó en extenso la siguiente bibliografía [42] [43] [44].

3.6 Reprogramación métodos originales

Las versiones básicas de NCA, desarrollada por J.C. Liao en sus documentos [26] [33] y descrita con anterioridad son un buen comienzo, y presentan un excelente comportamiento como se discutirá luego. Sin embargo presenta también algunas falencias en relación a su estabilidad y a la existencia de mínimos locales dadas las características del funcional. El trabajo comenzó reprogramando los métodos originales e incluyendo la técnica del gradiente del funcional en el algoritmo de optimización utilizando. Dicho procedimiento se torna básico a la hora de conocer el funcionamiento de NCA y todas las suposiciones y detalles que involucra la reconstrucción. Junto a la explicación del método se discutirán algunos aspectos relacionados con su uso y características.

3.6.1 NCA básico (NCAbasic)

La versión más básica de NCA es desarrollada inicialmente por J.C. Liao [26] y consiste en un método que permite reconstruir las señales de regulación y CS[34] de una red de regulación transcripcional. Lo anterior es equivalente a descomponer una matriz E de datos de expresión génica en dos matrices, A y P que resumen matricialmente la información anterior para diferentes sets de experimentos. El problema que se intenta resolver consiste en encontrar matrices A y P que minimicen el residuo o ajuste a los datos, entregado por la Ecuación 9.

$$\mathbb{P}: \quad \min_{A,P} \; J(E - A \cdot P)$$

$J(\cdot)$ debe ser escogida apropiadamente de acuerdo a las características del problema: por ejemplo, la norma Frobenius extendida a matrices. Como ya se ha comentado, la solución ha dicho problema no es única si no se imponen ciertas condiciones sobre la estructura de la matriz A , resumida en el espacio vectorial Z_A, definido como en la Ecuación 11. La estructura está definida por ciertas entradas en la

[34] Grados de influencias de reguladores sobre los genes.

matriz A que se conocen con certeza (*información a priori*), correspondientes a ceros atribuidos a conexiones inexistentes entre $TF's$ y genes.

NCAbasic busca matrices en dichos espacios vectoriales minimizando el residuo $\Gamma = E - A \cdot P$, utilizando un algoritmo de optimización alternada que explota las características convexas del problema. Como se explica en el capítulo anterior, en cada iteración se optimiza el funcional suponiendo constante una de las matrices, y dividiendo en diferentes problemas independientes el problema principal. En cada iteración se obtiene el valor de las matrices que minimiza el funcional utilizando el método del gradiente. El problema y el algoritmo es el siguiente, que se repetirá para un mayor orden y claridad.[35] Se escoge usar el cuadrado de la norma Frobenius en el funcional, la cual será utilizada en el resto del documento. .

Problema y algoritmo resolución NCAbasic

Si las matrices buscadas pertenecen a sub espacios vectoriales que cumplen con el criterio NCA, el problema a resolver será:

$$\mathbb{P}: \quad \min_{A,P} \ \|E - A \cdot P\|_F{}^2$$
$$s.t. \ A \in Z_A$$

En cada iteración se resolverá un problema equivalente al siguiente, donde se trabajan con vectores columnas:

$$\min_x \ \|b - A \cdot x\|_F{}^2 = (b - A \cdot x)^t \cdot (b - A \cdot x)$$

Como ya se vio, es posible obtener una expresión para el x^* que minimiza la expresión anterior dado aquel que hace cero el gradiente del funcional.

$$\boldsymbol{x^* = (A^t \cdot A)^{-1} A^t \cdot b}$$

Ecuación 28

El algoritmo de resolución que converge al mínimo de $\mathbb{P}$ es el siguiente, en la cual solo se estiman los parámetros distintos de cero en la estructura resumida en Z_A[36]:

Se genera una adivinación inicial para los parámetros a estimar de A y P, de forma que se respeten las restricciones en la estructura de A: $A^{(0)} \in Z_A$ y $P^{(0)} \in \mathbb{R}^{LxM}$.

1. Dado $A^{(0)}$, encontrar $P^{(1)}$ que resuelva:

[35] Notar que no es exactamente el mismo modelo de los puntos anteriores, ya que la otra versión permitía restricciones en la matriz P. Esto es equivalente a asumir que el sub espacio vectorial $Z_P = \mathbb{R}^{LxM}$.

[36] Los ceros, como siempre, se consideran información conocida, representando una conexión inexistente entre $TF's$ y genes.

$$\min_{P^{(1)}} \left\| E - A^{(0)} \cdot P^{(1)} \right\|_F^2$$

Esto se logra dividiendo el problema anterior en M problemas (uno para cada columna en E) de la forma ($k = 1, \dots, M$):

$$\min_{P_{ck}{}^{(1)}} \left\| E_{ck} - A^{(0)} \cdot P_{ck}{}^{(1)} \right\|_F^2$$

, donde E_{ck} es la columna k de E y P_{ck} es la columna k de P.

De esta manera, los problemas anteriores tienen una solución $\boldsymbol{P_{ck}{}^{(1)}}$ que puede ser obtenida mediante la expresión de la Ecuación 28.

$$\boldsymbol{P_{Rck}{}^{(1)} = \left(A^{(0)t} \cdot A^{(0)} \right)^{-1} A^{(0)t} \cdot E_{ck}}$$

2. Dado $P^{(1)}$, encontrar $A^{(1)}$ que resuelva:

$$\min_{A^{(1)}} \left\| E - A^{(1)} \cdot P^{(1)} \right\|_F^2$$
$$s.t.\ A^{(1)} \in Z_A$$

Esto es equivalente a resolver el problema siguiente, el que convenientemente está escrito en la misma forma que la Ecuación 17.

$$\min_{A^{(1)}} \left\| E^t - P^{(1)t} \cdot A^{(1)t} \right\|_F^2$$
$$s.t.\ A^{(1)} \in Z_A$$

La estrategia de resolución es similar, dividiendo el problema anterior en N problemas (uno para cada columna en E^t) de la forma ($i = 1, \dots, N$):

$$\min_{A_{Rri}{}^{(1)}} \left\| E_{ri}{}^t - P_{Ri}{}^{(1)t} \cdot A_{Rri}{}^{(1)t} \right\|_F^2$$

$$s.t.\ A_{Rri}{}^{(1)} \in Z_A$$

, donde $E_{ri}{}^t$ es la traspuesta de la fila i de E y $A_{Rri}{}^{(1)t}$ es la traspuesta de la fila i de A reducida; esto es la traspuesta de la fila i de A original, a la que se le han eliminado las entradas restringidas a cero en Z_A. $P_{Ri}{}^{(1)t}$ corresponde a la matriz P traspuesta reducida, a la cual se le han eliminado las columnas correspondientes a las entradas eliminadas en $A_{Rri}{}^{(1)t}$.

De esta manera, los problemas anteriores tienen una solución $\boldsymbol{A_{Rri}{}^{(1)}}$ que puede ser obtenida por la siguiente expresión (utilizando la Ecuación 28 una vez más).

$$\boldsymbol{A_{Rri}}^{(1)} = \left(\boldsymbol{P_{Ri}}^{(1)} \cdot \boldsymbol{P_{Ri}}^{(1)^{t}}\right)^{-1} \boldsymbol{P_{Ri}}^{(1)} \cdot \boldsymbol{E_{ri}}^{t}$$

3. Se repite el paso 2. y 3. Hasta que el funcional evaluado en $A^{(n)}$ y $P^{(n)}$, $\left\|E - A^{(n)} \cdot P^{(n)}\right\|_F^{\,2}$ sea menor que un cierto grado de tolerancia o el error halla convergido.

Finalmente las matrices A y P son normalizadas por una matriz X definida como en la Ecuación 27

Comentarios sobre NCAbasic

En resultados posteriores se analizará en mayor detalle el funcionamiento y capacidad de NCAbasic, sin embargo es conveniente comentar algunos aspectos propios del método que hay que tener en mente.

En primer lugar, dadas las características del funcional es posible la existencia de mínimos locales. De esta manera, de acuerdo de acuerdo al punto donde comience la iteración, el algoritmo puede no converger a un mínimo global. En la siguiente figura se expresa esquemáticamente lo anterior. El eje horizontal corresponde al espacio hipotético donde viven A y P y el vertical el error del residuo al ajustar dichas matrices a una matriz de datos en específico. Como se observa existe un mínimo global y otro local. El método parte en un punto cualquiera y se mueve en la dirección del gradiente de ambas matrices de forma alternada, por lo que en caso de partir en un punto cercano al mínimo local, el método podría converger a dicho punto[37].

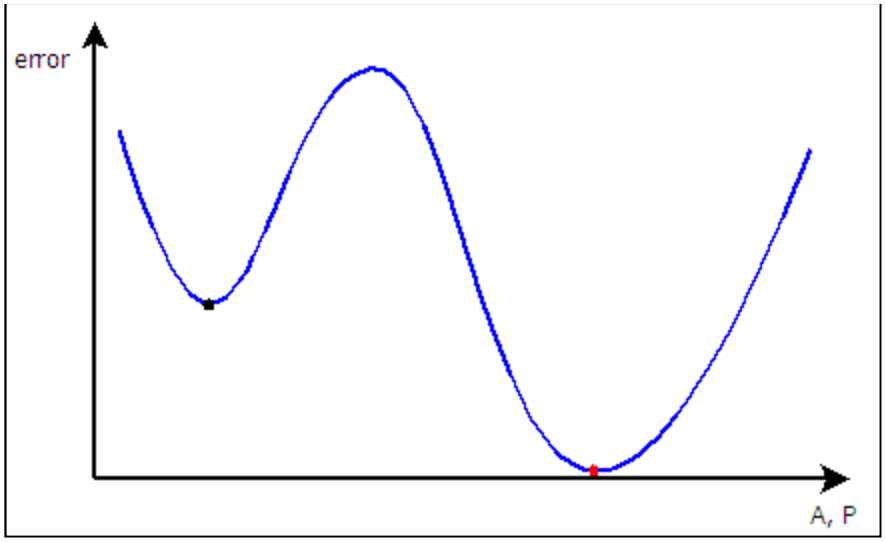


Figura 16: Existencia de mínimos locales.
Fuente: Elaboración propia.

[37] La existencia de mínimos locales está justificada en las características no lineales de la función a minimizar.

En la figura siguiente se muestra el resultado de una pequeña experiencia realizada con una red sintética de tamaño pequeño (13 genes, 4 $TF's$ y 5 experimentos). Se probaron 200 reconstrucciones con diferentes adivinaciones iniciales, y sin errores en los datos. La distribución del error obtenido se observa en la figura siguiente.

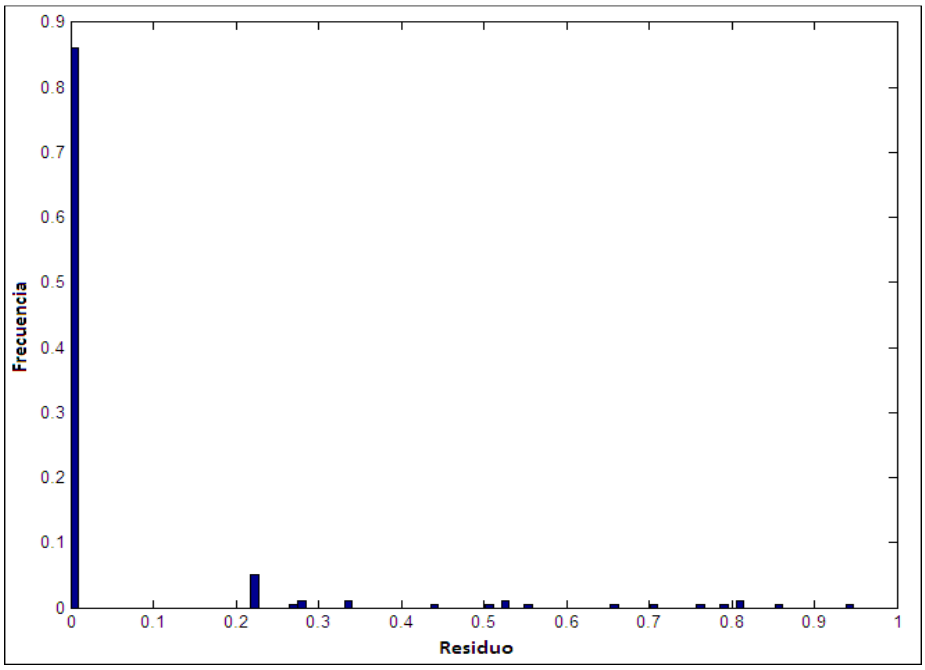


Figura 17: Distribución del residuo de ajuste obtenido con NCAbasic para 200 adivinaciones iniciales.
Fuente: Elaboración propia.

Claramente el mínimo global corresponde a la distribución más cercana a cero, mientras que el resto (aproximadamente el 15%) corresponden a convergencias a mínimos locales con un error considerable, asociado a una reconstrucción defectuosa.

Otro tema a considerar es que bajo ciertas circunstancias el problema puede tornarse mal condicionado. Independiente de que las matrices (o más exactamente, su estructura) cumplan con el criterio NCA, lo cual asegura la existencia, unicidad de la solución y en cierto sentido la estabilidad, bajo ciertos valores que adquiere la matriz en el proceso de iteración los errores en los datos pueden distorsionar fuertemente las estimaciones, haciendo que el método converja a resultados que poseen un gran error respecto a los parámetros reales. Como ya se ha mencionado, el mal condicionamiento es un problema común de los problemas inversos, en especial el referente a la estabilidad numérica de la solución, por lo que es necesario tener en mente dicho factor y considerarlo a la hora de interpretar los resultados. De todas formas existen varias técnicas que es posible implementar con el fin de superar dicho problema, haciendo más robusto el algoritmo. Una posibilidad es testear dicho problema[38] a medida que el

[38] El referente a la existencia de matrices mal condicionadas en las iteraciones.

algoritmo se mueve por diferentes matrices en el proceso de iteración, alterando los resultados en los casos de encontrar sistemas con matrices mal condicionadas. Otra alternativa que se discutirá más extensamente en otra sección es la técnica de regularización, que en palabras simples fuerza a algunas de las matrices a permanecer cerca de un punto y por lo tanto restringiendo la dirección del movimiento en las iteraciones.

3.6.2 NCA básico general (gNCAbasic)

La segunda versión de NCA, atribuida también al documento de J.C. Liao [33], es una versión extendida del método anterior, con la misma funcionalidad, pero que permite la inclusión de restricciones en la estructura de la matriz P también en adición a la estructura de la matriz A. Esta nueva información, resumida en el sub espacio vectorial Z_P es utilizada en conjunto a las restricciones en A con el fin de disminuir en parte los requerimientos del criterio NCA. El criterio ii) para NCAbasic impone condiciones sobre las matrices reducidas de A, A_{rj} $(j = 1, \dots, L)$, mientras que el criterio general (que incorpora la posibilidad de tener restricciones en P) las impone sobre las matrices reducidas que conforman la matriz reducida de G.

El problema que se intenta resolver es similar al de NCAbasic, estos es encontrar 2 matrices que descompongan la matriz de datos E y minimicen el ajuste a los datos. Esto es la expresión entregada por la Ecuación 9.

$$\mathbb{P}: \quad \min_{A,P} \; J(E - A \cdot P)$$

gNCAbasic busca matrices en los sub espacios vectoriales Z_A y Z_P , que resumen la información respecto a las restricciones de las entradas de las matrices, mientras minimiza el residuo $\Gamma = E - A \cdot P$. El problema y el algoritmo utilizando para resolver el problema es el siguiente.

Problema y algoritmo resolución gNCAbasic

Si las matrices buscadas pertenecen a sub espacios vectoriales que cumplen con el criterio NCA, el problema a resolver será:

$$\mathbb{P}: \quad \min_{A,P} \; \|E - A \cdot P\|_F{}^2$$
$$s.t. \; A \in Z_A \;\; \text{y} \;\; P \in Z_P$$

En cada iteración se resolverá un problema equivalente al siguiente, donde se trabaja con vectores columnas:

$$\min_x \; \|b - A \cdot x\|_F{}^2 = (b - A \cdot x)^t \cdot (b - A \cdot x)$$

Como ya se vio, es posible obtener una expresión para el x^* que minimiza la expresión anterior dado aquel que hace cero el gradiente del funcional, dado por la Ecuación 28

$$\boldsymbol{x^* = (A^t \cdot A)^{-1} A^t \cdot b}$$

El algoritmo de resolución que converge al mínimo de $\mathbb{P}$ es el siguiente, en el cual sólo se estiman los parámetros distintos de cero en la estructura resumida en Z_A y Z_P (los ceros se consideran información conocida, representando una conexión inexistente entre $TF's$ y genes o experimentos knock-out en P):

Se genera una adivinación inicial para los parámetros a estimar de A y P, de forma que se respeten las restricciones en la estructura de A: $A^{(0)} \in Z_A$ y $P^{(0)} \in Z_P$.

1. Dado $A^{(0)}$, encontrar $P^{(1)}$ que resuelva:

$$\min_{P^{(1)}} \left\| E - A^{(0)} \cdot P^{(1)} \right\|_F^2$$
$$s.t.\ P^{(1)} \in Z_P$$

Esto se logra dividiendo el problema anterior en M problemas (uno para cada columna en E) de la forma ($k = 1, \dots, M$):

$$\min_{P_{Rck}{}^{(1)}} \left\| E_{ck} - A_{Rk}{}^{(0)} \cdot P_{Rck}{}^{(1)} \right\|_F^2$$

$$s.t.\ P_{Rck}{}^{(1)} \in Z_P$$

, donde E_{ck} es la columna k de E y P_{Rck} es la columna k de P reducida; esto es la columna k de P original, a la que se le han eliminado las entradas restringidas a cero en Z_P (a fin de no estimarlas de nuevo). A_{Rk} corresponde a la matriz A reducida, a la cual se le han eliminado las columnas correspondientes a las filas eliminadas en P_{Rck}.

De esta manera, y dadas las condiciones NCA especificadas con anterioridad, los problemas anteriores tienen una solución dada por la Ecuación 28:

$$\boldsymbol{P_{Rck}{}^{(1)} = \left(A_{Rk}{}^{(0)^t} \cdot A_{Rk}{}^{(0)} \right)^{-1} A_{Rk}{}^{(0)^t} \cdot E_{ck}}$$

2. Dado $P^{(1)}$, encontrar $A^{(1)}$ que resuelva:

$$\min_{A^{(1)}} \left\| E - A^{(1)} \cdot P^{(1)} \right\|_F^2$$
$$s.t.\ A^{(1)} \in Z_A$$

Esto es equivalente a resolver el problema siguiente, el que convenientemente está escrito en la misma forma que la Ecuación 17.

$$\min_{A^{(1)}} \left\| E^t - {P^{(1)}}^t \cdot {A^{(1)}}^t \right\|_F^2$$
$$s.t.\ A^{(1)} \in Z_A$$

La estrategia de resolución es similar, dividiendo el problema anterior en N problemas (uno para cada columna en E^t) de la forma ($i = 1, \dots, N$):

$$\min_{{A_{Rri}}^{(1)}} \left\| {E_{ri}}^t - {P_{Ri}}^{(1)^t} \cdot {A_{Rri}}^{(1)^t} \right\|_F^2$$

$$s.t.\ {A_{Rri}}^{(1)} \in Z_A$$

, donde ${E_{ri}}^t$ es la traspuesta de la fila i de E y ${A_{Rri}}^{(1)^t}$ es la traspuesta de la fila i de A reducida; esto es la traspuesta de la fila i de A original, a la que se le han eliminado las entradas restringidas a cero en Z_A. ${P_{Ri}}^{(1)^t}$ corresponde a la matriz P traspuesta reducida, a la cual se le han eliminado las columnas correspondientes a las entradas eliminadas en ${A_{Rri}}^{(1)^t}$.

De esta manera, los problemas anteriores tienen una solución $\boldsymbol{{A_{Rri}}^{(1)}}$ que puede ser obtenida por la siguiente expresión (utilizando la Ecuación 28 una vez más).

$$\boldsymbol{{A_{Rri}}^{(1)} = \left({P_{Ri}}^{(1)} \cdot {P_{Ri}}^{(1)^t} \right)^{-1} {P_{Ri}}^{(1)} \cdot {E_{ri}}^t}$$

3. Se repite el paso 2. y 3. Hasta que el funcional evaluado en $A^{(n)}$ y $P^{(n)}$, $\left\| E - A^{(n)} \cdot P^{(n)} \right\|_F^2$ sea menor que un cierto grado de tolerancia o el error halla convergido.

Finalmente las matrices A y P son normalizadas por una matriz X definida como en la Ecuación 27.

Sobre gNCAbasic y restricciones en P

La incorporación de restricciones en P implica un set de suposiciones inherentes al método, que en algunos casos quizás son demasiado fuertes, pero que en contra parte entrega una gran ventaja al método de resolución. Imponiendo restricciones a P es posible trabajar con redes que violen en parte el criterio NCA y reparar dichos problemas con la inclusión de experimentos knock-out que restrinjan algunas de las entradas de P. El asunto ya fue discutido en parte en el punto 3.3 Criterios NCA y su implicancia, donde se señaló que de existir un problema en algún TF en específico que haga se viole el criterio NCA, es posible cumplirlo generando un experimento knock-out en el mismo gen.

Como un ejemplo ilustrativo se utilizará la red sintética de pequeño tamaño (5 genes y 3 TF's) resumida en la figura siguiente.

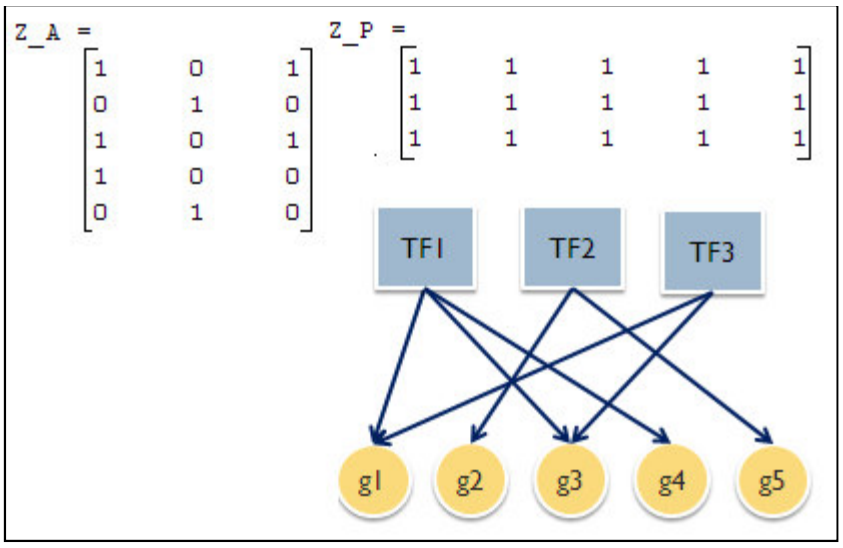
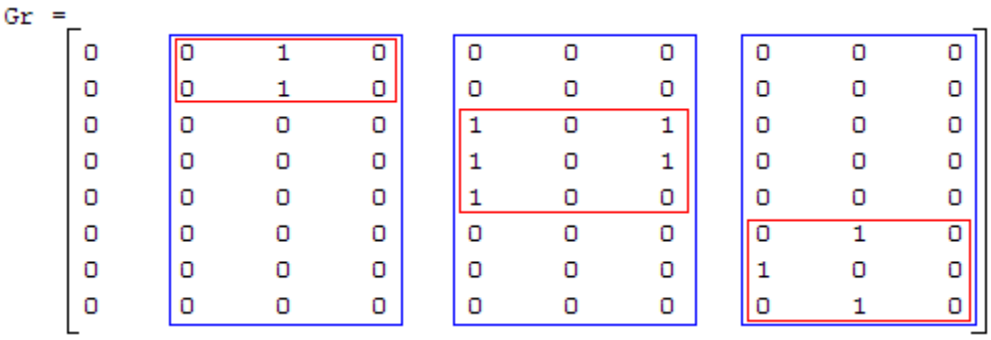


Figura 18: Red sintética experimentos knock-out.
Fuente: Elaboración propia.

Estas matrices representan los sub espacios vectoriales generados por las restricciones en la estructura de A y P. Los valores ceros son entradas conocidas, mientras que los valores 1 son parámetros a estimar. Nótese además que la estructura entregada en dicha matriz define la red dibujada (quien se conecta con quien) y que la estructura de P no posee restricciones. Es fácil al analizar las matrices reducidas de A o equivalentemente la matriz reducida de G que se viola el criterio NCA en el $TF3$.



En la matriz anterior en azul están las sub matrices de G_r (las cuales interesa obtener su rango) y en rojo las correspondientes matrices reducidas de A (basta eliminar la columna de ceros para obtener las mismas matrices). Debido a que los ceros que están alrededor no cambiarán el rango, se confirma la equivalencia entre ambos análisis cuando no hay restricciones en P. Se observa también que el $TF3$ presenta problemas, ya que en la primera matriz reducida aparece una columna de ceros extra en la correspondiente a dicho factor, por lo que el rango no puede ser 2 como exige el método para esta red particular. Como se comentó, esto es solucionable si se realiza un experimento knock-out en el gen con problemas. Consideramos ahora un estructura de P modificada, en la cual en el 5° experimento y en el $TF3$ se ha realizado una experiencia del tipo mencionada. Su estructura quedaría como se muestra, y la matriz reducida de G recalculada sería la siguiente.

```
Z_P =
[1  1  1  1  1]
[1  1  1  1  1]
[1  1  1  1  0]
```

```
Gr =
[0  0  1  0  0  0  0  0  0]
[0  0  1  0  0  0  0  0  0]
[0  0  0  0  1  0  1  0  0]
[0  0  0  0  1  0  1  0  0]
[0  0  0  0  1  0  0  0  0]
[0  0  0  0  0  0  0  0  1]
[0  0  0  0  0  0  0  1  0]
[0  0  0  0  0  0  0  0  0]
[0  0  0  0  0  0  0  0  0]
[0  0  1  0  0  0  0  0  0]
```

Como se observa, corresponde a la misma matriz anterior, pero con una fila extra producto del aporte del experimento knock-out. Dicha fila adiciona un factor extra que repara el problema con la condición ii) de NCA al estudiar las sub matrices que conforman G_r. Sin embargo ya no es equivalente hacer lo mismo con las matrices reducidas de A ya que estas seguirán violando el criterio.

Respecto a la naturaleza de los supuestos realizados en los experimentos knock-out, es necesario hacer varias aclaraciones. En una experiencia de este tipo se intenta mediante algún método de ingeniería genética bloquear la expresión de un gen en particular, en este caso del gen que codifica el TF de interés. Por supuesto dicho experimento provocaría que él TF no expresase, cuando en realidad lo que se está suponiendo en este caso es que la expresión relativa a un estado de referencia se mantiene constante. Sin embargo dicha salvedad puede obviarse si se llevan a cabo experimentos en cepas knock-out que tomen en cuenta este factor. En segundo lugar se asume que las conexiones de los demás TF's no se altera producto de la modificación genética realizada para lograr el knock-out. Si bien las actividades de los TF's pueden variar producto de la interacción de los mismos, las conexiones y los CS asociados no variarán en dicho experimento. Todo el efecto se traduce simplemente en cambios en las señales de regulación, pero no en las características del sistema, lo cual puede ser una suposición considerable pero necesaria para utilizar el modelo descrito.

A cambio de los supuestos anteriores, la incorporación de suposiciones en P entrega varias ventajas al momento de reconstruir una red. En primer lugar y como se demostró, permite el análisis de redes que violen el criterio NCA básico. Por otra parte, y en especial la condición ii), las condiciones NCA no presentan un fuerte fundamente físico-biológico, por lo que podrían ser violados por algunas redes. El combinar los datos de cepas wild-type[39] con mutantes permite analizar incluso ese tipo de redes y reconstruir sin problemas los parámetros del sistema. Por otra parte, incluso cuando no se tengan problemas con el criterio NCA, la información extra que entregan dichos datos es de gran importancia, permitiendo comparar las señales de regulación de cepas normales y mutantes. Finalmente, y dado que

[39] Experimentos en los cuales no se han realizado cambios del tipo genético.

la matriz A debiese ser estimada solo una vez (asumiendo cierto el punto anterior), la incorporación de datos provenientes de experimentos extras, ya sea de cepas mutantes u otras, debiese conducir rápidamente a resultados dada la menor cantidad de parámetros a estimar.

Se destaca también que los resultados, dada la estructura de los métodos obtenidos vía NCAbasic coinciden a los de gNCAbasic cuando no se imponen restricciones en P. De esta manera, todos los métodos siguientes utilizarán como base la estructura y teoría anterior, que se presenta más general y con la capacidad de abordar diferentes tipos de problemas.

Finalmente, y como se verá en las experiencias finales, el tamaño de la red es un punto relevante respecto al tiempo de resolución del problema. Comprobar el criterio NCA calculando la matriz G y sus correspondientes reducidas no es algo trivial, dado la gran cantidad de cálculos que esto involucra y el tamaño de la matriz G, que es mayor al de la red. En redes de tamaño real (que pueden superar los 3000 genes y 300 $TF's$) los tiempos de espera pueden ser considerables, y dada la menor cantidad de operaciones que necesita NCAbasic y el menor tamaño de las matrices creadas en el proceso, su uso puede ser más adecuado si no se dispone de los equipos y la memoria adecuada.

3.6.3 Sobre métodos de regularización

Los problemas de estabilidad inducidos por matrices mal condicionadas en sistemas de ecuación, y en particular en problemas inversos como el analizado, ha sido ampliamente estudiado, y se han generado diversas técnicas para solucionar dicho inconveniente.

Los métodos de regularización son utilizados en problemas inversos, como una manera de mejorar la estabilidad de las soluciones a perturbaciones en los datos. El método de regularización de Tikhonov [45] propuesto en el documento original [33], introduce un parámetro de regularización en el funcional a minimizar, con el fin de evitar problemas de sensibilidad producidos por matrices mal condicionadas. Esto es fácilmente aplicado al problema anterior, como se verá en lo siguiente.

3.6.4 NCA básico general con regularización (gNCAreg)

Con el fin de incluir un parámetro de regularización en NCA, J.C. Liao [33] en su documento modifica el funcional original incluyendo un parámetro de regularización en la matriz P. Dicha función corresponde a la siguiente expresión

$$J(E - A \cdot P) = \frac{1}{2} \cdot \|E - A \cdot P\|_F^{\ 2} + \frac{\lambda}{2} \cdot \|P\|_F^{\ 2}$$

Ecuación 29

Al minimizar dicho funcional modificado, se ajustan las matrices A y P a los datos mientras se castiga que los valores de P se hagan exageradamente grandes. El parámetro de regularización λ controla este compromiso entre minimizar ambos miembros de la ecuación. Si bien los resultados se comprometen en relación a la exactitud de la reconstrucción, la regularización asegura un método más estable y robusto.

El algoritmo de resolución es similar a los ya detallados, considerando por supuesto las diferencias en el funcional. El gradiente del funcional una vez más entrega una expresión para las matrices que minimizarán la expresión en cada iteración. Al minimizar en A el problema será equivalente a los ya vistos (ya que la expresión agregada no depende de este parámetro), por lo que es posible encontrar el mínimos haciendo uso de la Ecuación 17. Al momento de minimizar en P el asunto es diferente, en estos casos se resolverán problemas equivalentes al siguiente:

$$\min_x \quad f(x) = \frac{1}{2} \cdot \|b - A \cdot x\|_F{}^2 + \frac{\lambda}{2} \cdot \|x\|_F{}^2$$

$$\min_x \quad \frac{1}{2} \cdot (b - A \cdot x)^t \cdot (b - A \cdot x) + \frac{\lambda}{2} \cdot x^t x$$

Ecuación 30

El funcional se puede descomponer como sigue:

$$\frac{1}{2} \cdot [(b - A \cdot x)^t \cdot (b - A \cdot x) + \lambda \cdot x^t x] =$$

$$\frac{1}{2} \cdot [(b^t - x^t \cdot A^t) \cdot (b - A \cdot x) + \lambda \cdot x^t x] = \frac{1}{2} \cdot [b^t b - b^t A x - x^t A^t b + x^t A x + \lambda \cdot x^t x]$$

Pero como $b^t A x$ es una constante, $(b^t A x)^t = x^t A^t b$. Así se tiene:

$$f(x) = \frac{1}{2} \cdot [(b - A \cdot x)^t \cdot (b - A \cdot x) + \lambda \cdot b^t b] = \frac{1}{2} \cdot [b^t b - 2 b^t A x + x^t A x + \lambda \cdot x^t x]$$

Finalmente es posible obtener el gradiente de la función a minimizar siguiendo las reglas de derivación de vectores, formas lineales y formas cuadráticas detalladas en los anexos.

$$\frac{df(x)}{dx} = \nabla f(x) = \frac{1}{2} \cdot [0 - 2A^t b + 2A^t A x + 2\lambda x]$$

El mínimo estará dado en donde el gradiente vale cero, esto es:

$$\nabla f(x^*) = A^t b + A^t A x^* + \lambda x^* = 0$$

$$(A^t A + \lambda I) x^* = A^t b$$

$$\boldsymbol{x^* = (A^t A + \lambda I)^{-1} A^t \cdot b}$$

Ecuación 31

Se hace notar que aquí la expresión $A^t A + \lambda I$ es una matriz cuadrada, por lo que al cumplirse las condiciones NCA, la inversa existirá sin problemas.

Calculando la segunda derivada del funcional, se ve que efectivamente se trata de un mínimo.

$$\frac{d^2 f(x)}{d^2 x} = \mathcal{H} f(x) = 2(A^t A + \lambda I)$$

Ecuación 32

, que bajo ciertas condiciones generales es semi-definida positiva. Como antes, y dada las características del problema, en cada iteración serán necesarios resolver M de estos problemas para P respetando siempre aquellas entradas en las que se impone la presencia de un cero en la matriz.

El problema es entonces.

$$\mathbb{P}: \quad \min_{A,P} \ \frac{1}{2} \cdot \|E - A \cdot P\|_F{}^2 + \frac{\lambda}{2} \cdot \|P\|_F{}^2$$
$$s.t. \ A \in Z_A \ \text{ y } \ P \in Z_P$$

Ecuación 33

El algoritmo de resolución es el siguiente.

Se genera una adivinación inicial para los parámetros a estimar de A y P, de forma que se respeten las restricciones en la estructura de A y P: $A^{(0)} \in Z_A$ y $P^{(0)} \in Z_P$.

1. Dado $A^{(0)}$, encontrar $P^{(1)}$ que resuelva:

$$\min_{P^{(1)}} \frac{1}{2} \cdot \left\| E - A^{(0)} \cdot P^{(1)} \right\|_F{}^2 + \frac{\lambda}{2} \cdot \left\| P^{(1)} \right\|_F{}^2$$
$$s.t. \ P^{(1)} \in Z_P$$

   Esto se logra dividiendo el problema anterior en M problemas (uno para cada columna en E) de la forma ($k = 1, \dots, M$):

$$\min_{P_{Rck}{}^{(1)}} \ \frac{1}{2} \cdot \left\| E_{ck} - A_{Rk}{}^{(0)} \cdot P_{Rck}{}^{(1)} \right\|_F{}^2 + \frac{\lambda}{2} \cdot \left\| P_{Rck}{}^{(1)} \right\|_F{}^2$$

$$s.t. \ P_{Rck}{}^{(1)} \in Z_P$$

   , donde E_{ck} es la columna k de E y P_{Rck} es la columna k de P reducida; esto es la columna k de P original, a la que se le han eliminado las entradas restringidas a cero en Z_P (a fin de no estimarlas de nuevo). A_{Rk} corresponde a la matriz A reducida, a la cual se le han eliminado las columnas correspondientes a las filas eliminadas en P_{Rck}.

   De esta manera, y dadas las condiciones NCA, los problemas anteriores tienen una solución dada por la Ecuación 33:

$$\boldsymbol{P_{Rck}}^{(\mathbf{1})} = \left(\boldsymbol{A_{Rk}}^{(\mathbf{0})^{\boldsymbol{t}}} \cdot \boldsymbol{A_{Rk}}^{(\mathbf{0})} + \boldsymbol{\lambda} \cdot \boldsymbol{I}\right)^{-\mathbf{1}} \boldsymbol{A_{Rk}}^{(\mathbf{0})^{\boldsymbol{t}}} \cdot \boldsymbol{E_{ck}}$$

2. Dado $P^{(1)}$, encontrar $A^{(1)}$ que resuelva (no se ha considerado el término de regularización en P dado que al minimizar respecto a A dicho termino es irrelevante:

$$\min_{A^{(1)}} \left\| E - A^{(1)} \cdot P^{(1)} \right\|_F^{\,2}$$
$$s.t.\ A^{(1)} \in Z_A$$

Esto es equivalente a resolver el problema siguiente, el que convenientemente está escrito en la misma forma que la Ecuación 17.

$$\min_{A^{(1)}} \left\| E^t - {P^{(1)}}^t \cdot {A^{(1)}}^t \right\|_F^{\,2}$$
$$s.t.\ A^{(1)} \in Z_A$$

La estrategia de resolución es similar, dividiendo el problema anterior en N problemas (uno para cada columna en E^t) de la forma ($i = 1, \dots, N$):

$$\min_{{A_{Rri}}^{(1)}} \left\| {E_{ri}}^t - {P_{Ri}}^{(1)^t} \cdot {A_{Rri}}^{(1)^t} \right\|_F^{\,2}$$

$$s.t.\ {A_{Rri}}^{(1)} \in Z_A$$

, donde ${E_{ri}}^t$ es la traspuesta de la fila i de E y ${A_{Rri}}^{(1)^t}$ es la traspuesta de la fila i de A reducida; esto es la traspuesta de la fila i de A original, a la que se han eliminado las entradas restringidas a cero en Z_A. ${P_{Ri}}^{(1)^t}$ corresponde a la matriz P traspuesta reducida, a la cual se han eliminado las columnas correspondientes a las entradas eliminadas en ${A_{Rri}}^{(1)^t}$.

De esta manera, los problemas anteriores tienen una solución $\boldsymbol{A_{Rri}}^{(\mathbf{1})}$ que puede ser obtenida por la siguiente expresión (utilizando la Ecuación 28 una vez más).

$$\boldsymbol{A_{Rri}}^{(\mathbf{1})} = \left(\boldsymbol{P_{Ri}}^{(\mathbf{1})} \cdot \boldsymbol{P_{Ri}}^{(\mathbf{1})^{\boldsymbol{t}}}\right)^{-\mathbf{1}} \boldsymbol{P_{Ri}}^{(\mathbf{1})} \cdot \boldsymbol{E_{ri}}^{\boldsymbol{t}}$$

3. Se repite el paso 2. y 3. Hasta que el funcional evaluado en $A^{(n)}$ y $P^{(n)}$, $\left\| E - A^{(n)} \cdot P^{(n)} \right\|_F^{\,2}$ sea menor que un cierto grado de tolerancia o el error halla convergido.

Finalmente las matrices A y P son normalizadas por una matriz X definida como en la Ecuación 27.

Comentarios sobre gNCAreg

En el algoritmo anterior es intuitivo que se sacrifica exactitud en pos de la estabilidad del método. Dado el término extra del funcional los errores serán más elevados, y aun considerando sólo el residuo de ajuste, la distorsión introducida en la matriz P debido al término de regulación hará que la exactitud de la reconstrucción sea menor que con los métodos anteriores. En contraparte el método repara[40] los problemas producidos por matrices mal condicionadas en las iteraciones.

Sin embargo, la aparente necesidad matemática de castigar a la matriz P al alejarse de cero, implica la suposición de que pequeñas fluctuaciones en la matriz P de un experimento a otro explican de buena forma el fenómeno de regulación génica. Es necesario recalcar que las entradas de dicha matriz son el logaritmo de actividades relativas, por lo que los valores relativamente pequeños para dichas expresiones son justificados. En efecto, una actividad relativa de 100 (100 veces superior respecto al estado de referencia) entregaría un nivel de actividad log-relativa de tan solo 2.

Un punto importante es lo referente al parámetro λ que mide la importancia relativa respecto al residuo de ajuste que se da al término extra del funcional. Mientras más grande sea dicho valor mayor ponderación se dará al castigo de P y en cierto sentido el error de ajuste disminuirá. Si el parámetro de regularización disminuye los parámetros de P tendrán mayor libertad y el error de ajuste convergerá al obtenido con NCAbasic. En la figura siguiente se aprecia esquemáticamente dicho fenómeno. En el eje horizontal se muestra la norma de P (el segundo término del funcional), mientras que en el vertical el residuo de ajuste. Cada punto de la curva se obtiene con un λ distinto, y creciente en la medida que se mueve hacia la derecha por la curva.

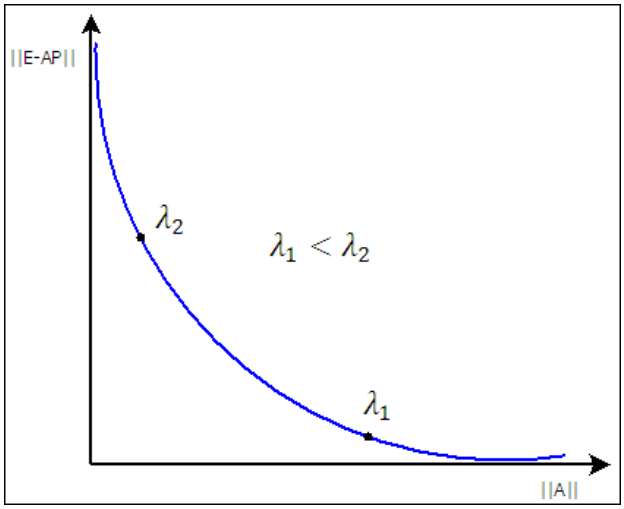


Figura 19: Efecto parámetro de regularización en el ajuste a los datos.
Fuente: Elaboración propia.

[40] Al menos en parte.

En las pruebas realizadas se utilizará un $\lambda = 0{,}1$, que empíricamente demuestra ser un buen compromiso entre ambos objetivos (lograr un buen ajuste y mejorar la estabilidad del método). En estricto rigor, y con el fin de obtener el máximo provecho del método de regularización, el parámetro debiese ser ajustado en todas las iteraciones; esto es, en cada iteración explorar la curva dada por la Figura 19 y analizar el λ óptimo que ofrezca el mejor compromiso entre ambos errores. De acuerdo al método L-curve [46], este parámetro será aquel que produzca la esquina de la gráfica anterior, por lo que dicho procedimiento necesitaría ser repetido en cada iteración, e incluso en cada sub sistema independiente producido al optimizar la matriz P. Por supuesto este método es intensivo en recursos computacionales por lo que el método se hará excesivamente lento, en especial para redes de gran tamaño. En el presente estudio se utilizará un parámetro fijo para una red particular, que no exagere la distorsión producida por el término de regularización.

3.6.5 Métodos programados

El objetivo final del análisis anterior es programar los métodos originales en MatLab, y probar su funcionamiento con diferentes redes sintéticas. El estudio de los fundamentos matemáticos, las técnicas de optimización y algoritmos de convergencia utilizados, la necesidad de normalización, la medición de errores, la existencia de problemas de estabilidad y de mínimos locales, permiten conocer los alcances y limitaciones del método original desde un punto de vista teórico, lo que a la hora de trabajar en forma práctica en la reconstrucción de redes permitirá identificar la posible fuentes de inconsistencias o los aspectos que sea necesario modificar o agregar para obtener mejores resultados. La programación de los métodos culmina esta primera etapa del trabajo, en donde las modificaciones principales respecto a los métodos originales consistieron en el uso del gradiente del funcional en las etapas de optimización, así como el uso de un parámetro de regularización global. Los métodos programados se resumen en el anexo correspondiente, y corresponden a los nombre nca_n.m (para NCAbasic), gnca_n.m (para gNCAbasic) y gnca_reg_n.m (para gNCAreg).

3.7 Modificación a métodos originales

Realizado el procedimiento de estudio, principalmente a nivel teórico y en algún sentido funcional, se identifican una serie de problemas que se detallan a continuación:

- **Existencia de mínimos locales:** Si bien en algunas redes (como la de la Figura 17) la distribución del residuo del error privilegiaba enormemente el mínimo global, otras redes estudiadas presentaban una distribución más dispersa. El error de ajuste se disparaba al caer en un mínimo local, y el ajuste a las matrices originales de A y P era pésimo, en especial este último. Incluso en redes como la de la Figura 17 existía la posibilidad de caer en un mínimo local, y si bien en un experimento sintético es posible con certeza saber que mínimo se ha alcanzado, no es así a la hora de enfrentarse a redes reales. Más aun, no basta con asumir que un error bajo corresponde al mínimo global, debido a que al adicionar error a la matriz de datos, el error de ajuste lógicamente tenderá a aumentar.

- El método de convergencia propuesto por J.C. Liao [26] impone que el error de ajuste debe superar una cierta barrera mínima para terminar con las iteraciones. Esto en la práctica no es posible, debido a que la presencia de errores en los datos e incluso la convergencia hacia un mínimo local hará que el error jamás alcance dicha cota.

Las soluciones propuestas para dichos problemas se detallan a continuación:

3.7.1 Sobre errores de convergencia

En relación al problema comentado referente a los errores de convergencia, se implementan 2 soluciones que modifican los métodos originales. La primera de ellas consiste en restringir el número de iteraciones hasta alcanzar la convergencia; el método sigue iterando hasta alcanzar dicho número y se queda con dichas matrices como resultados. Dicha solución presenta el inconveniente de no asegurar una convergencia al mínimo, dado que en algunos casos dicho punto puede alcanzarse en un número de iteraciones elevadas, que puede cambiar de una red a otra e incluso respecto a diferentes adivinaciones iniciales. Como se ve en la Figura 20 puede requerirse un gran número de iteraciones para alcanzar el mínimo global (la línea roja en el gráfico) o para converger a algún error. Por otra parte dicho método puede acortar el tiempo de espera en convergencias demasiado lentas, en donde la disminución del error de una iteración a otra es despreciable o poco significativa.

Otra opción para dicho problema es analizar directamente la convergencia del error. Las dos curvas observadas en la Figura 20 convergen a un error determinado. Si bien una llega a un error menor, las dos alcanzan su punto mínimo al cabo de una cierta cantidad de iteraciones (cuando la curva se hace horizontal). Estudiando esta convergencia al momento de ir iterando es posible obtener un método de parada más práctico y exacto. Es posible, por ejemplo, implementarlo comparando el error de 2 iteraciones consecutivas, y comprobar si el cambio obtenido es considerable, o equivalentemente si su diferencia es menor que cierto valor umbral.

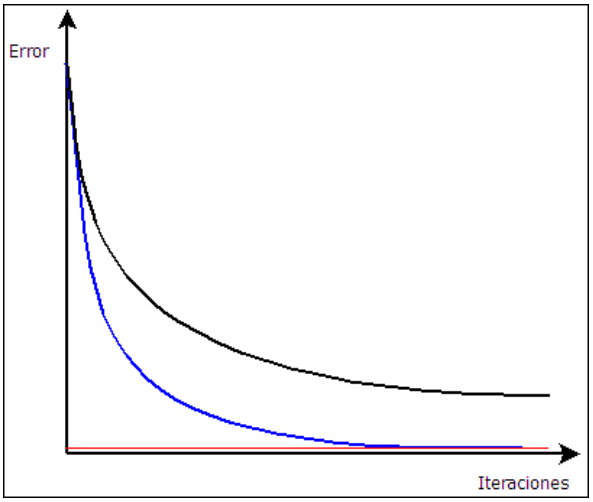


Figura 20: Convergencia de error en iteraciones.
Fuente: Elaboración propia.

3.7.2 Sobre mínimos locales

Más importante que lo anterior es el asunto referente a los mínimos locales ya comentado. La solución propuesta logra superar dicho inconveniente desarrolla una suerte de escaneo previo del comportamiento de las iteraciones respecto de diferentes adivinaciones iniciales. Mediante algunas pruebas simples, se comprobó que la convergencia hacia un punto en particular depende sólo de 2 factores, la red estudiada y la adivinación inicial. Como la intuición lo predice y tal cual en una función de una variable, el punto de iteración inicial definirá el mínimo alcanzado. Se debe recordar que el método se mueve en dirección del gradiente del funcional a partir de un punto inicial, por lo que si se parte del mismo punto siempre se obtendrá el mismo residuo.

Con lo anterior en mente, la modificación propuesta consiste en dividir cada método NCA en 2 partes. Un programa que itere y reconstruya las matrices y otro que se encargue de analizar el efecto del error frente a diferentes puntos de inicio de la iteración. Dicho programa genera una cantidad dada de adivinaciones, y estudia de qué forma evoluciona el error en un número bajo de iteraciones (para disminuir el tiempo de procesamiento). Analizado lo anterior, detecta el punto de inicio de iteración que llego al mínimo error, y lanza una vez más el programa de reconstrucción (esta vez sin límites de iteraciones con el fin de que el error converja) comenzando desde el punto identificado. Dicho procedimiento asegura que se converja al mínimo global con alta probabilidad. En la figura siguiente se observa la distribución del error de 100 pruebas de reconstrucción utilizando el mecanismo descrito. Si bien pareciera que la distribución es dispersa, la diferencia es despreciable dada la escala del gráfico.

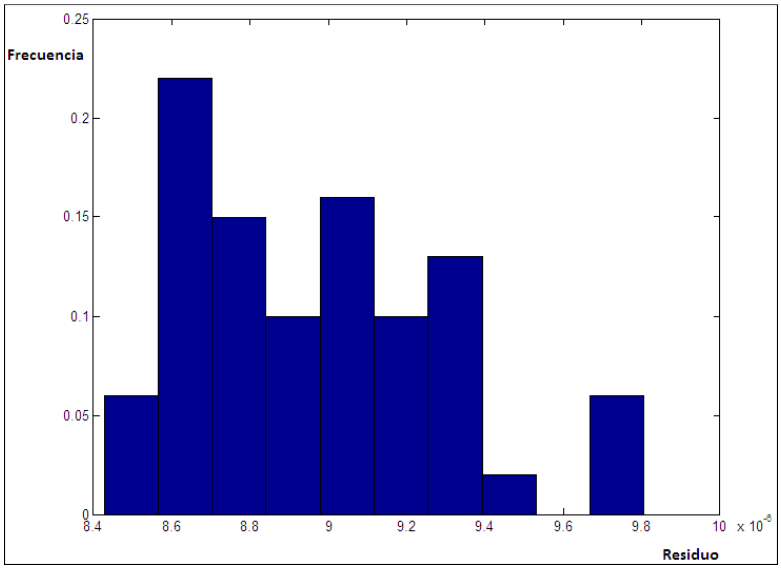


Figura 21: Frecuencia del residuo utilizando método modificado.
Fuente: Elaboración propia.

3.8 Nuevos métodos NCA

El análisis de los métodos originales y la experiencia obtenida mediante diferentes pruebas analizadas, así como el aporte realizado en la modificación parcial de dichos métodos, motiva la creación de nuevas funciones NCA. La idea de generar nuevos métodos NCA es atacar ciertas falencias que se detectan en la técnica original, no tanto dirigida a su trasfondo, si no a la posibilidad no explotada de hacer uso de otro tipo de información para obtener mejores resultados[41]. Dentro de los puntos que motivan la creación de nuevos métodos destacan:

- ¿Es posible utilizar una mayor información de los datos para obtener información más fidedigna del sistema?
- Generalmente no se realiza una sola medición de los resultados, si no que los experimentos se realizan en copias, y los datos finalmente utilizados corresponden al promedio de las mediciones. ¿Es posible analizar la varianza de dichas mediciones?
- ¿Es posible incorporar información externa a los datos?
- Si se tiene certeza del signo o magnitud de algún dato a reconstruir, ¿es posible utilizar dicha información?
- ¿Y si sólo se tiene una certeza parcial respecto a dicha información?
- ¿Se puede obtener una suerte de precisión de las reconstrucciones?
- ¿En qué datos reconstruidos es posible confirmar más?

En los párrafos siguientes se describe la motivación particular de cada método creado y se analiza en extenso su implementación.

3.9 Errores en los datos y NCA considerando confiabilidad

Un problema común en los problemas inversos es los errores a los que están sujetos los datos de entrada. Dichos datos[42]son utilizados para reconstruir las matrices A y P (o descomponer E en las mismas) estimando los parámetros desconocidos en éstas. Por supuesto, los errores en los datos se propagarán también a los parámetros reconstruidos, por lo que se hace imprescindible generar alguna técnica que logre disminuir esta distorsión..

Por lo general los datos de microarrays se obtienen en duplicados e incluso triplicados, y los datos utilizados para trabajar corresponden al promedio de los mismos. De la misma manera, es posible calcular la varianza de cada entrada de la matriz de datos, en función de las diferentes mediciones que se pudiesen haber realizado.

[41] En el sentido de su exactitud.
[42] En conjunto con otra información parcial.

Sean $E_{m\_1}, E_{m\_2}, \dots, E_{m\_D}$ las D repeticiones de los experimentos microarray en que se obtienen los datos. Sea además e_{ikd} la expresión del gen i en el experimento k en la repetición d de los experimentos ($i = 1, \dots, N; k = 1, \dots, M; d = 1, \dots, D$). Con estos datos es posible calcular el promedio de los datos y la varianza de los mismos. Esto es:

$$\bar{E} = E_p = \begin{bmatrix} \sum_{d=1}^{D} e_{ikf} \end{bmatrix} = \begin{bmatrix} \sum_{d=1}^{D} e_{11d} & \sum_{d=1}^{D} e_{12d} & \cdots \\ \vdots & \ddots & \\ \sum_{d=1}^{D} e_{N1d} & \cdots & \sum_{d=1}^{D} e_{NMd} \end{bmatrix} = \begin{bmatrix} \overline{e_{11}} & \overline{e_{12}} & \cdots \\ \vdots & \ddots & \\ \overline{e_{N1}} & & \overline{e_{NM}} \end{bmatrix}$$

Ecuación 34

$$Var(e_{ik}) = \frac{1}{D}\sum_{d=1}^{D} (\overline{e_{ik}} - e_{ikd})^2 \qquad \forall\, i = \{1, \dots, N\}, \qquad k = \{1, \dots, M\}$$

Ecuación 35[43]

Más aun, es posible definir la covarianza de los datos, aunque dicha expresión no se utilizará en el análisis.

$$Cov(e_{ik}, e_{i'k'}) = \frac{1}{D}\sum_{d=1}^{D} (\overline{e_{ik}} - e_{ikd}) \cdot (\overline{e_{i'k'}} - e_{i'k'd}) \qquad \forall\, i, i' = \{1, \dots, N\}, \qquad k, k' = \{1, \dots, M\}$$

Ecuación 36

Finalmente se considerará la expresión de todos los genes en un experimento particular. Se define entonces V_k la matriz de varianza-covarianza de los genes para el experimento k, $k = \{1, \dots, M\}$.

$$R_k = \begin{bmatrix} Var(e_{1k}) & Cov(e_{1k}, e_{2k}) & \dots & Cov(e_{1k}, e_{Nk}) \\ Cov(e_{2k}, e_{1k}) & Var(e_{2k}) & & \vdots \\ \vdots & & \ddots & \vdots \\ Cov(e_{Nk}, e_{1k}) & \dots & \dots & Var(e_{Nk}) \end{bmatrix}$$

Ecuación 37

Desde ahora no se considerará que las covarianzas entre datos es significativa, por lo que la matriz anterior se simplifica como sigue.

[43] Es posible utilizar también un estimador insesgado de la varianza, pero para el análisis construido es irrelevante.

$$R_k = \begin{bmatrix} Var(e_{1k}) & 0 & \dots & 0 \\ 0 & Var(e_{2k}) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & Var(e_{Nk}) \end{bmatrix}$$

Ecuación 38

Finalmente se define la matriz de desviaciones estándar de los genes para un experimento k, $k = \{1, \dots, M\}$ como sigue.

$$S_k = \begin{bmatrix} \sqrt{Var(e_{1k})} & 0 & \dots & 0 \\ 0 & \sqrt{Var(e_{2k})} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & V\sqrt{ar(e_{Nk})} \end{bmatrix}$$

Ecuación 39

El objetivo principal es utilizar los datos anteriores respecto a la variabilidad de la matriz de datos E con el fin de mejorar la estimación de los parámetros frente a errores.

3.9.1 Utilización de matrices de confiabilidad

La idea detrás del uso de matrices de confiabilidad sigue la línea descrita en el punto anterior. Se considera el problema genérico como el de la Ecuación 17.

$$\min_x \ \|b - A \cdot x\|_F{}^2$$

Al intentar ajustar la expresión $A \cdot x$ a los datos en b (que se asumirá es un vector de datos de dimensión N) es posible que algunos estén medidos con mayor exactitud que otros (menor varianza), o lo que es lo mismo, que la estimación del valor real dado por el promedio en algun vector de datos b sea más confiable. Si R_b es la matriz de varianza-covarianza del vector b, y considerando cero la covarianza entre datos:

$$R_b = \begin{bmatrix} Var(b_1) & 0 & \dots & 0 \\ 0 & Var(b_2) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & Var(b_N) \end{bmatrix}$$

Es posible definir una matriz de confiabilidad de los datos, definida por la matriz ${R_b}^{-1}$.

$$R_b{}^{-1} = \begin{bmatrix} \frac{1}{Var(b_1)} & 0 & \dots & 0 \\ 0 & \frac{1}{Var(b_2)} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{Var(b_N)} \end{bmatrix}$$

Ecuación 40

Sea entonces el siguiente problema, en que se ha modificado el funcional de la expresión anterior.

$$\min_x \ \|b - A \cdot x\|_F{}^2{}_{R_b{}^{-1}} = (b - A \cdot x)^t \cdot R_b{}^{-1} \cdot (b - A \cdot x)$$

Ecuación 41

Como se observa, $R_b{}^{-1}$ es una matriz de precisión. Mientras mayor varianza tenga un dato, menor precisión se le asociará a la coordenada de b. Analizando con mayor detalle el funcional de la Ecuación 41.

$$\|b - A \cdot x\|_F{}^2{}_{R_b{}^{-1}} = (b - A \cdot x)^t \cdot R_b{}^{-1} \cdot (b - A \cdot x)$$

$$= \begin{pmatrix} b_1 - A_{r1} \cdot x \\ b_2 - A_{r2} \cdot x \\ \vdots \\ b_N - A_{rN} \cdot x \end{pmatrix}^t \cdot \begin{bmatrix} Var(b_1)^{-1} & & 0 \\ & \ddots & \\ 0 & & Var(b_N)^{-1} \end{bmatrix} \cdot \begin{pmatrix} b_1 - A_{r1} \cdot x \\ b_2 - A_{r2} \cdot x \\ \\ b_N - A_{rN} \cdot x \end{pmatrix}$$

$$= (b_1 - A_{r1} \cdot x)^2 Var(b_1)^{-1} + (b_2 - A_{r2} \cdot x)^2 Var(b_2)^{-1} + \cdots + (b_N - A_{rN} \cdot x)^2 Var(b_N)^{-1}$$

Lo anterior equivale a decir que el funcional a minimizar corresponde a una suerte de ponderación de los ajustes a cada uno de los datos en b, donde los ponderadores corresponden a la confiabilidad dada al dato en particular. De esta forma si el dato presenta una baja confiabilidad (alta varianza en su medición), al momento de encontrar el vector x que minimiza el ajuste a los datos se dará menos importancia al ajuste de los datos menos precisos

3.9.2 NCA general con confiabilidad en los datos

Es necesario modificar el funcional de NCA con el fin de incluir la teoría antes descrita. El funcional propuesto para este nuevo método es el siguiente.

$$J(E - A \cdot P) = \|E - A \cdot P\|_F{}^2{}_{R^{-1}}$$

Al minimizar dicha expresión, se toma en cuenta en mayor grado el ajuste a los datos que con mayor confiabilidad están medidos. Aquí R^{-1}, y dado que E es una matriz de dimensión NxM se define como

una matriz de la misma dimensión, en donde en cada coordenada posee la confiabilidad de la coordenada respectiva de E. Se define entonces para el problema propuesto:

$R_{ri}{}^{-1}$: Matriz diagonal que posee los datos de confiabilidad de la fila i de E en su diagonal, formada con los datos de la fila i de R^{-1}.

$R_{ck}{}^{-1}$: Matriz diagonal que posee los datos de confiabilidad de la columna k de E en su diagonal, formada con los datos de la columna k de R^{-1}.

El algoritmo de resolución es similar a los ya detallados, considerando por supuesto las diferencias en el funcional. El gradiente del funcional entrega una expresión para las matrices que minimizarán la expresión en cada iteración. En cada iteración, para A y P se resolverán problemas equivalentes al de la Ecuación 41.

$$\min_x \ \|b - A \cdot x\|_F{}^2{}_{R_b{}^{-1}} = (b - A \cdot x)^t \cdot R_b{}^{-1} \cdot (b - A \cdot x)$$

El funcional se puede descomponer como sigue:

$$(b - A \cdot x)^t \cdot R_b{}^{-1} \cdot (b - A \cdot x) =$$

$$(b^t - x^t \cdot A^t) \cdot R_b{}^{-1} \cdot (b - A \cdot x) = b^t R_b{}^{-1} b - b^t R_b{}^{-1} A x - x^t A^t R_b{}^{-1} b + x^t A R_b{}^{-1} x$$

Pero como $b^t R_b{}^{-1} A x$ es una constante, $\left(b^t R_b{}^{-1} A x\right)^t = x^t A^t R_b{}^{-1} b$. Así se tiene:

$$f(x) = (b - A \cdot x)^t \cdot R_b{}^{-1} \cdot (b - A \cdot x) = b^t R_b{}^{-1} b - 2 b^t R_b{}^{-1} A x + x^t A R_b{}^{-1} x$$

Finalmente es posible obtener el gradiente de la función a minimizar siguiendo las reglas de derivación de vectores, formas lineales y formas cuadráticas detalladas en los anexos.

$$\frac{df(x)}{dx} = \nabla f(x) = 0 - 2A^t R_b{}^{-1} b + 2A^t R_b{}^{-1} A x$$

El mínimo estará en el punto donde el gradiente vale cero, esto es:

$$\nabla f(x^*) = 2A^t R_b{}^{-1} b + 2A^t R_b{}^{-1} A x = 0$$

$$\left(A^t R_b{}^{-1} A\right) x^* = A^t R_b{}^{-1} b$$

$$\boldsymbol{x^* = \left(A^t R_b{}^{-1} A\right)^{-1} A^t R_b{}^{-1} b}$$

Ecuación 42

Notemos que aquí la expresión $A^t R_b{}^{-1} A$ es una matriz cuadrada, por lo que al cumplirse las condiciones NCA, la inversa existirá sin problemas.

Calculando la segunda derivada del funcional, se ve que efectivamente se trata de un mínimo.

$$\frac{d^2 f(x)}{d^2 x} = \mathcal{H} f(x) = 2(A^t {R_b}^{-1} A)$$

Ecuación 43

Que bajo ciertas condiciones generales es semi-definida positiva. Dada las características del problema, en cada iteración serán necesarios resolver M de estos problemas para P y N para A respetando siempre aquellas entradas en las que se impone la presencia de un cero en la matriz.

El problema es entonces:

$$\mathbb{P}: \min_{A,P} \|E - A \cdot P\|_F{}^2{}_{R^{-1}}$$
$$s.t.\ A \in Z_A \ \text{ y } \ P \in Z_P$$

Ecuación 44

El algoritmo de resolución es el siguiente.

Se genera una adivinación inicial para los parámetros a estimar de A y P, de forma que se respeten las restricciones en la estructura de A y P: $A^{(0)} \in Z_A$ y $P^{(0)} \in Z_P$.

1. Dado $A^{(0)}$, encontrar $P^{(1)}$ que resuelva:

$$\min_{P^{(1)}} \left\|E - A^{(0)} \cdot P^{(1)}\right\|_F{}^2{}_{R^{-1}}$$
$$s.t.\ P^{(1)} \in Z_P$$

   Esto se logra dividiendo el problema anterior en M problemas (uno para cada columna en E) de la forma ($k = 1, \dots, M$):

$$\min_{{P_{Rck}}^{(1)}} \left\|E_{ck} - {A_{Rk}}^{(0)} \cdot {P_{Rck}}^{(1)}\right\|_F{}^2{}_{{R_{ck}}^{-1}}$$

$$s.t.\ {P_{Rck}}^{(1)} \in Z_P$$

   , donde E_{ck} es la columna k de E y P_{Rck} es la columna k de P reducida; esto es la columna k de P original, a la que se han eliminado las entradas restringidas a cero en Z_P. A_{Rk} corresponde a la matriz A reducida, a la cual se han eliminado las columnas correspondientes a las filas eliminadas en P_{Rck}.

   De esta manera y dadas las condiciones NCA, los problemas anteriores tienen una solución dada por la Ecuación 42:

$$\boldsymbol{P_{Rck}}^{(1)} = \left(\boldsymbol{A_{Rk}}^{(0)^t} \cdot \boldsymbol{R_{ck}}^{-1} \cdot \boldsymbol{A_{Rk}}^{(0)}\right)^{-1} \boldsymbol{A_{Rk}}^{(0)^t} \cdot \boldsymbol{R_{ck}}^{-1} \cdot \boldsymbol{E_{ck}}$$

2. Dado $P^{(1)}$, encontrar $A^{(1)}$ que resuelva:

$$\min_{A^{(1)}} \left\| E - A^{(1)} \cdot P^{(1)} \right\|_{F\ R^{-1}}^{2}$$
$$s.t.\ A^{(1)} \in Z_A$$

Esto es equivalente a resolver el problema siguiente, el que convenientemente está escrito en la misma forma que la Ecuación 17.

$$\min_{A^{(1)}} \left\| E^t - P^{(1)^t} \cdot A^{(1)^t} \right\|_{F\ R^{-1}}^{2}$$
$$s.t.\ A^{(1)} \in Z_A$$

La estrategia de resolución es similar, dividiendo el problema anterior en N problemas (uno para cada columna en E^t) de la forma ($i = 1, \dots, N$):

$$\min_{A_{Rri}{}^{(1)}} \left\| E_{ri}{}^t - P_{Ri}{}^{(1)^t} \cdot A_{Rri}{}^{(1)^t} \right\|_{F\ R_{ri}{}^{-1}}^{2}$$
$$s.t.\ A_{Rri}{}^{(1)} \in Z_A$$

, donde $E_{ri}{}^t$ es la traspuesta de la fila i de E y $A_{Rri}{}^{(1)^t}$ es la traspuesta de la fila i de A reducida; esto es la traspuesta de la fila i de A original, a la que se han eliminado las entradas restringidas a cero en Z_A. $P_{Ri}{}^{(1)^t}$ corresponde a la matriz P traspuesta reducida, a la cual se han eliminado las columnas correspondientes a las entradas eliminadas en $A_{Rri}{}^{(1)^t}$.

De esta manera, los problemas anteriores tienen una solución $\boldsymbol{A_{Rri}}^{(1)}$ que puede ser obtenida por la siguiente expresión (utilizando la Ecuación 28 una vez más).

$$\boldsymbol{A_{Rri}}^{(1)} = \left(\boldsymbol{P_{Ri}}^{(1)} \cdot \boldsymbol{R_{ri}}^{-1} \cdot \boldsymbol{P_{Ri}}^{(1)^t}\right)^{-1} \boldsymbol{P_{Ri}}^{(1)} \cdot \boldsymbol{R_{ri}}^{-1} \cdot \boldsymbol{E_{ri}}^{t}$$

3. Se repite el paso 2. y 3. Hasta que el funcional evaluado en $A^{(n)}$ y $P^{(n)}$, $\left\| E - A^{(n)} \cdot P^{(n)} \right\|_F^{\ 2}$ sea menor que un cierto grado de tolerancia o el error haya convergido.

Finalmente las matrices A y P son normalizadas por una matriz X definida como en la Ecuación 27.

3.9.3 NCA general con confiabilidad en los datos y regularización

Se propone un nuevo método NCA que combine la teoría de la regularización en conjunto a la confiabilidad de los datos. El funcional propuesto para este nuevo método es el siguiente:

$$J(E - A \cdot P) = \frac{1}{2} \cdot \|E - A \cdot P\|_F{}^2{}_{R^{-1}} + \frac{\lambda}{2} \cdot \|P\|_F{}^2$$

Ecuación 45

El algoritmo de resolución es similar a los ya detallados, considerando por supuesto las diferencias en el funcional. El gradiente del funcional entrega una expresión para las matrices que minimizarán la expresión en cada iteración. En cada iteración, para A y P se resolverán problemas equivalentes al siguiente.

$$\min_x \ \|b - A \cdot x\|_F{}^2{}_{R_b{}^{-1}} + \frac{\lambda}{2} \cdot \|x\|_F{}^2 = \frac{1}{2} \cdot (b - A \cdot x)^t \cdot R_b{}^{-1} \cdot (b - A \cdot x) + \frac{\lambda}{2} \cdot x^t x$$

Ecuación 46

El funcional se puede descomponer como sigue:

$$\frac{1}{2} \cdot \left[(b - A \cdot x)^t \cdot R_b{}^{-1} \cdot (b - A \cdot x) + \lambda \cdot x^t x\right] =$$

$$= \frac{1}{2} \cdot \left[b^t R_b{}^{-1} b - b^t R_b{}^{-1} A x - x^t A^t R_b{}^{-1} b + x^t A R_b{}^{-1} x + \lambda \cdot x^t x\right]$$

Pero como $b^t R_b{}^{-1} A x$ es una constante, $\left(b^t R_b{}^{-1} A x\right)^t = x^t A^t R_b{}^{-1} b$. Así se tiene:

$$f(x) = \frac{1}{2} \cdot \left[b^t R_b{}^{-1} b - 2 b^t R_b{}^{-1} A x + x^t A R_b{}^{-1} x + \lambda \cdot x^t x\right]$$

Finalmente es posible obtener el gradiente de la función a minimizar siguiendo las reglas de derivación de vectores, formas lineales y formas cuadráticas detalladas en los anexos.

$$\frac{df(x)}{dx} = \nabla f(x) = \frac{1}{2} \cdot \left[0 - 2A^t R_b{}^{-1} b + 2A^t R_b{}^{-1} A x + 2\lambda x\right]$$

El mínimo estará en el punto donde el gradiente vale cero, esto es:

$$\nabla f(x^*) = 2A^t R_b{}^{-1} b + 2A^t R_b{}^{-1} A x + 2\lambda x = 0$$

$$\left(A^t R_b{}^{-1} A + \lambda x\right) x^* = A^t R_b{}^{-1} b$$

$$\boldsymbol{x^* = \left(A^t R_b{}^{-1} A + \lambda I\right)^{-1} A^t R_b{}^{-1} b}$$

Ecuación 47

Se hace notar que aquí la expresión $A^t {R_b}^{-1} A + \lambda x$ es una matriz cuadrada, por lo que al cumplirse las condiciones NCA, la inversa existirá sin problemas.

Calculando la segunda derivada del funcional, se ve que efectivamente se trata de un mínimo.

$$\frac{d^2 f(x)}{d^2 x} = \mathcal{H} f(x) = 2(A^t {R_b}^{-1} A + \lambda I)$$

Ecuación 48

Que bajo ciertas condiciones generales es semi-definida positiva. Dada las características del problema, en cada iteración será necesario resolver M de estos problemas para P y N para A respetando por supuesto aquellas entradas en las que se impone la presencia de un cero en la matriz.

El problema es entonces:

$$\mathbb{P}: \quad \min_{A,P} \ \frac{1}{2} \cdot \|E - A \cdot P\|_F{}^2{}_{R^{-1}} + \frac{\lambda}{2} \cdot \|P\|_F{}^2$$
$$s.t. \ A \in Z_A \ \text{ y } \ P \in Z_P$$

Ecuación 49

El algoritmo de resolución es el siguiente:

Se genera una adivinación inicial para los parámetros a estimar de A y P, de forma que se respeten las restricciones en la estructura de A y P: $A^{(0)} \in Z_A$ y $P^{(0)} \in Z_P$.

1. Dado $A^{(0)}$, encontrar $P^{(1)}$ que resuelva:

$$\min_{P^{(1)}} \ \frac{1}{2} \cdot \left\|E - A^{(0)} \cdot \ P^{(1)}\right\|_F{}^2{}_{R^{-1}} + \frac{\lambda}{2} \cdot \left\|P^{(1)}\right\|_F{}^2$$
$$s.t. \ P^{(1)} \in Z_P$$

   Esto se logra dividiendo el problema anterior en M problemas (uno para cada columna en E) de la forma ($k = 1, \dots, M$):

$$\min_{{P_{Rck}}^{(1)}} \ \frac{1}{2} \cdot \left\|E_{ck} - {A_{Rk}}^{(0)} \cdot \ {P_{Rck}}^{(1)}\right\|_F{}^2{}_{{R_{ck}}^{-1}} + \frac{\lambda}{2} \cdot \left\|{P_{Rck}}^{(1)}\right\|_F{}^2$$

$$s.t. \ {P_{Rck}}^{(1)} \in Z_P$$

   , donde E_{ck} es la columna k de E y P_{Rck} es la columna k de P reducida; esto es la columna k de P original, a la que se han eliminado las entradas restringidas a cero en Z_P. A_{Rk} corresponde a la matriz A reducida, a la cual se han eliminado las columnas correspondientes a las filas eliminadas en P_{Rck}.

De esta manera, y dadas las condiciones NCA, los problemas anteriores tienen una solución dada por la Ecuación 47:

$$\boldsymbol{P_{Rck}}^{(1)} = \left(\boldsymbol{A_{Rk}}^{(0)^t} \cdot \boldsymbol{R_{ck}}^{-1} \cdot \boldsymbol{A_{Rk}}^{(0)} + \boldsymbol{\lambda} \cdot \boldsymbol{I}\right)^{-1} \boldsymbol{A_{Rk}}^{(0)^t} \cdot \boldsymbol{R_{ck}}^{-1} \cdot \boldsymbol{E_{ck}}$$

2. Dado $P^{(1)}$, encontrar $A^{(1)}$ que resuelva (no se considera el termino de regulación debido a que el argumento a optimizar no lo afecta):

$$\min_{A^{(1)}} \left\| E - A^{(1)} \cdot P^{(1)} \right\|_{F\ R^{-1}}^{2}$$
$$s.t.\ A^{(1)} \in Z_A$$

Esto es equivalente a resolver el problema siguiente, el que convenientemente está escrito en la misma forma que la Ecuación 17.

$$\min_{A^{(1)}} \left\| E^t - {P^{(1)}}^t \cdot {A^{(1)}}^t \right\|_{F\ R^{-1}}^{2}$$
$$s.t.\ A^{(1)} \in Z_A$$

La estrategia de resolución es similar, dividiendo el problema anterior en N problemas (uno para cada columna en E^t) de la forma ($i = 1, \dots, N$):

$$\min_{{A_{Rri}}^{(1)}} \left\| {E_{ri}}^t - {P_{Ri}}^{(1)^t} \cdot {A_{Rri}}^{(1)^t} \right\|_{F\ {R_{ri}}^{-1}}^{2}$$
$$s.t.\ {A_{Rri}}^{(1)} \in Z_A$$

, donde ${E_{ri}}^t$ es la traspuesta de la fila i de E y ${A_{Rri}}^{(1)^t}$ es la traspuesta de la fila i de A reducida; esto es la traspuesta de la fila i de A original, a la que se han eliminado las entradas restringidas a cero en Z_A. ${P_{Ri}}^{(1)^t}$ corresponde a la matriz P traspuesta reducida, a la cual se han eliminado las columnas correspondientes a las entradas eliminadas en ${A_{Rri}}^{(1)^t}$.

De esta manera, los problemas anteriores tienen una solución $\boldsymbol{A_{Rri}}^{(1)}$ que puede ser obtenida por la siguiente expresión (utilizando la Ecuación 28 de apartados anteriores).

$$\boldsymbol{A_{Rri}}^{(1)} = \left(\boldsymbol{P_{Ri}}^{(1)} \cdot \boldsymbol{R_{ri}}^{-1} \cdot \boldsymbol{P_{Ri}}^{(1)^t}\right)^{-1} \boldsymbol{P_{Ri}}^{(1)} \cdot \boldsymbol{R_{ri}}^{-1} \cdot \boldsymbol{E_{ri}}^{t}$$

3. Se repite el paso 2. y 3. Hasta que el funcional evaluado en $A^{(n)}$ y $P^{(n)}$, $\left\| E - A^{(n)} \cdot P^{(n)} \right\|_F^{\ 2}$ sea menor que un cierto grado de tolerancia o el error haya convergido.

Finalmente las matrices A y P son normalizadas por una matriz X definida como en la Ecuación 27.

Algunas observaciones

Una detalle importante respecto a los métodos NCA propuestos es la necesidad de un input extra en adición a los datos resumidos en E. En estos casos es necesario incluir la varianza de la medición de los datos, una suerte de exactitud de estas mediciones. Si bien la matriz E puede corresponder al promedio de las mediciones, a una medición particular u otro estadístico de los parámetros reales, es necesario incluir forzosamente datos respecto a la variación de los mismos. Si bien el objetivo de estos métodos es obtener resultados más confiables, o mejorar las estimaciones obtenidos, hay que considerar este requerimiento extra a la hora de considerar su utilización.

3.10 Regularización general. Suposiciones a priori en A y P

Como se especificó en el método de regularización, introducido como una manera de superar los problemas de estabilidad propios de los problemas inversos, su idea principal era mantener las entradas de P cerca de cero. Esto puede ser extendido a una nueva funcionalidad del método NCA. Al imponer la regularización en P, implícitamente se realizaba la suposición de que las coordenadas de la matriz se encontraban cerca de cero. Dicho proceso se puede generalizar e imponer que diferentes coordenadas de P se encuentren cerca de ciertos valores, que no necesariamente deben ser el mismo. Incluso sería posible imponer esto en la matriz A creando una suerte de suposiciones iniciales para los parámetros en estudio. Por supuesto, el objetivo es combinar esta información preliminar con la información obtenida de los datos, a fin de lograr buenos estimadores de los verdaderos parámetros, y que por ende ajusten mejor la reconstrucción de las matrices.

Dichas suposiciones deben partir de un análisis previo, y deben estar fundamentadas debido a estudios anteriores o a sospechas particulares de los valores de algunos parámetros.

La implementación es como sigue. Sea el problema genérico entregado por la Ecuación 17, donde como siempre b es un vector columna de datos, A una matriz de parámetros fijos y x la incógnita, que busca ajustar de la mejor forma posible $A \cdot x$ a b.

$$\min_{x} \|b - A \cdot x\|_F^{\,2}$$

Sea además x_b[44] una matriz de suposiciones de x, donde en cada coordenada guarda la suposición de la coordenada correspondiente de x. De esta manera es posible implementar lo anterior mediante el planteamiento siguiente, similar al realizado para NCA con regularización.

[44] No confundir este sub-índice b de x con el vector b de los datos.

$$\min_x \ f(x) = \frac{1}{2} \cdot \|b - A \cdot x\|_F^{\ 2} + \frac{\lambda}{2} \cdot \|x - x_b\|_F^{\ 2}$$

Ecuación 50

De esta manera, al minimizar dicha expresión se estará buscando disminuir la diferencia respecto a los datos, mientras al mismo tiempo se desea mantener a x cercano a su suposición. Por supuesto dicho compromiso, tal cual en el método con regularización, lo controla el parámetro λ. Mientras mayor es, más importancia de dará a mantener x cerca de su suposición en la minimización, debido a que más peso se dará a dicha parte del funcional en la minimización.

La resolución de problemas como el anterior es simple, y puede ser resuelto aplicando el mismo método que en los casos anteriores. El funcional se puede descomponer como sigue:

$$f(x) = \frac{1}{2} \cdot \|b - A \cdot x\|_F^{\ 2} + \frac{\lambda}{2} \cdot \|x - x_b\|_F^{\ 2}$$
$$= \frac{1}{2} \cdot [(b - A \cdot x)^t \cdot (b - A \cdot x) + \lambda \cdot (x - x_b)^t (x - x_b)]$$

$$= \frac{1}{2} \cdot [b^t b - b^t A x - x^t A^t b + x^t A x + \lambda \cdot x^t x + \lambda \cdot {x_b}^t x_b - \lambda \cdot {x_b}^t x - \lambda \cdot x^t x_b]$$

Pero como $b^t Ax$ es una constante, $(b^t Ax)^t = x^t A^t b$. De la misma manera ${x_b}^t x = x^t x_b$. Así se tiene:

$$f(x) = \frac{1}{2} \cdot [(b - A \cdot x)^t \cdot (b - A \cdot x) + \lambda \cdot b^t b]$$
$$= \frac{1}{2} \cdot [b^t b - 2b^t A x + x^t A x + \lambda \cdot x^t x + \lambda \cdot {x_b}^t x_b - 2\lambda \cdot x^t x_b]$$

Finalmente es posible obtener el gradiente de la función a minimizar siguiendo las reglas de derivación de vectores, formas lineales y formas cuadráticas detalladas en los anexos.

$$\frac{df(x)}{dx} = \nabla f(x) = \frac{1}{2} \cdot [0 - 2A^t b + 2A^t A x + 2\lambda x - 2\lambda x_b]$$

El mínimo estará en el punto donde el gradiente vale cero, esto es:

$$\nabla f(x^*) = -A^t b + A^t A x^* + \lambda x^* - \lambda x_b = 0$$

$$(A^t A + \lambda I) x^* = A^t b + \lambda x_b$$

$$\boldsymbol{x^* = (A^t A + \lambda I)^{-1} (A^t b + \ \lambda x_b)}$$

Ecuación 51

Se hace notar que aquí la expresión $A^t A + \lambda I$ es una matriz cuadrada, por lo que al cumplirse las condiciones NCA, la inversa existirá sin problemas.

Calculando la segunda derivada del funcional, se ve que efectivamente se trata de un mínimo.

$$\frac{d^2 f(x)}{d^2 x} = \mathcal{H} f(x) = 2(A^t A + \lambda I)$$

Ecuación 52

Que bajo ciertas condiciones generales es semi-definida positiva.

Es posible ahora extender lo anterior a las matrices que se desea reconstruir. Sean A_b y P_b las matrices que contienen información respecto a las suposiciones de A y P respectivamente. A_b y P_b tendrán las mismas dimensiones que sus contrapartes y en cada coordenada tendrán la suposición de la misma coordenada de su homóloga correspondiente. La idea básica tras este planteamiento es similar a la anterior. Cada uno de los sub-problemas que se resuelven es similar al tratado con anterioridad, por lo que dicho enfoque es válido para un método NCA extendido a suposiciones. Una vez más, la reconstrucción de los parámetros tendrá 2 componentes, uno dado por los datos, y otro por las suposiciones, que en conjunto entregarán el estimador del parámetro. Por supuesto y como se verá existirán 2 parámetros que controlen la relación anterior, uno para cada matriz reconstruida.

Es útil sin embargo, mencionar algo respecto a los métodos de estimación en 2 partes o bayesianos [35], y su directa analogía con lo propuesto con anterioridad.

3.10.2 NCA general con suposiciones (agNCAr)

Se propone un nuevo método NCA que entregue la posibilidad de incluir suposiciones a priori de los parámetros, basados en información y sospechas justificadas. El funcional propuesto para este nuevo método es el siguiente:

$$J(E - A \cdot P) = \frac{1}{2} \cdot \|E - A \cdot P\|_F^{\,2} + \frac{\lambda_P}{2} \cdot \|P - P_b\|_F^{\,2} + \frac{\lambda_A}{2} \cdot \|A - A_b\|_F^{\,2}$$

Ecuación 53

Como se observa, existen ahora 2 parámetros que regulan el grado de compromiso que se adjudica a las suposiciones en relación a los datos, λ_P y λ_A. Si bien el problema luce diferente al propuesto en la Ecuación 50, es posible en cada iteración encontrar expresiones equivalentes.

El algoritmo de resolución es similar a los ya detallados, y consistente en un método de optimización alternado, en el cual minimizando alternadamente A y P se converge al mínimo del funcional. El gradiente del funcional entrega una expresión para las matrices que minimizarán la expresión en cada iteración, como se detalla en la Ecuación 51. EL problema a resolver es entonces:

$$\mathbb{P}: \min_{A,P} \quad \frac{1}{2} \cdot \|E - A \cdot P\|_F{}^2 + \frac{\lambda_P}{2} \cdot \|P - P_b\|_F{}^2 + \frac{\lambda_A}{2} \cdot \|A - A_b\|_F{}^2$$
$$s.t.\ A \in Z_A \ \text{ y } \ P \in Z_P$$

Ecuación 54

El algoritmo de resolución es el siguiente.

Se genera una adivinación inicial para los parámetros a estimar de A y P, de forma que se respeten las restricciones en la estructura de A y P: $A^{(0)} \in Z_A$ y $P^{(0)} \in Z_P$.

1. Dado $A^{(0)}$, encontrar $P^{(1)}$ que resuelva:

$$\min_{P^{(1)}} \ \frac{1}{2} \cdot \left\|E - A^{(0)} \cdot P^{(1)}\right\|_F{}^2 + \frac{\lambda_P}{2} \cdot \left\|P^{(1)} - P_b\right\|_F{}^2$$
$$s.t.\ P^{(1)} \in Z_P$$

   Como se aprecia, es un problema similar al dado por la Ecuación 50. Se ha omitido el término de suposición de A debido a que al minimizar en P dicho término no es relevante. La resolución del problema se logra dividiendo el problema anterior en M problemas (uno para cada columna en E) de la forma ($k = 1, \dots, M$):

$$\min_{P_{Rck}{}^{(1)}} \frac{1}{2} \cdot \left\|E_{ck} - A_{Rk}{}^{(0)} \cdot P_{Rck}{}^{(1)}\right\|_F{}^2 + \frac{\lambda_P}{2} \cdot \left\|P_{Rck}{}^{(1)} - P_{b_{Rck}}\right\|_F{}^2$$
$$s.t.\ P_{Rck}{}^{(1)} \in Z_P$$

   , donde E_{ck} es la columna k de E y P_{Rck} es la columna k de P reducida; esto es la columna k de P original, a la que se han eliminado las entradas restringidas a cero en Z_P. De la misma manera $P_{b_{Rck}}$ es la columna k de P_b reducida, definida de la misma manera que las reducidas de P. A_{Rk} corresponde a la matriz A reducida, a la cual se han eliminado las columnas correspondientes a las filas eliminadas en P_{Rck}.

   De esta manera y dadas las condiciones NCA, los problemas anteriores tienen una solución dada por la Ecuación 50.

$$\boldsymbol{P_{Rck}{}^{(1)} = \left(A_{Rk}{}^{(0)^t} \cdot A_{Rk}{}^{(0)} + \lambda_P \cdot I\right)^{-1} \left(A_{Rk}{}^{(0)^t} \cdot E_{ck} + \lambda_P \cdot P_{b_{Rck}}\right)}$$

2. Dado $P^{(1)}$, encontrar $A^{(1)}$ que resuelva (no se considera el termino de regulación debido a que el argumento a optimizar no lo afecta):

$$\min_{A^{(1)}} \left\| E - A^{(1)} \cdot P^{(1)} \right\|_F^2 + \frac{\lambda_A}{2} \cdot \left\| A^{(1)} - A_b \right\|_F^2$$
$$s.t.\ A^{(1)} \in Z_A$$

Igual que antes, el problema es similar al dado por la Ecuación 50. Esto es equivalente a resolver el problema siguiente, el que convenientemente está escrito en la misma forma que la Ecuación 17.

$$\min_{A^{(1)}} \left\| E^t - P^{(1)^t} \cdot A^{(1)^t} \right\|_F^2 + \frac{\lambda_A}{2} \cdot \left\| A^{(1)^t} - A_b{}^t \right\|_F^2$$
$$s.t.\ A^{(1)} \in Z_A$$

La estrategia de resolución es similar, dividiendo el problema anterior en N problemas (uno para cada columna en E^t) de la forma ($i = 1, \dots, N$):

$$\min_{A_{Rri}{}^{(1)}} \left\| E_{ri}{}^t - P_{Ri}{}^{(1)^t} \cdot A_{Rri}{}^{(1)^t} \right\|_F^2 + \frac{\lambda_A}{2} \cdot \left\| A_{Rri}{}^{(1)^t} - A_{b_{Rri}}{}^t \right\|_F^2$$
$$s.t.\ A_{Rri}{}^{(1)} \in Z_A$$

, donde $E_{ri}{}^t$ es la traspuesta de la fila i de E y $A_{Rri}{}^{(1)^t}$ es la traspuesta de la fila i de A reducida; esto es la traspuesta de la fila i de A original, a la que se han eliminado las entradas restringidas a cero en Z_A. $A_{b_{Rri}}{}^t$ es la traspuesta de la fila i de A_b reducida, definida de la misma manera que para las reducidas de A. $P_{Ri}{}^{(1)^t}$ corresponde a la matriz P traspuesta reducida, a la cual se han eliminado las columnas correspondientes a las entradas eliminadas en $A_{Rri}{}^{(1)^t}$.

De esta manera, los problemas anteriores tienen una solución $\boldsymbol{A_{Rri}{}^{(1)}}$ que puede ser obtenida por la siguiente expresión (utilizando la Ecuación 50 una vez más).

$$\boldsymbol{A_{Rri}{}^{(1)} = \left(P_{Ri}{}^{(1)} \cdot P_{Ri}{}^{(1)^t} + \lambda_A \cdot I \right)^{-1} \left(P_{Ri}{}^{(1)} \cdot E_{ri}{}^t + \lambda_A \cdot A_{b_{Rri}}{}^t \right)}$$

3. Se repite el paso 2. y 3. Hasta que el funcional evaluado en $A^{(n)}$ y $P^{(n)}$, $\left\| E - A^{(n)} \cdot P^{(n)} \right\|_F^2$ sea menor que un cierto grado de tolerancia o el error haya convergido.

Finalmente las matrices A y P son normalizadas por una matriz X definida como en la Ecuación 27.

Comentarios sobre el método

El método antes descrito modifica en gran medida al método original, al permitir incluir una extensa fuente de información en la reconstrucción. Sin embargo y pese al gran potencial que esta herramienta puede entregar, es necesario considerar algunos aspectos técnicos y de implementación. En primer lugar este método requiere una gran cantidad de información. En ocasiones se puede tener una idea respecto al valor de un parámetro y escasa información respecto al resto, por lo que dicha limitante podría perjudicar su implementación. Más aun, los datos en las matrices se encuentran normalizados, lo que

indica que el valor en sí no es lo que entrega información, sino su magnitud relativa a las otras entradas y su signo. De esta forma los datos deben entregarse en cierto sentido normalizados, con el fin de evitar que suposiciones mal planteadas perjudiquen la reconstrucción proveniente de la información atribuida a los datos experimentales. Estos y otros puntos serán analizados experimentalmente y discutidos con mayor profundidad en puntos posteriores.

Es interesante además la posibilidad de relajar los ceros impuestos en A, e imponer una suposición en su lugar, como una manera de explorar redes alternativas y el comportamiento de los métodos.

3.11 Grado de certeza en las suposiciones de A y P

Del punto anterior, y dada la analogía del método de suposiciones a la estimación de 2 partes, surge la idea y la necesidad de trabajar con parámetros que manejen la certeza de las suposiciones que se puedan hacer para las matrices que se requiere reconstruir. En forma análoga a la confiabilidad que se atribuya a los datos, es posible asignar el mismo parámetro a las suposiciones de los valores a estimar.

Lo anterior se fundamenta en alguno de los puntos discutidos en cgNCAr y en la necesidad de asignar distinta ponderación en el proceso de reconstrucción a aquellas entradas en que se tiene una mayor certeza respecto a su valor, o a la suposición de su valor. De la misma manera, es posible no tener certeza respecto a la suposición de algunas entradas, en cuyo caso asignar el mismo grado de confiabilidad que a las suposiciones más certeras sería un error que se proyectaría en las reconstrucciones.

Con lo anterior en mente, se definen entonces las matrices D y B como las matrices de variabilidad de las suposiciones de A_b y P_b respectivamente. Esto es, matrices de las mismas dimensiones que las de las suposiciones, donde en cada coordenada poseen una suerte de desviación estándar de la misma suposición. A mayor sea el valor de dicha desviación, menor certeza se tendrá respecto a la validez de la suposición propuesta en las matriz de suposiciones.

El problema genérico que es necesario resolver y que implementa lo anterior es el siguiente (donde como siempre b es un vector columna y x se considerará un vector de dimensión T)

$$\min_x \ f(x) = \frac{1}{2} \cdot \|b - A \cdot x\|_F^{\ 2} + \frac{\lambda}{2} \cdot \|x - x_b\|_F^{\ 2}{}_{H^{-1}}$$

Ecuación 55

, donde H es la matriz de varianza de las suposiciones de x (resumidas en x_b). Dicha matriz corresponderá a una matriz diagonal, en la cual estarán resumidas la varianza[45] de las suposiciones de x. Así H_{tt} resumirá la varianza de la suposición de la coordenada t de x. H^{-1} corresponderá entonces a una matriz diagonal de confiabilidad de las suposiciones de x, en donde cada coordenada estará definida

[45] O si se quiere, la exactitud de la suposición.

como la inversa de la correspondiente en H. Se hace notar la analogía con el método de confiabilidad propuesto para los datos y discutido con anterioridad.

Analizando con mayor detalle el funcional de la Ecuación 55, y definiendo por conveniencia H^{-1} como sigue:

$$H^{-1} = \begin{bmatrix} \frac{1}{Var(x_{b_1})} & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & \frac{1}{Var(x_{b_T})} \end{bmatrix}$$

Ecuación 56

, donde $Var(x_{b_t})$ corresponde a la variabilidad de la suposición de la coordenada t de x[46]. La parte del funcional correspondiente al término relacionado con la suposición puede descomponerse como:

$$\|x - x_b\|_F{}^2{}_{H^{-1}} = (x - x_b)^t \cdot H^{-1} \cdot (x - x_b)$$

$$= \begin{pmatrix} x_1 - x_{b_1} \\ x_2 - x_{b_2} \\ \vdots \\ x_T - x_{b_T} \end{pmatrix}^t \cdot \begin{bmatrix} Var(x_{b_1})^{-1} & & 0 \\ & \ddots & \\ 0 & & Var(x_{b_T})^{-1} \end{bmatrix} \cdot \begin{pmatrix} x_1 - x_{b_1} \\ x_2 - x_{b_2} \\ \vdots \\ x_T - x_{b_T} \end{pmatrix}$$

$$= (x_1 - x_{b_1})^2 \cdot ds(x_{b_1})^{-1} + (x_2 - x_{b_2})^2 \cdot Var(x_{b_2})^{-1} + \cdots + (x_T - x_{b_T})^2 \cdot Var(x_{b_T})^{-1}$$

Como se puede apreciar, al minimizar el funcional de la Ecuación 55, aquellas suposiciones con mayor certeza (menor variabilidad) tendrán un mayor peso respecto a las con menor confiabilidad. Luego, el funcional a minimizar puede verse como un término de ajuste a los datos y otro de ajuste a las suposiciones, que corresponde a una suma ponderada de los ajustes de cada coordenada de x. Por ende, el minimizar la expresión, el método tenderá a forzar en mayor medida a estar cerca de su suposición a aquellas coordenadas que tengan mayor confiabilidad, y por ende un mayor peso en el funcional a minimizar. No se debe olvidar también que el compromiso entre el primer y segundo término de la función está controlado por el parámetro λ, independiente del grado de confiabilidad que se le asigne a las suposiciones.

Con el fin de implementar lo anterior al método NCA es necesario realizar algunas definiciones adicionales respecto a las matrices de suposiciones antes presentadas. En conjunto con la definición entregada para D y B, sean:

> D_{ri}: Matriz de variabilidad. Matriz diagonal que posee los datos de confiabilidad de la fila i de A en su diagonal, formada con los datos de la fila i de D.

[46] Se ahondará más respecto al significado de la varianza de una suposición en breve.

B_{ck}: Matriz de variabilidad. Matriz diagonal que posee los datos de confiabilidad de la columna k de P en su diagonal, formada con los datos de la columna k de B.

${D_{ri}}^{-1}$: Matriz de confiabilidad. Corresponde a una matriz diagonal que entrega la confiabilidad de las suposiciones de la fila i de A. Cada entrada corresponde a las reciprocas de las entradas de D_{ri}.

${B_{ck}}^{-1}$: Matriz de confiabilidad. Corresponde a una matriz diagonal que entrega la confiabilidad de las suposiciones de la columna k de P. Cada entrada corresponde a las reciprocas de las entradas de B_{ck}.

, que corresponderán a las matrices utilizadas en el proceso de iteración NCA, donde cada sub-problema será reducido a una expresión similar a la de la Ecuación 55[47].

3.11.1 Definición y cálculo de confiabilidad de suposiciones

Otro punto importante a considerar de forma previa a implementar el método, se relaciona con los valores que se deben ingresar en las matrices de confiabilidad o variabilidad, que a simple vista parecen totalmente arbitrarios e independientes. Por supuesto esto no es así, ya que si bien el valor en sí no dice nada (son ponderadores) la diferencia relativa entre las confiabilidades de una y otra suposición si es relevante; incluso el valor puede tornarse relevante, si es que se utiliza en algún cálculo o en otro método como se verá luego. En el caso de la confiabilidad de los datos, estos eran obtenidos directamente como un estadístico de las mediciones, por lo que dicha preocupación no era relevante. En este caso la interpretación de la variabilidad de las suposiciones no es tan directa.

En relación a lo anterior, es posible interpretar cada suposición como una distribución de probabilidad. El promedio de dicha distribución correspondería a la suposición en sí, mientras que la dispersión de la misma indicaría su variabilidad[48]. En la figura siguiente se esquematiza dicha idea para una coordenada cualquiera de x de la Ecuación 55.

[47] En lo anterior, las expresiones corresponden a matrices diagonales, por lo que se ha asumido que no existe covarianza entra las suposiciones. Luego, la variabilidad de una observación no está relacionada con la de otra.
[48] O grado de inexactitud.

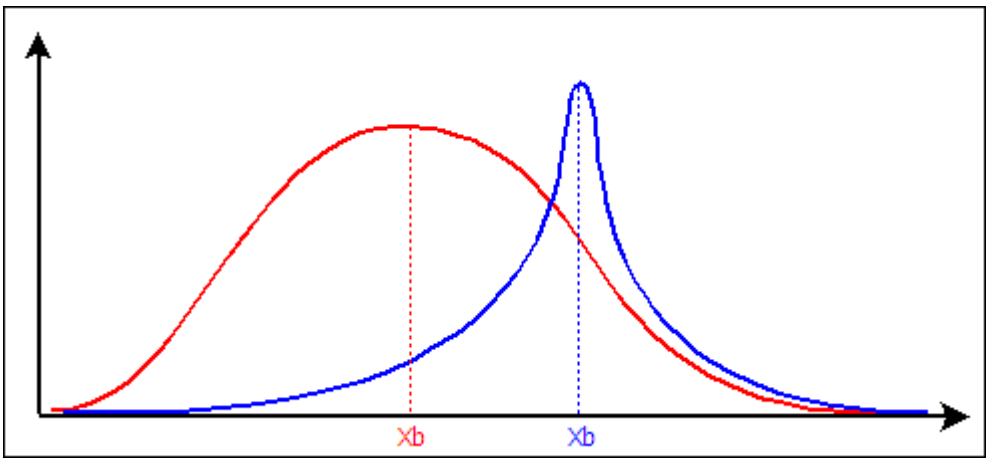


Figura 22: Distribución de suposiciones.
Fuente: Elaboración propia.

Es posible observar 2 suposiciones para coordenadas cualquieras de x. La primera, en rojo, posee una alta variabilidad, ya que su curva de distribución es bastante ancha respecto al promedio (posee una alta varianza). En cambio la suposición en azul parece ser una suposición más confiable, en el sentido de que su distribución parece menos dispersa en torno al promedio (existe una mayor probabilidad de que el valor real se encuentre cerca de la suposición).

Con el fin de trabajar con lo anterior, se supondrá que cada suposición realizada se distribuirá según una distribución normal de media igual a la suposición particular que se esté utilizando, y de varianza (o desviación estándar) tal que su curva represente el grado de certeza que se tenga respecto a la suposición. Esto es:

$$X_b \longrightarrow N(x_b, {\sigma_{x_b}}^2)$$

Ecuación 57

, donde X_b representa la distribución y x_b en minúscula la media de la distribución (la suposición en si considerada).

Con esto en mente, las distribuciones efectivamente tendrán la forma especificada en la Figura 22 y por ende será simétrica. Más aún, para una distribución normal, se tendrá que con un 95% de probabilidad es posible encontrar la suposición en un intervalo como el siguiente, a una distancia de 1,96 desviaciones estándar de la suposición.

$$P\left(X_b \in \left[x_b - 1{,}96 \cdot \sigma_{x_b}\, , x_b + 1{,}96 \cdot \sigma_{x_b}\right]\right) = 0{,}95$$

Ecuación 58

Utilizando un planteamiento inverso, se puede encontrar la varianza de una distribución normal X_b con media x_b tal que con un 95% de probabilidad la suposición se encuentre en cierto intervalo rodeando a la suposición x_b. Así por ejemplo, si para una suposición cualquiera es aceptada una flexibilidad 30%, el

objetivo es encontrar una desviación estándar σ_{x_b} tal que con un 95% de posibilidades la suposición se ubique en el intervalo definido por: $[x_b(1-0.3)\,,\ x_b(1+0.3)]$. En otras palabras se desea encontrar una desviación estándar de la desviación que asegure que el 95% de la variabilidad permitida para la suposición se ubique en dicho intervalo. Si se disminuye el intervalo[49], la desviación estándar debiese disminuir. Genéricamente, sea f_{x_b} el grado de flexibilidad admitido para la suposición x_b, se busca encontrar una desviación estándar para una normal de media x_b, tal que:

$$P(\,X_b \in \left[x_b\left(1-f_{x_b}\right),\ x_b\left(1+f_{x_b}\right)\right] = 0{,}95$$

Esto se puede lograr por igualación de términos con la Ecuación 58, por lo que se tendrá:

$$x_b(1+f) = x_b + 1{,}96 \cdot \sigma_{x_b}$$

$$\sigma_{x_b} = \frac{x_b \cdot f_{x_b}}{1{,}96}$$

Ecuación 59

Se observa que efectivamente, a medida que la flexibilidad disminuye, la varianza también lo hace, asemejándose más a la curva azul de la Figura 22.

Es posible resumir la información de la flexibilidad de cada suposición en una matriz F_x que en cada coordenada guarde la flexibilidad admitida para la coordenada correspondiente de X. Luego con esto y la Ecuación 59 es posible calcular la variabilidad (desviación estándar) y con esta la varianza de cada medición con el fin de construir la matriz H.

Esto es fácilmente aplicable al método NCA. En matrices F_A y F_P se tendrán resumidas las flexibilidades admitidas para las suposiciones de las coordenadas correspondientes de A y P. Con esto es posible obtener la desviación estándar para cada suposición y construir las matrices D y B. El esquema siguiente especifica lo anterior.

[49] Se disminuye la flexibilidad o se acota el intervalo en donde se permite este la mayor parte de la variabilidad de la suposición.

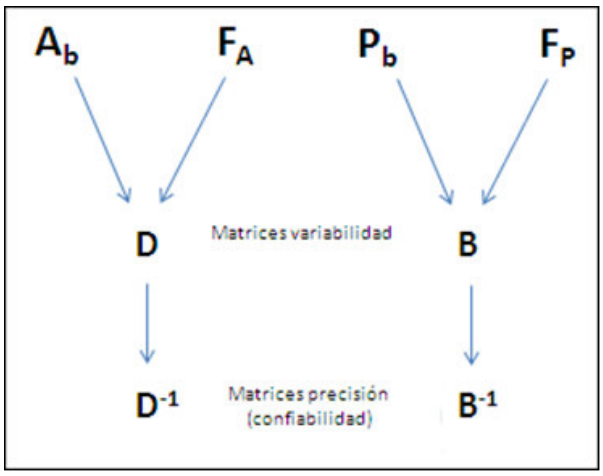


Figura 23: Esquema construcción matrices precisión de suposiciones.
Fuente: Elaboración propia.

3.11.2 NCA general con confiabilidad en suposiciones (acgNCAreg).

Se propone un nuevo método NCA que incorpore la teoría planteada con anterioridad, permitiendo combinar las suposiciones con un factor de confiabilidad de las mimas, que permita mejorar los resultado de la reconstrucción y discriminar entre suposiciones con mayor valor que otras.. El funcional propuesto para este nuevo método es el siguiente.

$$J(E - A \cdot P) = \frac{1}{2} \cdot \|E - A \cdot P\|_F{}^2 + \frac{\lambda_P}{2} \cdot \|P - P_b\|_F{}^2{}_{B^{-1}} + \frac{\lambda_A}{2} \cdot \|A - A_b\|_F{}^2{}_{D^{-1}}$$

Ecuación 60

Como se muestra a continuación, al minimizar el funcional anterior de acuerdo al procedimiento utilizando en los demás métodos NCA, se resolverán problemas equivalentes al de la Ecuación 55.

$$\min_x \ f(x) = \frac{1}{2} \cdot \|b - A \cdot x\|_F{}^2 + \frac{\lambda}{2} \cdot \|x - x_b\|_F{}^2{}_{H^{-1}}$$

El mínimo de la expresión anterior puede ser obtenido utilizando el método del gradiente del funcional ya descrito. Esta puede ser descompuesta de la siguiente forma:

$$f(x) = \frac{1}{2} \cdot \|b - A \cdot x\|_F{}^2 + \frac{\lambda}{2} \cdot \|x - x_b\|_F{}^2{}_{H^{-1}}$$

$$= \frac{1}{2} \cdot [(b - A \cdot x)^t \cdot (b - A \cdot x) + \lambda \cdot (x - x_b)^t \cdot H^{-1} \cdot (x - x_b)]$$

$$= \frac{1}{2} \cdot [b^t b - b^t A x - x^t A^t b + x^t A x + \lambda \cdot x^t H^{-1} x + \lambda \cdot x_b{}^t H^{-1} x_b - \lambda \cdot x_b{}^t H^{-1} x - \lambda \cdot x^t H^{-1} x_b]$$

Pero como $b^t Ax$ es una constante, $(b^t Ax)^t = x^t A^t b$. De la misma manera ${x_b}^t H^{-1} x = x^t H^{-1} x_b$. Así se tiene:

$$f(x) = \frac{1}{2} \cdot [(b - A \cdot x)^t \cdot (b - A \cdot x) + \lambda \cdot b^t b]$$
$$= \frac{1}{2} \cdot [b^t b - 2b^t Ax + x^t Ax + \lambda \cdot x^t H^{-1} x + \lambda \cdot {x_b}^t H^{-1} x_b - 2\lambda \cdot x^t H^{-1} x_b]$$

Finalmente es posible obtener el gradiente de la función a minimizar siguiendo las reglas de derivación de vectores, formas lineales y formas cuadráticas detalladas en los anexos.

$$\frac{df(x)}{dx} = \nabla f(x) = \frac{1}{2} \cdot [0 - 2A^t b + 2A^t Ax + 2\lambda H^{-1} x - 2\lambda H^{-1} x_b]$$

El mínimo estará en el punto donde el gradiente vale cero, esto es:

$$\nabla f(x^*) = -A^t b + A^t Ax^* + \lambda H^{-1} x^* - \lambda H^{-1} x_b = 0$$

$$(A^t A + \lambda H^{-1}) x^* = A^t b + \lambda H^{-1} x_b$$

$$\boldsymbol{x^* = (A^t A + \lambda H^{-1})^{-1} (A^t b + \lambda H^{-1} x_b)}$$

Ecuación 61

Se hace notar que aquí la expresión $A^t A + \lambda H^{-1}$ es una matriz cuadrada, por lo que al cumplirse las condiciones NCA, la inversa existirá sin problemas.

Calculando la segunda derivada del funcional, se ve que efectivamente se trata de un mínimo.

$$\frac{d^2 f(x)}{d^2 x} = \mathcal{H} f(x) = 2(A^t A + \lambda H^{-1})$$

Ecuación 62

Que bajo ciertas condiciones generales es semi-definida positiva.

El algoritmo de resolución es similar a los ya detallados, y consistente en un método de optimización alternado, en el cual minimizando alternadamente A y P se converge al mínimo del funcional. El gradiente del funcional entrega una expresión para las matrices que minimizarán la expresión en cada iteración, como de detalla en la Ecuación 62. EL problema a resolver es entonces:

$$\mathbb{P}: \quad \min_{A,P} \quad \frac{1}{2} \cdot \|E - A \cdot P\|_F^{\,2} + \frac{\lambda_P}{2} \cdot \|P - P_b\|_{F\,B^{-1}}^{\,2} + \frac{\lambda_A}{2} \cdot \|A - A_b\|_{F\,D^{-1}}^{\,2}$$
$$s.t. \;\; A \in Z_A \;\; \text{y} \;\; P \in Z_P$$

Ecuación 63

El algoritmo de resolución es el siguiente.

Se genera una adivinación inicial para los parámetros a estimar de A y P, de forma que se respeten las restricciones en la estructura de A y P: $A^{(0)} \in Z_A$ y $P^{(0)} \in Z_P$.

1. Dado $A^{(0)}$, encontrar $P^{(1)}$ que resuelva:

$$\min_{P^{(1)}} \ \frac{1}{2} \cdot \left\| E - A^{(0)} \cdot P^{(1)} \right\|_F^2 + \frac{\lambda_P}{2} \cdot \left\| P^{(1)} - P_b \right\|_{F\ B^{-1}}^2$$
$$s.t.\ P^{(1)} \in Z_P$$

   Como se puede observar, es un problema similar al dado por la Ecuación 55; se ha omitido el término de suposición de A debido a que al minimizar en P dicho término no es relevante. La resolución del problema se logra dividiendo el problema anterior en M problemas (uno para cada columna en E) de la forma ($k = 1, \dots, M$):

$$\min_{{P_{Rck}}^{(1)}} \frac{1}{2} \cdot \left\| E_{ck} - {A_{Rk}}^{(0)} \cdot {P_{Rck}}^{(1)} \right\|_F^2 + \frac{\lambda_P}{2} \cdot \left\| {P_{Rck}}^{(1)} - P_{b_{Rck}} \right\|_{F\ {B_{Rck}}^{-1}}^2$$
$$s.t.\ {P_{Rck}}^{(1)} \in Z_P$$

   , donde E_{ck} es la columna k de E y P_{Rck} es la columna k de P reducida; esto es la columna k de P original, a la que se han eliminado las entradas restringidas a cero en Z_P. De la misma manera $P_{b_{Rck}}$ es la columna k de P_b reducida, definida de la misma manera que las reducidas de P. ${B_{Rck}}^{-1}$ corresponde a la versión reducida de ${B_{ck}}^{-1}$. A_{Rk} corresponde a la matriz A reducida, a la cual se han eliminado las columnas correspondientes a las filas eliminadas en P_{Rck}.

   De esta manera, y dadas las condiciones NCA, los problemas anteriores tienen una solución dada por la Ecuación 61.

$$\boldsymbol{{P_{Rck}}^{(1)} = \left({A_{Rk}}^{(0)^t} \cdot {A_{Rk}}^{(0)} + \lambda_P \cdot {B_{Rck}}^{-1} \right)^{-1} \left({A_{Rk}}^{(0)^t} \cdot E_{ck} + \lambda_P \cdot {B_{Rck}}^{-1} \cdot P_{b_{Rck}} \right)}$$

2. Dado $P^{(1)}$, encontrar $A^{(1)}$ que resuelva (no se considera el termino de regulación debido a que el argumento a optimizar no lo afecta):

$$\min_{A^{(1)}} \left\| E - A^{(1)} \cdot P^{(1)} \right\|_F^2 + \frac{\lambda_A}{2} \cdot \left\| A^{(1)} - A_b \right\|_{F\ D^{-1}}^2$$
$$s.t.\ A^{(1)} \in Z_A$$

Igual que antes, el problema es similar al dado por la Ecuación 55. Esto es equivalente a resolver el problema siguiente, el que convenientemente está escrito en la misma forma que la Ecuación 17.

$$\min_{A^{(1)}} \left\| E^t - P^{(1)^t} \cdot A^{(1)^t} \right\|_F^2 + \frac{\lambda_A}{2} \cdot \left\| A^{(1)^t} - A_b{}^t \right\|_{F\ D^{-1}}^2$$
$$s.t.\ A^{(1)} \in Z_A$$

La estrategia de resolución es similar, dividiendo el problema anterior en N problemas (uno para cada columna en E^t) de la forma ($i = 1, \dots, N$):

$$\min_{A_{Rri}{}^{(1)}} \left\| E_{ri}{}^t - P_{Ri}{}^{(1)^t} \cdot A_{Rri}{}^{(1)^t} \right\|_F^2 + \frac{\lambda_A}{2} \cdot \left\| A_{Rri}{}^{(1)^t} - A_{b\,Rri}{}^t \right\|_{F\ D_{Rri}{}^{-1}}^2$$
$$s.t.\ A_{Rri}{}^{(1)} \in Z_A$$

, donde $E_{ri}{}^t$ es la traspuesta de la fila i de E y $A_{Rri}{}^{(1)^t}$ es la traspuesta de la fila i de A reducida; esto es la traspuesta de la fila i de A original, a la que se han eliminado las entradas restringidas a cero en Z_A. $A_{b\,Rri}{}^t$ es la traspuesta de la fila i de A_b reducida, definida de la misma manera que para las reducidas de A. $D_{Rri}{}^{-1}$ es la versión reducida de $D_{ri}{}^{-1}$. $P_{Ri}{}^{(1)^t}$ corresponde a la matriz P traspuesta reducida, a la cual se han eliminado las columnas correspondientes a las entradas eliminadas en $A_{Rri}{}^{(1)^t}$.

De esta manera, los problemas anteriores tienen una solución $\boldsymbol{A_{Rri}{}^{(1)}}$ que puede ser obtenida por la siguiente expresión (utilizando la Ecuación 50 una vez más).

$$\boldsymbol{A_{Rri}{}^{(1)} = \left(P_{Ri}{}^{(1)} \cdot P_{Ri}{}^{(1)^t} + \lambda_A \cdot D_{Rri}{}^{-1} \right)^{-1} \left(P_{Ri}{}^{(1)} \cdot E_{ri}{}^t + \lambda_A \cdot D_{Rri}{}^{-1} \cdot A_{b\,Rri}{}^t \right)}$$

3. Se repite el paso 2. y 3. Hasta que el funcional evaluado en $A^{(n)}$ y $P^{(n)}$, $\left\| E - A^{(n)} \cdot P^{(n)} \right\|_F^2$ sea menor que un cierto grado de tolerancia o el error haya convergido.

Finalmente las matrices A y P son normalizadas por una matriz X definida como en la Ecuación 27.

3.12 NCA General

Es posible proponer un nuevo método NCA que incorpore todas las ampliaciones descritas en puntos anteriores. Así, NCA General pretende incorporar en una sola herramienta todas las funcionalidades antes descritas, a fin de combinar el efecto de todas ellas en la reconstrucción de las matrices. El funcional propuesto para tal propósito es:

$$J(E - A \cdot P) = \frac{1}{2} \cdot \|E - A \cdot P\|_F{}^2{}_{R^{-1}} + \frac{\lambda_P}{2} \cdot \|P - P_b\|_F{}^2{}_{B^{-1}} + \frac{\lambda_A}{2} \cdot \|A - A_b\|_F{}^2{}_{D^{-1}}$$

Ecuación 64

Como siempre, al minimizar el funcional anterior de acuerdo al procedimiento utilizando en los demás métodos NCA, se resolverán problemas equivalentes al de la ecuación siguiente.

$$\min_x \ f(x) = \frac{1}{2} \cdot \|b - A \cdot x\|_F{}^2{}_{R_b{}^{-1}} + \frac{\lambda}{2} \cdot \|x - x_b\|_F{}^2{}_{H^{-1}}$$

Ecuación 65

El mínimo de la expresión anterior puede ser obtenido utilizando el método del gradiente del funcional ya descrito. Esta puede ser descompuesta de la siguiente forma:

$$f(x) = \frac{1}{2} \cdot \left[(b - A \cdot x)^t \cdot R_b{}^{-1} \cdot (b - A \cdot x) + \lambda \cdot (x - x_b)^t \cdot H^{-1} \cdot (x - x_b)\right]$$

$$= \frac{1}{2} \cdot \left[b^t R_b{}^{-1} b - b^t R_b{}^{-1} A x - x^t A^t R_b{}^{-1} b + x^t A R_b{}^{-1} x + \lambda \cdot x^t H^{-1} x + \lambda \cdot x_b{}^t H^{-1} x_b - \lambda \cdot x_b{}^t H^{-1} x - \lambda \cdot x^t H^{-1} x_b\right]$$

Pero como $b^t R_b{}^{-1} A x$ es una constante, $\left(b^t R_b{}^{-1} A x\right)^t = x^t A^t R_b{}^{-1} b$. De la misma manera $x_b{}^t H^{-1} x = x^t H^{-1} x_b$. Así se tiene:

$$f(x) = \frac{1}{2} \cdot \left[b^t R_b{}^{-1} b - 2 b^t R_b{}^{-1} A x + x^t R_b{}^{-1} A x + \lambda \cdot x^t H^{-1} x + \lambda \cdot x_b{}^t H^{-1} x_b - 2\lambda \cdot x^t H^{-1} x_b\right]$$

Finalmente es posible obtener el gradiente de la función a minimizar siguiendo las reglas de derivación de vectores, formas lineales y formas cuadráticas detalladas en los anexos.

$$\frac{df(x)}{dx} = \nabla f(x) = \frac{1}{2} \cdot \left[0 - 2A^t R_b{}^{-1} b + 2A^t R_b{}^{-1} A x + 2\lambda H^{-1} x - 2\lambda H^{-1} x_b\right]$$

El mínimo estará en el punto donde el gradiente vale cero, esto es:

$$\nabla f(x^*) = -A^t R_b{}^{-1} b + A^t R_b{}^{-1} A x^* + \lambda H^{-1} x^* - \lambda H^{-1} x_b = 0$$

$$\left(A^t R_b{}^{-1} A + \lambda H^{-1}\right) x^* = A^t R_b{}^{-1} b + \lambda H^{-1} x_b$$

$$\boldsymbol{x^* = \left(A^t {R_b}^{-1} A + \lambda H^{-1}\right)^{-1}\left(A^t {R_b}^{-1} b + \lambda H^{-1} x_b\right)}$$

Ecuación 66

Se hace notar que aquí la expresión $A^t {R_b}^{-1} A + \lambda H^{-1}$ es una matriz cuadrada, por lo que al cumplirse las condiciones NCA, la inversa existirá sin problemas.

Calculando la segunda derivada del funcional, se ve que efectivamente se trata de un mínimo.

$$\frac{d^2 f(x)}{d^2 x} = \mathcal{H} f(x) = 2(A^t {R_b}^{-1} A + \lambda H^{-1})$$

Ecuación 67

Que bajo ciertas condiciones generales es semi-definida positiva.

El algoritmo de resolución es similar a los ya detallados, y consistente en un método de optimización alternado, en el cual minimizando alternadamente A y P se converge al mínimo del funcional. El gradiente del funcional entrega una expresión para las matrices que minimizarán la expresión en cada iteración, como se detalla en la Ecuación 66. EL problema a resolver es entonces:

$$\mathbb{P}: \quad \min_{A,P} \quad \frac{1}{2} \cdot \|E - A \cdot P\|_F{}^2{}_{R^{-1}} + \frac{\lambda_P}{2} \cdot \|P - P_b\|_F{}^2{}_{B^{-1}} + \frac{\lambda_A}{2} \cdot \|A - A_b\|_F{}^2{}_{D^{-1}}$$
$$s.t. \ A \in Z_A \ \text{ y } \ P \in Z_P$$

Ecuación 68

El algoritmo de resolución es el siguiente:

Se genera una adivinación inicial para los parámetros a estimar de A y P, de forma que se respeten las restricciones en la estructura de A y P: $A^{(0)} \in Z_A$ y $P^{(0)} \in Z_P$.

1. Dado $A^{(0)}$, encontrar $P^{(1)}$ que resuelva:

$$\min_{P^{(1)}} \ \frac{1}{2} \cdot \left\|E - A^{(0)} \cdot P^{(1)}\right\|_F{}^2{}_{R^{-1}} + \frac{\lambda_P}{2} \cdot \left\|P^{(1)} - P_b\right\|_F{}^2{}_{B^{-1}}$$
$$s.t. \ P^{(1)} \in Z_P$$

   Como se observa, es un problema similar al dado por la Ecuación 65. Se ha omitido el término de suposición de A debido a que al minimizar en P dicho término no es relevante. La resolución del problema se logra dividiendo el problema anterior en M problemas (uno para cada columna en E) de la forma ($k = 1, \dots, M$):

$$\min_{P_{Rck}{}^{(1)}} \frac{1}{2} \cdot \left\| E_{ck} - A_{Rk}{}^{(0)} \cdot P_{Rck}{}^{(1)} \right\|^2_{F \; R_{ck}{}^{-1}} + \frac{\lambda_P}{2} \cdot \left\| P_{Rck}{}^{(1)} - P_{b_{Rck}} \right\|^2_{F \; B_{Rck}{}^{-1}}$$

$$s.t. \; P_{Rck}{}^{(1)} \in Z_P$$

, donde E_{ck} es la columna k de E y P_{Rck} es la columna k de P reducida; esto es la columna k de P original, a la que se han eliminado las entradas restringidas a cero en Z_P. De la misma manera $P_{b_{Rck}}$ es la columna k de P_b reducida, definida de la misma manera que las reducidas de P. $B_{Rck}{}^{-1}$ corresponde a la versión reducida de $B_{ck}{}^{-1}$. A_{Rk} corresponde a la matriz A reducida, a la cual se han eliminado las columnas correspondientes a las filas eliminadas en P_{Rck}.

De esta manera, y dado las condiciones NCA, los problemas anteriores tienen una solución dada por la Ecuación 66.

$$\boldsymbol{P_{Rck}{}^{(1)} = \left(A_{Rk}{}^{(0)^t} R_{ck}{}^{-1} A_{Rk}{}^{(0)} + \lambda_P B_{Rck}{}^{-1} \right)^{-1} \left(A_{Rk}{}^{(0)^t} R_{ck}{}^{-1} E_{ck} + \lambda_P B_{Rck}{}^{-1} P_{b_{Rck}} \right)}$$

2. Dado $P^{(1)}$, encontrar $A^{(1)}$ que resuelva (no se considera el termino de regulación debido a que el argumento a optimizar no lo afecta):

$$\min_{A^{(1)}} \left\| E - A^{(1)} \cdot P^{(1)} \right\|^2_{F \; R^{-1}} + \frac{\lambda_A}{2} \cdot \left\| A^{(1)} - A_b \right\|^2_{F \; D^{-1}}$$
$$s.t. \; A^{(1)} \in Z_A$$

Igual que antes, el problema es similar al dado por la Ecuación 65. Esto es equivalente a resolver el problema siguiente, el que convenientemente está escrito en la misma forma que la Ecuación 17.

$$\min_{A^{(1)}} \left\| E^t - P^{(1)^t} \cdot A^{(1)^t} \right\|^2_{F \; R^{-1}} + \frac{\lambda_A}{2} \cdot \left\| A^{(1)^t} - A_b{}^t \right\|^2_{F \; D^{-1}}$$
$$s.t. \; A^{(1)} \in Z_A$$

La estrategia de resolución es similar, dividiendo el problema anterior en N problemas (uno para cada columna en E^t) de la forma ($i = 1, \dots, N$):

$$\min_{A_{Rri}{}^{(1)}} \left\| E_{ri}{}^t - P_{Ri}{}^{(1)^t} \cdot A_{Rri}{}^{(1)^t} \right\|^2_{F \; R_{ri}{}^{-1}} + \frac{\lambda_A}{2} \cdot \left\| A_{Rri}{}^{(1)^t} - A_{b_{Rri}}{}^t \right\|^2_{F \; D_{Rri}{}^{-1}}$$

$$s.t. \; A_{Rri}{}^{(1)} \in Z_A$$

, donde $E_{ri}{}^t$ es la traspuesta de la fila i de E y $A_{Rri}{}^{(1)^t}$ es la traspuesta de la fila i de A reducida; esto es la traspuesta de la fila i de A original, a la que se han eliminado las entradas restringidas

a cero en Z_A. ${A_b}_{Rri}{}^t$ es la traspuesta de la fila i de A_b reducida, definida de la misma manera que para las reducidas de A. ${D_{Rri}}^{-1}$ es la versión reducida de ${D_{ri}}^{-1}$. ${P_{Ri}}^{(1)^t}$ corresponde a la matriz P traspuesta reducida, a la cual se han eliminado las columnas correspondientes a las entradas eliminadas en ${A_{Rri}}^{(1)^t}$.

De esta manera, los problemas anteriores tienen una solución $\boldsymbol{{A_{Rri}}^{(1)}}$ que puede ser obtenida por la siguiente expresión (utilizando la Ecuación 66 una vez más).

$$\boldsymbol{{A_{Rri}}^{(1)} = \left({P_{Ri}}^{(1)} {R_{ri}}^{-1} {P_{Ri}}^{(1)^t} + \lambda_A {D_{Rri}}^{-1}\right)^{-1} \left({P_{Ri}}^{(1)} {R_{ri}}^{-1} {E_{ri}}^t + \lambda_A {D_{Rri}}^{-1} {A_{b_{Rri}}}^t\right)}$$

3. Se repite el paso 2. y 3. Hasta que el funcional evaluado en $A^{(n)}$ y $P^{(n)}$, ${\left\|E - A^{(n)} \cdot P^{(n)}\right\|_F}^2$ sea menor que un cierto grado de tolerancia o el error haya convergido.

Finalmente las matrices A y P son normalizadas por una matriz X definida como en la Ecuación 15.

3.13 Un nuevo enfoque: Interpolación optimal

Con los métodos desarrollados, surge inmediatamente la idea de entregar las reconstrucciones, correspondientes a las estimaciones de los parámetros con un grado de confiabilidad o exactitud, o lo que es lo mismo, con la varianza o desviación estándar correspondiente. El método desarrollado hasta ahora es un método variacional, en el cual se resuelve (varias veces en cada iteración) el problema genérico de la Ecuación 65. Esto es:

$$\min_x \ f(x) = \frac{1}{2} \cdot {\|b - A \cdot x\|_F}^2{}_{{R_b}^{-1}} + \frac{\lambda}{2} \cdot {\|x - x_b\|_F}^2{}_{H^{-1}}$$

Como se vio, una posibilidad equivalente de interpretar el problema anterior consiste en encontrar un estimador de x, $\hat{x}$ que ajuste de la mejor forma posible $A\hat{x}$ a los datos, y que a la vez permanezca cercano a la suposición construida para el estimador. Luego, el problema anterior es equivalente al de encontrar un estimador de x que se comporte bien y cumpla algunas características deseadas. Esto último se refiere a que, por ejemplo, sea insesgado y posea varianza mínima. Sin embargo no es directo obtener la varianza del estimador de este análisis, dado que no se ha hecho suposición alguna sobre la distribución de los errores.

Sin embargo, mediante un enfoque de interpolación optimal, es posible resumir la familia de estimadores lineales del problema anterior y obtener propiedades importantes que no son directas con el enfoque convencional.

Es posible demostrar que el mejor estimador insesgado lineal (BLUE) o el estimador de Gauss – Markov que permite ajustar $A \cdot \hat{x}$ a los datos y permanecer cerca de x_b es aquel que cumple:

1. Se puede escribir de la forma:

$$\hat{x} = x_b + K \cdot (b - Ax_b)$$

Ecuación 69

2. Es insesgado:

$$E(\hat{x}) = x_{true}$$

Ecuación 70

3. La varianza de $\hat{x}$ es mínima.

Luego, el objetivo es encontrar K que logre que lo anterior se cumpla.

Por otra parte, es posible analizar las expresiones obtenidas para los mínimos de los problemas como el de la Ecuación 65 (los estimadores o soluciones obtenidos por dicho método) y demostrar que efectivamente pueden ser reescritos como en la Ecuación 66. En efecto, teníamos que:

$$x^* = \left(A^t {R_b}^{-1} A + \lambda H^{-1}\right)^{-1} \left(A^t {R_b}^{-1} b + \lambda H^{-1} x_b\right)$$

Trabajando dicha expresión:

$$x^* = \left(A^t {R_b}^{-1} A + \lambda H^{-1}\right)^{-1} \left(A^t {R_b}^{-1} b + \lambda H^{-1} x_b + A^t {R_b}^{-1} A x_b - A^t {R_b}^{-1} A x_b\right)$$

$$= \left(A^t {R_b}^{-1} A + \lambda H^{-1}\right)^{-1} \left(\left(A^t {R_b}^{-1} A + \lambda H^{-1}\right) x_b + A^t {R_b}^{-1} b - A^t {R_b}^{-1} A x_b\right)$$

$$= \left(A^t {R_b}^{-1} A + \lambda H^{-1}\right)^{-1} \left(A^t {R_b}^{-1} A + \lambda H^{-1}\right) \cdot x_b + \left(A^t {R_b}^{-1} A + \lambda H^{-1}\right)^{-1} A^t {R_b}^{-1} (b - A x_b)$$

$$\boldsymbol{x^* = x_b + \left(A^t {R_b}^{-1} A + \lambda H^{-1}\right)^{-1} A^t {R_b}^{-1} (b - A x_b)}$$

Ecuación 71

De donde se identifica el término K buscado.

$$K = \left(A^t {R_b}^{-1} A + \lambda H^{-1}\right)^{-1} A^t {R_b}^{-1}$$

Ecuación 72

En la relación anterior, y de acuerdo al método de optimización optimal, la matriz de varianza - covarianza del estimador corresponderá a uno de los factores de la expresión anterior.

$$Var(x^*) = \left(A^t {R_b}^{-1} A + \lambda H^{-1}\right)^{-1}$$

Ecuación 73

3.13.1 Confiabilidad de las reconstrucciones

En base a lo anterior, es posible definir expresiones para la confiablidad de las matrices reconstruidas (de A y P) para los diferentes métodos creados. Las expresiones generales de la precisión para las matrices serían las siguientes, tomando como referencia el problema resuelto de la Ecuación 68:

$$Var\left(\widehat{P_{Rck}}\right) = \left({A_{Rk}}^t {R_{ck}}^{-1} A_{Rk} + \lambda {B_{Rck}}^{-1}\right)^{-1}$$

Ecuación 74

$$Var\left(\widehat{A_{Rri}}\right) = \left(P_{Ri} {R_{ri}}^{-1} {P_{Ri}}^t + \lambda {D_{Rri}}^{-1}\right)^{-1}$$

Ecuación 75

En donde las matrices dependientes de A y P a la derecha de las expresiones, corresponden a las obtenidas en la última iteración del método respectivo. De esta forma, se definen las matrices de varianza de las reconstrucciones como:

$$Var\left(\hat{P}\right) = Q_P$$

Ecuación 76

$$Var\left(\hat{A}\right) = Q_A$$

Ecuación 77

, donde Q_P posee en cada columna la varianza de la columna respectiva de P definida por la Ecuación 74, y Q_A definida de manera análoga. De esa manera, cada coordenada de dichas matrices tendrá la varianza de la reconstrucción estimador de la coordenada respectiva de las matrices.

Por supuesto, dado que las ecuaciones anteriores son validadas para el caso general (GgNCAreg), es posible adaptar las mismas expresiones para los demás métodos NCA. Así, por ejemplo, en el caso de querer obtener las expresiones respectivas para NCAbasic, se utilizan las mismas expresiones anteriores, pero considerando los parámetros λ_A y λ_P, y las matrices R, B y D como las correspondientes neutras, esto es, las que al ser utilizadas en el método correspondiente no producen cambios, reduciendo sus resultados a los métodos básicos. Esto es, considerar R, B y D como las matrices unitarias, y λ_A y λ_P iguales a cero. Así, para NCAbasic las correspondientes versiones para las varianzas de cada fila y cada columna de las reconstrucciones de A y P respectivamente son:

$$Var\left(\widehat{P_{Rck}}\right) = \left({A_{Rk}}^t A_{Rk}\right)^{-1}$$

Ecuación 78

$$Var\left(\widehat{A_{Rri}}\right) = \left(P_{Ri}{P_{Ri}}^{t}\right)^{-1}$$

Ecuación 79

Con el fin de incluir la teoría anterior en los método ya creados, se desarrollaron funciones que tomando los datos de la reconstrucción, entregan para cada método la correspondiente expresión de varianza de las reconstrucciones.

3.14 Resumen de métodos NCA

Con el fin de aclarar y organizar la información, en la siguiente tabla se resumen los métodos NCA creados y reproducidos. Se entrega en cada caso el nombre, nombre resumido (usado para mayor facilidad de escritura en adelante), y su descripción.

Nombre	Nombre Resumido	Descripción
NCA básico	nca_n	NCA básico sin modificaciones.
gNCA básico	gnca_n	gNCA básico sin modificaciones.
gNCA regularización	gnca_reg_n	gNCA con regularización en P. Se utiliza un parámetro de regularización global.
NCA básico	NCAbasic	NCA básico modificado. Converge a mínimo global.
gNCA básico	gNCAbasic	gNCA básico modificado. Converge a mínimo global.
gNCA regularización	gNCAbasic_reg	gNCA con regularización en P. Converge a mínimo global.
gNCA con confiabilidad	cgNCAreg	gNCA modificado para incluir confiabilidad de los datos. Converge a mínimo global.
gNCA con suposiciones	agNCAreg	gNCA modificado para incluir suposiciones en los parámetros a reconstruir. Converge a mínimo global.
gNCA con confiabilidad en suposiciones	acgNCAreg	gNCA modificado para incluir suposiciones en los parámetros y confiabilidad de las mismas.
gNCA General	GgNCAreg	gNCA modificado. Permite trabajar con confiabilidad de los datos, suposiciones y su confiabilidad. Converge a mínimo global.

Tabla 1: Resumen métodos NCA.
Fuente: Elaboración propia.

3.15 Pruebas sintéticas

Como se describe en las metodologías, una vez establecida la teoría, creados y programados los métodos NCA originales y la nueva funcionalidad, se procede a realizar una serie de pruebas que permiten establecer el alcance, capacidad y limitaciones de los métodos de reconstrucción, así como en las diferentes situaciones que cada método se comportaba mejor que otro. De la misma forma estas pruebas sirven de motivación para el desarrollo de nuevas herramientas, algunas desarrolladas y otras discutidas y planteadas como objetivos de nuevos trabajos.

Las pruebas se dividen según su complejidad en 3 grupos, y en cada uno de estos se realizan pruebas orientadas a diferentes funcionalidades, divididas en niveles. En el primer grupo de pruebas, orientado a exponer bajo diferentes condiciones la reconstrucción de las redes de regulación, se testea en extenso cada uno de los métodos NCA reproducidos y creados con el fin de analizar su comportamiento y capacidad. Se comentan también los diferentes programas que fueron creados con el fin de testear los métodos. En el segundo se realizaron pruebas generales respecto a varias situaciones llamadas "extremas" y de interés particular, mientras que en el tercero, se resume un conjunto de pruebas realizadas para motivar el inicio de un nuevo enfoque de trabajo.

3.16 Pruebas sintéticas: Grupo 1

Este grupo de pruebas sintéticas representa el fuerte de los test que se realizan, y se divide en 5 niveles, cada uno con diferentes pruebas orientadas a probar diferentes funcionalidades de los métodos NCA. El procedimiento utilizado es el descrito en la metodología de pruebas con redes sintéticas, discutido en puntos anteriores. Estos es, crear redes sintéticas de diferentes tamaños, de las cuales se conoce con certeza todos los parámetros de la red: conexiones y $CS's$, las señales de regulación de los $TF's$ y los datos de microarrays que dicha configuración general (matriz E). Luego se prueban de diferentes maneras los métodos, ya sea perturbando con errores la matriz de datos, incluyendo suposiciones, confiabilidad, etc., y comparando la exactitud y comportamiento de las reconstrucciones.

3.16.1 Nivel 1

Este primer nivel tiene como objetivo probar con redes sintéticas de carácter simple (pequeñas a medianas en tamaño) el funcionamiento de los 3 métodos más básicos. Se pretende comparar la velocidad de reconstrucción y el efecto de la regularización en P.

Experiencia 1: Red Simple

En esta primera experiencia se utiliza una red pequeña de 9 genes, 4 TF y 5 experimentos, con una densidad de un 65% de ceros en la estructura de A[50]. Con el fin de resumir lo anterior, se expresará de ahora en adelante como una red de 9x4x5x65, y se asumirá que se creó cumpliendo el criterio NCA. El objetivo es utilizar nca_n empleando la matriz de datos E sin errores, con el fin observar el comportamiento de los métodos y analizar las diferentes alternativas de fin de iteraciones discutidas con anterioridad. Como una manera de orientar el procedimiento, se presentan las matrices utilizadas, donde como siempre A_{on} y P_{on} corresponde a las matrices reales A y P que se desean reconstruir (que definen el sistema) y la matriz E_r que corresponde a los datos reales de microarrays que se obtendrían de dicho sistema en un experimento teóricamente perfecto.

[50] Con densidad de ceros se refiere al porcentaje total de ceros en la estructura respecto al total de entradas de la matriz.

```
Pon =
    0.0126    0.5464   -0.0470   -0.0934    0.0099
   -0.1406    0.1303    0.0241    0.2078    0.1255
    0.1945    0.0739   -0.1997    0.2360    0.3071
    0.2265   -0.1983   -0.4457    0.3814    0.1688

Aon =
   -1.2324    2.1276    1.7400         0
    3.3315    0.9603    1.6548         0
         0         0         0   -1.0794
    0.8552         0         0    0.2566
         0    1.5409         0    1.7971
    2.5141   -0.6288   -0.4798    2.4801
         0         0    1.0850         0
         0         0         0    1.5457
   -0.4685         0         0         0

Er =
    0.0238   -0.2675   -0.2382    0.9680    0.7890
    0.2287    2.0678   -0.4638    0.2789    0.6617
   -0.2445    0.2141    0.4811   -0.4117   -0.1822
    0.0689    0.4164   -0.1545    0.0179    0.0518
    0.1904   -0.1556   -0.7637    1.0056    0.4967
    0.5886    0.7645   -1.1428    0.4670    0.2173
    0.2111    0.0802   -0.2167    0.2561    0.3332
    0.3502   -0.3065   -0.6889    0.5895    0.2609
   -0.0059   -0.2560    0.0220    0.0438   -0.0047
```

Se prueba en primer lugar la reconstrucción utilizando la matriz E_r sin errores, obteniendo los ajustes que se observa en los siguientes gráficos:

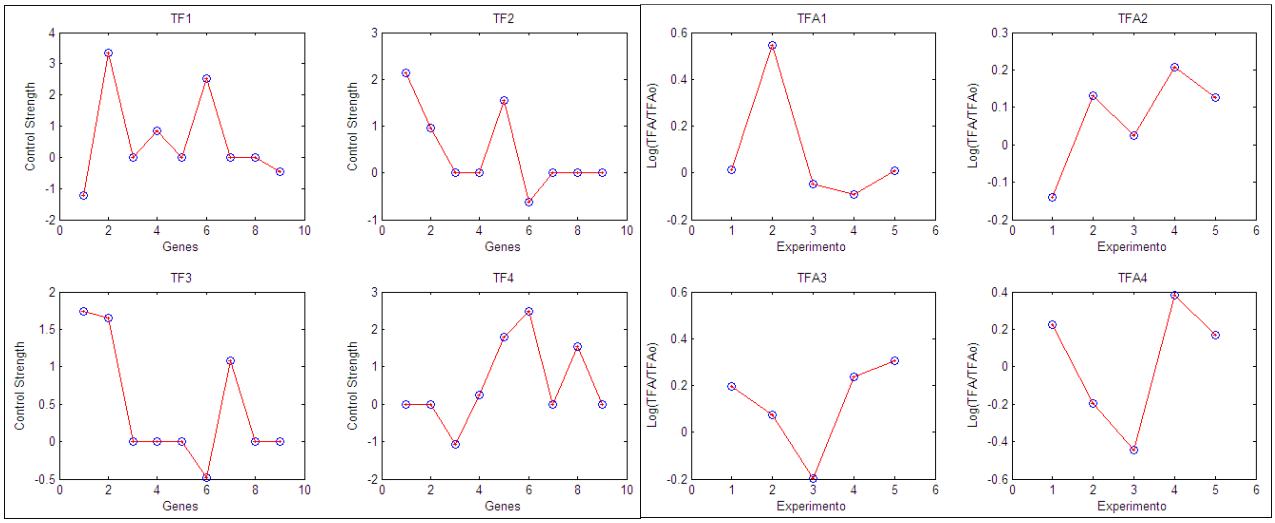


Figura 24: Ajuste reconstrucción. Pruebas G1N1E1.
Fuente: Elaboración propia.

La forma gráfica de visualizar el ajuste mostrado en la Figura 24 es utilizado en buena parte de las pruebas (aunque quizá no mostrando todos los TF en redes de gran tamaño). En los gráficos de dicha figura se observa el resumen del ajuste para la matriz A en la izquierda y para la matriz P en la derecha. Para cada caso existen 4 gráficos (uno para cada TF del sistema). En el caso del ajuste de A la información de cada gráfico resume la información de una de sus columnas (cada columna corresponde a un TF). Para cada TF se muestra gráficamente el valor numérico CS en cada experimento. En el caso del ajuste de A la información de cada gráfico corresponde a una de las filas de P. En ellos se resume la actividad relativa del TF para cada experimento (notar la rotulación de los ejes). Los puntos azules

corresponden a los valores reales de los parámetros de las matrices (los resumidos en A_{on} y P_{on}), mientras que las líneas rojas unen los puntos reconstruidos. Mientras más en el centro de los puntos este el pico de las curvas formadas por las líneas rojas (que corresponde a las estimaciones de los parámetros vía NCA), mejor será el ajuste. En la figura se ve un ajuste prácticamente perfecto de los parámetros estimados del sistema respecto a los reales. Este tipo de análisis gráfico será utilizado a lo largo de todas las pruebas.

Otra opción para visualizar el ajuste de las reconstrucciones, y que es el más extensamente utilizado, consiste en utilizar los errores numéricos definidos en el punto 2.7.2 Medición de errores en pruebas sintéticas[51].

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
nca_n	1,99E-09	0,00	2,60E-11	0,01	9,84E-06	0,00	9,84E-06	0,00

Tabla 2: Resumen errores utilizando nca_n. Pruebas G1N1E1.
Fuente: Elaboración propia.

Como se puede apreciar, el error porcentual de reconstrucción es nulo en este caso. Un punto a tener en consideración es la interpretación del MSE[52] definidos para el ajuste de las matrices A y P. Es interesante comparar dicho valor entre métodos (distintos NCA para la misma redes), pero no entre experiencias distintas, ya que el tamaño de la red puede distorsionar dicho valor independiente del ajuste existente. De esta manera, con el fin de comparar distintas redes es mejor utilizar el error porcentual promedio de cada ajuste.

Es posible repetir el mismo proceso de iteración, y se observa que en ocasiones se obtienen resultados como el detallado en la siguiente tabla.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
nca_n	1,32E+00	81,81	1,80E+02	9927,83	1,47E-01	15,13	1,47E-01	15,13

Tabla 3: Mínimo local Red 1 utilizando nca_n. Pruebas G1N1E1.
Fuente: Elaboración propia.

En este caso, los errores son considerables en las matrices reconstruidas, especialmente el observado en la matriz P. El error de ajuste cae también casi 5 órdenes de magnitud, lo que como se ha comentado se asocia a la convergencia del método a un mínimo local. Esto se aprecia también al estudiar la velocidad de convergencia de los métodos y el criterio de detención utilizado. En el primer caso, la convergencia se alcanza en poco más de 500 iteraciones, y el criterio de detención es la convergencia al error umbral

[51] Esta es el tipo de tabla utilizada en todo el resto del documento para resumir las reconstrucciones. En la 1° columna se muestra el método utilizado. En la 4 siguientes el MSE y el error porcentual medio de ajuste de cada matriz reconstruida a su contraparte real, según la Ecuación 18 y Ecuación 20. Finalmente se muestra el error de ajuste (la norma de la diferencia de las matrices) y el error porcentual medio entre el modelo reconstruido, y los datos medidos y los reales.
[52] Mean square error.

utilizado; en este caso, $1 \cdot 10^{-5}$. Por supuesto, al no tener errores en la matriz de datos, el error de convergencia podría haber disminuido arbitrariamente hasta cualquier cota, pero el umbral impuesto es una buena combinación de exactitud y de tiempo de espera. (Las 500 iteraciones se alcanzan en cerca de 2 segundos). Por supuesto el número de iteraciones dependerá del error umbral mínimo escogido y del tamaño de la red, lo mismo el tiempo de procesamiento de las iteraciones.

En el segundo caso (el del mínimo local resumido en la tabla anterior), el número de iteraciones para alcanzar una convergencia es de casi 5.500, y el tiempo de espera aumenta proporcionalmente. En este caso no se alcanza nunca el error mínimo umbral, ya que se utiliza el segundo criterio (la convergencia del error) para el término del método. Empíricamente se observa que las iteraciones quedaron atrapadas en un mínimo local, ya que no existe forma de hacer disminuir más el error de ajuste. Como una forma de apreciar el gran error de ajuste en la matriz P, se detalla su gráfico de ajuste. Los errores se aprecian a simple vista.

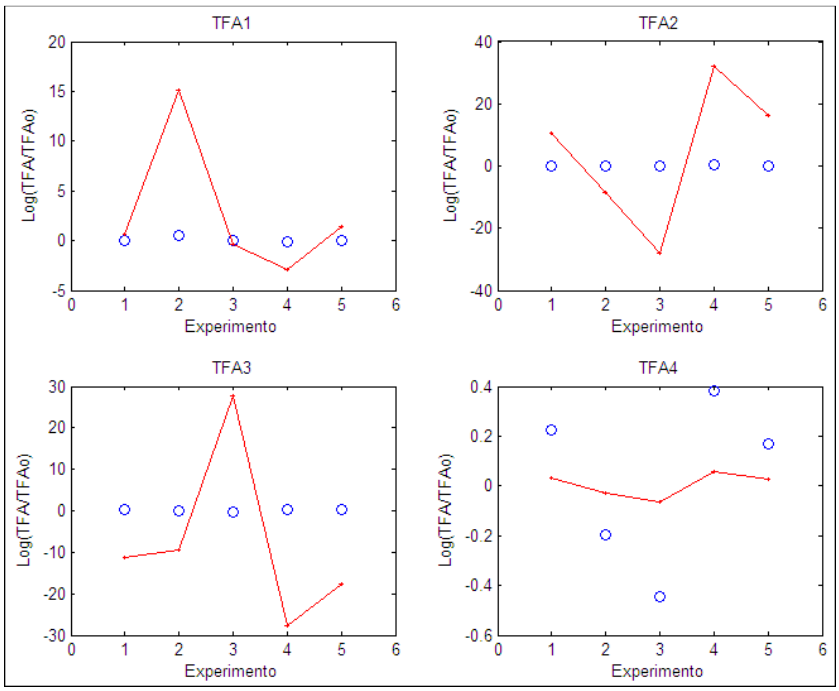


Figura 25: Ajuste erróneo en P producto de un mínimo local. Pruebas G1N1E1.
Fuente: Elaboración propia.

Con el fin de analizar la convergencia a mínimos locales y su real implicancia, se crea un programa denominado *disterror_NCA.m* que permite analizar la distribución del error utilizando distintos métodos NCA en base a 100 reconstrucciones de la red.

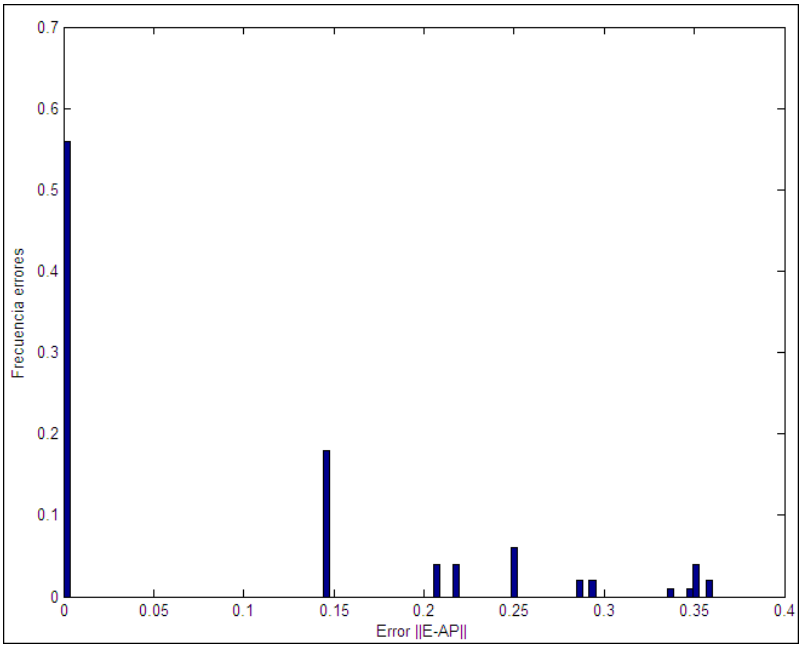


Figura 26: Frecuencia errores reconstrucción NCA. Red 1. Pruebas G1N1E1.
Fuente: Elaboración propia.

Como se observa, poco más del 55% de las adivinaciones iniciales conducen al óptimo global en la red estudiada. El resto corresponde a distintos óptimos locales en los cuales queda atrapada la convergencia. La segunda barra con mayor frecuencia corresponde de hecho al mínimo alcanzado con anterioridad. Es necesario destacar que la frecuencia de los errores dependerá de la red estudiada. Se muestra un ejemplo similar en la Figura 17, donde la frecuencia del error asignado al mínimo global era de un 85%.

En la figura siguiente se presenta un nuevo gráfico para otra red estudiada, en donde se demuestra que efectivamente la distribución del error parece depender de la red en particular que se está analizando.

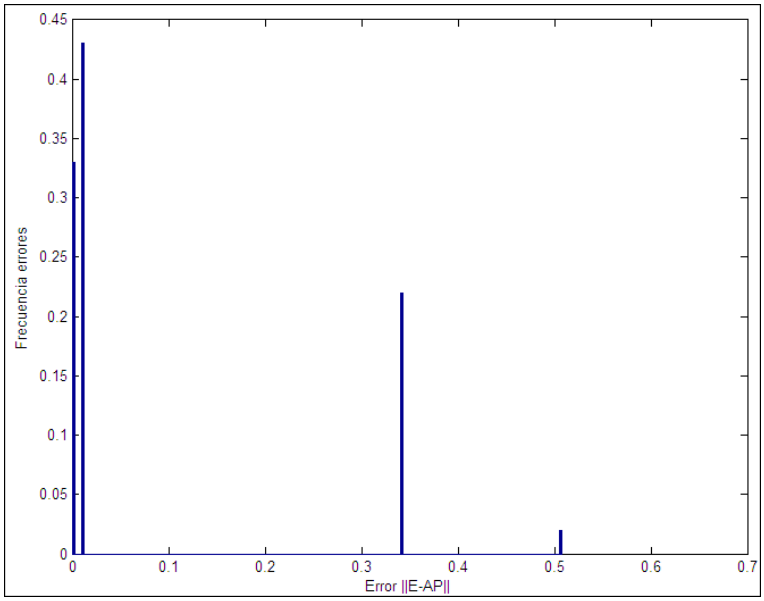


Figura 27: Frecuencia errores reconstrucción NCA. Red 2. Pruebas G1N1E1.
Fuente: Elaboración propia.

Otros puntos relevantes a comentar son los siguientes:

- En 2 de los casos estudiados, el error de ajuste perfecto correspondiente al mínimo global presentó la máxima frecuencia. Si bien esto asegura que con mayor probabilidad se obtendrá el mínimo global en la reconstrucción, no permite una construcción sin incertidumbre en redes desconocidas.
- En las reconstrucciones erróneas (las correspondientes a mínimos locales, por lo general las entradas de la matriz P resultan ser exageradamente grandes en relación a su valor real (basta ver el error de ajuste a dicha matriz). Luego, las técnicas de regularización junto con aumentar la estabilidad numérica del proceso, podrían ayudar a evitar la convergencia a este tipo de mínimos.

Experiencia 2: Red de mayor tamaño.

En este caso se trabajó con una red de mayor tamaño, de 25x6x8x70. Se realizaron pruebas similares a las anteriores utilizando nca_n, observando una reconstrucción rápida y de gran exactitud utilizando la matriz E sin errores. En la figura y tabla siguiente se muestra el resultado de dicha reconstrucción. Como se observa, son resultados análogos a los obtenidos con una red pequeña.

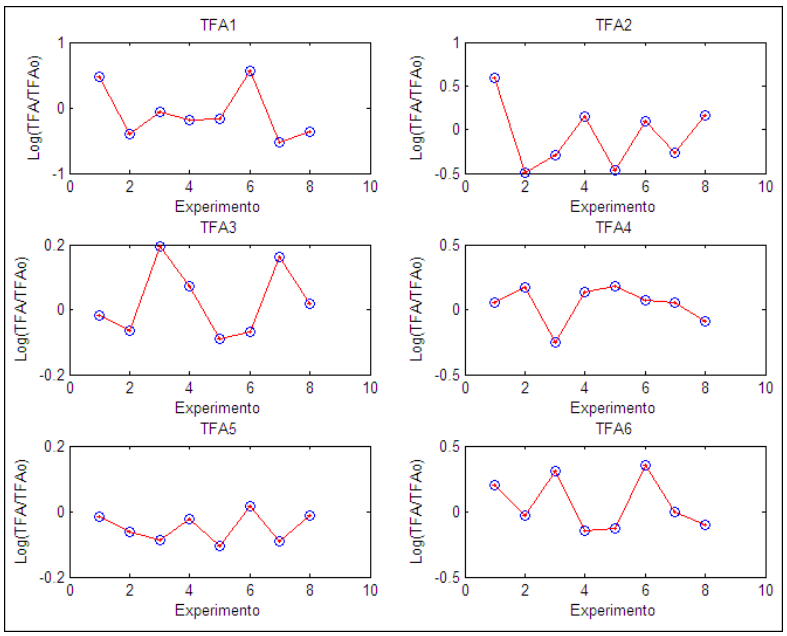


Figura 28: Ajuste gráfico reconstrucción. Pruebas G1N1E2.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
nca_n	2,09E-10	0,00	2,72E-12	0,00	9,82E-06	0,00	9,82E-06	0,00

Tabla 4: Resumen errores utilizando nca_n. Pruebas G1N1E2.
Fuente: Elaboración propia.

Se realizó además un test con el fin de estudiar la distribución del error, y se obtuvieron resultados similares a los de las redes más pequeñas. Sin embargo, en este caso la distribución del error parece ser algo más insensible frente a las diferentes adivinaciones iniciales.

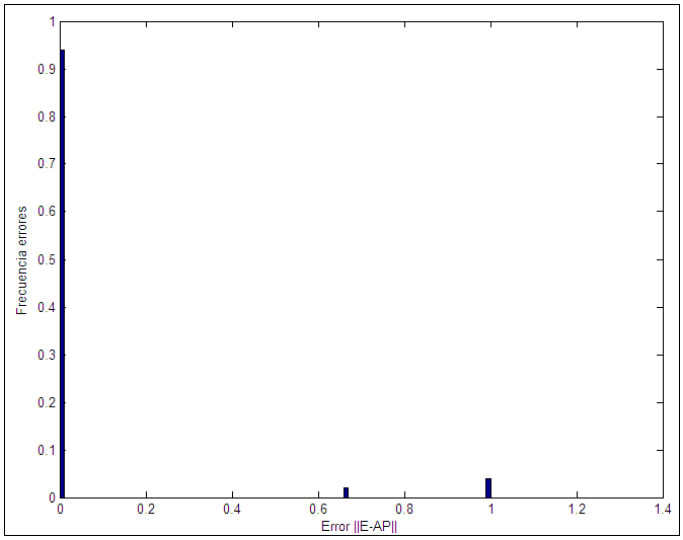


Figura 29: Distribucón de los errores de ajuste frente a diferentes adivinaciones. Pruebas G1N1E2.
Fuente: Elaboración propia.

Lo anterior se repite para 2 redes adicionales, de tamaños similares, y los resultados fueron los mismos. En ambos casos más del 90% de la distribución de los errores correspondían al mínimo global. Si bien no se puede aseverar con certeza, el resultado pareciese indicar que el punto de convergencia se hace más insensible a adivinaciones iniciales a medida que la red crece. Otro punto relevante a destacar es que para un mismo umbral de error mínimo, con redes más grandes es requerido un menor número de iteraciones para alcanzarlo. Lo contrario ocurre en la convergencia a mínimos locales (donde el criterio de termino de iteraciones es la convergencia del error) el número de iteraciones necesarias aumenta considerablemente.

Experiencia 3: Red grande

En esta experiencia se testea una red de tamaño grande, de 94x18x20x70. En este caso las matrices que definen el sistema son exageradamente grandes, y no existe una manera fácil de visualizarla y observar las conexiones entre los $TF's$ y los genes.

Se utiliza una vez más nca_n para reconstruir la red, y se puede comprobar una reconstrucción perfecta una vez alcanzado el mínimo global del método. Es interesante destacar una vez más que el número de iteraciones necesarias para alcanzar el ajuste perfecto disminuye con el aumento del tamaño de la red (se necesitan tan solo 130 iteraciones, en contraste con las cerca de 500 necesarias para una red pequeña). Dicho efecto puede explicarse por la mayor cantidad de información que se obtiene de una matriz de datos E proporcionalmente más grande, por lo que el salto (entre las sucesiones de matrices que convergen a la de mejor ajuste) en dirección del funcional de una iteración a otra es más largo, como se ve esquemáticamente en la figura siguiente.

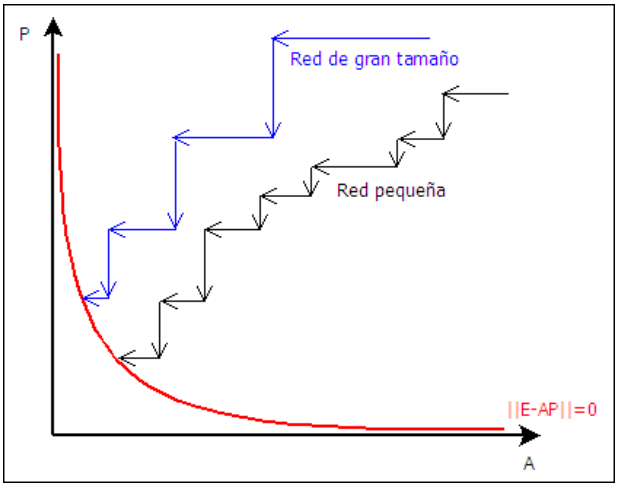


Figura 30: Representación esquematica del numero de iteraciones. Pruebas G1N1E3.
Fuente: Elaboración propia.

En la tabla siguiente se resumen los resultados de la reconstrucción. Se destaca que independiente del tamaño de la red, dichas mediciones del error de ajuste siguen siendo comparables de un experimento a otro, ya que sus magnitudes son independientes del tamaño del experimento considerado[53].

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
nca_n	1,63E-12	0,00	4,14E-14	0,00	9,53E-06	0,00	9,53E-06	0,00

Tabla 5: Resumen de errores utilizando nca_n. Pruebas G1N1E3.
Fuente: Elaboración propia.

Analizando el gráfico de distribución de errores, se puede apreciar un efecto similar al de las redes de tamaño medio. El error es bastante insensible a las diferentes adivinaciones iniciales, por lo que se privilegia en la mayoría de los casos el mínimo global. Se observa también que los errores de ajuste correspondiente a mínimos globales son grandes en magnitud en relación a las experiencias anteriores, lo que se explica por el mayor número de entradas que posee una red de mayor tamaño.

[53] En el caso del error porcentual medio por supuesto.

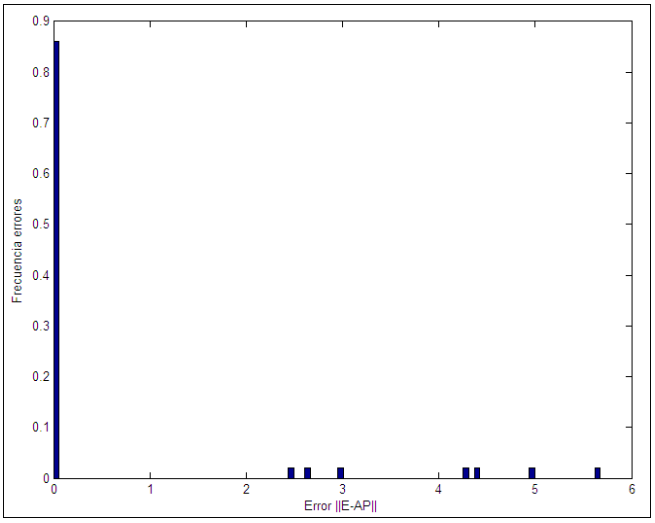


Tabla 6: Distribución del error. Pruebas G1N1E3.
Fuente: Elaboración propia.

Experiencia 4: Pruebas con gNCA (gnca_n)

Con el fin de comprobar el funcionamiento de gnca_n (la versión básica de gNCA) se utilizan 3 redes de tamaños similares a las utilizadas en las experiencias anteriores, esto es, una red pequeña (red 1) (9x4x5x75), una mediana (red 2) (25x6x7x75) y una red grande (red 3) (94x18x20x85). La diferencia principal es que las redes propuestas no son compatibles con el criterio NCA básico (el que impone restricciones sobre las matrices reducidas de A). Las siguientes matrices corresponden a las que definen la red 1, y como se ve en la matriz A_{on}, esta no cumple el criterio NCA ya que el $TF4$ controla un subconjunto de los genes que contrala el $TF3$.

```
Aon =
    0    1.4106    3.1178         0
-0.6269         0         0         0
    0    1.4670         0         0
    0         0    0.3463    0.0560
 2.5128    0.0518         0         0
    0    1.0706    0.8556    0.5848
    0         0    0.6624    2.3592
 1.1141         0         0         0
    0         0    0.0181         0

Pon =
 0.0127    0.0410    0.0832   -0.1507    0.0699
-0.0236    0.1176   -0.0367   -0.3214   -0.4588
 0.3210   -0.0944    0.0355    0.1371   -0.0551
 0.1948   -0.2569    0.3065   -0.2136         0
```

Pero como se observa también en la matriz P existe un experimento knock-out en el experimento 5 del

$TF4$ (el mismo con problemas), por lo que las matrices si cumplirán el criterio NCA general[54]. Luego, es posible aplicar gnca_n a las redes anteriores[55].

Es posible observar que los resultados concuerdan con bastante exactitud a lo obtenido en las experiencias anteriores con nca_n, si bien numéricamente difirieron levemente. El error medio porcentual de ajuste de A y P era pequeño al 4° o 5° decimal, mientras que con gnca_n esto disminuye al 2°. Por otra parte la diferencia es escasa, y los resultados son analíticamente equivalentes, por lo que ambos métodos funcionalmente se comportan de manera similar. La única diferencia se observa en el tiempo de procesamiento y en el número de iteraciones, que aumentan bastante. Si bien en la red pequeña y mediana el tiempo de espera a la convergencia fue similar al obtenido con nca_n, en la red de mayor tamaño, y dada sus dimensiones, la diferencia se hizo apreciable. En esta las iteraciones eran bastante lentas, debido principalmente a las operaciones y cálculos extras que es necesario realizar con gnca_n[56].

De todas formas, y pese a este último inconveniente, no se destaca la utilidad de gNCA básico como una herramienta de análisis que permita trabajar con redes, que incluso no cumplan el criterio NCA básico. Los resultados obtenidos son bastante similares y de la misma validez que los obtenidos con NCA básico, pero la posibilidad de trabajar con redes más genéricas tiene una importancia elevada. Como una forma de visualizar esto, se forzará a NCA básico (nca_n) a reconstruir la red 1 de este experimento, aún cuando no cumple con el criterio NCA correspondiente. Los resultados se resumen en la tabla siguiente utilizando NCA y gNCA.ñ

Como se observa, los errores cometidos al utilizar nca_n con redes no compatibles son considerables, de ahí la importancia de comprobar y respetar dichos criterios.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
nca_n	3,25E-01	25,40	2,44E-02	48,29	9,86E-06	0,00	9,86E-06	0,00
gnca_n	4,10E-08	0,01	1,14E-09	0,02	1,89E-05	0,01	1,89E-05	0,01

Tabla 7: Resumen reconstrucción red 1. Pruebas G1N1E4.
Fuente: Elaboración propia.

Por otra parte, es posible apreciar que pese al gran error cometido en la estimación de las matrices A y P, el ajuste del conjunto $(A \cdot P)$ a los datos es bastante preciso. Se obtiene un residuo de orden de magnitud 10^{-6}, y por ende el error porcentual de ajuste es también sumamente bajo. Se omiten los resultados obtenidos con las redes más grandes, ya que pese al tiempo de procesamiento, al considerar la matriz de datos sin error los resultados son cualitativamente similares.

En lo siguiente, y dada la similitud de los datos, cuando existan experiencias en que no se utilicen experimentos knock-out, se utilizará la versión básica de NCA dada su mayor rapidez.

[54] El que impone condiciones sobre las matrices reducidas de G.
[55] Las redes 2 y 3 tendrán características similares.
[56] En particular y como se mencionó, el cálculo extra que significa computa la matriz G y sus reducidas.

Experiencia 5: Regularización en P

La siguiente experiencia tiene como objetivo comprobar el funcionamiento de gNCA con regularización (gnca_reg_n) cuando la matriz de datos se considera sin errores. Si bien el objetivo del método de regularización es mejorar la estabilidad de la reconstrucción frente errores en los datos, es de interés analizar también su comportamiento, y compararlo respecto al observado con los métodos NCA ya probados.

El primer punto a destacar es que el error de convergencia alcanzado con gnca_reg_n en ningún caso (incluso sin considerar errores en E) alcanza el umbral mínimo de error. El método termina de iterar debido a la convergencia del error, que de acuerdo al tamaño de la red y al parámetro de regularización se alcanza en diferentes puntos[57]. Esto resume en parte el efecto del término de regularización en la reconstrucción. Existe un compromiso entre minimizar el residuo de ajuste, y mantener a la matriz P cerca de cero, por lo que independientemente de si la matriz de datos tiene errores, el ajuste no será tan exacto como con nca_n y gnca_n. En la siguiente tabla se compara la reconstrucción utilizando los 3 métodos para una red de tamaño pequeño, 9x4x5x75.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
nca_n	1,99E-09	0,00	2,60E-11	0,01	9,84E-06	0,00	9,84E-06	0,00
gnca_n	1,98E-09	0,00	2,62E-11	0,01	9,84E-06	0,00	9,84E-06	0,00
gnca_reg_n	3,96E-06	0,145025	4,55E-07	0,93864	0,0008194	0,00	0,0008194	0,00

Tabla 8: Resumen de errores, red pequeña. Pruebas G1N1E5.
Fuente: Elaboración propia.

Como se puede observar, la reconstrucción utilizando los primeros 2 métodos es indiferenciable. Respecto a gnca_reg_n, se puede apreciar como efectivamente el error de ajuste a los datos es menor, y el error de reconstrucción si bien es bajo, es mayor al de los otros métodos (cercano al 1% en P). Los datos no poseen errores, por lo que obviamente la distorsión es atribuida el afecto de la regularización.

A medida que el parámetro de regularización aumente, el efecto distorsionador en el ajuste a los datos aumentará, por lo que tal como ya se ha comentado, es necesario encontrar un equilibrio entre el término de regularización y el término de residuo (el ajuste a los datos). Al menos en las experiencias realizadas en este documento, se preferirá que domine el correspondiente al residuo, dado que propagará menos errores a las matrices A y P.

En la figura siguiente se presenta el gráfico de distribución de los errores utilizando gnca_reg_n y nca_n para la misma red.

[57] En el 4° decimal por lo general, aunque esto está fuertemente condicionado por el tamaño de la red.

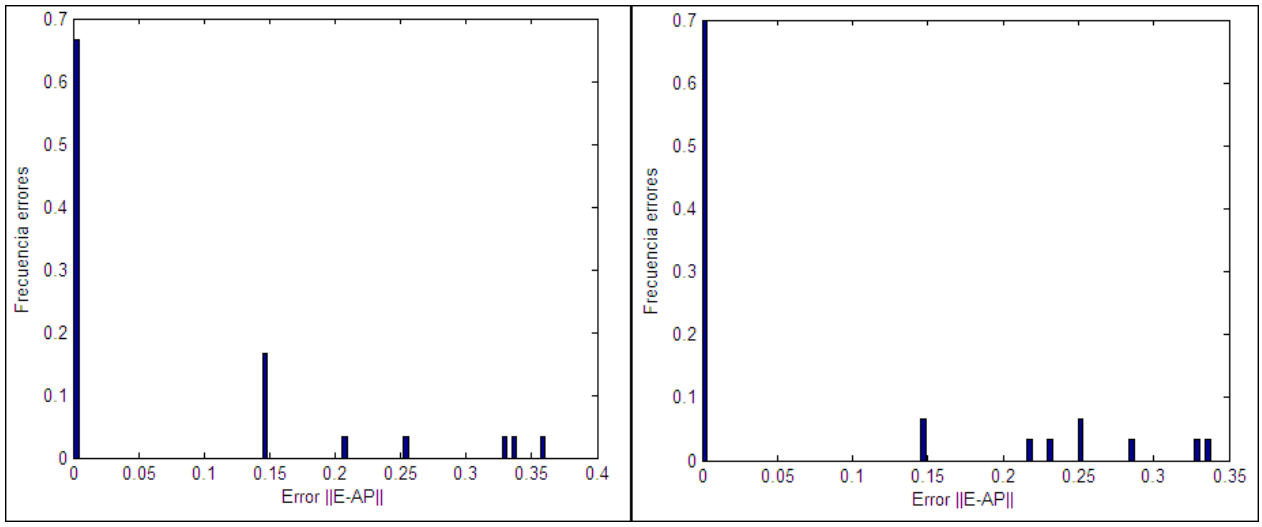


Figura 31: Distribución del error. Pruebas G1N1E5.
A la izquierda se resume el resultado utilizando nca_n, a la derecha con g_nca_reg.
Fuente: Elaboración propia.

Como se observa, la distribución del residuo parece ser independiente del método utilizando, por lo que al menos en el caso estudiado (sin errores en los datos), la regularización no parece eliminar el efecto de mínimos locales. Se termina concluyendo la relativa exactitud del método, que pese a entregar reconstrucciones menos exactas que sus contrapartes sin regularización, los resultados son aun de gran calidad. Se espera ver la verdadera utilidad y el efecto de este método en los experimentos con errores, en los cuales la estabilidad numérica se torna relevante.

3.16.2 Nivel 2

En este nivel se pretende realizar experiencias con errores en las matrices de datos. Con el fin de reproducir los errores experimentales en las redes sintéticas, se adiciona a cada entrada errores blancos no correlacionados. Específicamente, a cada coordenada de la matriz E se le suma un valor aleatorio de distribución normal, con media cero y desviación estándar igual a cierto porcentaje de un parámetro p. Dicho parámetro corresponderá al promedio de las entradas de E. De esta forma, los errores adicionados vendrán de una misma distribución, mimetizando el posible error de medición. En la figura siguiente se muestra esquemáticamente la distribución de dicho error. Ambas poseen el mismo parámetro p, pero los porcentajes de este (que serán igual a la varianza de la distribución), variarán.

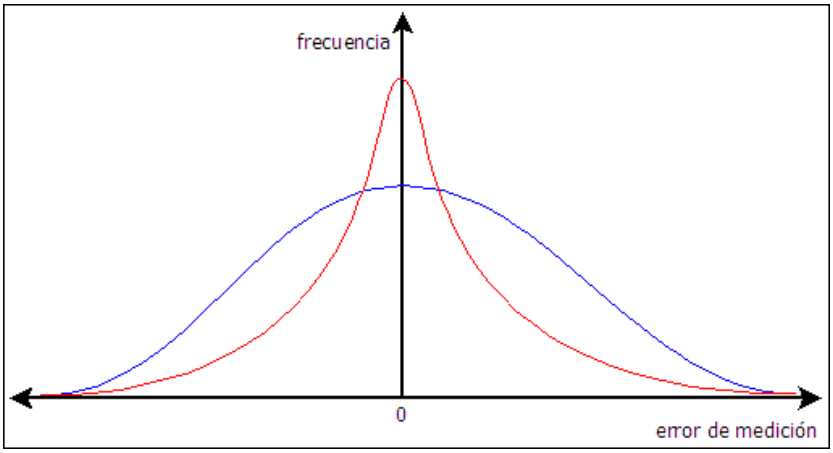


Figura 32: Distribución del error de los datos. Pruebas G1N2.
Fuente: Elaboración propia.

La distribución del error de medición de la curva roja tiene una menor varianza, por lo que en promedio, adiciona menos error a los datos en comparación con la correspondiente a la curva azul.

Como un ejemplo, considérese la siguiente matriz de datos de una red sintética (en la que se mide la expresión de 14 genes en 6 experimentos). Dichos datos son los reales, obtenidos hipotéticamente mediante un experimento 100 por ciento confiable y preciso. Por supuesto, a la hora de medir la expresión de los genes, los datos estarán contaminados por ruido blanco proveniente de errores en la medición. Esto es, se obtendrá una matriz E_m que diferirá de su contraparte real.

Er =

-0,0386	0,5264	-0,2297	-0,1329	-0,0159	0,1002
0,0082	0,0428	0,0374	0,0129	-0,0147	0,0474
-0,0370	0,0657	-0,0059	0,0082	-0,0273	-0,0663
-0,0025	0,0344	-0,0150	-0,0087	-0,0010	0,0065
-0,0184	0,1388	-0,0515	-0,0212	-0,0110	0,0311
0,0047	0,0246	0,0215	0,0074	-0,0085	0,0273
-0,0445	-0,2310	-0,2021	-0,0698	0,0796	-0,2559
-0,0302	0,4111	-0,1794	-0,1038	-0,0124	0,0783
-0,0032	0,0453	-0,0155	0,0079	0,0139	-0,0068
-0,0002	0,0030	-0,0013	-0,0008	-0,0001	0,0006
-0,0489	-0,0170	0,0625	0,0888	-0,0413	0,0256
0,0247	-0,0439	0,0040	-0,0055	0,0182	0,0443
-0,0275	-0,2580	-0,1233	-0,0673	0,0340	-0,1800
-0,0279	0,0496	-0,0045	0,0062	-0,0206	-0,0500

Em =

-0,0389	0,5252	-0,2305	-0,1320	-0,0167	0,0989
0,0083	0,0417	0,0379	0,0127	-0,0159	0,0470
-0,0373	0,0644	-0,0065	0,0068	-0,0263	-0,0652
-0,0027	0,0348	-0,0160	-0,0085	-0,0017	0,0067
-0,0177	0,1381	-0,0530	-0,0214	-0,0100	0,0319
0,0042	0,0225	0,0207	0,0075	-0,0074	0,0271
-0,0436	-0,2315	-0,2021	-0,0695	0,0792	-0,2559
-0,0306	0,4107	-0,1792	-0,1048	-0,0115	0,0784
-0,0023	0,0451	-0,0161	0,0078	0,0147	-0,0072
-0,0007	0,0021	-0,0010	-0,0001	0,0010	0,0012
-0,0480	-0,0169	0,0621	0,0895	-0,0403	0,0245
0,0253	-0,0430	0,0041	-0,0051	0,0186	0,0434
-0,0276	-0,2593	-0,1231	-0,0679	0,0331	-0,1802
-0,0276	0,0486	-0,0057	0,0062	-0,0207	-0,0496

Comparando ambas matrices, utilizando los criterios de errores ya detallados, se puede tener una idea de la magnitud de error de medición. El error porcentual promedio de la matriz E_m respecto a su contraparte real es de 23,8%, mientras que la matriz de errores porcentuales por coordenada es la que se muestra a continuación.

0,86	0,22	0,36	0,67	4,77	1,35
0,79	2,36	1,24	1,93	7,73	0,83
0,94	2,02	8,55	17,44	3,79	1,66
6,05	1,24	6,91	1,99	60,48	2,74
3,57	0,47	2,85	1,25	9,54	2,41
10,54	8,54	3,61	0,66	13,29	0,68
2,01	0,24	0,01	0,41	0,49	0,02
1,45	0,12	0,13	1,00	7,37	0,15
29,84	0,42	4,22	2,09	5,73	6,60
200,16	28,48	20,73	93,06	1224,48	114,25
1,77	0,59	0,59	0,81	2,39	4,12
2,41	2,16	2,75	7,34	1,80	2,03
0,33	0,52	0,13	0,98	2,49	0,14
1,12	1,89	28,10	0,59	0,61	0,93

Como se observa, cada coordenada posee un error diferente, pero que proviene de una misma distribución. Algo destacable es que las coordenadas con valores numéricos más pequeños, tendrán en promedio mayor error, debido a la alta precisión y sensibilidad que se requeriría de un instrumento[58]. El método propuesto para adicionar error a la matriz de datos mimetiza estos aspectos de los errores de medición.

Experiencia 1: Reconstrucción con errores pequeños

En esta experiencia, se testea el funcionamiento de NCA y gNCA básico sin modificaciones (nca_n y gnca_n), frente a la presencia de errores pequeños en la matriz de datos (del orden de 1% a 6% de error porcentual medio).

Se utiliza una red de tamaño pequeño (9x4x5x75), y se generan diferentes matrices de datos con error de acuerdo al siguiente detalle.

Matriz de datos	%(Er-Em)
E_{m1}	1,00
E_{m2}	2,00
E_{m4}	4,03
E_{m6}	6,05
E_{m10}	10,02

Tabla 9: Resumen de errores de las matrices de datos utilizadas. Pruebas G1N2E1.
Fuente: Elaboración propia.

[58] Se ahondará más en este punto en experiencias posteriores.

En primer lugar se utiliza nca_n con el fin de reconstruir las matrices de interés, utilizando como datos las matrices (y sus errores asociados) descritas en la Tabla 9. Los resultados fueron los siguientes.

% Error	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	MSE(Er-Em)	||Em-AP||	||Er-AP||
0	7,39E-08	0,00	1,30E-11	0,02	0,00E+00	9,53E-06	9,53E-06
1	2,35E-03	0,68	7,09E-08	1,51	0,00E+00	1,81E-03	1,81E-03
2	7,23E-04	0,66	1,32E-07	1,10	0,00E+00	3,18E-03	3,18E-03
4,03	2,72E-03	1,14	4,45E-07	3,98	0,00E+00	6,15E-03	6,15E-03
6,05	4,68E-03	2,43	3,49E-06	9,86	0,00E+00	9,17E-03	9,17E-03
10,02	9,53E-02	3,80	3,22E-06	6,20	0,00E+00	1,53E-02	1,53E-02

Tabla 10: Resumen de reconstrucción utilizando nca_n. Pruebas G1N2E1.
Fuente: Elaboración propia.

Como se observa, los errores en la matriz de datos se ven reflejados en la calidad y exactitud de la reconstrucción. Es necesario destacar que por lo general, y como se observa, los errores de los datos se distribuyen en las matrices reconstruidas, y por lo general los errores individuales de éstas serán menores al error porcentual medio de los datos. No obstante, el método se muestra bastante robusto al generar reconstrucciones bastante exactas pese al error de los datos. Además y tal cual parece indicar la tabla anterior, ciertas configuraciones de errores en la matriz de datos (independiente de error global) o cierta concentración de errores en solo algunas componentes, pueden propagar un mayor error a las matrices reconstruidas. Este es el caso de Em_6 y Em_{10}, donde pese a ser mayor el error en la segunda, el error propagado a la matriz P es mayor. Es posible analizar las matrices de errores porcentuales de ambas mediciones.

E_6 =

23.2719	4.1351	9.9590	2.2668	0.4689
0.4258	3.3288	14.7889	3.9208	0.6586
0.6432	0.4709	0.5970	0.3307	2.6166
2.4664	1.4447	0.3084	0.0906	1.0182
118.7271	5.1004	11.1549	5.2848	9.6754
0.8276	2.8279	13.7530	0.2407	1.7705
1.5123	8.4152	3.6849	2.3080	1.0708
0.0236	0.7088	2.4325	0.0637	0.0870
1.1026	0.2520	7.3670	0.1252	0.5163

E_10 =

6.0169	4.1799	190.2737	7.4346	3.0834
0.4794	1.9837	45.2941	1.0313	1.5227
1.3002	1.7701	2.9531	2.0170	28.5860
5.0867	0.4557	3.8989	2.6438	4.1315
25.5093	3.2189	14.9549	3.5957	2.3221
4.1669	11.9167	34.5392	4.8246	0.4171
0.9315	10.9712	0.2332	0.1308	1.6491
2.3993	0.1093	7.7334	0.3862	0.5064
0.3631	0.4875	4.3385	1.0406	0.3047

Se aprecia que en ambos casos, cerca del 42% del error porcentual medio es explicado por tan solo el error de una coordenada de la matriz de datos. Más aún, si dicha coordenada se torna relevante en la reconstrucción de una coordenada de A o P con bajo valor numérico, el error porcentual adicionado a dicha coordenada será considerablemente mayor por un efecto de precisión. Esto es lo que sucede en la matriz P reconstruida utilizando la matriz de datos con un 6% de error, en donde la mayor parte del error en P es explicado por una coordenada de bajo valor numérico mal reconstruida.

En la siguiente figura se puede apreciar el ajuste gráfico para la reconstrucción utilizando la matriz de datos con un 10% de error.

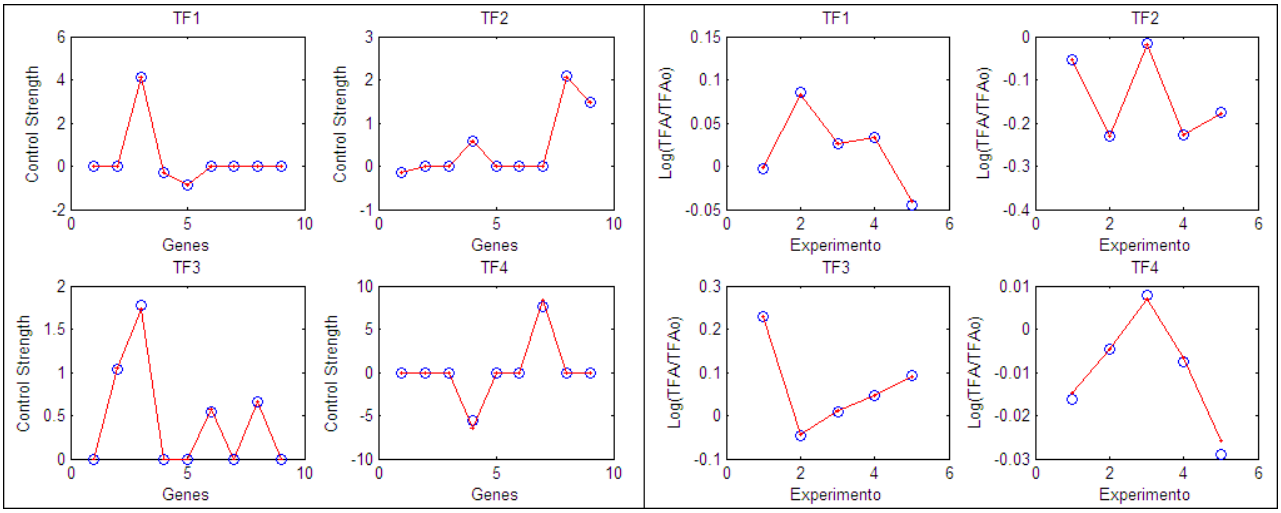


Figura 33: Ajuste gráfico de la reconstrucción utilizando nca_n. Pruebas G1N2E1.
Fuente: Elaboración propia.

Se observa que el ajuste es bastante exacto aunque es necesario destacar que lo anterior sirve solo como análisis cualitativo. Si se observa, la escala de los ejes es diferente para cada TF, por lo que tan lejos estén los puntos que unen la línea roja (reconstrucción) de los verdaderos valores (puntos azules) no habla directamente de la magnitud del error en comparación al resto de los $TF's$. Por ejemplo, en el experimento 5 del $TF4$, se observa un error considerable, cuando el realidad este es de solo un 10%. El error mayor lo posee en la matriz P el $TF1$ en el experimento 1, que por escala del gráfico, no se alcanza a apreciar. Esto, sin embargo es apreciado cualitativamente en las 2 matrices siguientes. La primera es el error porcentual de reconstrucción de cada coordenada para la matriz P, y la segunda la matriz de concentración de los errores para los gráficos anteriores.

```
E_P =
   38,59    3,19    4,46    2,56    7,95
    0,22    1,05    7,46    0,93    0,68
    0,54    3,84   11,35    0,60    3,15
    8,35    0,09    9,18    8,89   10,96
```

```
CE_P =
   31,11    2,57    3,59    2,07    6,41
    0,18    0,85    6,02    0,75    0,55
    0,43    3,09    9,15    0,48    2,54
    6,73    0,07    7,40    7,17    8,83
```

El experimento anterior fue repetido con otras matrices de datos (con los mismo errores), obteniendo los siguientes resultados:

% Error	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
0,00	7,86E-08	0,00	1,38E-11	0,02	9,83E-06	0,01	9,83E-06	0,01
1,00	7,20E-04	0,33	3,45E-08	1,02	1,21E-03	0,79	1,41E-03	0,57
2,01	2,84E-02	2,12	3,08E-07	2,05	2,18E-03	1,32	2,49E-03	0,84
4,02	1,08E-02	1,93	3,00E-07	2,64	3,59E-03	2,38	3,97E-03	1,44
6,04	1,47E-01	6,05	1,84E-06	6,50	6,11E-03	3,27	6,71E-03	2,56
10,02	3,75E-02	5,71	2,89E-06	7,25	7,95E-03	20,45	1,11E-02	4,17

Tabla 11: Resumen reconstrucción utilizando nca_n. Red diferente. Pruebas G1N2E1.
Fuente: Elaboración propia.

Se observa que los resultados coinciden con el análisis anterior, y se comportan incluso de mejor forma. También es necesario destacar que los errores en las reconstrucciones son diferentes (aunque cercanos) a los anteriores, pese a que el error porcentual en los datos es el mismo. Esto, tal cual se mencionó, es un indicio de que la concentración de los errores en la matriz[59] de datos es tanto o más relevante que el error porcentual medio en sí, y lo que realmente condiciona el resultado de la reconstrucción. Se desarrollará una experiencia para estudiar esto en mayor detalle.

Un último aspecto a considerar, es que las reconstrucciones anteriores correspondieron todas a mínimos globales, aun cuando los métodos convergían en ocasiones a mínimos locales. Lo mismo fue probado también con gnca_n, cuyos resultados se presentan a continuación. Se observa que pese a ser muy similares, algunas coordenadas presentan leves diferencias, por ejemplo, el error de reconstrucción de P con un 4% de error en los datos. Dichas diferencias se explican debido a que el método, en ocasiones converge a mínimos con diferentes ajustes, pero cuyas diferencias son tan ínfimas (al 5° o 6° decimal) que no son apreciables en los gráficos de distribución de los errores. Esto se asocia a errores numéricos de aproximación, y dichos errores se trasmiten a las matrices reconstruidas, existiendo pequeñas diferencias. Lo anterior puede ser eliminado en parte utilizando las funciones modificadas que son insensibles a la adivinación inicial. De todas maneras, y para fines prácticos, pueden considerarse reconstrucciones similares. En adelante, al comparar métodos se detallarán los resultados de nca_n y gnca_n, por lo que de haber diferencias (y suponiendo que no hay restricciones en la matriz P) se debe tener en consideración que los métodos deberían entregar el mismo resultado, y que las diferencias se deben a efectos de aproximación numérica.

[59] De que formase distribuyen estos errores en la matriz de datos, y cuanto aporta cada coordenada al error porcentual medio.

% Error	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	MSE(Er-Em)	||Em-AP||	||Er-AP||
0,00	7,49E-08	0,00	1,32E-11	0,02	0,00E+00	9,60E-06	9,60E-06
1,00	7,20E-04	0,33	3,45E-08	1,02	7,68E-08	1,21E-03	1,41E-03
2,01	2,84E-02	2,12	3,08E-07	2,05	2,43E-07	2,18E-03	2,49E-03
4,02	1,13E-02	1,97	3,23E-07	2,84	6,40E-07	3,59E-03	3,99E-03
6,04	1,47E-01	6,05	1,84E-06	6,50	1,82E-06	6,11E-03	6,71E-03
10,02	3,75E-02	5,71	2,89E-06	7,25	4,16E-06	7,95E-03	1,11E-02

Tabla 12: Resumen reconstrucción utilizando gnca_n. Pruebas G1N2E1.
Fuente: Elaboración propia.

Experiencia 2: Errores en la matriz de datos utilizando gnca_reg_n

El objetivo de esta experiencia es probar el funcionamiento del método no modificado de NCA con regularización (gnca_reg_n), utilizando la misma red y matrices de errores que las especificadas en el experimento resumido en la experiencia anterior.

% Error	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
0,00	3,18E-05	0,11	1,68E-08	0,62	3,30E-04	0,25	3,30E-04	0,25
1,00	3,60E-04	0,22	4,05E-08	0,57	1,31E-03	0,92	1,45E-03	0,38
2,01	3,15E-02	2,23	3,80E-07	3,05	2,27E-03	1,72	2,60E-03	1,20
4,02	8,94E-03	1,82	2,51E-07	1,53	3,67E-03	2,72	3,92E-03	0,97
6,04	1,56E-01	6,27	2,26E-06	7,90	6,17E-03	4,16	6,98E-03	3,11
10,02	3,10E-02	5,66	3,80E-06	8,37	8,00E-03	18,81	1,14E-02	4,81

Tabla 13: Resumen reconstrucción utilizando gnca_reg_n y diferentes errores en los datos. Pruebas G1N2E2.
Fuente: Elaboración propia.

Como se observa, cualitativamente el comportamiento es similar al obtenido con nca_n y gnca_n. El error en la reconstrucción aumenta en función de cuán grande es el error en los datos, pero por lo general este se dispersa entre las matrices reconstruidas. Luego, el error de reconstrucción es en general menor al de los datos. Tanto el error de ajuste a los datos reales como a los medidos aumenta con el error de los datos. Comparando los errores de reconstrucción medios de ambos métodos[60], se observa que en general es ligeramente menor en el método de regularización (en prácticamente todos los errores de datos menores a 4%) y prácticamente igual o ligeramente mayor con errores mayores. Esto puede ser un indicio del término de regularización, que pese a distorsionar los datos, cuando existen errores el efecto puede ser el contrario y disminuir la propagación del error a las matrices reconstruidas. Cabe destacar también que el tiempo de espera es considerablemente mayor al usar regularización. Una vez más, las reconstrucciones anteriores fueron obtenidas forzando al método a alcanzar el mínimo global.

[60] El mostrado y los obtenidos con nca_n y gnca_n (Tabla 11 y Tabla 12)

Es interesante analizar también lo que sucede con el error porcentual de ajuste a la matriz medida (7° columna) al aumentar el error en los datos. Si bien, el error nunca se escapa de forma abrupta al subir gradualmente el error de la medición, al pasar del 6% de error al 10%, el error de ajuste aumenta considerablemente. En este caso particular, una vez más el fenómeno es explicado por errores anormales en algunas coordenadas de la matriz de datos, que concentran la mayor parte del error global. Como se observa en la columna siguiente, el error de ajuste a los datos reales aun es bastante bueno.

En los siguientes gráficos se muestra un análisis de la distribución del error cuando se utiliza la matriz de datos con 4% de error en la matriz anterior para nca_n y gnca_n.

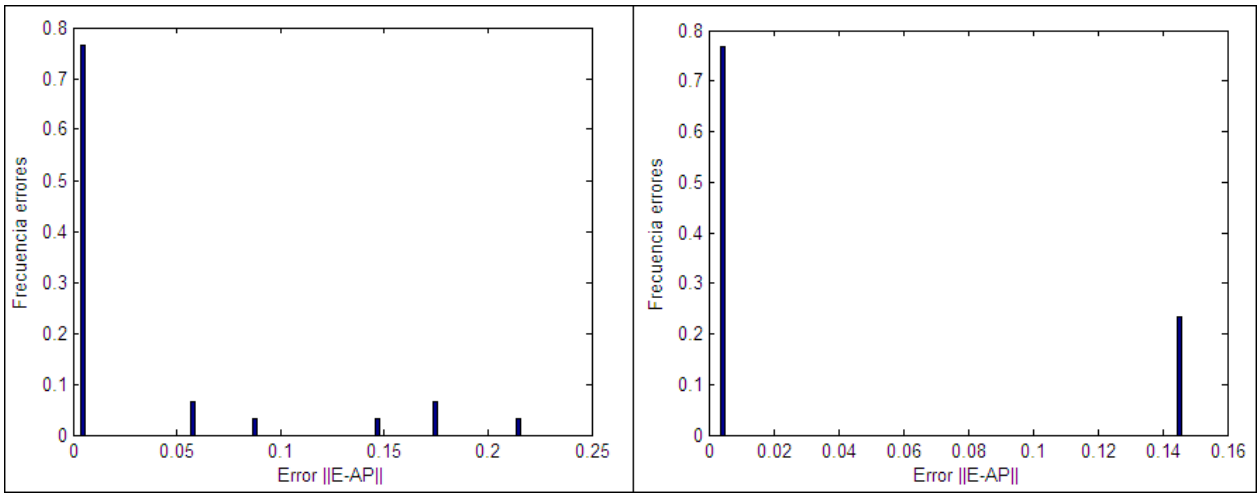


Figura 34: Distribución del error. Pruebas G1N2E2
A la isquierda al utilizatr nca_n, a la derecha gnca_reg.
Fuente: Elaboración propia.

Como se observa, la distribución del error se hace más estrecha con el uso del término de regularización, efecto que no se observaba al trabajar con la matriz de datos sin errores.

Experiencia 3: Reconstrucción en red de mayor tamaño y con un 30% error

En la siguiente experiencia se describen los resultados de reconstrucción al usar una red de mayor tamaño (35x8x9x80) y un 30% de error en la matriz de datos. En la tabla siguiente se presentan los resultados de la reconstrucción utilizando los 3 métodos originales.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
nca_n	2,26E-02	17,27	3,55E-06	6,32	9,19E-02	15,33	8,21E-02	12,34
gnca_n	2,26E-02	17,26	3,54E-06	6,33	9,19E-02	15,33	8,21E-02	12,34
gnca_reg_n	2,42E-02	17,21	3,79E-06	6,66	9,19E-02	15,33	8,22E-02	12,14

Tabla 14: Resumen de la reconstrucción utilizando 3 métodos. Pruebas G1N2E3.
Fuente: Elaboración propia.

Como se observa, los resultados sorprenden en relación a la exactitud del método, comparándolos con los errores observados en la matriz de datos. En especial el error de la matriz P es extremadamente bajo, lo que habla de la gran capacidad de la reconstrucción. Como antes, el método con regularización presenta un comportamiento similar, con valores de errores ligeramente inferiores en la matriz A y superiores en la matriz P producto del término de regularización. Como una manera de interpretar de mejor forma la reconstrucción, se presentan a continuación las matrices de errores porcentuales de reconstrucción de A y P.

E_A=

1,19	0	0	0	0	0	0	0
0	0	0	0	0	4,88	0	0
0	0	4,34	0	0	0	0	0
0,73	26,36	0	0	6,02	0	0	0,89
0	22,45	14,03	1,62	0	0	3,04	0
1,53	4,59	0	0	0	0	0	25,66
1,13	0	0	5,51	0	0	0	0
1,49	0	4,19	0	0	0	0	0
0,63	0	0	0	0	0	0	0
0	0	0	0	0	2,21	2,16	3,36
0	0,63	0	23,05	0	0	0	0
3,12	0	0	1,72	41,58	0	0	0
0,77	0	0	2,09	0	0	0	0
0	0	0	3,23	0,96	0	0	1,61
0	0	1,12	7,04	0	2,96	0	0,92
0	0	0,27	0	0	1,21	0	8,78
0	0	0	0	6,63	0	0	0
0	0	0	0	3,20	0	2,90	0
0	0	0	31,59	0	10,64	0	0
0	0	3,43	0	0	0	0	0
2,35	0	0	0	0	0	0	0
0	0	0	7,39	0	2,20	0	0
0,25	0	0	0	13,04	0	1,70	0
0,65	0	0	0	0	0	527,24	0
0,93	0	0	0	0	0	9,54	20,57
0	0	0	160,51	0	0	0,09	0,10
0	0	0	0	0	93,01	0	2,65
0	0	0	0	0	0	10,28	0
0	12,44	3,30	4,61	0	0	0	5,33
0	0	2,16	0	36,54	3,81	0	0
8,64	0	0	0	0	0	3,31	0
0	0	0	29,84	0	0	0	0
0	0	0	0	2,61	0	19,29	0
1,85	61,48	0	0	0	0	0	0
0	0	0	0	0	0	1,06	0,64

E_P=

1,06	0,01	0,87	1,55	11,46	0,57	6,88	3,44	63,74
11,66	2,78	3,38	29,18	15,25	10,69	13,96	6,74	5,43
4,01	6,13	2,27	8,07	45,62	3,85	2,70	3,68	3,33
2,61	12,49	1,17	6,58	4,75	4,65	3,02	1,52	0,69
0,17	1,68	1,67	2,63	4,40	4,17	1,06	0,61	7,01
1,04	6,35	14,73	5,41	4,28	1,23	3,50	3,50	1,17
0,72	1,17	3,44	1,41	0,91	10,95	0,96	4,00	0,51
2,52	0,07	3,85	0,03	0,70	8,95	2,54	40,24	0,15

Es posible apreciar varias cosas al respecto. La matriz A, que presenta un error porcentual medio de cerca del 17%, presenta una entrada que concentra la mayor parte de dicho error. Exactamente, el 39,1% del error porcentual medio se atribuye a la entrada destacada. Eliminándola del análisis, el error

porcentual bajaría a tan sólo un 10,64%, pese al 30% de error en los datos. Inspeccionando los errores de ajuste de cada coordenada, se ve que en la mayor parte de las mismas (en ambas matrices) la reconstrucción es en extremo exacta, lo que es un indicio de la gran robustez y capacidad de la técnica. Centrándose específicamente en la matriz A, se puede apreciar que la entrada que concentra la mayor parte del error, es también una con un valor numérico real sumamente bajo (cercano a cero). Luego, el error de reconstrucción, debido a un efecto de precisión tiende a ser mayor en coordenadas que presentan dicha característica[61]. Estos casos son fácilmente identificados al encontrar coordenadas en las matrices de reconstrucción que concentren una parte importante del error. En redes reales por supuesto, este análisis no es factible, pero si es posible tener mayor precaución en el grado de validez que se da a coordenadas que de por sí, fueron reconstruidas asignándoles un valor numérico bajo. Otra forma de interpretar lo anterior, es que el error en la matriz de datos se distribuye de forma más o menos uniforme en la reconstrucción. Obviamente las entradas de menor valor recibirán una mayor perturbación, traducido en el efecto comentado con anterioridad. Esto es por supuesto, válido también para la matriz P.

Es posible también analizar el ajuste gráfico para la reconstrucción de la matriz P. Si bien esto entrega una apreciación cuantitativa, se ve que el ajuste es bastante exacto, y que la tendencia de la actividad de los $TF's$ se reproduce perfectamente.

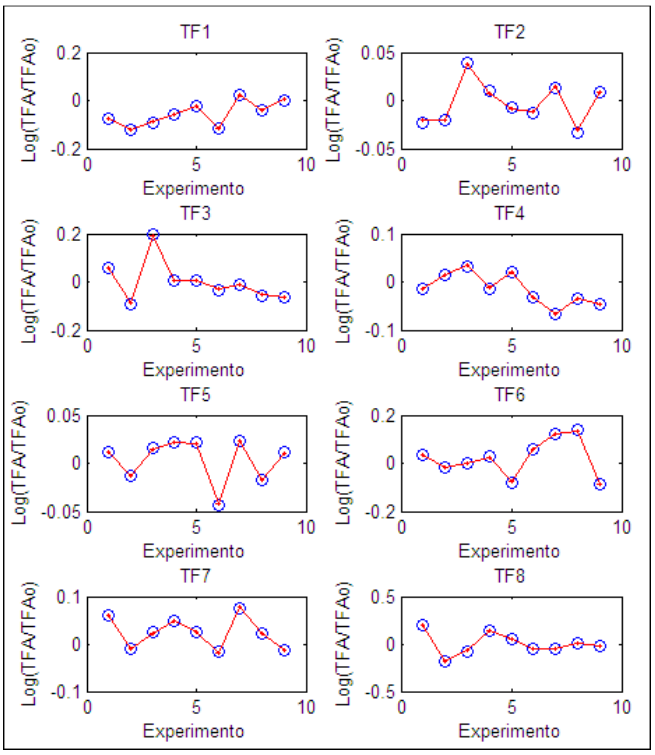


Figura 35: Ajuste gráfico $\boldsymbol{TFA's}$ reconstruidos. Pruebas G1N2E3.
Fuente: Elaboración propia.

[61] Es más difícil ajustar el verdadero valor a coordenadas que requieren una mayor precisión, por lo que el ajuste será más deficiente en dichos casos.

Como un análisis extremo, se comprueba el funcionamiento de los métodos utilizando errores considerables en la matriz de datos (un 100% de error porcentual medio). El resumen de los resultados se presenta en la siguiente tabla:

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
nca_n	6,10E-01	87,44	4,93E-05	51,16	2,91E-01	53,94	2,78E-01	46,65
gnca_n	6,02E-01	87,64	4,92E-05	51,53	2,91E-01	53,91	2,78E-01	46,67
gnca_reg_n	6,29E-01	87,08	5,04E-05	50,51	2,91E-01	54,16	2,78E-01	46,80

Tabla 15: Resumen de la reconstrucción frente a errores extremos. . Pruebas G1N2E3.
Fuente: Elaboración propia.

Como se puede observar, los errores de la reconstrucción son considerables, pero aun así menores a los provenientes de los datos. Es interesante analizar que a medida que crece el error en los datos, el ajuste a los mismos, y a la matriz de datos real se hace más insensible al método[62]. Es algo que también se observa al analizar los datos con un 30% de error. Una vez más la matriz reconstruida de A presenta el mayor error.

Si bien no se detallan, al observar las matrices de errores porcentuales por coordenadas de las matrices reconstruidas, se puede apreciar que aún cuando algunas coordenadas presentan inexactitudes groseras, el 57,6% de las entradas reconstruidas presentan un error porcentual inferior al 20%, y un 39,7% inferior al 10%. Esto habla en promedio de una gran exactitud de la reconstrucción, pese a la gran omisión de la medición de los datos. El gran error porcentual medio de A es explicado por ciertas coordenadas que concentran la mayor parte del mismo. Exactamente, 2 coordenadas de la matriz explican el 50% del 87,18% de error porcentual medio observado. Si no se considera la mayor de éstas en el análisis, el parámetro disminuye a cerca del 60%, equiparándose aún más con lo observado en la matriz P. Una vez más y tal cual se ha comentado, las entradas con los errores más groseros corresponden a entradas con bajo valor numérico, que relativamente reciben una mayor perturbación.

El caso de la matriz P es similar. Si bien el error de reconstrucción es elevado, es posible apreciar inspeccionando la matriz respectiva que más del 60% de las entradas presentan desajuste menores al 20%. Más aun, sólo una coordenada concentra el 55,7% del error porcentual medio, el cual disminuye a 22,7% al obviar dicha entrada. Es destacable la gran exactitud de la mayor parte de los parámetros estimados, pese a la gran perturbación a la que son sometidos los datos.

Finalmente, y como una manera de apreciar gráficamente el desajuste de algunos parámetros, se presenta un esquema de ajuste de la matriz P, donde en esta caso particular es posible apreciar de buena forma los errores en algunas coordenadas.

[62] Si se observan las columnas respectivas, prácticamente no varian.

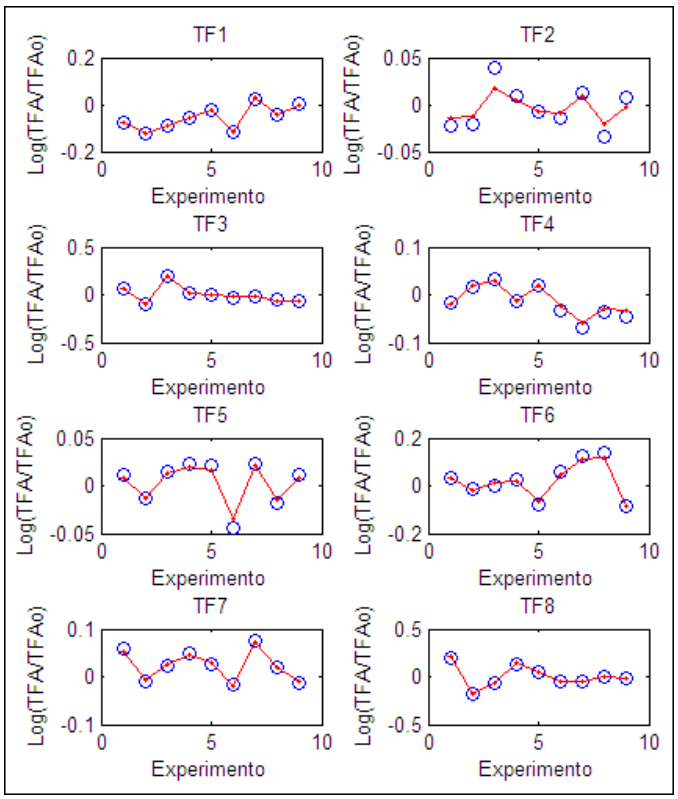


Figura 36: Ajuste gráfico $\boldsymbol{TFA's}$ reconstruidos con un error del 100%. Pruebas G1N2E3.
Fuente: Elaboración propia.

<u>Experiencia 4: Prueba en reconstrucción con matrices de datos con diferentes concentraciones de errores</u>

Una pregunta que surge respecto a las reconstrucciones anteriores, es el efecto en la reconstrucción cuando las matrices de datos presentan diferentes concentraciones de errores. Es lógico pensar que si bien 2 matrices de datos poseen los mismos errores porcentuales medios, el efecto será diferente si en una de ellas el error se encuentra concentrado mayoritariamente en algunas coordenadas, que distribuido de forma más uniforme como se esperaría en presencia de error experimental.

En esta experiencia se utilizan 2 matrices de datos de una red de tamaño pequeño (17x4x5x78), con similares errores porcentuales medios, pero diferentes concentraciones del mismo. Las matrices de errores porcentuales y de concentración de errores se muestran a continuación:

E1 =

0,13	115,23	1,18	3,42	4,63
4,49	0,73	93,32	3,26	25,62
23,38	1,72	3,26	0,66	0,41
1,41	4,32	0,32	8,98	1,49
1,56	6,97	80,53	0,01	12,19
1,51	0,21	2,61	2,65	0,06
5,00	2,02	6,29	21,84	2,25
47,47	372,85	28,43	217,91	10,19
345,07	1,50	98,05	10,91	0,89
45,53	427,00	21,49	73,96	45,45
2,01	6,67	0,73	57,01	1,32
33,33	0,04	3,76	2,44	0,01
5,77	34,60	4,82	4,76	3,61
0,30	1,28	0,58	1,12	1,49
0,55	7,39	10,80	24,59	1,63
9,79	0,12	8,20	0,82	0,60

C1 =

0,01	4,76	0,05	0,14	0,19
0,19	0,03	3,86	0,13	1,06
0,97	0,07	0,13	0,03	0,02
0,06	0,18	0,01	0,37	0,06
0,06	0,29	3,33	0,00	0,50
0,06	0,01	0,11	0,11	0,00
0,21	0,08	0,26	0,90	0,09
1,96	15,40	1,17	9,00	0,42
14,26	0,06	4,05	0,45	0,04
1,88	17,64	0,89	3,06	1,88
0,08	0,28	0,03	2,36	0,05
1,38	0,00	0,16	0,10	0,00
0,24	1,43	0,20	0,20	0,15
0,01	0,05	0,02	0,05	0,06
0,02	0,31	0,45	1,02	0,07
0,40	0,00	0,34	0,03	0,02

E2 =

1,18	4,40	0,32	1,65	0,21
0,87	0,30	13,08	0,13	1571,77
0,52	0,14	1,10	0,18	0,02
0,06	2,72	0,38	9,31	66,86
1,67	0,28	14,61	1,18	3,19
1,07	0,53	3,20	0,10	0,09
0,48	0,32	1,22	0,10	0,62
18,99	234,54	3,18	24,84	5,80
111,17	0,99	29,86	1,86	1,83
6,82	139,55	2,35	4,20	11,98
0,16	52,67	6,46	4,02	0,01
3,54	0,64	0,63	0,74	0,57
1,09	19,56	3,02	9,10	2,26
0,55	0,21	0,06	1,19	0,29
1,22	3,43	4,87	5,38	0,56
0,30	0,19	11,67	0,48	1,01

C2 =

0,05	0,18	0,01	0,07	0,01
0,04	0,01	0,54	0,01	64,48
0,02	0,01	0,05	0,01	0,00
0,00	0,11	0,02	0,38	2,74
0,07	0,01	0,60	0,05	0,13
0,04	0,02	0,13	0,00	0,00
0,02	0,01	0,05	0,00	0,03
0,78	9,62	0,13	1,02	0,24
4,56	0,04	1,22	0,08	0,07
0,28	5,72	0,10	0,17	0,49
0,01	2,16	0,27	0,16	0,00
0,15	0,03	0,03	0,03	0,02
0,04	0,80	0,12	0,37	0,09
0,02	0,01	0,00	0,05	0,01
0,05	0,14	0,20	0,22	0,02
0,01	0,01	0,48	0,02	0,04

Como se aprecia por inspección en la primera matriz de datos, la entrada con el error más grosero concentra el 17,6% del error porcentual medio total, y las 2 siguientes el 15,4% y el 14,2% respectivamente. El resto del error se reparte de manera más o menos uniforme entre las coordenadas restantes. Así, cerca del 47% está concentrada por dichas 3 coordenadas. En la segunda matriz de datos en cambio, una sola entrada concentra el 64,4% del error porcentual medio. El objetivo es analizar el efecto de esto en las matrices reconstruidas.

En la tabla siguiente se observa el efecto en la reconstrucción.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	MSE(Er-Em)	%(Eo-Em)	||Em-AP||	||Er-AP||
nca_n	4,14E-03	6,32	2,03E-05	11,32	2,86E-05	30,26	3,09E-02	3,66E-02
nca_n	1,22E+00	38,71	6,27E-04	40,24	3,79E-03	30,47	2,71E-01	5,07E-01

Tabla 16: Resumen de la reconstrucción frente a diferentes concentraciones de errores. Pruebas G1N2E4. La primera fila resume los resultados para un error porcentual poco concentrado, mientras la segunda lo hace para un error altamente concentrada. Ambas matrices de datos poseen el mismo error porcentual medio. Fuente: Elaboración propia.

3.16.3 Nivel 3

En el siguiente conjunto de experiencias se pretende comenzar a testear el funcionamiento de los nuevos métodos creados, y compararlos con los métodos NCA originales. En particular se comprueba la dinámica del uso de datos de la confiablidad de los datos en la reconstrucción, como una forma de mejorar los resultados frente a errores en los datos.

Un punto importante al respecto es que desde ahora en adelante se utilizarán los métodos originales modificados (NCAbasic, gNCAbasic, gNCAreg), que han sido alterados con un algoritmo de pre-búsqueda, que asegura la convergencia a un mínimo global[63]. De esta forma se evitan errores en las reconstrucciones inducidas por la convergencia a mínimos locales, que no es el interés de esta sección.

Con el fin de simular los datos necesarios para este tipo de experiencias, se definirán 2 matrices nuevas: la matriz promedio de la medición de datos (E_p), y la matriz de varianza y desviación estándar de la medición de los datos ($R\ y\ S$).

Las matrices anteriores se definen en función de diferentes mediciones (experimentos) de la matriz E real. Por ejemplo, es posible realizar diferentes experimentos de microarray (cada uno corresponderá a una medición particular), y definir las matrices anteriores en función de dichas matrices. Más exactamente, sea E_{m_f} la medición (sujeta a errores experimentales por supuesto) f de la matriz de datos reales con $f = \{1, 2, \dots, F\}$. Se define entonces de la misma manera que en 3.9 Errores en los datos y NCA considerando confiabilidad, las siguientes matrices:

$$E_p = \frac{\sum_{t=1}^{F} E_{m\_f}}{F}$$

Ecuación 80

$$R = \frac{\sum_{t=1}^{F} \left(E_{m\_f} - E_p\right)^2}{F}$$

Ecuación 81

$$S = \sqrt{R}$$

Ecuación 82

Donde se ha hecho un abuso de notación, y las funciones aplicadas a las matrices se aplican a cada componente de ésta por separado.

[63] Funcionalmente, y asumiendo que alcanzan el mínimo global, se comportan de la misma manera que los métodos antes testeados.

A modo de ejemplo se considerará una red de tamaño pequeño (16x4x5x78) y 5 mediciones con error de medición porcentual promedio del 35% respecto a los valores reales de la matriz de datos. La matriz siguiente representa la segunda de dichas mediciones, junto a su matriz de errores porcentuales.

Em_2 =

-0,0799	-0,1709	0,0695	0,0158	0,1286
-0,0620	-0,0614	0,0610	0,0573	-0,0937
-0,1028	-0,1015	0,1020	0,0953	-0,1552
-0,1370	-0,2919	0,1184	0,0263	0,2188
-0,1511	-0,3199	0,1590	0,0740	0,1178
0,3321	0,1147	-0,0719	-0,0119	-0,1293
0,1540	-0,1510	0,0267	-0,0147	-0,0085
-0,1768	-0,1750	0,1753	0,1635	-0,2676
-0,0055	-0,0027	0,0018	0,0002	0,0023
-0,0049	-0,0102	0,0040	0,0008	0,0079
-0,0003	0,0004	-0,0003	0,0002	0,0006
-0,3069	0,1815	-0,0177	0,0242	0,0479
-0,5629	0,5518	-0,0977	0,0543	0,0318
-0,1728	0,3694	-0,0967	0,0107	-0,0890
0,0604	-0,0655	0,0125	-0,0060	-0,0006
-0,5260	0,4963	-0,0794	0,0590	0,0141

E =

0,50	0,09	0,30	2,63	0,25
0,26	0,33	0,55	0,11	0,04
0,23	0,08	0,32	0,13	0,13
0,09	0,13	0,01	0,26	0,12
0,08	0,01	0,06	0,29	0,08
0,03	0,22	0,39	0,21	0,07
0,04	0,07	0,27	1,44	3,12
0,16	0,17	0,13	0,16	0,01
5,22	5,18	26,91	2,14	4,82
6,32	3,15	1,27	7,16	5,67
11,16	322,74	418,01	1619,25	332,11
0,00	0,22	1,16	1,00	0,83
0,03	0,11	0,00	0,43	0,35
0,01	0,03	0,02	1,31	0,27
0,45	0,40	0,74	5,32	2,46
0,04	0,05	0,51	0,12	1,58

La matriz siguiente, junto a su respectiva matriz de error representa la matriz promedio de las mediciones. Se puede apreciar que el error porcentual medio ha disminuido a 23,34%.

Ep =

-0,0803	-0,1709	0,0696	0,0156	0,1284
-0,0618	-0,0611	0,0611	0,0572	-0,0938
-0,1026	-0,1015	0,1019	0,0949	-0,1555
-0,1370	-0,2917	0,1185	0,0262	0,2190
-0,1509	-0,3200	0,1590	0,0740	0,1177
0,3322	0,1149	-0,0715	-0,0117	-0,1291
0,1543	-0,1511	0,0267	-0,0149	-0,0087
-0,1766	-0,1750	0,1752	0,1638	-0,2677
-0,0056	-0,0027	0,0012	-0,0001	0,0023
-0,0048	-0,0100	0,0040	0,0009	0,0073
-0,0004	-0,0001	0,0000	-0,0002	0,0001
-0,3069	0,1811	-0,0176	0,0243	0,0481
-0,5629	0,5524	-0,0979	0,0545	0,0320
-0,1728	0,3693	-0,0968	0,0106	-0,0894
0,0606	-0,0654	0,0125	-0,0057	-0,0007
-0,5257	0,4966	-0,0797	0,0592	0,0144

E =

0,03	0,08	0,42	1,21	0,06
0,12	0,11	0,35	0,22	0,04
0,02	0,02	0,23	0,24	0,03
0,08	0,04	0,10	0,76	0,04
0,04	0,03	0,06	0,35	0,02
0,04	0,03	0,08	1,33	0,05
0,12	0,04	0,02	0,22	0,15
0,02	0,20	0,07	0,02	0,02
2,89	4,31	18,02	133,30	2,84
4,06	1,13	1,70	4,09	1,64
6,89	52,08	140,24	1464,23	4,85
0,00	0,01	1,95	0,51	0,38
0,02	0,01	0,22	0,02	0,35
0,00	0,03	0,13	0,00	0,09
0,15	0,20	0,73	0,17	12,62
0,01	0,03	0,14	0,28	0,22

Dicha disminución se explica, de hecho, por el promedio de las diferentes mediciones. El promedio de cada entrada tenderá a ser un mejor estimador que una medición en particular[64] (y de menor varianza), por lo que es normal que el error disminuya en dicha matriz. En este ejemplo particular hay que destacar también que una de las entradas concentra más del 70% del error porcentual medio), y si bien la mayor parte de las coordenadas disminuye su error en la matriz de datos promedios, el efecto es menor en dicha entrada. Es posible observar también la matriz de varianza de las mediciones. Si bien sus valores numéricos son pequeños, se debe considerar que los valores de la matriz de los datos también lo son, y las varianzas de la matriz R están en dichas unidades al cuadrado. Además, relativamente hablando, se

[64] Esto es fácilmente demostrable. Si se tiene una muestra aleatoria, el promedio, bajo suposiciones generales, es un estimador de menor varianza que un elementos particular de dicha muestra (pese a ser ambos insesgado).

pueden observar datos con varianzas (una suerte de inexactitud de la medición) hasta 4 veces mayores de una entrada a otra.

R =

0,0003	0,0001	0,0002	0,0005	0,0002		0,86	0,22	0,36	0,67	4,77	1,35
0,0002	0,0002	0,0002	0,0001	0,0001		0,79	2,36	1,24	1,93	7,73	0,83
0,0001	0,0001	0,0003	0,0003	0,0003		0,94	2,02	8,55	17,44	3,79	1,66
0,0002	0,0001	0,0002	0,0004	0,0001		6,05	1,24	6,91	1,99	60,48	2,74
0,0002	0,0002	0,0002	0,0002	0,0003		3,57	0,47	2,85	1,25	9,54	2,41
0,0001	0,0002	0,0002	0,0003	0,0002		10,54	8,54	3,61	0,66	13,29	0,68
0,0005	0,0001	0,0001	0,0002	0,0003		2,01	0,24	0,01	0,41	0,49	0,02
0,0002	0,0003	0,0002	0,0001	0,0002		1,45	0,12	0,13	1,00	7,37	0,15
0,0001	0,0002	0,0005	0,0002	0,0002		29,84	0,42	4,22	2,09	5,73	6,60
0,0002	0,0001	0,0001	0,0001	0,0005		200,16	28,48	20,73	93,06	1224,48	114,25
0,0001	0,0002	0,0002	0,0002	0,0003		1,77	0,59	0,59	0,81	2,39	4,12
0,0001	0,0003	0,0005	0,0001	0,0002		2,41	2,16	2,75	7,34	1,80	2,03
0,0001	0,0004	0,0004	0,0002	0,0004		0,33	0,52	0,13	0,98	2,49	0,14
0,0002	0,0004	0,0002	0,0001	0,0002		1,12	1,89	28,10	0,59	0,61	0,93
0,0001	0,0002	0,0003	0,0002	0,0001							
0,0001	0,0003	0,0002	0,0005	0,0002							

Experiencia 1: Efecto en redes pequeñas

Una vez más se comienza explorando el efecto de la reconstrucción con cgNCA utilizando redes pequeñas, y comparando su funcionamiento con los métodos modificados NCAbasic, gNCAbasic, gNCAreg. La red utilizada es de tamaño (16x4x5x77) y se generaron 5 mediciones con un 45% de error porcentual medio. La matriz promedio de los datos (E_p) presentó un error de 24,3%.

En la tabla siguiente se resumen los resultados de la reconstrucción utilizando 4 métodos NCA diferentes.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,83E-04	0,48	5,54E-08	1,46	2,86E-03	9,50	2,93E-03	3,30
gNCAbasic	1,83E-04	0,48	5,54E-08	1,46	2,86E-03	9,50	2,93E-03	3,30
gNCAreg	1,91E-04	0,53	1,18E-08	0,96	2,94E-03	9,71	2,83E-03	2,07
cgNCA	1,64E-04	0,61	4,80E-08	1,47	2,96E-03	9,56	3,04E-03	2,73

Tabla 17: Resumen de la reconstrucción utilizando E_p . Pruebas G1N3E1.
Fuente: Elaboración propia.

Como se observa, lo resultado obtenidos con los 3 métodos ya testeados coinciden con los obtenidos con anterioridad. Los resultados de NCAbasic y gNCAbasic coinciden, y en este caso particular gNCAreg presenta un error levemente menor en P producto de la característica de los datos (los datos reales de la matriz P se ubican cerca de cero en promedio, por lo que la distorsión del termino de regularización es menor que la causada por el error de los datos). En el caso de cgNCA, el error aumenta ligeramente en P pero lo hace en mayor proporción en la matriz A. Esto es algo no esperable dada la teoría, y tiene su explicación en la naturaleza de los datos. Como se mencionó con anterioridad, el considerar el promedio de diferentes mediciones, el error de la matriz resultante es menor que el de las mediciones. Y es lo que

precisamente se quiere conseguir en un experimento al usar el promedio como un estimador del valor real que se intenta medir. La media muestral surge como un estimador lógico, que bajo ciertas condiciones es insesgado, por lo que su media tenderá al valor real. Si se ve desde otro aspecto, no es más que una consecuencia de la ley de los grandes números. Es posible también analizar la matriz del promedio de los datos y de su varianza, y se identifica que aquellos valores con una mayor varianza no presentan un error del todo grosero en la matriz de datos promedios respecto a su valor real. En otras palabras, pese a que la varianza de la medición puede ser grande en alguna entrada en particular, la varianza del promedio de dicha medición puede ser bastante más pequeña. Luego, incluir la varianza en la reconstrucción puede distorsionar los datos, que de por sí, ya están próximos al valor real. Obviamente lo anterior supone que el ruido es blanco, y que en promedio todas las entradas reciben el mismo tipo de error de medición.

Nótese también, que pese al gran error de los datos, la reconstrucción presenta un error despreciable con todos los métodos, lo que ciertamente abala la capacidad y robustez del mecanismo. Esto es más bien la regla cuando no existen coordenadas anómalas que concentran la mayor parte del error de las reconstrucciones. Además, una vez más se observa el efecto ya comentado, donde el error porcentual de ajuste a los datos reales es menor en todos los casos al mismo error al considerar el ajuste a la medición utilizada.

Como una manera de verificar lo anterior, se generan otras mediciones (con el fin de obtener $\boldsymbol{E_{p2}}$) con las mismas características, y se procede a reconstruir los parámetros utilizando una vez más los 4 métodos. Los resultados son los siguientes.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	3,30E-04	0,89	4,26E-08	1,39	3,52E-03	10,24	3,00E-03	1,46
gNCAbasic	2,63E-04	0,85	3,97E-08	1,33	3,52E-03	10,23	3,00E-03	1,33
gNCAreg	1,12E-04	0,69	4,08E-08	1,07	3,60E-03	10,12	3,06E-03	1,75
cgNCA	1,86E-04	0,92	4,10E-08	1,61	3,78E-03	10,60	3,34E-03	1,86

Tabla 18: Resumen de la reconstrucción utilizando $\boldsymbol{E_{p2}}$. Pruebas G1N3E1.
Fuente: Elaboración propia.

Como se observa, los resultados son similares a los obtenidos con anterioridad.

Experiencia 2: Prueba con redes de mayor tamaño

En esta experiencia se repite el experimento anterior, pero utilizando una red de tamaño medio (54x16x17x80) y una red de tamaño grande (115x27x28x78). En cada red su utilizaron 3 mediciones con un error del 40% para calcular la matriz de datos promedios y la de varianza R.

La red de tamaño medio genera una matriz de datos promedios con un error del 22,4%. Los resultados de la reconstrucción se resumen en la siguiente tabla.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,30E-01	4,91	2,57E-06	13,44	1,25E-01	26,77	1,19E-01	6,97
gNCAbasic	1,31E-01	4,88	2,57E-06	13,50	1,25E-01	26,75	1,19E-01	6,97
gNCAreg	1,19E-01	5,04	2,57E-06	13,45	1,25E-01	26,59	1,20E-01	6,95
cgNCA	2,78E-01	7,09	4,00E-06	18,16	1,52E-01	23,91	1,44E-01	9,91

Tabla 19: Resumen de la reconstrucción en red de tamaño medio. Pruebas G1N3E2.
Fuente: Elaboración propia.

Se observa algo similar a lo comentado con anterioridad. La reconstrucción de los métodos básicos es similar, y en este caso, el efecto de la regularización aumenta levemente el valor del error promedio en A. De la misma manera, el método que incorpora la matriz de varianza de las mediciones distorsiona considerablemente los resultados comparándolo a los métodos anteriores, resultado atribuible al efecto comentado ya en la sección anterior. De todas maneras, y en todos los casos, el error de reconstrucción es inferior al error de los datos.

En la red de tamaño grande, la matriz de datos promedios presenta un error porcentual de 25,6%. Los resultados de la reconstrucción son los siguientes.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	4,94E+00	8,81	2,64E-06	8,12	4,09E-01	13,19	3,83E-01	14,08
gNCAbasic	4,62E+00	8,62	2,61E-06	8,15	4,10E-01	13,15	3,83E-01	14,03
gNCAreg	4,87E+00	8,67	2,63E-06	8,01	4,10E-01	12,92	3,84E-01	14,13
cgNCA	4,05E+01	11,62	4,16E-06	10,73	5,14E-01	15,03	4,96E-01	18,26

Tabla 20: Resumen de la reconstrucción en red de tamaño grande. Pruebas G1N3E2.
Fuente: Elaboración propia.

Si bien el algoritmo demora más en converger dada la dimensión de la red, los resultados son similares a los ya comentados. Una diferencia radica en que en esta ocasión, el ajuste a los datos medidos fue levemente mejor que el ajuste a los reales. En redes pequeñas no se había observado dicho fenómeno, por lo que es interesante analizar si este comportamiento tiene que ver algo con el tamaño de la red.

Una reflexión útil en esta instancia, y que servirá para aclarar lo realizado en las próximas experiencias, tiene que ver con el algoritmo para crear las redes sintéticas y mimetizar el error adicionado a los datos producto de las mediciones experimentales. El procedimiento general, como ya se ha explicado, consiste en generar una red de un tamaño definido, definiendo aleatoriamente los parámetros para A y P. Esto a su vez define una matriz de datos E real. Dicha matriz es perturbada adicionando ruido blanco, a fin de obtener un error promedio dado. El punto relevante, es que a medida que crece la red, la matriz E tendrá cada vez más entradas, y por ende, la probabilidad que una de ellas presente un valor numérico bajo aumento. El error adicionado proviene de la misma distribución, por lo que en estos casos, las entradas con bajo valor numérico recibirán relativamente un mayor error, y de esta manera concentrarán la mayor parte del error total objetivo de la matriz. Esto no es deseado, ya que como se ha

visto, altas concentraciones de errores alteran los resultados, por lo que en adelante, los métodos de creación de redes son modificados tomando esto en consideración.

Experiencia 3: Prueba con varias redes

En la siguiente experiencia, y como una forma de corroborar las conclusiones halladas hasta el momento, se procede a realizar una experiencia que testeará de forma automática el funcionamiento de los métodos en 30 redes diferentes (aunque del mismo tamaño). Si bien la mayor parte de las redes se comportaron de manera normal (de acuerdo a lo ya explicado, y consistentemente con las conclusiones y explicaciones obtenidas), algunas redes presentan comportamientos anormales en la reconstrucción.

A continuación se describirá lo encontrado, y las explicaciones propuestas.

Las 30 redes generadas presentaron un tamaño medio (22x5x6x75), y el error adicionado a las matrices de datos fue escogido aleatoriamente. Dentro de ellas, los casos interesantes fueron los siguientes.

Red 2: En esta red la matriz de datos medidos posee un error porcentual medio de 21%. Los resultados de la reconstrucción son los siguientes.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	9,20E-05	0,66	2,07E-07	5,26	8,14E-03	4,34	6,60E-03	3,51
gNCAbasic	9,19E-05	0,66	2,06E-07	5,22	8,14E-03	4,34	6,60E-03	3,49
gNCAreg	9,46E-05	0,67	2,69E-07	5,89	8,24E-03	4,37	7,04E-03	3,77
cgNCA	1,06E-04	0,83	2,21E-07	16,33	9,41E-03	4,15	7,80E-03	6,60

Tabla 21: Resumen de la reconstrucción. Red 2.Pruebas G1N3E3.
Fuente: Elaboración propia.

Como se observa, si bien los 3 primeros métodos se comportan de buena manera, el error introducido a la matriz P al utilizar cgNCA aumenta casi al triple. Analizando los datos la explicación sale a la luz, y tiene mucho que ver con el comentario realizado al final de la sección anterior. La matriz de datos medidos posee una entrada que concentra el 60,4% del 21% de error porcentual de la medición. Analizando el dato con dicho problema, se ve que corresponde a una entrada con bajo valor numérico, y que por lo tanto, en la simulación recibe un error anormal que concentra la mayor parte del error objetivo. Y más grave que eso hay otro factor que propaga aún más el error a la reconstrucción, y es el que la varianza de dicha medición es numéricamente baja (casi un orden de magnitud menor a la del resto de las observaciones). Teóricamente, una entrada mal medida debiese tener una alta varianza, con el objetivo de ser poco considerada con el método cgNCA, pero en este caso ocurre lo contrario. La entrada que concentra la mayor parte de error presenta la varianza más baja (lo que equivale a una alta precisión), por lo que se le da mayor importancia en la reconstrucción, empeorando los resultados. El porqué de esta anomalía no tiene que ver con el método, sino con la simulación realizada al adicionar los errores a los datos. Una vez más se recalca, que como regla general, hay que tener precaución (a nivel teórico y experimental) con los datos medidos que presentan bajo valor numérico y con las reconstrucciones de

las mismas características. En el primer caso, la precisión del instrumento puede adicionar un gran error a dicha entrada, que no necesariamente va a estar correlacionado con una alta varianza, mientras que en el segundo, la precisión inherente de la reconstrucción puede adicionar mayor error a dichas estimaciones.

Es posible, de hecho, repetir la reconstrucción anterior modificando artificialmente la coordenada de la matriz de varianza correspondiente al dato con problemas, haciendo que realmente represente dicha inexactitud. Los resultados son los siguientes, donde claramente se observa la diferencia, y donde se aprecia incluso una mejor reconstrucción. Esto último es esperable, en este caso el efecto de distorsión introducido por R es dominado por la mayor exactitud que se gana al considerar en menor medida el dato mal medido.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	9,19E-05	0,65	2,09E-07	4,89	8,14E-03	4,33	6,61E-03	3,43
gNCAbasic	9,19E-05	0,65	2,10E-07	4,94	8,14E-03	4,33	6,61E-03	3,44
gNCAreg	1,12E-04	0,72	2,55E-07	5,47	8,23E-03	4,28	6,82E-03	3,66
cgNCA	1,14E-04	0,85	2,13E-07	2,81	9,35E-03	4,65	7,70E-03	3,09

Tabla 22: Resumen de la reconstrucción. Red 2 modificada. Pruebas G1N3E3.
Fuente: Elaboración propia.

Red 5: Este caso corresponde a una red, en donde a pesar de ser bajo el error de la mediciones (aproximadamente un 6%), el error en la reconstrucción es considerable en ambas matrices. Un resumen de las reconstrucciones se presenta en la siguiente tabla:

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,95E+01	12,87	1,43E-06	19,53	3,18E-02	4,21	3,03E-02	3,09
gNCAbasic	2,00E+01	13,05	1,48E-06	19,94	3,18E-02	4,21	3,04E-02	3,10
gNCAreg	1,84E+01	12,62	1,36E-06	18,64	3,19E-02	4,20	3,02E-02	3,07
cgNCA	1,78E+01	12,01	1,24E-06	18,97	3,43E-02	4,54	3,00E-02	3,42

Tabla 23: Resumen de la reconstrucción. Red 5. Pruebas G1N3E3.
Fuente: Elaboración propia.

Analizando la matriz de datos, esta se presenta bastante normal. La entrada que concentra la mayor parte del error no supera el 11% de este, y como se ve, el problema no sólo se asocia al método que utiliza la matriz R sino que es más generalizado. Una hipótesis consiste en asumir que la forma de los datos (la distribución de la concentración de los errores principalmente) podría influir en los resultados, por lo que fueron generadas otras matrices promedios con el mismo error experimental. Se observa sin embargo que si bien algunas matrices de datos entregaban resultados de reconstrucción más acordes (menores o iguales al error de los datos), la reconstrucción no es tan confiable como en otros casos.

No se pudo llegar a una explicación satisfactoria, pero todo parece indicar que las características de la red producen sub-problemas mal condicionados (o que con mayor probabilidad son mal condicionados), y que dicho efecto se trasmite y amplifica en el desarrollo del algoritmo bi-alternado de optimización utilizado.

Red 6, 17 y 27: En la red 6 el error de los datos alcanza un 18,5% promedio, en donde una de las entradas concentra el 35% del error total. Los resultados de la reconstrucción son los siguientes, donde en un principio se confunden con el efecto discutido en le red 5 al observar la reconstrucción de A.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	5,21E-03	21,69	1,66E-06	4,50	4,21E-02	13,41	3,04E-02	4,53
gNCAbasic	5,21E-03	21,69	1,67E-06	4,50	4,21E-02	13,40	3,04E-02	4,53
gNCAreg	5,25E-03	23,80	1,72E-06	4,45	4,22E-02	13,35	3,02E-02	4,42
cgNCA	3,07E-03	22,33	2,33E-06	5,11	4,50E-02	13,79	3,50E-02	4,92

Tabla 24: Resumen de la reconstrucción. Red 6. Pruebas G1N3E3.
Fuente: Elaboración propia.

Sin embargo, la explicación en este caso es más sencilla, y tiene que ver con el comentario final realizado para la red 2 de esta sección. Analizando la reconstrucción de la matriz A, se puede apreciar que una sola coordenada (una con bajo valor numérico real y reconstruido) concentra el 90% del error de la reconstrucción. De hecho, eliminando dicha entrada, el error disminuye a un 2,2%.

El caso de la red 17 y 27 es análogo.

Red 18: En la siguiente red se observa un fenómeno bastante anormal respecto a los resultados obtenidos. Como se aprecia en la Tabla 25, el error adicionado a las matrices reconstruidas al usar el método con regularización es en extremo grande. Más aun, al comparar las características de la reconstrucción de los métodos restantes, se aprecia un ajuste bastante bueno. La matriz de datos no presenta características del todo anormales (concentración elevada de error en una sola componente por ejemplo), por lo que inequívocamente el problema se encuentra en el parámetro de regularización. En este caso particular las coordenadas de la matriz P no se encuentran cerca de cero, por lo que la distorsión producida por el término de regularización es relevante. Es algo a tener en cuenta al usar este método

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	4,10E-04	4,00	3,68E-07	2,21	1,56E-02	7,27	1,38E-02	3,36
gNCAbasic	4,10E-04	4,01	3,69E-07	2,22	1,56E-02	7,27	1,38E-02	3,36
gNCAreg	2,12E-01	218,87	3,34E-03	178,56	6,03E-02	21,99	6,04E-02	19,79
cgNCA	3,64E-04	3,48	3,40E-07	2,46	1,73E-02	7,19	1,54E-02	3,89

Tabla 25: Resumen de la reconstrucción. Red 18. Pruebas G1N3E3.
Fuente: Elaboración propia.

Experiencia 4: Errores anormales

En este caso se supone que en una red de tamaño medio (19x5x6x77), una de las mediciones presenta un error anormal en una de las coordenadas (más allá del error experimental que proviene de una misma distribución para todos los datos). Se trabaja con 2 matrices promedios de datos: la primera (E_{p1}) proveniente de 2 mediciones con errores experimentales como antes, mientras que en la segunda (E_{p2}), una de las 2 mediciones es la que contiene el error convencional, y en la otra se ha adicionado de manera artificial un error anormal en el experimento 6[65]. Esto por supuesto se verá reflejado en la varianza de esta segunda matriz promedio. Sin embargo, en una experiencia real no se sabrá que dicho experimento presenta errores anormales, por lo que utilizar la información contenida en la matriz de varianza R se puede tornar importante.

Los resultados al utilizar la matriz promedio 1 son los siguientes, que no muestran mucha diferencia con los ya extensamente revisados.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	6,01E-02	5,21	2,48E-05	7,22	2,12E-02	8,33	1,69E-02	3,45
gNCAbasic	6,01E-02	5,21	2,47E-05	7,22	2,12E-02	8,33	1,69E-02	3,44
gNCAreg	6,18E-02	5,45	2,58E-05	7,98	2,12E-02	8,53	1,72E-02	3,90
cgNCA	7,02E-02	7,66	3,74E-05	11,85	3,57E-02	7,88	3,62E-02	7,35

Tabla 26: Resumen de la reconstrucción utilizando $\boldsymbol{E_{p1}}$. Pruebas G1N3E4.
Fuente: Elaboración propia.

Utilizando la segunda matriz de datos promedios los resultados obtenidos son los siguientes:

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,06E-01	18,80	3,41E-04	23,09	1,74E-01	31,01	2,31E-01	22,29
gNCAbasic	1,06E-01	18,79	3,41E-04	23,14	1,74E-01	30,99	2,31E-01	22,29
gNCAreg	1,10E-01	18,57	3,42E-04	22,78	1,74E-01	31,45	2,31E-01	22,79
cgNCA	6,20E-02	6,32	1,77E-04	10,39	2,37E-01	32,31	1,76E-01	9,92

Tabla 27: Resumen de la reconstrucción utilizando $\boldsymbol{E_{p2}}$. Pruebas G1N3E4.
Fuente: Elaboración propia.

Es interesante analizar, que si bien la segunda matiz de datos presenta un error del 32% (casi 6 veces la de la primera), los resultados de la reconstrucción son mejores relativamente hablando. Es cierto que fueron superiores (aunque levemente) a los obtenidos con los restantes métodos en la prueba anterior, pero hay que considerar que se está trabajando con datos con niveles de errores totalmente diferente. Incluso los resultados se mostraron más exactos que los obtenidos con cgNCA en la primera experiencia. Lo más interesante es que en estos casos, cgNCA se muestra superior a los otros métodos, al ofrecer una reconstrucción más precisa y acertada. De hecho, el error en A disminuye 3 veces, mientras que el de P a la mitad.

[65] Suponiendo, por ejemplo, que en dicho experimento hay problemas con la maquinaria.

Como se puede comprobar, cgNCA es capaz de utilizar la varianza de los datos para dar mayor peso en la reconstrucción a aquellos daros que se midieron de manera más exacta. Si bien cuando todas las entradas en promedio reciben el mismo error, el método no aporta mucho, es interesante su potencial a la hora de trabajar con errores anormales en algunas mediciones.

3.16.4 Nivel 4

El objetivo de este cuarto nivel de pruebas es analizar el funcionamiento de los métodos que utilizan suposiciones y la confiabilidad asignada a éstas en la reconstrucción de los parámetros. Como ya se mencionó en el punto 3.10 Regularización general. Suposiciones a priori en $\boldsymbol{A}$ y $\boldsymbol{P}$, *l*a motivación es simple; se quiere generar un método que permita incluir información a priori respecto al valor y distribución de los parámetros, con el fin de utilizar dicha información en conjunto a la entregada por los datos, para reconstruir los parámetros de interés. Con este fin se definen matrices de suposiciones de las matrices A y P, que en cada coordenada poseen la suposición de la coordenada respectiva de la matriz a reconstruir. Dada la forma del funcional, el método de regularización propuesto con anterioridad es un caso particular del método más general propuesto aquí, en donde la suposición para la matriz P corresponde a 0 en todas las coordenadas.

De la misma manera, es necesario calibrar en cierta medida los parámetros λ_P y λ_A, que de acuerdo a la Ecuación 53 indicará el peso relativo o importancia dada a mantener la reconstrucción en un punto cercano a su suposición. En adelante se utilizará $\lambda_P = \lambda_A = 0{,}1$, de la misma manera que en los experimentos de regularización.

Experiencia 1: Una prueba simple de suposiciones en $\boldsymbol{A}$

En esta experiencia se trabaja con una red de tamaño medio (22x5x6) y varias opciones de suposiciones para la matriz A:

- A_{br}: Suposición igual a la matriz A real.
- A_b: Suposición con entradas cercanas a la reales, pero alteradas.
- A_{bn}: Lo mismo que la matriz A_b pero normalizando las suposiciones.

Una pregunta que surge inmediatamente al trabajar con este tipo de técnicas, es si es necesario generar suposiciones para todas las entradas (que en ocasiones pueden ser muchas) y a la forma de entregar dicha información (normalizada por ejemplo). Respecto a la primera pregunta, el método propuesto, efectivamente necesita que sean ingresadas suposiciones en todas las entradas a reconstruir. Más adelante se verá una forma de relajar ésto utilizando la confiabilidad de las suposiciones.

Tal cual al comparar la matriz A o P reconstruida con sus valores reales, es posible obtener el error promedio de suposición, definido como el error porcentual medio entre la suposición y el valor real de

los parámetros[66]. En base a eso, se adelanta que dicho error influirá directamente en los resultados obtenidos. Dada una matriz de datos medida, NCAbasic y los demás métodos tendrán un cierto nivel de ajuste en A y P, que podría en ciertas situaciones ser mejorado por el método de regularización o el de confiabilidad en suposiciones. Si el error de suposición en alguna de las matrices es superior al error de ajuste de los otros métodos utilizando solo la información de los datos, es presumible que el error de la reconstrucción aumentará al utilizar agNCAreg. Al contrario, si dicho error es menor, los resultados deberían mejorar. Se verá esto en el desarrollo de las experiencias con varios ejemplos.

En la siguiente tabla se resume el resultado de la reconstrucción con los diferentes métodos utilizando una matriz de datos con un error elevado (97,9%) y la matriz de suposición para A igual a A_{br}. En el caso de P se utilizará como suposición la matriz de ceros, reproduciendo el efecto de regularización.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,17E-03	27,64	1,34E-05	7,95	6,26E-02	33,68	4,68E-02	18,08
gNCAbasic	2,39E-03	27,88	1,43E-05	8,20	6,26E-02	33,69	4,70E-02	18,14
gNCAreg	2,72E-03	25,99	1,51E-05	8,66	6,27E-02	33,69	4,70E-02	18,31
agNCAreg	1,19E-04	6,94	3,94E-06	4,35	6,92E-02	35,26	3,48E-02	7,00

Tabla 28: Resumen de la reconstrucción utilizando A_{br}. Pruebas G1N4E1.
Fuente: Elaboración propia.

Como se puede observar, y dada la magnitud del error en la matriz de datos, la reconstrucción con los 3 métodos convencionales entrega errores considerables (aunque relativamente mucho menores a los de los datos), y con un comportamiento de los métodos similar al ya identificado y discutido. En el caso de agNCAreg se aprecia el efecto de la suposición de los parámetros en A. El error en dicha matriz, e incluso en la matriz P disminuyó considerablemente, por lo que el método efectivamente forzó a los resultados a permanecer cerca de la suposición establecida. No obstante este es el caso ideal (cuando se conoce con certeza los valores de la matriz real), y válido sólo para ejemplificar la utilidad del método.

Es posible apreciar también que pese a tener una suposición perfecta, el error en la reconstrucción no desaparece del todo. Esto es lógico, y se debe a que los estimadores de los parámetros son una combinación de los datos y de la suposición, por ende el error residual es asociado al correspondiente a los datos.

Otro punto de interés se obtiene al comparar los resultados de gNCAreg y agNCAreg, en donde debido a que la suposición para P es la matriz nula, se reproduce el efecto de regularización de gNCAreg. No obstante los resultados de P reconstruido difieren, y son mejor utilizando agNCAreg. La explicación de este fenómeno tiene que ver con la geometría del problema, y de los valores reales de la red, por lo que no es algo que pueda ser testeado ex-ante. Específicamente, el método de resolución converge alternadamente hacia el mínimo global, por lo que en cierto sentido al mejorar el valor reconstruido de la matriz A, esto afectará también a P que se mueve en conjunto con A al punto de convergencia. De qué forma lo afectará, dependerá de las características del problema.

[66] Por supuesto este error de suposición es solo concebible en un marco de pruebas sintéticas.

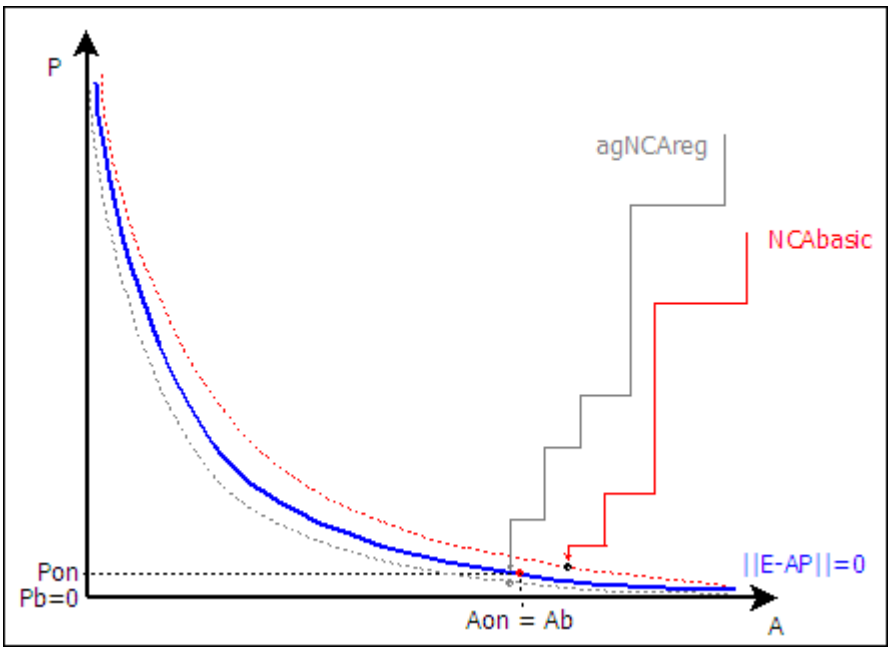


Figura 37: Esquema gráfico de convergencia utilizando suposiciones para los datos. Pruebas G1N4E1.
Fuente: Elaboración propia.

Gráficamente lo explicado se resume en la figura anterior. La línea roja representa la curva de iso-error a la cual converge NCAbasic para una matriz de datos fija. Como se ve, dicho valor está lejos de los valores reales (el punto rojo sobre la línea azul de iso-error cero). agNCAreg en cambio obliga a la reconstrucción de A a acercarse a su valor real (la suposición en el caso anterior), convergiendo al punto gris sobre la línea del mismo color. Como se ve, y dada la geometría del problema, en este caso el P reconstruido también varía, acercándose al valor real y disminuyendo su error. No obstante, podría ser posible también que se alejara del valor real al intentar posicionar a A cerca la suposición.

El análisis anterior se repite ahora considerando la misma matriz de datos, pero utilizando las suposiciones para A dadas por A_b y A_{bn}. Dichas matrices poseen un error de suposición del 31,8% y 27,2% respectivamente. Los resultados se resumen en las siguientes tablas.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,39E-03	27,88	1,43E-05	8,20	6,26E-02	33,69	4,70E-02	18,14
gNCAbasic	2,08E-03	27,58	1,31E-05	7,90	6,26E-02	33,67	4,68E-02	18,07
gNCAreg	2,71E-03	25,99	1,51E-05	8,66	6,27E-02	33,69	4,70E-02	18,31
agNCAreg	1,18E-01	23,88	7,51E-05	11,67	3,21E-01	52,71	3,11E-01	27,52

Tabla 29: Resumen de la reconstrucción utilizando A_b. Pruebas G1N4E1.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,36E-03	27,83	1,41E-05	8,15	6,26E-02	33,69	4,69E-02	18,13
gNCAbasic	2,37E-03	27,84	1,42E-05	8,15	6,26E-02	33,69	4,69E-02	18,13
gNCAreg	1,80E-03	26,14	1,17E-05	7,71	6,27E-02	33,68	4,66E-02	18,05
agNCAreg	1,07E-01	29,12	7,79E-05	11,69	3,02E-01	52,39	2,90E-01	27,12

Tabla 30: Resumen de la reconstrucción utilizando $\boldsymbol{A_{bn}}$. Pruebas G1N4E1.
Fuente: Elaboración propia.

Los resultados saltan a la vista, en relación a las diferencias obtenidas con una suposición perfecta. Utilizando A_b el error de ajuste de A disminuye, pero en contraste el de P aumenta en la misma proporción. El caso con A_{bn}, pese a tener un error menor de suposición que su versión normalizada, presenta un error más considerable en el ajuste de A. La razón no es clara, pero al parecer se debe a suposiciones particulares realizadas en A_{bn} (que pueden ser en parte diferentes a las de A_b al haberla normalizado). Si se analiza dicha matriz y los resultados de la reconstrucción en A, se ve que A_{bn} pese a tener un error global menor que A_b presenta entradas particulares con un error porcentual mayor de suposición (aunque obviamente tendrá otras con menor error). Luego, son estas entradas las que distorsionan los resultados de la reconstrucción, concentrando un mayor error en dichas entradas. De hecho, en la matriz A reconstruida utilizando A_b como suposición, la entrada con mayor error presenta un 144% de desajuste, concentrando el 19% del error, mientras que utilizando A_{bn} el error en dicha entrada aumenta 291%, concentrando el 33,3% del error total de ajuste. Obviando esta entrada, el error global disminuye a 20%, por lo que claramente dicho componente está distorsionando los resultados. Luego, es un punto a considerar al utilizar matrices de suposiciones en las reconstrucciones. Como comentario final, hay que considerar también que en la experiencia anterior, la exactitud de las suposiciones tiene un nivel de error similar al obtenido en las reconstrucciones con los métodos estándar.

Experiencia 2: Otro análisis con una red más pequeña

En la siguiente prueba se trabaja con una red pequeña (9x3x4x80) con el fin de analizar de mejor manera el efecto de las suposiciones. La red es la presentada en las siguientes matrices.

Aon =

0	0	0,12
0	0	-0,43
0	1,17	0
0	0,31	0
0	1,51	0
0	0	1,01
0,60	0	0
1,40	0	0
0	0	3,30

Er =

0,0064	0,0128	0,0117	-0,0019
-0,0227	-0,0455	-0,0414	0,0066
-0,0138	-0,2343	-0,0593	-0,0725
-0,0037	-0,0630	-0,0160	-0,0195
-0,0179	-0,3031	-0,0767	-0,0938
0,0532	0,1066	0,0970	-0,0155
0,0046	-0,0368	-0,0038	-0,0168
0,0108	-0,0857	-0,0089	-0,0392
0,1748	0,3502	0,3187	-0,0510

Pon =

0,0077	-0,0613	-0,0063	-0,0280
-0,0118	-0,2001	-0,0507	-0,0619
0,0529	0,1060	0,0965	-0,0154

Se trabaja además con 2 matrices de datos medidos, una con un error del 29,9% (E_{m1}) y otra con un error mayor de 61,8% (E_{m2}). Respecto a la suposiciones, se utilizan 3 diferentes como en el caso anterior: una correspondiente al valor real de los parámetros (A_{br}), otra una suposición con cierto grado de error (16,8%) (A_b), y la tercera su versión normalizada (error del 13,9%) (A_{bn}).

En primer lugar se presentan los resultados de la reconstrucción utilizando la matriz de suposición A_{br}, en conjunto con los datos E_{m1} y E_{m2}.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAreg	2,27E-03	7,70	9,27E-06	12,37	1,95E-02	68,77	2,66E-02	14,28
agNCAreg	9,12E-05	1,51	1,36E-05	13,35	2,69E-02	49,71	2,50E-02	10,54

Tabla 31: Resumen de la reconstrucción utilizando $\boldsymbol{A_{br}}$ y $\boldsymbol{E_{m1}}$. Pruebas G1N4E2.
Fuente: Elaboración propia

.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAreg	1,68E-02	29,02	1,65E-05	20,75	3,36E-02	39,76	5,49E-02	40,78
agNCAreg	4,52E-04	5,10	1,70E-05	19,69	5,41E-02	38,87	3,10E-02	19,91

Tabla 32: Resumen de la reconstrucción utilizando $\boldsymbol{A_{br}}$ y $\boldsymbol{E_{m2}}$. Pruebas G1N4E2.
Fuente: Elaboración propia.

De la Tabla 31 se puede observar que el error de ajuste disminuye considerablemente en la matriz A, y aumenta levemente en la matriz P al usar datos con un caso 30% de error. De la misma manera, el ajuste a la matriz de datos real mejora en casi 4 puntos porcentuales. El aumento identificado en P tiene relación con la explicación ya entregada en la sección anterior, y la convergencia conjunta de ambas variables, lo que en ocasiones puede distorsionar un elemento al tratar de mantener el otro cerca de la suposición.

Al utilizar la matriz de datos con un error mayor (61,8%), el efecto es similar. Dada la magnitud del error de los datos, NCAbasic por ejemplo, comete un 29% de error al reconstruir la matriz de conexiones. Utilizando la suposición, dicho error disminuye a un 5%. Incluso la reconstrucción de P mejora levemente en este caso.

Las siguientes tablas describen la misma información anterior, pero esta vez utilizando A_b y A_{bn}, y nuevamente ambas matrices de datos.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAreg	2,26E-03	7,69	9,97E-06	12,09	1,94E-02	55,52	2,71E-02	14,01
agNCAreg	1,08E-02	9,92	2,54E-05	14,99	6,36E-02	47,60	6,13E-02	16,11

Tabla 33: Resumen de la reconstrucción utilizando $\boldsymbol{A_b}$ y $\boldsymbol{E_{m1}}$. Pruebas G1N4E2.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAreg	1,68E-02	29,02	1,65E-05	20,92	3,36E-02	40,50	5,49E-02	40,93
agNCAreg	1,12E-02	14,32	2,89E-05	20,14	7,91E-02	39,10	6,40E-02	24,61

Tabla 34: Resumen de la reconstrucción utilizando $\boldsymbol{A_b}$ y $\boldsymbol{E_{m2}}$. Pruebas G1N4E2.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAreg	2,26E-03	7,68	9,47E-06	11,95	1,95E-02	52,03	2,66E-02	13,95
agNCAreg	1,02E-02	9,59	2,56E-05	14,97	6,14E-02	47,54	5,92E-02	15,83

Tabla 35: Resumen de la reconstrucción utilizando $\boldsymbol{A_{bn}}$ y $\boldsymbol{E_{m1}}$. Pruebas G1N4E2.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAreg	1,68E-02	29,02	1,65E-05	20,77	3,36E-02	39,92	5,49E-02	40,79
agNCAreg	1,04E-02	13,23	2,99E-05	20,18	7,77E-02	39,21	6,16E-02	23,48

Tabla 36: Resumen de la reconstrucción utilizando $\boldsymbol{A_{bn}}$ y $\boldsymbol{m}$. Pruebas G1N4E2.
Fuente: Elaboración propia.

Es interesante analizar los resultados de las tablas anteriores, que concuerdan con la teoría y las conclusiones realizadas previas al análisis experimental. Como se ve, al utilizar la matriz de datos con menor error, las reconstrucciones con los métodos convencionales rondan el 7,7% (Tabla 33 y Tabla 35). Luego, al utilizar una suposición como A_b o A_{bn} cuyos errores de suposición rondan el 15% (mayor al 7,7% de reconstrucción básica), el efecto será simplemente distorsionar aún más los datos, como se observa en el aumento del error de ajuste de ambas matrices. Al utilizar en cambio E_{m2} que presenta un error mayor en los datos, la reconstrucción con los métodos básicos ronda el 29%. Luego, las

suposiciones presentan un menor error, por lo que debiera mejorar la reconstrucción al utilizar agNCAreg. Esto es efectivo, y en ambas experiencias se observa que los errores disminuyen considerablemente. En este caso, contrario al de la experiencia 1, la matriz normalizada de suposiciones se comportó levemente mejor que la no normalizada.

Es necesario destacar que el análisis anterior de comparación de errores de reconstrucción y de suposiciones, sólo se puede realizar en estas experiencias artificiales, ya que sería imposible en redes reales. La lección más bien va por el lado de tener cuidado a la hora de usar esta herramienta, en el sentido de su gran sensibilidad a la confianza o exactitud que tenga la información a priori utilizada.

Experiencia 3: Suposiciones en A y P.

En esta experiencia se utiliza la misma red anterior, con el fin de visualizar el efecto de imponer también suposiciones en la matriz P. Para esto se utiliza la misma matriz de suposiciones A_b de la experiencia anterior y 3 matrices de suposiciones para la matriz de señales de regulación: Las 2 primeras con suposiciones tan solo en algunas entradas, y la tercera con suposiciones en todas ellas. Esto es posible de realizar (al contrario del caso de A) debido a que por construcción, los valores de P están ubicados cerca de cero, por lo que no imponer una suposición y dejarla por defecto igual a cero, es simplemente recobrar el efecto de regularización en dichos parámetros.

P_{b1} posee un error de suposición de 53%, P_{b2} uno del 67%, mientras que P_{b3} será igual a la matriz P real. Los resultados se resumen en las 6 tablas siguientes, en donde para cada matriz de suposición se utilizó la matriz de datos E_{m1} y luego E_{m2}, definidas de la misma manera que en la experiencia anterior. Todas estas pruebas utilizan además la matriz A_b del punto anterior.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAreg	2,26E-03	7,68	9,41E-06	11,93	1,95E-02	51,61	2,66E-02	13,94
agNCAreg	1,07E-02	9,90	2,02E-05	14,45	6,31E-02	57,57	6,02E-02	15,63

Tabla 37: Resumen de la reconstrucción utilizando $\boldsymbol{P_{b1}}$ y $\boldsymbol{E_{m1}}$. Pruebas G1N4E3.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAreg	1,68E-02	29,02	1,65E-05	20,75	3,36E-02	39,81	5,49E-02	40,77
agNCAreg	1,11E-02	14,31	2,66E-05	20,12	7,87E-02	38,98	6,38E-02	24,96

Tabla 38: Resumen de la reconstrucción utilizando $\boldsymbol{P_{b1}}$ y $\boldsymbol{E_{m2}}$. Pruebas G1N4E3.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAreg	2,26E-03	7,68	9,50E-06	11,93	1,95E-02	51,60	2,67E-02	13,94
agNCAreg	1,07E-02	9,87	1,92E-05	14,59	6,25E-02	47,51	5,98E-02	15,88

Tabla 39: Resumen de la reconstrucción utilizando $\boldsymbol{P_{b2}}$ y $\boldsymbol{E_{m1}}$. Pruebas G1N4E3.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAreg	1,68E-02	29,02	1,65E-05	20,77	3,36E-02	39,87	5,49E-02	40,79
agNCAreg	1,11E-02	14,32	1,93E-05	19,73	7,83E-02	39,08	6,14E-02	24,56

Tabla 40: Resumen de la reconstrucción utilizando $\boldsymbol{P_{b2}}$ y $\boldsymbol{E_{m2}}$. Pruebas G1N4E3.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAreg	2,26E-03	7,68	9,48E-06	11,95	1,95E-02	51,98	2,66E-02	13,95
agNCAreg	1,06E-02	9,85	1,43E-05	13,56	6,23E-02	58,48	5,89E-02	15,03

Tabla 41: Resumen de la reconstrucción utilizando $\boldsymbol{P_{br}}$ y $\boldsymbol{E_{m1}}$. Pruebas G1N4E3.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAreg	1,68E-02	29,02	1,65E-05	20,81	3,36E-02	40,07	5,49E-02	40,83
agNCAreg	1,11E-02	14,29	1,75E-05	19,09	7,80E-02	35,82	6,15E-02	24,25

Tabla 42: Resumen de la reconstrucción utilizando $\boldsymbol{P_{br}}$ y $\boldsymbol{E_{m2}}$. Pruebas G1N4E3.
Fuente: Elaboración propia.

Como se observa, los resultados no difieren significativamente de los obtenidos utilizando una suposición de P basada en la matriz de ceros, lo que en parte parece abalar el uso del método de regularización si se asume que efectivamente la hipótesis de que las coordenadas de P son cercanas a cero es válida. Numéricamente no es demasiado el error que se comete.

Experiencia 4: Confiabilidad suposiciones

En esta experiencia se testea el funcionamiento del método acgNCAreg, que permite incluir en el procedimiento información respecto a la confiabilidad de las suposiciones propuestas para los parámetros. Como ya se ha comentado, dicha confiabilidad viene dada por la varianza de cada suposición, asumiendo que la misma se distribuye según una función normal con media igual a la estimación, y una varianza calculada de acuerdo a la Ecuación 59 vía las matrices de flexibilidad de las suposiciones, F_A y F_P.

La prueba se realiza utilizando la misma red que las pruebas anteriores, una suposición por defecto en P y la siguiente matriz de suposición A_b y su correspondiente matriz de flexibilidad F_A.

Ab =

0	0	0,12
0	0	-0,50
0	1,30	0
0	0,30	0
0	1,00	0
0	0	4,00
0,55	0	0
0,60	0	0
0	0	2,50

F_A =

0	0	3
0	0	5
0	10	0
0	2	0
0	20	0
0	0	300
5	0	0
100	0	0
0	0	50

Se recuerda que la flexibilidad de cada suposición, corresponde al intervalo centrado en dicho valor en el cual el 95% de las veces es posible encontrar el verdadero valor. Así flexibilidades bajas harán muy pequeño dicho intervalo, asociando dicha suposición a una con gran certeza, y por ende pequeña varianza. Como se observa, la suposición con valor 0,12 presenta una flexibilidad de 3%, por lo que con probabilidad 95% es posible encontrar el valor en el intervalo $0{,}12 \pm 3\%$. La matriz de varianza de la suposición en A asociada a dichas flexibilidades es la siguiente:

D =

0	0	0,002
0	0	0,013
0	0,066	0
0	0,003	0
0	0,102	0
0	0	6,122
0,014	0	0
0,306	0	0
0	0	0,638

Como se observa, la suposición con menor flexibilidad (mayor certeza o menor varianza), presenta una varianza 4 órdenes de magnitud inferior a la con menor certeza.

Se comienza testeando el funcionamiento de los 5 métodos, utilizando la suposición anterior para A y como matriz D la matriz de unos. De esta forma, se reproduce el resultado de cgNCAreg, al dar a todas las entradas el mismo peso en la minimización. Como antes, se presentan los resultados utilizando E_{m1}, con un error de 29,9%, y luego utilizando los datos dados por E_{m2} con un error del 61,8%. Los resultados se resumen en las 2 tablas siguientes.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAreg	2,26E-03	7,69	9,97E-06	12,09	1,94E-02	55,52	2,71E-02	14,01
agNCAreg	5,41E-01	41,29	4,45E-05	16,62	3,32E-01	84,16	3,34E-01	41,90
acgNCAreg	5,41E-01	41,29	4,45E-05	16,62	3,32E-01	84,16	3,34E-01	41,90

Tabla 43: Resumen de la reconstrucción utilizando $\boldsymbol{A_b}$ y $\boldsymbol{E_{m1}}$. Pruebas G1N4E4.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAreg	1,68E-02	29,02	1,65E-05	20,72	3,36E-02	39,64	5,49E-02	40,75
agNCAreg	5,37E-01	38,68	6,38E-05	22,51	3,55E-01	68,14	3,34E-01	44,93
acgNCAreg	5,37E-01	38,68	6,38E-05	22,51	3,55E-01	68,14	3,34E-01	44,93

Tabla 44: Resumen de la reconstrucción utilizando $\boldsymbol{A_b}$ y $\boldsymbol{E_{m2}}$. Pruebas G1N4E4.
Fuente: Elaboración propia.

Lo primero que se observa es que acgNCAreg reproduce los resultados de agNCAreg si se le asigna peso unitario a todas las suposiciones. Más precisamente, si se les asignan a todas el mismo peso relativo. Además como se aprecia, el efecto de la suposición, con ambas matrices de datos, aumenta ampliamente el error de ajuste en comparación a los métodos normales. El error de suposición de A_b es de un 50%, por lo que no es sorprendente dicho efecto dado que los métodos convencionales tienen de por si un erro de ajuste menor. Además, y como es presumible, analizando la reconstrucción se puede apreciar que es una de las entradas la que concentra la mayor parte de dicho error de ajuste. Finalmente es necesario destacar un efecto interesante de los datos anteriores. La Tabla 43 resume los resultados para una matriz de datos con un error inferior a la usada en la reconstrucción resumida en la Tabla 44, por lo que los errores en ésta debieran aumentar. Y es precisamente lo que pasa con NCAbasic y los demás métodos similares. El error de ajuste de A por ejemplo sube de aproximadamente un 7,7% a un 29% por efecto de datos de menor calidad. Sin embargo el ajuste vía agNCA disminuye, lo que demuestra una vez más el efecto de las suposiciones, y su funcionamiento dependiente de la magnitud relativa entre el error de suposición y el de ajuste a los datos.

Las siguientes tablas resumen los resultados bajo las mismas condiciones, pero utilizando la matriz de varianzas de las suposiciones antes descrita.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAreg	2,27E-03	7,70	9,27E-06	12,37	1,95E-02	68,77	2,66E-02	14,28
agNCAreg	5,41E-01	41,29	4,45E-05	16,62	3,32E-01	84,16	3,34E-01	41,90
acgNCAreg	3,54E-01	35,47	2,78E-05	15,23	2,74E-01	79,20	2,76E-01	36,35

Tabla 45: Resumen de la reconstrucción utilizando $\boldsymbol{A_b}$, $\boldsymbol{E_{m1}}$ y $\boldsymbol{D}$. Pruebas G1N4E4.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAreg	1,68E-02	29,02	1,65E-05	20,75	3,36E-02	39,76	5,49E-02	40,78
agNCAreg	5,37E-01	38,68	6,38E-05	22,51	3,55E-01	68,14	3,34E-01	44,93
acgNCAreg	3,35E-01	34,55	3,58E-05	21,08	2,96E-01	65,93	2,70E-01	41,16

Tabla 46: Resumen de la reconstrucción utilizando $\boldsymbol{A_b}$, $\boldsymbol{E_{m2}}$ y $\boldsymbol{D}$. Pruebas G1N4E4.
Fuente: Elaboración propia.

Como se observa, en ambos casos el error disminuye al dar más importancia en la función de ajuste a aquellas suposiciones mejor medidas. En efecto, el error en la matriz de datos con menor error disminuye a un 35%, y la segunda a casi un 34%. Sin embargo, las magnitudes de los errores aún son elevadas, producto de la gran distorsión e inexactitud que posee la suposición para A.

Es posible analizar las matrices reconstruidas vía cgNCAreg y acgNCAreg, y compararlas componente a componente para la matriz de datos E_{m1}. Como se aprecia de la matriz de flexibilidad de la suposición para la matriz A, 4 componentes presentan una baja flexibilidad (y por ende se asocian a buenas suposiciones). Relativo al nivel de error de ajuste que dichas coordenadas obtuvieron utilizando cgNCAreg, 3 de las 4 disminuyeron su error porcentual, lo que es de esperar dada las características del método, y también porque se forzará de sobremanera a mantener las coordenadas de la reconstrucción cerca de esa buena suposición. Lo extraño y destacable es que las suposiciones con mayor flexibilidad (y por ende menor confiabilidad) también disminuyeron el error. De las 3 con mayores flexibilidades, 2 presentaron un mejor ajuste. Este efecto es algo nuevo, y no esperado en la definición del método original.

Si se analiza con calma, la explicación sigue la misma lógica de las demás experiencias, y se relaciona con el bajo error de los datos para E_{m1}. Como se observa, NCAbasic es capaz de reconstruir la matriz A con un error promedio del 7%, por lo que es capaz de obtener una buena reconstrucción dado el error de los

datos. Luego, el 41% de error obtenido con cgNCAreg se explica en su mayor parte por la distorsión ocasionada por la mala suposición, y el hecho de asignarle igual peso a todas ellas. Luego, el uso de la matriz D tiene 2 efectos: por una parte, la alta confiabilidad de algunas entradas obliga a las reconstrucciones a permanecer cerca de esos valores, porque si dichos valores a priori eran efectivamente correctos (o al menos realmente tan confiables como el parámetro de flexibilidad indica), el error de dichas entradas debiera disminuir. Pero por otra parte, las entradas con más flexibilidad (al importar menos relativamente en la función a minimizar) son dominadas por el efecto de reconstrucción de los datos, y por ende dejan de estar sujetas al efecto distorsionador de las suposiciones inexactas. Luego es predecible que, al menos en este caso donde el error de los datos no se extiende en demasía a la reconstrucción, que dichas entradas también disminuyan su error de ajuste.

En la Tabla 46 se presentan los mismos resultados discutidos anteriormente para la matriz de datos E_{m2}. Se observa un efecto similar, sólo que en este caso, y dado que los datos son de peor calidad, las entradas con alta flexibilidad pasarán a estar dominadas por un efecto que las desajusta tanto o más que el efecto de la suposición errónea. A pesar de esto se hace notar que ambos métodos de suposición disminuyeron su ajuste en A (aunque levemente), en relación al primer caso. Y por supuesto, aun mantenerlo es de gran relevancia dado el gran aumento de error en la medición de E.

En relación a la discusión anterior, se realiza una nueva reconstrucción con otra matriz de flexibilidad de las suposiciones de A, que exagera la flexibilidad de las entradas más imprecisas. De esta manera se quiere demostrar el segundo efecto antes descrito, que se piensa contribuirá a mejorar aún más la reconstrucción de las entradas en las que no se tenga una buena suposición. Específicamente, las flexibilidades de 2 "malas" suposiciones fueron aumentadas en un orden de magnitud. Los resultados se resumen en la tabla siguiente, utilizando la matriz de datos E_{m1} primero y luego E_{m2}.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAbasic	2,26E-03	7,69	9,79E-06	12,01	1,94E-02	53,57	2,69E-02	13,97
gNCAreg	2,26E-03	7,69	9,97E-06	12,09	1,94E-02	55,52	2,71E-02	14,01
agNCAreg	5,41E-01	41,29	4,45E-05	16,62	3,32E-01	84,16	3,34E-01	41,90
acgNCAreg	5,64E-02	20,83	3,21E-05	15,23	1,08E-01	64,56	1,15E-01	20,54

Tabla 47: Resumen de la reconstrucción utilizando $\boldsymbol{A_b}$, $\boldsymbol{E_{m1}}$ y $\boldsymbol{D}$ modificada. Pruebas G1N4E4.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAbasic	1,68E-02	29,03	1,65E-05	21,14	3,36E-02	41,11	5,49E-02	41,18
gNCAreg	1,68E-02	29,02	1,65E-05	20,75	3,36E-02	39,76	5,49E-02	40,78
agNCAreg	5,37E-01	38,68	6,38E-05	22,51	3,55E-01	68,14	3,34E-01	44,93
acgNCAreg	6,29E-02	23,04	4,49E-05	20,80	1,30E-01	50,52	1,19E-01	28,06

Tabla 48: Resumen de la reconstrucción utilizando $\boldsymbol{A_b}$, $\boldsymbol{E_{m2}}$ y $\boldsymbol{D}$ modificada. Pruebas G1N4E4.
Fuente: Elaboración propia.

Los resultados son claros, el efecto de la mayor flexibilidad elimina la exigencia a dichas entradas de permanecer cerca de su suposición (disminuye el castigo en la función) y por ende la reconstrucción vía los datos es la que domina. De todas formas, una cota inferior para el error que se puede alcanzar en dichas entradas es el error obtenido vía NCAbasic, valor al que se convergerá al aumentar la flexibilidad de dichas entradas. Se destaca además que una buena suposición puede superar dicha cota inferior al error, por lo que las potencialidades del método no dejan de ser interesantes.

Utilizando valores en extremo altos para las mismas entradas que antes, se comprueba que efectivamente el valor de dichas entradas converge al obtenido vía NCAbasic, pero no necesariamente el error global, que podría incluso aumentar levemente. Es necesario además destacar un punto interesante respecto a lo mismo, y que se pudo comprobar al utilizar diferentes redes (se omiten los resultados en esta ocasión por ser cualitativamente similares a los anteriores) y aumentar la flexibilidad de algunas suposiciones poco confiables. Si bien efectivamente los errores en dichas entradas disminuyeron, algunas entradas con buenas suposiciones aumentaron su error. Aun cuando no fueron todas, sí se apreció en entradas particulares, por lo que en algunas cosas el error global podría aumentar (pese a la disminución del error en las entradas restantes). Más aun, las entradas con baja flexibilidad que experimentaron variación fueron solamente las correspondientes a las mismas columnas de A a las que se les aumento la flexibilidad. El resto permaneció inalterado. Este efecto es extraño, y contrario a lo esperado dado el comportamiento iterativo y alternado del método de reconstrucción. Se comentará más sobre esto en otras experiencias.

Experiencia 5: Pruebas con una red de mayor tamaño

En este caso se busca trabajar con una red de mayor tamaño, con el fin de verificar los comportamientos identificados anteriormente. La red elegida fue de tamaño medio (22x5x6x80), y se utilizaron 2 matrices de datos, E_{m1} y E_{m2} con errores de 24,1% y 45% respectivamente. Nuevamente se trabaja únicamente con suposiciones en A, utilizando la suposición estándar para la matriz P. La matriz de conexiones real, la suposición y su matriz de flexibilidad asociada se resumen a continuación.

Aon =

0	0	0	9,526	0
0,030	0	0	0	0,488
0,984	0	0	0	0
0	-0,598	0,945	0	3,488
0	0	0	0,684	0,595
0	0	0	-8,298	0
0	0	0	0	1,897
0	2,447	3,359	0	0
0	0	2,807	-1,496	0
-0,209	0	0	0	0
0	0	-0,807	0	0
0	0	0,490	0,253	0
0	0,135	0	0	2,524
0	0,573	0	-1,974	5,034
1,854	2,162	0	0	0
1,962	1,282	0	2,624	0
0	0	-0,793	0	0
0	0	0	0	-2,641
0	0	0	6,680	0
0	0	0	0	-3,385
3,125	0	0	0	0
-0,747	0	0	0	0

F_Ab =

0	0	0	3	0
500	0	0	0	10
20	0	0	0	0
0	1000	2	0	40
0	0	0	3	10
0	0	0	20	0
0	0	0	0	5
0	10	5	0	0
0	0	20	10	0
1000	0	0	0	0
0	0	1000	0	0
0	0	10	5	0
0	100	0	0	30
0	30	0	3	10
4	60	0	0	0
60	5	0	40	0
0	0	400	0	0
0	0	0	0	15
0	0	0	1000	0
0	0	0	0	50
50	0	0	0	0
80	0	0	0	0

Ab =

0	0	0	11	0
1	0	0	0	0,5
1	0	0	0	1
0	-2	1	0	1
0	0	0	0,4	0,4
0	0	0	-10	0
0	0	0	0	2
0	2	5	0	0
0	0	3	-1	0
-0,209	0	0	0	0
0	0	2	0	0
0	0	0,5	0,4	0
0	1	0	0	5
0	0,6	0	-3	5
4	2	0	0	0
2	2	0	2	0
0	0	-2	0	0
0	0	0	0	3
0	0	0	5	0
0	0	0	0	-5
4	0	0	0	0
-5	0	0	0	0

Como se aprecia, la mayor parte de las combinaciones se rescatan en los resultados; en primer lugar existen suposiciones erróneas (la entrada (2,1) por ejemplo) que son identificadas como inexactas dada la gran flexibilidad que se les asocia. Por otra parte, existen buenas suposiciones que son asociadas como tal en relación a su baja flexibilidad. En contraparte, existen también suposiciones buenas que poseen una flexibilidad considerable, y suposiciones malas con grado de flexibilidad no tan alto como debería (de todas formas estos casos intermedios no son extremos, en el sentido de que no se exagera ni lo alto, ni lo bajo de las flexibilidades).

Las siguientes tablas resumen los resultados de la reconstrucción para E_{m1} y E_{m2}.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,09E-03	4,68	1,23E-06	2,06	2,57E-02	9,29	2,33E-02	5,21
gNCAbasic	1,09E-03	4,68	1,23E-06	2,04	2,57E-02	9,28	2,33E-02	5,24
gNCAreg	1,32E-03	4,73	1,44E-06	2,31	2,58E-02	9,22	2,33E-02	5,64
agNCAreg	7,97E-01	103,69	5,31E-04	72,55	6,90E-01	112,69	6,88E-01	184,72
acgNCAreg	7,50E-01	65,10	2,60E-04	64,67	4,36E-01	66,07	4,34E-01	126,93

Tabla 49: Resumen de la reconstrucción utilizando $\boldsymbol{E_{m1}}$. Pruebas G1N4E5.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,13E-01	22,29	2,09E-05	15,11	6,92E-02	58,44	7,22E-02	70,45
gNCAbasic	1,13E-01	22,29	2,09E-05	15,12	6,92E-02	58,44	7,22E-02	70,44
gNCAreg	1,03E-01	22,15	2,03E-05	15,08	6,92E-02	58,24	7,20E-02	69,16
agNCAreg	7,99E-01	104,82	4,79E-04	66,86	6,86E-01	148,70	6,90E-01	156,06
acgNCAreg	7,08E-01	70,16	2,41E-04	63,03	4,38E-01	101,39	4,38E-01	116,47

Tabla 50: Resumen de la reconstrucción utilizando $\boldsymbol{E_{m2}}$. Pruebas G1N4E5.
Fuente: Elaboración propia.

Una observación a realizar en este caso, es respecto al gran error global que posee la matriz de suposición de A, dado por una parte por la gran cantidad de entradas de esta (efecto que se exagera aún más al crecer la matriz) y a que en este caso las suposiciones fueron creadas de una manera más realista[67]. Esta es la razón de la magnitud de los errores de reconstrucción, al considerar que A_b tiene un erro de suposición de 188%. Obviando esto, se puede apreciar claramente el efecto en la reconstrucción del uso de la matriz D, donde el error, en comparación al obtenido con agNCAreg disminuye considerablemente. Sin embargo, analizando con más detenimiento las matrices reconstruidas, la realidad no es tan mala como parece. En todos los casos, más de la mitad del error de ajuste en A es explicado por una sola entrada, esta es la que posee valor numérico cercano a cero, en donde por un asunto de precisión se comete irremediablemente un mayor error. Además, dicha coordenada tiene una

[67] En el sentido que no se plantea tener una suposición acertada para cada componente.

suposición errada, lo que explica en parte el gran error obtenido por cgNCAreg, y la brusca disminución al utilizar acgNCAreg (si se observa, la flexibilidad asignada a dicha entrada es alta).

Como una manera de seguir testeando lo anterior, se utiliza otra matriz de flexibilidad en donde se exagera la asignada a la entrada con problemas, a fin de utilizar un en mayor medida la información de los datos y disminuir el error. El resultado es el siguiente para la matriz de datos E_{m1} :

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,09E-03	4,68	1,23E-06	2,06	2,57E-02	9,29	2,33E-02	5,21
gNCAbasic	1,09E-03	4,68	1,24E-06	2,03	2,57E-02	9,28	2,33E-02	5,30
gNCAreg	1,12E-03	4,65	1,19E-06	2,28	2,58E-02	9,18	2,32E-02	6,02
agNCAreg	7,97E-01	103,69	5,31E-04	72,55	6,90E-01	112,69	6,88E-01	184,72
acgNCAreg	7,45E-01	36,99	2,41E-04	63,71	4,32E-01	50,43	4,30E-01	121,52

Tabla 51: Resumen de la reconstrucción utilizando $\boldsymbol{E_{m1}}$, flexibilidad exagerada. Pruebas G1N4E5.
Fuente: Elaboración propia.

Verificando las matrices reconstruidas, se puede ver que el error en la entrada con problemas fue disminuido de más del 1000% a cerca de un 50%, lo que claramente se aprecia en el error global. Una vez más, tan solo la columna 1 (donde se varía la flexibilidad de una entrada) altera sus resultados respecto a lo obtenido con la matriz de confiabilidad original. De todas maneras los cambios no son significativos.

Experiencia 6: Imponer entradas solo en algunas entradas

La discusión anterior se resume en 2 puntos principalmente:

- La dificultad de generar suposiciones para la mayor parte de los valores de A (o P) sin comprometer en gran medida el error global de suposición de A_b.
- El uso de la flexibilidad como una suerte de control, que permita indicar al método cuando creer más a los datos o cuando utilizar más la suposición en la reconstrucción de las matrices.

De esta manera, es posible utilizar el método de una forma alternativa: imponer suposiciones sólo en algunas entradas de las matrices (suposiciones que con una probabilidad elevada estén bien), y dejar el resto de las suposiciones libres, asignándoles para ello una alta flexibilidad.

Para esta prueba se utiliza la misma red anterior, y una matriz de suposición que tenga suposiciones "decentes" en 3 coordenadas, y por ende una flexibilidad baja en la misma. El resto tendrá valores suposiciones malas, y una flexibilidad de 1.000. (3 órdenes de magnitud superior a la flexibilidad de las buenas suposiciones). Los resultados son los siguientes (utilizando E_m).

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,10E-03	4,69	1,24E-06	2,07	2,57E-02	9,29	2,33E-02	5,22
gNCAbasic	1,09E-03	4,69	1,24E-06	2,06	2,57E-02	9,29	2,33E-02	5,22
gNCAreg	1,32E-03	4,73	1,44E-06	2,30	2,58E-02	9,22	2,33E-02	5,64
agNCAreg	5,57E+00	160,56	8,26E-04	90,44	7,81E-01	91,32	7,84E-01	203,73
acgNCAreg	4,92E-02	19,29	1,63E-05	13,52	7,25E-02	18,97	7,21E-02	21,31

Tabla 52: Resumen de la reconstrucción. Flexibilidad 1.000. Pruebas G1N4E6.
Fuente: Elaboración propia.

Los resultados son bastante clarificadores. Con agNCAreg se comete un gran error de reconstrucción en ambas matrices, debido al elevado error de suposición incluido en A_b. Por supuesto la matriz P también se contagia con esto en el algoritmo. Utilizando la matriz D obtenida en base a la matriz de flexibilidad antes descrita, los resultados mejoran increíblemente. En este caso se combina lo bueno de ambos métodos: la suposición por una parte, que de ser adecuada obligará a las reconstrucciones a quedarse cerca de ellas, y la información de los datos por otra, que pese a tener errores es una mejor fuente de información que una mala suposición. Es posible mejorar aún más el resultado, utilizando flexibilidades crecientes que hagan converger las reconstrucciones de los parámetros libres a las obtenidas vía NCAbasic, que utiliza sólo la información de los datos.

Las siguientes tablas resumen los resultados al utilizar un factor 1'.000 en las entradas con malas suposiciones, luego de 1.000.000.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,10E-03	4,69	1,26E-06	2,05	2,57E-02	9,28	2,33E-02	5,29
gNCAbasic	1,09E-03	4,68	1,23E-06	2,05	2,57E-02	9,28	2,33E-02	5,24
gNCAreg	1,22E-03	4,73	1,35E-06	2,39	2,57E-02	9,30	2,36E-02	4,88
agNCAreg	5,57E+00	160,56	8,26E-04	90,44	7,81E-01	91,32	7,84E-01	203,73
acgNCAreg	1,59E-02	9,61	5,74E-06	4,84	4,02E-02	11,29	3,87E-02	8,29

Tabla 53: Resumen de la reconstrucción. Flexibilidad 10.000. Pruebas G1N4E6.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,10E-03	4,69	1,25E-06	2,05	2,57E-02	9,28	2,33E-02	5,28
gNCAbasic	1,09E-03	4,69	1,23E-06	2,06	2,57E-02	9,29	2,33E-02	5,21
gNCAreg	1,36E-03	4,75	1,45E-06	2,31	2,58E-02	9,23	2,34E-02	5,56
agNCAreg	5,57E+00	160,56	8,26E-04	90,44	7,81E-01	91,32	7,84E-01	203,73
acgNCAreg	2,44E-02	10,24	4,63E-06	4,17	4,26E-02	11,89	4,06E-02	8,17

Tabla 54: Resumen de la reconstrucción. Flexibilidad 1.000.000. Pruebas G1N4E6.
Fuente: Elaboración propia.

Como se observa, el ajuste mejora aún más. Sin embargo, al pasar de la flexibilidad 10.000 a 1.000.000, el error aumenta levemente en A. Si bien los valores con alta flexibilidad se acercan a los de NCAbasic, no llegan nunca a dicho punto, debido a la perturbación causada por las suposiciones. Luego, al aumentar aún más la flexibilidad de las suposiciones libres, llegará un punto en que el error no disminuirá más, si no que se observarán efectos como el anterior. Por supuesto, este es un caso extremo, en que la reconstrucción vía los métodos estándares es bastante buena. En la siguiente experiencia se repite lo anterior con una matriz de datos con errores más groseros.

Experiencia 7: Una matriz de datos con mayor error

Para este caso se repite el mismo experimento anterior, variando la flexibilidad de las entradas sin suposiciones en A_b, pero utilizando una matriz de datos con un error de la misma magnitud que el error de suposición. Los resultados, utilizando una flexibilidad de 1.000, 10.000 y 1.000.000 respectivamente son los siguientes.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	7,20E-02	28,76	8,18E-05	25,26	1,49E-01	47,69	1,51E-01	99,42
gNCAbasic	7,20E-02	28,76	8,19E-05	25,26	1,49E-01	47,69	1,51E-01	99,42
gNCAreg	7,62E-02	28,95	8,49E-05	24,86	1,49E-01	47,84	1,51E-01	100,17
agNCAreg	5,54E+00	161,48	8,11E-04	91,31	7,98E-01	96,67	7,89E-01	230,14
acgNCAreg	8,10E-02	25,26	9,13E-05	32,97	1,72E-01	52,58	1,55E-01	104,25

Tabla 55: Resumen de la reconstrucción. Flexibilidad 1.000. Pruebas G1N4E7.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	7,24E-02	28,77	8,26E-05	25,20	1,49E-01	47,69	1,51E-01	99,57
gNCAbasic	7,20E-02	28,76	8,19E-05	25,26	1,49E-01	47,69	1,51E-01	99,42
gNCAreg	7,51E-02	28,89	8,47E-05	25,05	1,49E-01	47,82	1,51E-01	99,86
agNCAreg	5,54E+00	161,48	8,11E-04	91,31	7,98E-01	96,67	7,89E-01	230,14
acgNCAreg	6,62E-02	18,03	7,39E-05	26,89	1,57E-01	49,98	1,46E-01	84,83

Tabla 56: Resumen de la reconstrucción. Flexibilidad 10.000. Pruebas G1N4E7.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	7,23E-02	28,77	8,22E-05	25,21	1,49E-01	47,69	1,51E-01	99,51
gNCAbasic	7,22E-02	28,76	8,21E-05	25,22	1,49E-01	47,69	1,51E-01	99,49
gNCAreg	7,78E-02	29,03	8,36E-05	24,42	1,49E-01	47,91	1,51E-01	100,42
agNCAreg	5,54E+00	161,48	8,11E-04	91,31	7,98E-01	96,67	7,89E-01	230,14
acgNCAreg	6,72E-02	19,47	8,03E-05	25,95	1,56E-01	48,79	1,45E-01	85,67

Tabla 57: Resumen de la reconstrucción. Flexibilidad 1.000.000. Pruebas G1N4E7.
Fuente: Elaboración propia.

El efecto es similar a la experiencia anterior. Sin embargo, dada la magnitud del error en este caso, los resultados con alta flexibilidad en las entradas sin suposiciones logran incluso mejorar los resultados de NCAbasic.

Para finalizar esta sección, y tal cual se comentó, se observa el efecto de que algunas entradas pueden aumentar también su error al aumentar la flexibilidad. Sin embargo en matrices grandes este efecto es poco dominante, ya que al haber muchos valores, la probabilidad de que el error inducido a una entrada domine el error global es menor que en matrices pequeñas.

3.16.5 Nivel 5

En este nivel final de las pruebas sintéticas de grupo 1 se pretende testear el funcionamiento general del método (GgNCAreg), que incorpora en conjunto todas las posibilidades antes descritas y testeadas, para luego interpretar el asunto discutido en la sección 3.13.1 Confiabilidad de las reconstrucciones, respecto a la posibilidad de entregar las reconstrucciones con niveles de precisión que permitan comparar los resultados (sin conocer la matriz real que se debiera obtener como es lo común en las pruebas realizadas).

Se comienza realizando algunas pruebas generales, en las que es posible incorporar suposiciones y confiabilidades del error en conjunto, y se analizará también un caso en que se demuestra su real utilidad. Luego se analizarán algunas experiencias en que los datos son entregados con niveles de varianza, permitiendo una explicación alternativa de las reconstrucciones.

Experiencia 1: Red pequeña. GgNCAreg.

En este caso se trabaja con una red pequeña (10x4x5x80), una matriz de datos E_p[68] con un error promedio de 14,6% y una matriz de suposición para A con su correspondiente matriz de flexibilidad. Los resultados resumidos son los siguientes. En P se utilizará la suposición por defecto.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,68E-02	11,14	1,14E-05	7,58	1,96E-02	20,18	2,09E-02	5,23
gNCAbasic	1,57E-02	10,62	1,02E-05	7,12	1,96E-02	20,35	2,08E-02	5,06
gNCAreg	1,42E-02	10,21	9,16E-06	6,56	1,96E-02	20,41	2,05E-02	4,88
cgNCAreg	1,49E-02	8,48	1,00E-05	6,34	2,23E-02	24,42	2,18E-02	4,15
agNCAreg	3,18E-01	24,14	1,96E-04	26,89	2,34E-01	39,69	2,26E-01	19,99
acgNCAreg	3,44E-01	25,62	2,06E-04	25,31	2,87E-01	42,10	2,78E-01	23,28
GgNCAreg	9,66E-03	5,24	5,76E-06	4,70	2,30E-02	23,34	2,08E-02	4,57

Tabla 58: Resumen de la reconstrucción utilizando GgNCAreg. Pruebas G1N5E1.
Fuente: Elaboración propia.

[68] El promedio de varias mediciones, por lo cual existirá también la correspondiente matriz de varianza de la misma (R).

Como se observa, el ajuste de los métodos estándar no es adecuado (es cercano al error de los datos), por lo que, y dado el error menor obtenido con cgNCA, se presume un problema con los datos. Es interesante que en este caso, el usar la matriz R mejora la reconstrucción, dado el tipo de medición errónea particular de esta experiencia, donde efectivamente se dan errores anormales. (Como lo visto en la experiencia 4 del nivel 3). En el caso de las experiencias de suposiciones, la reconstrucción no es del todo adecuada, aun utilizando la matriz D. Notar también que en este caso, A_b posee un error de ajuste del 26%. Es posible intentar mejorar lo anterior utilizando una matriz de flexibilidad con mayor valor a aquellas entradas en las que las suposiciones no tienen mucha fundamentación, lo que será testeado más adelante. Respecto a GgNCAreg, el efecto es sorprendente; en este caso, al combinar la matriz R con la suposición de los datos, los errores en A disminuyen drásticamente en relación a los obtenidos con NCAbasic, y el error tanto $de\ A$ como $de\ \ P$ presenta un mejor ajuste.

En la siguiente tabla se presenta la misma experiencia, pero generando una nueva matriz de datos promedios, con un error de 32,2% mayor al caso anterior.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	3,04E-02	18,12	4,34E-05	15,53	3,72E-02	9,13	4,81E-02	22,25
gNCAbasic	2,86E-02	17,49	3,83E-05	14,48	3,72E-02	9,23	4,80E-02	21,95
gNCAreg	2,78E-02	17,13	3,61E-05	13,95	3,73E-02	9,18	4,79E-02	21,63
cgNCAreg	3,44E-02	19,78	8,04E-05	20,05	4,10E-02	9,84	5,41E-02	25,20
agNCAreg	3,22E-01	24,97	2,33E-04	27,43	2,32E-01	24,15	2,32E-01	24,35
acgNCAreg	3,45E-01	25,73	2,43E-04	25,89	2,85E-01	29,46	2,84E-01	28,79
GgNCAreg	1,52E-02	7,84	2,72E-05	9,94	5,26E-02	16,68	4,90E-02	9,78

Tabla 59: Resumen de la reconstrucción utilizando GgNCAreg y otra matriz de datos. Pruebas G1N5E1.
Fuente: Elaboración propia.

Se observa un efecto similar al caso anterior, con la diferencia que dado los datos, la matriz R distorsiona levemente la reconstrucción utilizando cgNCAreg[69]. Además, una vez más el efecto de las suposiciones es distorsionador (obteniéndose errores similares a los obtenidos al utilizar la matriz de datos E_p). Respecto a GgNCAreg, una vez más GgNCAreg mejora sustancialmente los datos, a pesar de que la matriz R por sí sola no tiene un efecto de mejora de la reconstrucción. Es un efecto extraño, donde pareciera que el efecto de dicha matriz mejora el aporte de las suposiciones.

En la siguiente tabla se resume la misma reconstrucción de la Tabla 58, pero utilizando mayor nivel de flexibilidad para las suposiciones inexactas.

[69] Al parecer en este caso no hay datos anormalmente medidos, por lo que el promedio tiende a los valores reales.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,68E-02	11,13	1,14E-05	7,57	1,96E-02	20,19	2,09E-02	5,23
gNCAbasic	1,61E-02	10,82	1,07E-05	7,30	1,96E-02	20,29	2,09E-02	5,13
gNCAreg	1,42E-02	10,23	9,20E-06	6,57	1,96E-02	20,40	2,05E-02	4,89
cgNCAreg	1,49E-02	8,49	1,01E-05	6,34	2,23E-02	24,42	2,18E-02	4,15
agNCAreg	3,18E-01	24,14	1,96E-04	26,89	2,34E-01	39,69	2,26E-01	19,99
acgNCAreg	2,29E-02	11,78	1,61E-05	7,16	1,54E-01	37,91	1,47E-01	16,60
GgNCAreg	1,04E-02	5,32	5,98E-06	4,79	2,30E-02	23,33	2,09E-02	4,61

Tabla 60: Resumen de la reconstrucción utilizando GgNCAreg. Aumento de flexibilidad. Pruebas G1N5E1.
Fuente: Elaboración propia.

Se aprecia que el resultado condiciona sólo los resultados de acgNCAreg, donde lógicamente se mejora la reconstrucción. El resultado de los demás métodos, incluido GgNCAreg, no varía significativamente.

Experiencia 2: Pruebas con red de mayor tamaño

En este caso, se utiliza la misma red del experimento 7 de la sección anterior, en la que se generaron 3 mediciones de E, con el fin de obtener E_p y R. Los resultados son los siguientes.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	3,29E-03	4,15	1,89E-06	3,75	3,23E-02	345,83	3,45E-02	39,93
gNCAbasic	3,30E-03	4,15	1,89E-06	3,75	3,23E-02	345,86	3,45E-02	39,99
gNCAreg	3,76E-03	4,17	1,97E-06	4,10	3,24E-02	355,71	3,50E-02	40,46
cgNCAreg	7,24E-03	4,10	2,00E-06	5,19	3,79E-02	348,03	3,68E-02	49,84
agNCAreg	5,58E+00	161,27	8,67E-04	94,79	7,86E-01	506,48	7,86E-01	220,77
acgNCAreg	4,39E-02	19,64	1,22E-05	13,79	7,71E-02	98,05	7,20E-02	52,56
GgNCAreg	7,72E-03	7,54	2,28E-06	5,19	3,80E-02	344,70	3,69E-02	50,14

Tabla 61: Resumen de la reconstrucción. Pruebas G1N5E2.
Fuente: Elaboración propia.

Como se observa, los resultados son similares a los anteriores. Los 4 primeros métodos se comportan de acuerdo a lo esperado, y se destaca sorprendentemente el bajo error de las reconstrucciones, pese al error de casi 60% de la matriz E_p. En base a dicho pequeño error, lógicamente la suposición distorsiona en gran medida la reconstrucción. En contraparte, acgNCAreg mejora sustancialmente dicho error, al aumentar la libertad de las entradas con suposiciones no claras. Una vez más sorprende la capacidad de GgNCAreg, que mejora aún más el ajuste del método comentado con anterioridad, dada la combinación del efecto de la matriz de confiabilidad de los datos.

3.17 Pruebas sintéticas: Grupo 2

En este grupo de pruebas se pretende testear algunos puntos de interés del método y en ocasiones, casos extremos en los que es necesario analizar el funcionamiento de los métodos. En estas experiencias se utilizan principalmente los métodos estándar, ya que se desea analizar el comportamiento de los métodos en general, no de las nuevas funcionalidades propuestas (que ya fueron analizadas en extenso en la sección anterior).

3.17.1 Nivel 1: Redes gigantes

En este apartado se desea analizar el efecto de los métodos en redes de gran tamaño, que se asemejan en gran medida a las redes biológicas reales. Se testea el funcionamiento y rapidez de convergencia de NCAbasic, a datos con distintos niveles de error, como una forma de comprobar la capacidad de la técnica.

Experiencia 1: Red de gran tamaño y ausencia de error en los datos.

La red que se utiliza es de gran tamaño (500x44x45x85), y sus valores son generados aleatoriamente como es costumbre. Es necesario destacar que si bien las iteraciones para alcanzar el mínimo disminuyen, las mismas se tornan exageradamente lentas dada la gran cantidad de datos a procesar. En la siguiente tabla se resumen los resultados, donde se observa que pese al tamaño y al tiempo invertido, la reconstrucción es perfecta.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,54E-13	0,00	1,95E-16	0,00	6,54E-06	0,00	6,54E-06	0,00

Tabla 62: Resumen de la reconstrucción de una red grande utilizando NCAbasic. Pruebas G2N1E1.
Fuente: Elaboración propia.

Es posible forzar aún más la técnica, utilizando una matriz de datos con error de medición. Una observación relevante, tal cual se dijo anteriormente, es que al inducir un error proveniente de la misma distribución a la matriz de datos, los valores numéricamente más pequeños recibirán la mayor parte del error, concentrando prácticamente todo el error objetivo. Esto, obviamente y por un asunto probabilístico, se incrementa al crecer la red. Luego, se filtrará como antes los valores más pequeños, con el fin de agregar solo un error habitual a dichas coordenadas.

La matriz de datos utilizada presenta un error de un 20%, y ninguna coordenada domina el error total. Los resultados de la reconstrucción son los siguientes.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,15E-03	6,39	6,08E-07	4,32	1,53E+00	32,18	8,47E-01	10,23

Tabla 63: Resumen de la reconstrucción de una red grande utilizando NCAbasic. Datos con error. Pruebas G2N1E1.
Fuente: Elaboración propia.

Como se observa, pese al tamaño de la red la reconstrucción sigue mostrándose robusta. Con el fin de analizar el funcionamiento de las restantes funciones, se reconstruye la misma red anterior (utilizando la matriz con errores) utilizando gNCAbasic. Los resultados demuestran que el tiempo total de espera aumenta considerablemente. Si bien las iteraciones no son muy lentas, al momento de calcular la matriz G, y sus correspondientes reducidas, el tiempo de procesamiento es considerable, 26 segundos para la red anterior. Obviamente dicha matriz presenta dimensiones inmensas, y el efecto se amplificará en redes de mayor tamaño. Los resultados de la reconstrucción son los siguientes, que son similares como es de esperar a los anteriores:

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
gNCAbasic	1,15E-03	6,38	6,07E-07	4,32	1,53E+00	32,15	8,47E-01	10,24

Tabla 64: Resumen de la reconstrucción de una red grande utilizando gNCAbasic. Pruebas G2N1E1.
Fuente: Elaboración propia.

Experiencia 2: Red de tamaño real.

En este caso se utilizará una red real, de tamaño aproximado a las existentes en levaduras: Esto es 6270 genes y 130 $TF's$. Se utiliza en primer lugar NCAbasic para reconstruir la red. Los resultados son los siguientes, donde se destaca que el algoritmo demoró cerca de 2 horas en completar su cometido.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	8,10E-16	0,00	1,54E-18	0,00	7,38E-06	0,00	7,38E-06	0,00

Tabla 65: Resumen de la reconstrucción de una red de tamaño real. Pruebas G2N1E2.
Fuente: Elaboración propia.

Una vez más se observa una reconstrucción perfecta pese al tamaño de la red. Por supuesto el gran tiempo de espera se debe al tipo de algoritmo, en donde se realiza una exploración preliminar, repitiendo el algoritmo 30 veces. Si se modifica el método para no realizar dicha exploración, la convergencia se da en poco más de 5 minutos, donde cada iteración demora cerca de 5 segundos, más cerca de 1 minuto inicial en analizar la compatibilidad NCA. Por supuesto gNCA y los otros métodos que trabajan con un mayor número de operaciones, y analizando en un inicio las matrices reducidas de G, requerirán un mayor tiempo aun. De hecho, al intentar construir la matriz anterior, esta no puede ser construida debido a problemas de memoria. De esta manera, y para utilizar dichos métodos se recomienda, y se hace necesaria la utilización de equipos más potentes y con mayor capacidad.

Experiencia 3: Red gigante y uso de otros métodos.

Para la siguiente experiencia se repite los resultados obtenidos en la experiencia 1, utilizando la misma red y otros métodos NCA de interés. La matriz presenta un error promedio de 20% como antes. Los resultados se muestran en la siguiente tabla:

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,15E-03	6,39	6,08E-07	4,32	1,53E+00	32,18	8,47E-01	10,23
gNCAbasic	1,15E-03	6,39	6,08E-07	4,32	1,53E+00	32,18	8,47E-01	10,23
gNCAreg	4,3E-03	7,4	3,03E-06	5,2	1,64E+00	41,18	9,57E-00	13,43
cgNCA	2,15E-03	6,87	5,48E-07	3,74	1,53E+00	32,55	8,47E-01	9,8

Tabla 66: Resumen de reconstrucción de una red de tamaño grande utilizando varios métodos. Pruebas G2N1E2. Fuente: Elaboración propia.

Se omiten en este caso los métodos que utilizan suposiciones, dada la dimensión de la red. Es necesario recalcar, sin embargo, que tal cual se ilustro en resultados anteriores, es perfectamente posible realizar un análisis imponiendo suposiciones solo en algunas componentes cuando la red es de un tamaño considerable. En vista de los resultados, se puede apreciar que coincide con lo encontrado a pequeña escala. gNCAreg adiciona una distorsión que se ve, en promedio, reflejada en los errores de las reconstrucciones, mientras que cgNCA se comporta de mucho mejor forma. En este caso, y dada las características de los datos, la reconstrucción utilizando este método es bastante precisa, y la de P incluso levemente mejor a la obtenida con NCAbasic.

Las experiencias anteriores son un buen indicio de que la estabilidad y robustez encontrada en los casos analizados parece extenderse de buena manera a redes de tamaños cercanos a los reales.

3.17.2 Nivel 2: Análisis por partes de redes que no cumplen NCA

Uno de los puntos críticos del método es su gran dependencia respecto a cumplir con el criterio NCA. Si bien el primer criterio no significa problemas, las condiciones ii) y iii) presentan graves problemas, y carecen de sentido biológico. Son más bien restricciones de carácter matemático, que aseguran la existencia y unicidad (al menos en el sentido de normalización) de la solución. La restricción ii) evita que los nodos controlados por un TF correspondan a un sub-conjunto de los controlados por otro, mientras que para la iii) es una condición necesaria que el número de experimentos M sea mayor o igual al número de $TF's$ de la red, L. Obviamente, en redes de gran tamaño, dicha condición es en extremo difícil de cumplir, dada la gran cantidad de experiencias que sería necesario realizar.

Esto implica la necesidad de generar una alternativa de solución, que permita aplicar NCA a redes que no cumplan aquella restricción. La estrategia lógica consiste en particionar la red, considerando sub-grupos de $TF's$ (obviamente la idea es que cada sub-grupo presenta una cantidad de $TF's$ igual o menor a la de experimentos que se posee), y los genes asociados a dichos reguladores. Si la red puede particionarse de forma independiente, esto es, si existe sub-grupos que se puedan formar que controlen diferentes genes, es posible llevar a cabo el análisis en cada parte, y obtener resultados igualmente confiables que en el caso normal. En el caso contrario, donde existen genes comunes entre las particiones, se tendrá un error inducido en cada sub-red de manera que habrá que lidiar de alguna forma con dicha fuente de error.

A modo de ejemplo, considerar la siguiente red hipotética, en la que por ejemplo, se tendrán solo 3 experimentos (la segunda restricción del criterio NCA seguirá siendo cumplida en esta sección).

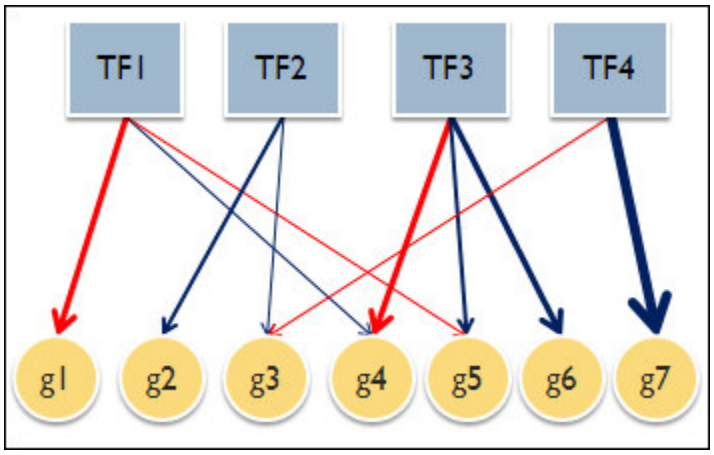


Figura 38: Red hipotética a particionar. Pruebas G2N2.
Fuente: Elaboración propia.

Si se analiza con calma la red, es posible particionarla de manera independiente, ya que los $TF's$ 2 y 4 controlan un conjunto de genes distinto al que controlan los 1 y 3. Esto es:

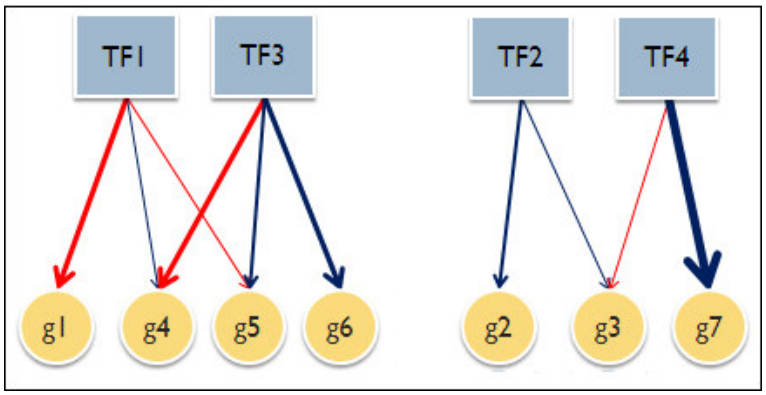


Figura 39: Particiones de la red propuesta. Pruebas G2N2.
Fuente: Elaboración propia.

Luego, es posible aplicar NCA sobre cada una de estas sub redes, uniendo luego los resultados y obtener la misma información que aplicando NCA sobre la red completa si se tuviese uno o dos experimentos más.

El caso más común será cuando la red no es particionable de manera independiente y obviamente introducirá errores en la red. Este es el caso de la siguiente red, por ejemplo, que es muy similar a la anterior, salvo que el TF controla al gen 7 en lugar del gen 5, y por ende ambas particiones dejan de ser independientes, poseen al gen 7 en común. Aun así, es posible definir diferentes particiones, tomando un grupo de $TF's$ y todos los genes que controlan. Esto necesariamente implicará que algunos genes

tendrán menos TF's en estas sub-redes que los reales, lo cual inducirá un cierto error en la reconstrucción.

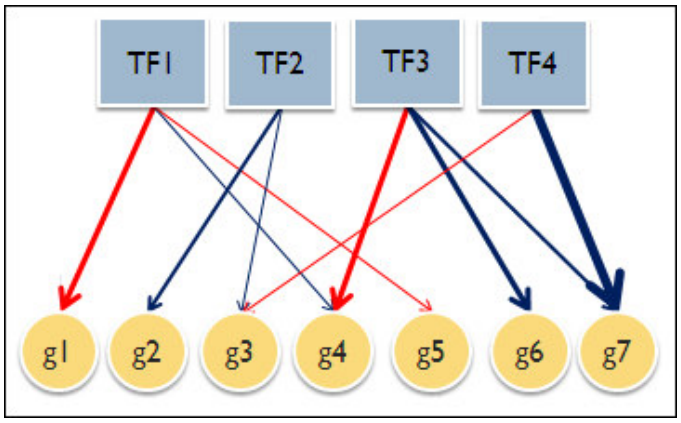


Figura 40: Red no particionable de manera perfecta. Pruebas G2N2.
Fuente: Elaboración propia.

Posibles particiones de la red anterior se muestran en la siguiente figura:

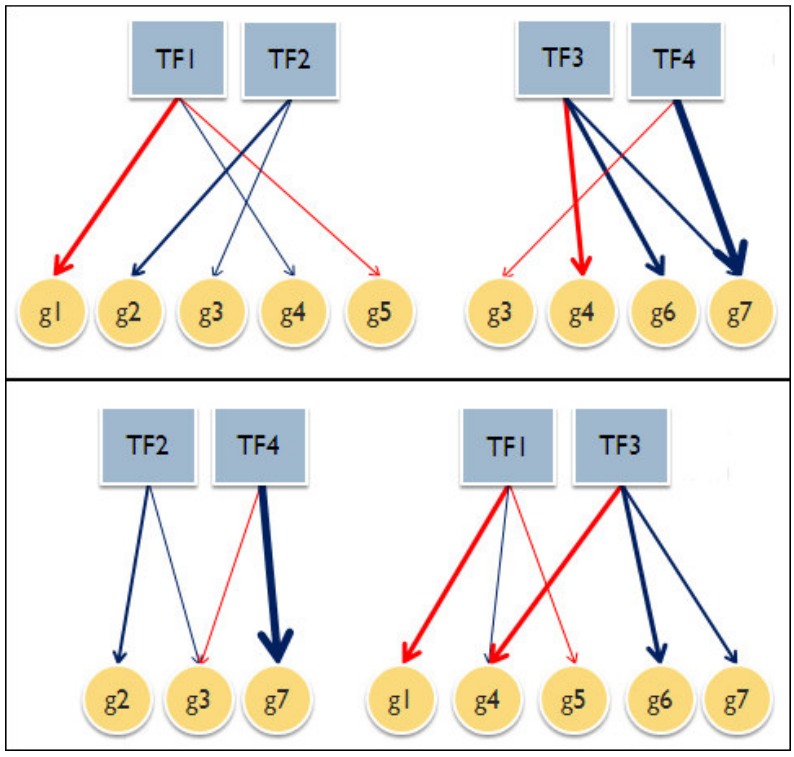


Figura 41: Posibles particiones de la red propuesta. Pruebas G2N2.
Fuente: Elaboración propia.

Obviamente la segunda es una mejor partición, en el sentido que posee menos nodos genéticos repetidos.

Experiencia 1: Prueba con red perfectamente separable

Para esta experiencia se utilizará la misma red de la Figura 38, en la que los valores para los $CS's$ de las conexiones se han generado al azar. De la misma manera, se consideraron 2 experimentos, y los valores respectivos de la matriz real P fueron generados aleatoriamente.

Al intentar utilizar NCAbasic, el método no permite continuar debido al problema de compatibilidad con el criterio NCA. Es de todas formas posible forzar al método, e intentar la reconstrucción, obteniendo los resultados siguientes al trabajar con la matriz de datos sin error.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	4,15E-02	8,25	3,46E-06	3,82	2,55E-05	0,51	2,55E-05	0,51

Tabla 67: Resumen reconstrucción red particionable**.** Pruebas G2N2E1.
Fuente: Elaboración propia.

Se observa un error de ajuste pequeño, pero si se piensa que la matriz carece de errores, el resultado es bastante elevado. La siguiente tabla resume el resultado al forzar el método y considerar una matriz con un error de un 25%, donde lógicamente el error de la reconstrucción aumenta aún más.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	7,77E-02	111,25	1,24E-05	74,22	5,77E-02	1310,89	4,81E-05	0,97

Tabla 68: Resumen reconstrucción red particionable y error en los datos**.** Pruebas G2N2E1.
Fuente: Elaboración propia.

Es posible, como se comentó antes, dividir la red anterior en 2 redes independientes, una conformada por los $TF's$ 1 y 3, y otra por los 3 y 4 junto a los genes que estos regulan, tal cual la Figura 39. El paso siguiente es aplicar NCAbasic sobre cada una de estas sub-redes, que sí cumplen el criterio NCA, y luego unir los datos de ambas con el fin de obtener todos los parámetros de la red original. En esta ocasión no hay datos cruzados entre las redes, por lo que la red debiese ser reconstruida de buena manera. Los resultados, una vez reconstruidas ambas redes por separado vía NCAbasic, son los siguientes.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,53E-07	0,12	1,34E-09	0,02	1,05E-05	0,00	1,05E-05	0,00

Tabla 69: Resumen reconstrucción utilizando particiones de la red**.** Pruebas G2N2E1.
Fuente: Elaboración propia.

El ajuste es prácticamente perfecto, y similar al obtenido al agregar un experimento extra a los datos (haciendo la red compatible con el criterio NCA), y reconstruir la red completa utilizando NCAbasic. La diferencia obviamente, será sólo que en esta nueva reconstrucción se tendrá una fila extra en P, pero las otras entradas reconstruyen los mismos parámetros del análisis parcial.

Experiencia 2: Red no separable

Una pregunta lógica surge del análisis anterior. ¿Qué sucede si las redes no son separables? En este caso, y tal cual se comentó con anterioridad, se tendrán nodos comunes sin importar que subconjunto de nodos se considere. Al respecto existen varias opciones:

- Una primera opción es dividir el conjunto de nodos reguladores en subconjuntos disjuntos de tamaño menor o igual al número de experimento, de forma aleatoria. Por ejemplo tomar algunas de las pariciones mostradas en la Figura 41. Sin embargo, y tal cual se desprende de ese simple análisis visual, existirán particiones que tengan menos genes en común.

- Otra opción es proceder por inspección, y considerar una partición como las comentadas que minimice los genes en común. Es necesario considerar, sin embargo, que las particiones mostradas en la Figura 41 para esa red en particular no son las únicas. También podría ser posible particionar la red en 2 subconjuntos de 1 y 3 elementos, e incluso en 3 subconjuntos de 1, 1 y 2 elementos. Las posibilidades, aun en la pequeña red analizada, son bastantes.

En vista a lo anterior, se hace ver que el análisis previo de la red, e incluso, el uso de otros softwares de análisis de redes que permitan particionar de forma inteligente la red propuesta se torna relevante. A modo de ejemplo, en la siguiente tabla se presenta el resultado de la reconstrucción utilizando NCAbasic en las 2 particiones mostradas con anterioridad (y luego uniendo los resultados está claro[70]).

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	1,33E-05	4,21	5,6E-05	2,1	4,05E-04	1,34	1,05E-05	0,34

Tabla 70: Resumen reconstrucción utilizando partición de red mínima (1 gen en común)**.** Pruebas G2N2E2.
Fuente: Elaboración propia.

Método	MSE(Aon-A)	%(Aon-A)	MSE(Pon-P)	%(Pon-P)	||Em-AP||	%(Em-AP)	||Er-AP||	%(Er-AP)
NCAbasic	4,53E-04	7,5	1,34E-03	3,53	1,05E-02	3,05	4,05E-04	1,45

Tabla 71: Resumen reconstrucción utilizando otra partición de red (2 genes en común)**.** Pruebas G2N2E2.
Fuente: Elaboración propia.

Como se observa, al usar una partición que minimice el número de genes comunes, se obtienen niveles de reconstrucción aceptables (menores que al proceder violando el criterio NCA), aunque de todas maneras elevados considerando el nulo error de los datos. El punto de interés, es que al contrastar dichos resultados con los presentados en la Tabla 71, se observa que en éste último caso los errores son mayores.

Inspeccionando las matrices reconstruidas en cada sub-red, se aprecia que las estimaciones para los parámetros repetidos en ambas particiones no siempre coinciden (dada la deficiencia de datos), por lo

[70] Para las reconstrucciones redundantes (que se obtienen en ambos métodos), se utiliza el promedio.

que mientras más sean los elementos en comunes en las sub-redes, mayores inconsistencias a la hora de reconstruir dichos parámetros de obtendrán.

Como regla general, el procedimiento consistirá en analizar la red a particiones, y elegir una partición que minimice el número de genes en comunes. Luego, sobre dichas redes es posible aplicar NCA, y luego de cierta forma unir los datos para obtener la visión global.

3.17.2 Nivel 3: Otras experiencias

En esta sección se desea estudiar y comentar respecto a un aspecto del desarrollo no tocado con anterior. La confiabilidad de las reconstrucciones. Como se introdujo en la sección 3.13, al visualizar el problema de reconstrucción mediante un enfoque de interpolación optimal, es posible obtener expresiones para la confiabilidad de los parámetros reconstruidos. El propósito de esta sección es discutir al respecto, e interpretar los resultados obtenidos mediante la teoría y el procedimiento ya desarrollado.

Experiencia 1: Análisis de precisión

En primer lugar se desea analizar las formas de la Ecuación 74 y la Ecuación 75, correspondientes a las expresiones para las varianzas de las estimaciones de las columnas de P y las filas de A respectivamente. Las mismas expresiones pueden ser escritas en forma de precisiones, que corresponden simplemente al inverso de dichas expresiones, y que como se verá, presenta mejores propiedades al momento del análisis.

$$\boldsymbol{Prec\left(\widehat{P_{Rck}}\right) = A_{Rk}{}^{t} R_{ck}{}^{-1} A_{Rk} + \lambda B_{Rck}{}^{-1}}$$

Ecuación 83

$$Prec\left(\widehat{A_{Rri}}\right) = P_{Ri} R_{ri}{}^{-1} P_{Ri}{}^{t} + \lambda D_{Rri}{}^{-1}$$

Ecuación 84

De esta forma, la precisión de las estimaciones corresponde a la suma de, por una parte a la precisión de las observaciones (termino $A_{Rk}{}^{t} R_{ck}{}^{-1} A_{Rk}$) y por otra la precisión de la información a priori ($\lambda B_{Rck}{}^{-1}$). Luego, si se considera que en ausencia de información a priori el segundo término de las expresiones anteriores se anula, al adicionar al análisis este tipo de información, la precisión de la reconstrucción necesariamente aumentará. Desde otro punto de vista se puede pensar que en un principio solo se posee información a priori del sistema (lado derecho de las sumas). Luego, al incluir la información experimental en la reconstrucción, y asumiendo que la expresión dada por $A_{Rk}{}^{t} R_{ck}{}^{-1} A_{Rk}$ es una matriz semi-definida positiva, la precisión necesariamente aumentará.

Notar sin embargo, que lo anterior no es válido necesariamente para la expresión de la varianza, entre otras cosas debido a la presencia de covarianzas entre los estimadores.

Si se analiza la Ecuación 74, es posible apreciar que la varianza de una columna de los estimadores de la matriz P, corresponde a una matriz cuadrada de varianza-covarianza. Con el fin de simplificar el análisis se obviarán las varianzas cruzadas, y se supondrá que las covarianzas son cero. De esta manera los únicos términos relevantes de dichas ecuaciones corresponderán a la diagonal. De la misma forma, en la expresión entregada para las matrices de precisión, solo se considerará como relevante la diagonal de la matriz.

Siguiendo con el análisis anterior, es posible definir la precisión de la reconstrucción de una columna de P, por ejemplo, como la traza de la matriz $\boldsymbol{Prec}\left(\widehat{P_{Rck}}\right)$[71]. De la misma manera, la precisión de la reconstrucción de la matriz P completa corresponderá a la suma de las precisiones de cada una de sus columnas. Gráficamente, es posible apreciar el efecto en la precisión de la matriz reconstruida en la siguiente figura.

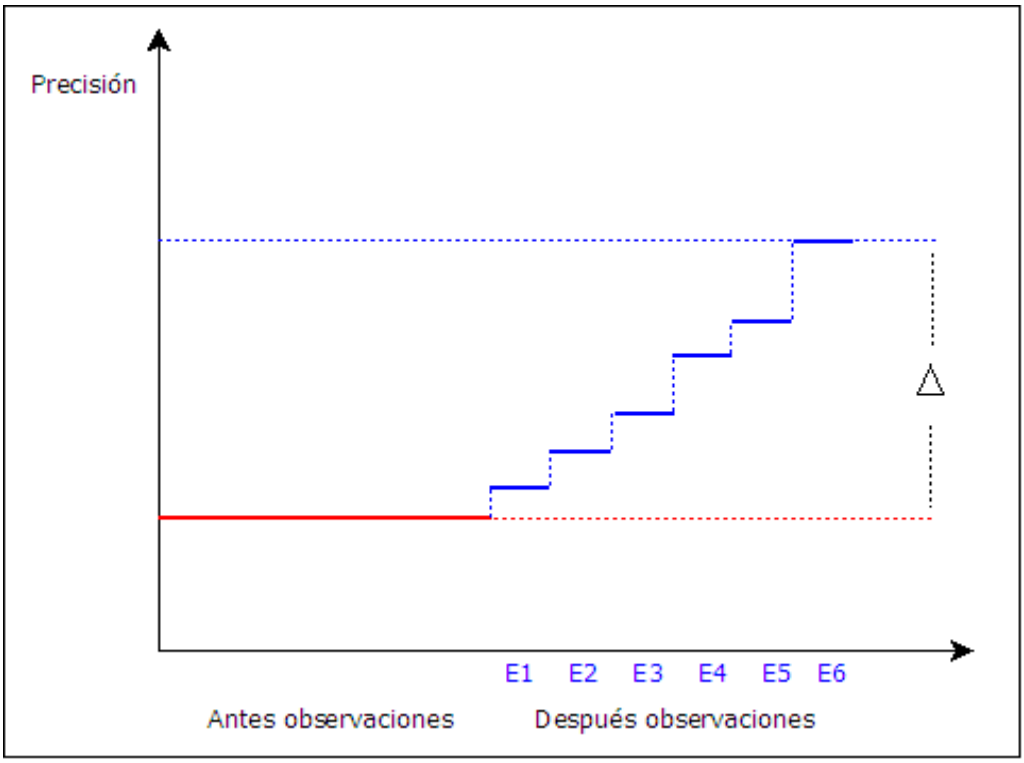


Figura 42: Análisis de precisión de las reconstrucciones. Pruebas G3N3.
Fuente: Elaboración propia.

Tal cual se desprende del análisis realizada previamente de la Ecuación 83, previo al uso de la información experimental (observaciones), solo se tiene la precisión otorgada por la información a priori del experimentador (la suma de las trazas de las matrices $\lambda {B_{Rck}}^{-1}$). Al utilizar las observaciones, la precisión aumentará de acuerdo a dicha expresión, hasta el punto marcado por la línea punteada azul

[71] Asumiendo como se indicó, que dicha matriz es diagonal.

dada por la suma de las trazas de las expresiones ${A_{Rk}}^t {R_{ck}}^{-1} A_{Rk} + \lambda {B_{Rck}}^{-1}$. De esta manera, cada columna de P (experimento) aumentará la precisión de la estimación de dicha matriz, en forma similar a lo especificado en la escalera (un escalón para cada experimento) de la Figura 42.

Experiencia 2: Aclaración sobre varianza de las reconstrucciones

En la siguiente experiencia se pretende aclarar la interpretación de las matrices de varianza (y por ende de las de precisión) obtenidas de las reconstrucciones.

A modo de ejemplo, se considera una red simple de tamaño pequeño (17x4x5x80) y una matriz de datos promedio con un 16% de error porcentual promedio con su respectiva matriz de varianza, con la cual se pretende contrastar los resultados. Es posible reconstruir las matrices utilizando NCAbasic, y obtener expresiones para las matrices de varianza de las reconstrucciones dadas por la Ecuación 76 y la Ecuación 77. En este caso no se trabajarán con suposiciones con el fin de simplificar los cálculos, por lo que para cada columna o fila de las matrices reconstruidas se calcularán las varianzas respectivas utilizando la Ecuación 78 y Ecuación 79, que obvian el termino de suposiciones a priori.

Las matrices de errores porcentuales obtenidas en las reconstrucciones son las siguientes:

```
E_A =
[     0    0.9253    3.1941         0
      0    0.2124         0         0
 1.1286         0         0    2.3755
      0         0         0    3.5842
      0    0.0018         0         0
 0.6586    0.6451         0         0
      0         0    0.0999         0
 0.9181    0.8660         0         0
 0.2810         0         0         0
 1.1197         0         0    1.0149
 0.8428         0         0         0
 9.2690    9.8367         0         0
      0         0    0.0786    6.8575
      0    0.4301         0         0
 1.2743         0         0         0
 0.8643         0         0         0
      0         0    0.2223         0]
E_P =
[1.6135    0.5762    0.3199    0.3280    0.9009
 0.7123    1.3212    0.8207    0.7537    0.7402
 0.1854    0.5487   20.9493    0.4809    0.4073
 0.1865    4.5088    3.9292    2.7968    4.0153]
```

Al analizar las matrices de varianza de las reconstrucciones se obtienen los siguientes resultados:

```
QA =
    0    0.0000    0.0001         0
    0    0.0000         0         0
0.0006         0         0    0.0032
    0         0         0    0.0000
    0    0.0002         0         0
0.0009    0.0001         0         0
    0         0    0.0000         0
0.0007    0.0001         0         0
0.0022         0         0         0
0.0019         0         0    0.0045
0.0012         0         0         0
0.0027    0.0001         0         0
    0         0    0.0003    0.0010
    0    0.0001         0         0
0.0012         0         0         0
0.0008         0         0         0
    0         0    0.0001         0
QP =
  1.0e-006 *
    0.0196    0.0043    0.0161    0.0057    0.0134
    0.2890    0.2734    0.9207    0.7442    0.2865
    0.0817    0.0285    0.6275    0.3546    0.0455
    0.5596    0.1061    0.4403    0.0007    0.1130
```

Analizando por inspección las matrices, se ve que no existe necesariamente una correlación entre los errores obtenidos en esta experiencia y la varianza de las coordenadas respectivas. En efecto, las entradas con mayor error en la matriz A presentan una varianza intermedia en relación a las demás entradas. La explicación es clara. Mientras que el error obtenido es una suerte de exactitud para la observación particular estudiada, la varianza presenta un enfoque más general, entregando la exactitud de los estimadores mirando los mismos como variable aleatoria (lo que en realidad es simplemente un enfoque diferente para el mismo problema). En otras palabras, pese a que entradas con baja varianza pueden tener errores altos para esta medición en particular, su reconstrucción es estadísticamente más precisa al tomar diferentes errores en los datos.

Es necesario tener esto en mente a la hora de interpretar los resultados dados por dichas matrices.

La raíz (componente a componente) de las matrices anteriores representa las desviaciones estándar de los estimadores para las matrices, por lo que es posible incluso generar intervalos de confianza para los resultados entregados.

3.18 Redes alternativas: Un nuevo enfoque

A continuación se pretende describir un nuevo enfoque motivado en algunos de los resultados y limitaciones identificadas en los métodos NCA. En particular, se destaca la limitante impuesta por el gran conocimiento e información que implica imponer desde un comienzo una estructura para la matriz A; esto implica saber de antemano las conexiones válidas que existen entre $TF's$ y genes y asignar un grado de validez inequívoco para que la reconstrucción y el uso del método sea válido. Si bien hay mecanismos que se conocen con mayor exactitud y están mejores definidos que otros, no siempre es

posible obtener el patrón de conexión de todo el genoma. Por otra parte, aun cuando la información pueda existir, esta puede encontrarse distribuida en diferentes fuentes y bases de datos, e incluso darse el caso que algunas referencias no coincidan exactamente.

Es por eso que surge la idea de desarrollar un método alternativo, que sea compatible con el análisis NCA y que pueda entregar una guía para solucionar, o al menos entregar una vía alternativa de análisis o interpretación cuando existan problemas como el anterior.

En los párrafos siguientes se describirá la motivación y teoría tras el método de análisis propuesto.

3.18.1 Experimentos para detectar unión TF-gen.

Existen diversos métodos experimentales utilizados en la detección de proteínas unidas al DNA. En particular, uno bien documentado corresponde al análisis Chip-CHIP [47]. Específicamente, el análisis consiste en un estudio de la unión de proteínas a ciertas secuencias del DNA. Una vez identificados los $TF's$ relevantes para un organismo particular, información ampliamente disponible en la literatura[72] [48] [49], se procede a crear cepas mutantes que contengan marcadores específicos en dichas proteínas. Luego se procede a hacer al análisis en sí, mediante la tecnología ChIP[73] para cada una de las cepas mutantes creadas, y las regiones promotoras del DNA a las cuales se unieron los reguladores son identificadas mediante hibridación a microarrays que contengan un set de las secuencias promotoras del organismo. Por supuesto, y tal cual se explica en [37], dadas las propiedades de los sistemas biológicos estudiados, el esperable ruido en los experimentos microarrays, etc., es indispensable la utilización de un modelo de errores, con el fin de interpreta de mejor forma los resultados, y obtener un análisis probabilístico de los mismos. Luego, se generaliza la utilización de p-value, que define la confianza con la cual los resultados son atribuibles efectivamente a una unión valida de un regulador con un promotor[74]. De esta forma es posible reconstruir las redes de regulación para un organismo particular, indicando que genes se relacionan con un regulador en específico.

Es necesario recalcar también una serie de dificultades y problemas con el método anterior, que fundamentan en parte la motivación anteriormente descrita y la necesidad de relajar el criterio de información previo de la red requerido por NCA. En primer lugar, los procedimientos experimentales no son precisos, y muchas veces parte importante de la información no es obtenida completamente. A modo de ejemplo, y tal cual se describe en [37], si bien el análisis parte intentando identificar el papel de muchos $TF's$, al intentar crear las cepas mutantes para cada uno de ellos el tema se complica. En esta etapa es difícil obtener cepas viables[75] para todos los $TF's$, por lo que algunos irremediablemente son dejados de lado. Y más importante aún, es posible que la inclusión de los marcadores afecte la actividad e incluso la afinidad de los reguladores, por lo que los resultados pueden presentar importantes alteraciones.

[72] No así las relaciones específicas de los $TF's$ y los genes.
[73] Chromatin inmunoprecipitation.
[74] Generalmente es utilizado un p-value de 0,001.
[75] Que efectivamente expresen el TF marcado.

3.18.2 Criterio de redes equivalentes

J.C. Liao desarrolló en uno de sus trabajos un criterio que no ha sido comentado con anterioridad, el correspondiente a un criterio que permite distinguir entre 2 redes propuestas para un problema particular.

En primer lugar, comenzando con una definición del asunto, se dirá que 2 redes son indistinguibles si al utilizarlas para descomponer una matriz de datos dada, el residuo es similar. Como se vio con anterioridad, dada una matriz de datos E y una cierta estructura resumida en $\{Z_{A1}\ ,\ Z_{P1}\}$ que cumpla con el criterio NCA, ésta puede ser descompuesta en 2 matrices, $A_1 \in Z_{A1}$ y $P_1 \in Z_{P1}$ de forma esencialmente única (con un residuo mínimo). Sin embargo podría existir otra estructura, por ejemplo $\{Z_{A2}\ ,\ Z_{P2}\}$ (donde por ejemplo, algunas conexiones entre TF's y genes sean diferentes) que también cumpla con el criterio NCA, y por ende acepte una descomposición esencialmente única en $A_2 \in Z_{A2}$ y $P_2 \in Z_{P2}$. Por supuesto, el ajuste no será nunca del todo perfecto, y existirá en cada reconstrucción un cierto residuo mínimo. Esto es:

$$\Gamma_1 = E - A_1 \cdot P_1$$

$$\Gamma_2 = E - A_2 \cdot P_2$$

Si Γ_1 y Γ_2 son iguales, se dice que las redes son indistinguibles, mientras que en caso contrario las redes pueden distinguirse una de otra.

El criterio desarrollado para saber cuándo 2 redes son distinguibles es el siguiente.

Teorema 2: Capacidad de distinguir entre redes.

Sean 2 estructuras a priori $\{Z_{A1}\ ,\ Z_{P1}\}$ y $\{Z_{A2}\ ,\ Z_{P1}\}$ representando 2 redes que satisfacen el criterio gNCA. Para unas matrices $A_1 \in Z_{A1}$ y $P_1 \in Z_{P1}$, y si la matriz reducida $A_{1_{Rj}}$ o P^t_{1Rj} tiene rango L para al menos un $j, j = \{1, \dots, L\}$, entonces A_1 y P_1 pueden ser distinguible de cualquier otra red $A_2 \in Z_{A2}$ y $P_2 \in Z_{P2}$.

En este caso, $A_{1_{Rj}}$ se define como la matriz A_1 a la cual se han eliminado las filas correspondientes a los elementos no nulos en la columna j de la estructura de Z_{A2}. Lo mismo en el caso de P^t_{1Rj}.

En otras palabras, con el criterio anterior se puede saber si una red es distinguible de otra al momento del compararlas. De serlo, sería interesante analizar si la nueva configuración se ajusta mejor a los datos, en cuyo caso, podría quizás ofrecer una mejor explicación al problema. En el caso de redes indistinguibles, el teorema asegura que existirá una matriz invertible Y tal que:

$$A_2 = A_1 \cdot Y$$

$$P_2 = Y^{-1} \cdot P_1$$

En adelante se supondrá que solo la estructura de $\boldsymbol{A}$ puede variar de una red a otra (no se supondrán restricciones diferentes a la estructura de $\boldsymbol{P}$, $\boldsymbol{Z_P}$), por lo que el teorema anterior será equivalente a comprobar que alguna matriz reducida $\boldsymbol{A_{1_{Rj}}}$ tiene rango $\boldsymbol{L}$.

3.18.2 Simulated Annealing (Recocido simulado)

La técnica de recocido simulado es un algoritmo de búsqueda meta-heurística para problemas de optimización global. En otras palabras, el objetivo es buscar una buena aproximación del óptimo global de una función en un espacio de búsqueda extenso. Dicha técnica es aplicable a una extensa variedad de problemas, y se propone en lo siguiente aplicarlo al problema NCA identificado.

Sin entrar en detalles, el método de recocido simulado parte de una solución factible del problema[76] y perturba dicho punto aleatoriamente, comprobando el valor del funcional en la nueva ubicación. El método se desplaza de ubicación (o cambia de estado) si es que el nuevo punto ofrece una solución mejor[77], o si por efectos del azar el método así lo decide. El punto principal de la técnica es que las perturbaciones son locales y los saltos son permitidos aun cuando la solución sea peor a la actual (siempre habrá una pequeña probabilidad). Además serán más probables y con mayor rapidez al inicio del proceso, por lo que es posible escapar de posibles mínimos locales en donde el sistema se halla quedado estancado, permitiendo la convergencia al mínimo global del problema.

El nombre de la técnica proviene del método de recocido del acero (con el fin de obtener un material más resistente), en el cual el material es calentado a altas temperaturas y enfriado de forma controlada, permitiendo la formación de cristales de mayor tamaño y reduciendo así sus defectos. Físicamente, al calentar el material se permite que los átomos se muevan libremente y escapen de mínimos locales de energía, alcanzando finalmente un estado de energía menor, más estable y de mayor dureza y resistencia en este caso.

3.18.3 Definición de la técnica de recocido simulado

Un problema de recocido simulado está compuesto por los siguientes elementos:

- Un conjunto de soluciones factibles S, que corresponde a las soluciones admisibles para el problema.
- Una función de costos o energía $J(s)$, con $s \in S$.
- Una función de entornos N de S en partes de S. En otras palabras, una regla que defina cuales son los elementos de S a los cuales se puede acceder estando en un $s \in S$ particular.
- Una sucesión T_k de estado o "temperatura[78]", que normalmente disminuirá de una iteración a otra.

[76] Un mínimo local por ejemplo, aunque otros puntos factibles también son permitidos.
[77] Si el funcional aumento o disminuye, de acuerdo al tipo de problema que se trate.
[78] Haciendo analogía al método de recocido del acero, en el cual la temperatura es disminuida de forma constante para lograr el equilibrio térmico.

- Un criterio de movimiento probabilístico, definido por una función $\pi(T_k, \Delta J)$, donde , ΔJ es la diferencia de energía entre 2 estados adyacentes, s y s', $\Delta J = J(s') - J(s)$. Luego, la probabilidad de transición dependerá del estado actual del sistema y de la diferencia de energía de los puntos comparados.

El algoritmo de la heurística es simple. En la iteración k del método, estando en $s \in S$ se propone un candidato $s' \in S$ al que se pueda acceder mediante la función de entorno N. Luego se calcula $\Delta J = J(s') - J(s)$, y de acuerdo al criterio de movimiento probabilístico se decide si se salta al nuevo punto o no. Normalmente $\pi(T_k, \Delta J)$ es mayor que 1 independiente de T_k si $\Delta C < 0$, y disminuye exponencialmente mientras mayor sea el cambio de energía y menor la temperatura si es que $\Delta C > 0$. Al finalizar dicha iteración, el estado del sistema T_k es actualizado a T_{k+1} de acuerdo a alguna regla establecida.

Por lo general se utilizarán las siguientes expresiones para $\pi(T_k, \Delta J)$ y T_k:

$$\pi(T_k, \Delta J) = \exp\left(\frac{-\Delta J}{T_k}\right)$$

Ecuación 85

$$T_k = \alpha \cdot T_{k-1}$$

Ecuación 86

$$T_k = \frac{1}{1 + \beta \cdot T_{k-1}}$$

Ecuación 87

La Ecuación 86 corresponde a un enfriamiento mediante un factor geométrico, en donde $\alpha < 1$ y muy cercano a 1. En la Ecuación 87 en cambio β es un real positivo cercano a cero.

Otra opción más compleja a implementar en este tipo de técnicas, es el método conocido como monótono, que difiere algo del algoritmo anterior en el sentido de que la temperatura o estado del sistema se actualiza sólo cuando se ha alcanzado un número de iteraciones fijas K o se ha producido un número máximo de saltos A. Al terminar dichas iteraciones se supone se ha alcanzado un equilibrio a la temperatura T_k, por lo que dicho parámetro se actualiza. El factor K se aumenta levemente cada vez que la temperatura disminuye, asumiendo que a temperaturas menores se requiere un mayor tiempo para alcanzar un equilibrio.

Finalmente existe otra versión conocida como método no monótono, en el cual se da la oportunidad a la temperatura de aumentar en un momento determinado, si es que algún indicador hace suponer que la temperatura ha disminuido mucho, y el sistema aún se encuentra estancado en un punto cercano a un mínimo local.

La figura siguiente describe el algoritmo anteriormente descrito, en el cual la temperatura es disminuida en cada iteración. Las flechas representan saltos de un punto a otro del sistema. Las iteraciones no mostradas fueron momentos en los cuales ninguna de las alternativas propuestas ofreció posibilidades de cambio del punto actual, por lo cual el sistema permaneció sin alteraciones

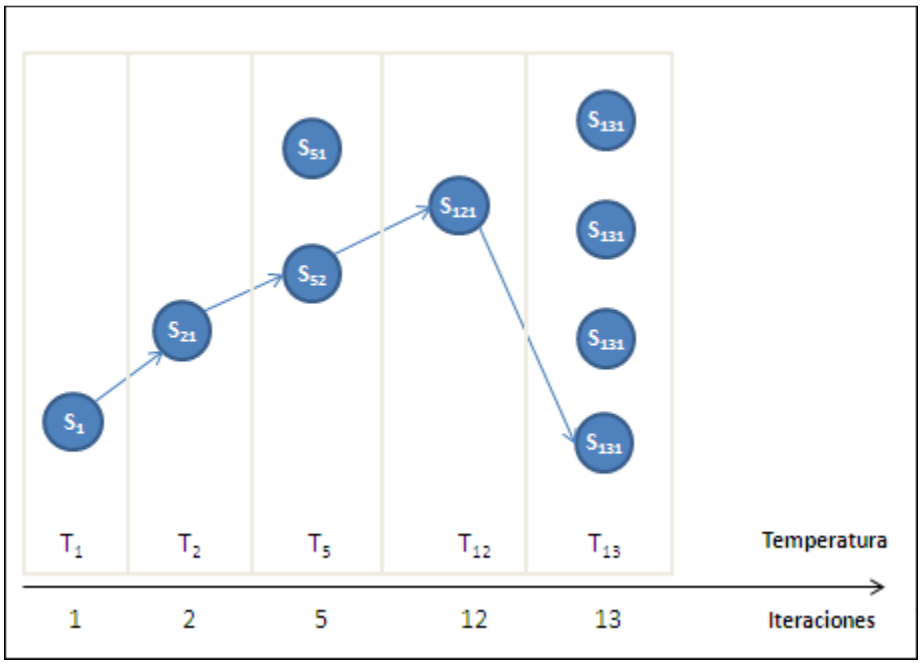


Figura 43: Representación gráfica del algoritmo de recocido simulado.
Fuente: Elaboración propia.

3.18.4 Aplicación a NCA: Implementación y comentarios.

La aplicación de la teoría anterior al problema NCA se logra cambiando la orientación del problema. Normalmente lo que se intenta es disminuir el error de ajuste de los datos a ciertas matrices, restringidas por una estructura dada (la estructura de A resumida en Z_A). En este caso lo que se hará será definir una función de energía que depende de la estructura Z_A escogida[79], que en este caso, se interpretará como una matriz de conectividad con ceros y unos, indicando que TF está unido con que gen. Esto es, utilizando por ejemplo NCAbasic para un Z_P y E dados y fijos (aunque podrían usarse otros métodos NCA de la misma manera), la función de energía se define como.

$$J(Z_A) = \min_{A,P} \ \|E - A \cdot P\|_F^{\,2}$$
$$s.t.\ A \in Z_A \ y \ P \in Z_P$$

Ecuación 88

[79] No del ajuste de una estructura fija a diferentes tipos de datos. En este caso se considerará fija la matriz de datos.

Esto define inmediatamente el espacio de soluciones factibles S:

$$S = \{Z_A \mid Z_A \in M^{NxL} \ \ y\ es\ una\ estructura\ que\ cumple\ con\ el\ criterio\ NCA\}$$

Ecuación 89

De esta manera, y con esto en mente, es posible definir un modelo de recocido simulado que permita explorar la aplicación de redes alternativas a los datos de interés. Algunos puntos de interés serán tocados en lo siguiente.

Sobre la función de entorno y variaciones de la matriz de conectividad

La definición de la función de entorno N para este caso es algo más complejo, por lo que no se dará una expresión funcional para la misma. Su construcción está motivada en la composición de las estructuras Z_A. En primer lugar se debe destacar que de admitir cualquier configuración como función de entorno, la combinación de opciones sería abrumante[80], por lo que es necesario restringir de alguna forma las posibilidades permitidas. Como ya se ha mencionado, la redes reales son altamente densas en ceros (entre el 70% - 80%) por lo que la combinación de estructuras se restringe en este aspecto. Una función de entorno lógica no debiera permitir que al pasar de un estado a otro la densidad de ceros varía considerablemente. Se hace notar también que esto es equivalente a restringir el espacio de soluciones factibles S antes definido, ya que a partir de una configuración inicial no será posible alcanzar todas las posibilidades (en particular las altamente densas en unos).

En segundo lugar es posible definir 3 posibles cambios a realizar a una configuración o matriz de conectividad dada. Es posible eliminar una conexión entre un TF y un gen (eliminar un 1 de la matriz de configuración), agregar una conexión (agregar un 1) o permutar una conexión (permutar un 1 de posición hacia un lugar donde había un cero en Z_A, dejando un cero en el lugar original). Dichos cambios puntuales definirán de hecho el concepto de entorno en esta aplicación. Así, partiendo desde un espacio de conectividad particular Z_A, se dirá que el entorno de dicho espacio corresponderá a todos los espacios que pertenezcan a S dado como en la Ecuación 89, y a los que es posible acceder mediante un cambio puntual como los detallados anteriormente. Aún más, se exigirá que dichos cambios sean solo dentro de una misma fila o columna en el caso de las permutaciones.

Además de las 2 restricciones anteriores, se definirá el concepto de entradas vetadas en la estructura, lo que ayudará a disminuir aún más la cantidad de combinaciones de redes posibles. Los espacios vetados en una matriz de conectividad corresponderán a 2 tipos, que se nombrarán como ceros fijos y unos fijos. Los primeros corresponden a relaciones inexistentes entre un gen y un TF, que con alta certeza se sabe

[80] Dado el tamaño que puede tomar A, y por ende la cantidad de combinaciones que pueden existir. Por ejemplo, en el caso de una red de 10 genes y 4 reguladores, se tendrá un total de 40 entradas en la matriz de conectividad. Si se acepta que el 20% de dichas entradas corresponden a 1's (conexiones validad) y el resto a ceros, es necesario ubicar 8 ceros en 40 espacios. Dejando de lado por un momento que las posibles redes deben cumplir el criterio NCA, se poseen $40 \cdot 39 \cdot 38 \cdot 37 \cdot 36 \cdot 35 \cdot 34 \cdot 32$ combinaciones, más de 1 millón de millones de posibilidades.

es de esta manera. Los segundos en cambio, corresponden a conexiones que con un alto grado de confiabilidad existen, y no se tiene duda de aquello. La motivación es simple: si bien de acuerdo con lo explicado con anterioridad no es posible otorgar una gran certeza a una matriz de conectividad dada, si existen entradas puntuales de dichas matrices[81], correspondiente a la existencia o no existencia de conexiones, que son conocidas extensamente y por ende la confiabilidad de dicha información es elevada. Luego, no tendría sentido perturbar dichas entradas al momento de modificar la estructura Z_A vía la técnica propuesta. Esto es lo que se llamará entradas o coordenadas vetadas.

A modo de resumen se definen entonces los siguientes cambios puntuales que conducen a estructuras entornos de Z_A.

1. Agregar una conexión. Corresponde a agregar una conexión a la estructura de Z_A con el fin de generar ${Z_A}'$, cuidando se cumpla lo siguiente:

    - La estructura resultante pertenece a S y por ende respeta el criterio NCA.
    - Al agregar la conexión la densidad de ceros de la matriz no disminuye considerablemente[82].
    - La conexión no se agrega a entradas vetadas (ceros fijos).

2. Eliminar una conexión. Corresponde a eliminar una conexión a la estructura de Z_A con el fin de generar ${Z_A}'$, cuidando se cumpla lo siguiente:

    - La estructura resultante pertenece a S y por ende respeta el criterio NCA.
    - Al eliminar la conexión la densidad de ceros de la matriz no aumenta considerablemente.
    - La conexión no se agrega a entradas vetadas (unos fijos).

3. Permutar una conexión. Corresponde a permutar una conexión en la estructura de Z_A con el fin de generar ${Z_A}'$, cuidando se cumpla lo siguiente:

    - El cambio fue realizado en la misma fila o columna respecto a donde estaba originalmente la conexión.
    - La estructura resultante pertenece a S y por ende respeta el criterio NCA.
    - La conexión no se permuta a entradas vetadas (ceros fijos) ni se permuta desde entradas vetadas (unos fijos).

En la figura siguiente se representan esquemáticamente dichos cambios. En rojo se representan las entradas vetadas de la matriz original, mientras que en azul se representan los cambios.

[81] Conexiones puntuales gen-TF.
[82] En específico, se permitirá agregar (de forma neta) solo un número específico de conexiones.

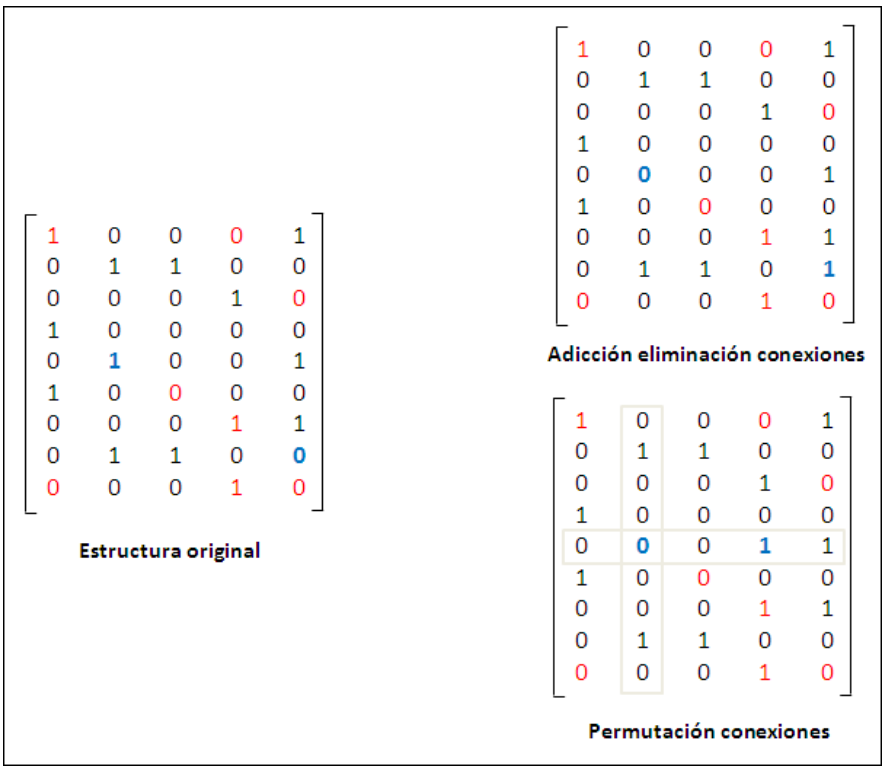


Figura 44: Representación esquemática de los cambios puntuales en la estructura de conexión.
Fuente: Elaboración propia.

Con el fin de comprobar el efecto en la función $J(\cdot)$ de los cambios puntuales en la estructura antes descritos, se desarrolla una pequeña experiencia que pretende representar gráficamente dichas perturbaciones. Para dicho fin se utiliza una red pequeña (12x3x4x80) y se procederá a agregar, eliminar y permutar conexiones de forma aleatoria en varias ocasiones, con el fin de analizar la distribución del error generado. Para dicho efecto se crea un programa que de forma automatizada permite en primer lugar probar agregando conexiones de forma aleatoria a la matriz, y luego a 3 filas y 3 columnas objetivos. En cada caso el procedimiento se desarrolla varias veces, con el fin de mostrar distribuciones del error obtenido. El programa repite el mismo procedimiento para el caso de eliminar y permutar conexiones.

En la figura siguiente se resumen los gráficos de distribución en el caso de agregar conexiones puntuales a la matriz base. Cabe destacar que todos los cambios son aplicados a una estructura base, por lo cual los esquemas representan la distribución de los errores de las estructuras entornos de la base considerada.

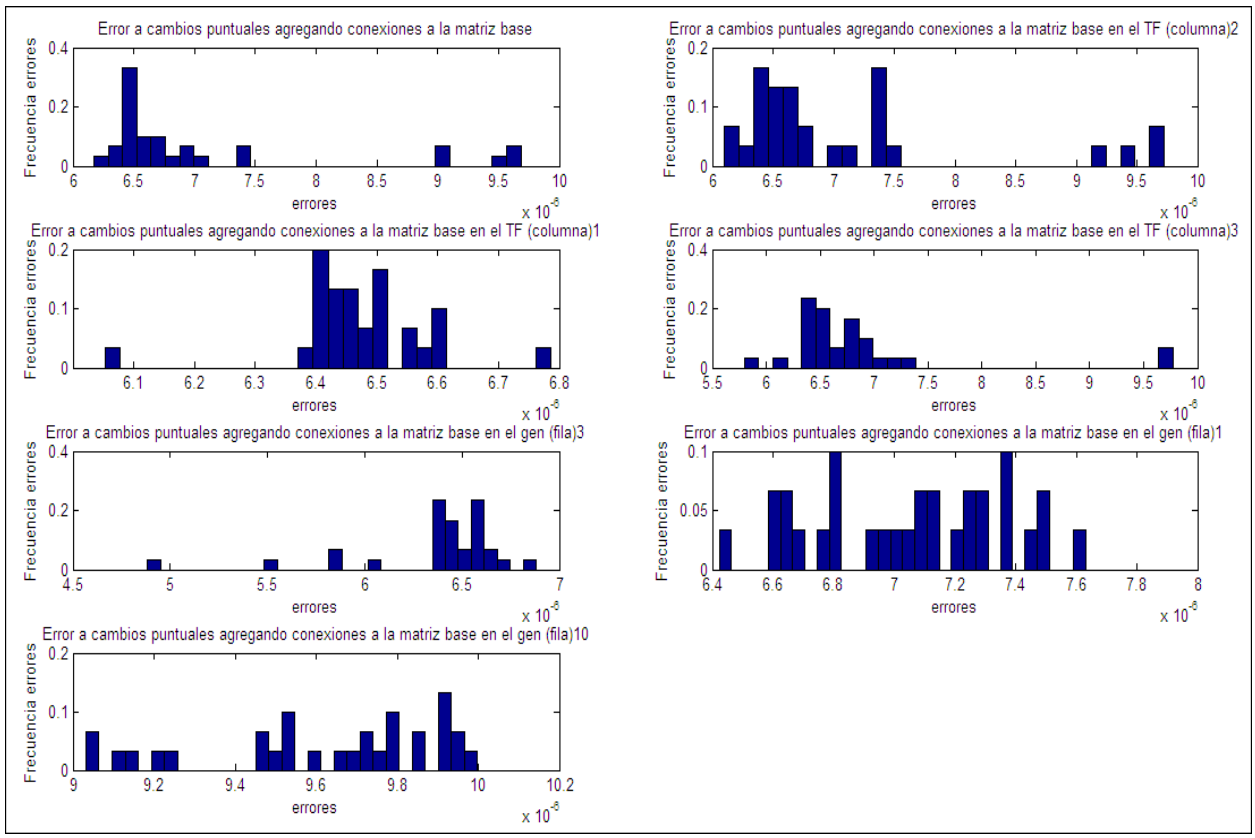


Figura 45: Distribución de los errores a cambios puntuales en la estructura. Agregar conexiones.
Fuente: Elaboración propia.

El error original de la matriz es del orden de 10^{-6} por lo que en todos los casos la distribución del error se mantiene en dicha magnitud. En otras palabras, el error no aumenta en demasía a cambios puntuales de este tipo. Antes de entrar en detalle, se debe aclarar que este efecto es esperado, y consistente con lo ya concluido en apartados anteriores. Dado la capacidad y robustez del método, NCA es capaz de llevar a un peso cercano a cero las entradas que fueron declaradas como una conexión[83], por lo que el error cometido al agregar entradas de más no es considerable[84]. Analizando los diferentes gráficos se ve que la distribución de todos se encuentra en torno a $6 \cdot 10^{-6}$, a excepción de la distribución de los errores cuando las conexiones son agregadas en la fila 10. En este caso la distribución está centrada $9 \cdot 10^{-6}$, y si bien la diferencia es pequeña, hace pensar en la existencia de sectores más sensibles a cambios puntuales de los parámetros. Además, y particularmente en el caso donde las conexiones fueron agregadas en la fila 3, se puede observar la presencia de errores inferiores a $6 \cdot 10^{-6}$, que pueden ser atribuidos a redes que en un principio, parecen ajustarse mejor a los datos que la red que los generó. La implicancia de esto y su validez serán discutidas en otras secciones, pero se adelanta que si bien se puede tener una red con un gran ajuste, no se puede descartar la existencia de otras redes con ajuste

[83] Cuando en realidad no existía.
[84] En modificaciones puntuales lógicamente. Al agregar muchas conexiones se perderá la estructura del problema, y NCA no será capaz de llevar a cero el peso de dichas entradas.

similar (incluso mejor), y lo más importante, la interpretación biológica que se le pueda dar a dichos resultados.

En la figura siguiente se resume un experimento similar al anterior, pero esta vez eliminando conexiones.

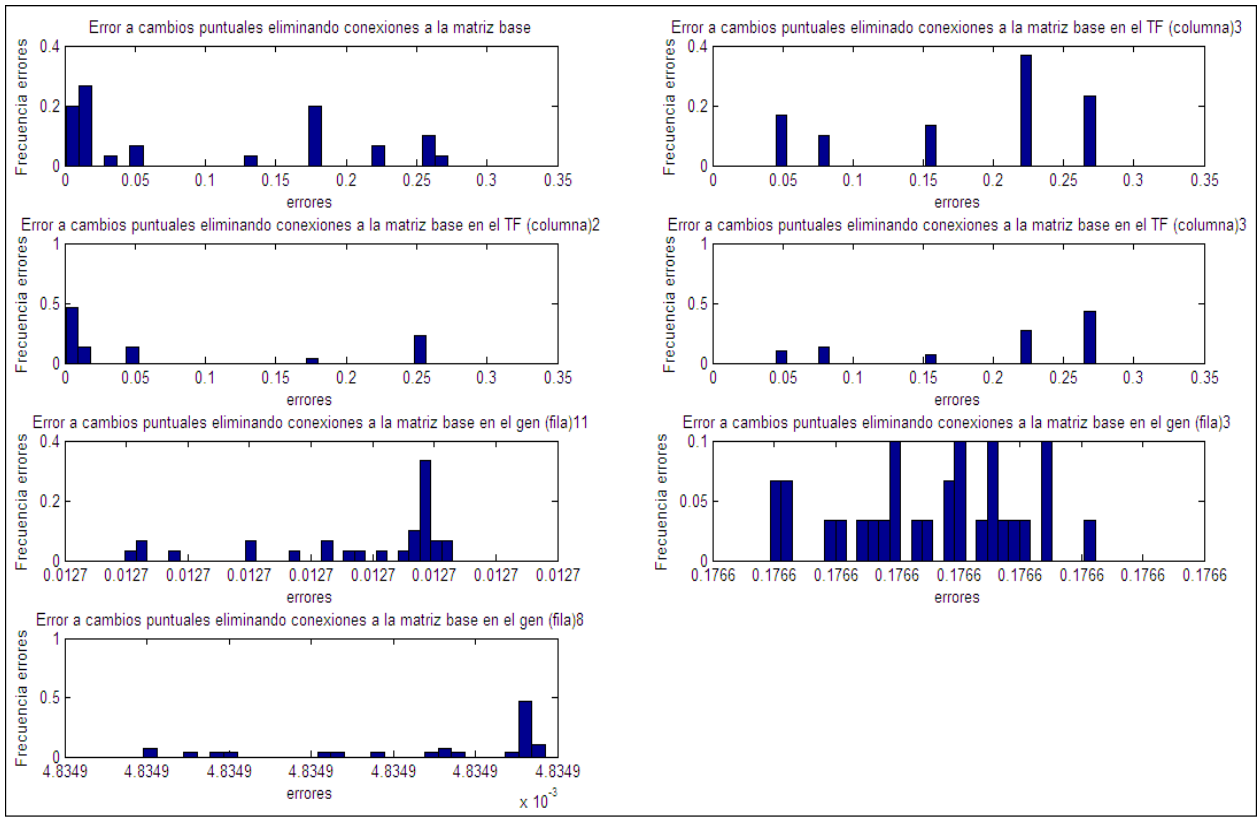


Figura 46: Distribución de los errores a cambios puntuales en la estructura. Eliminar conexiones.
Fuente: Elaboración propia.

Se observa que en este caso la distribución del error es considerablemente mayor. Las redes son mucho más sensibles a omitir conexiones validas que a sobrestimarlas, ya que dada la estructura del método NCA los ceros en la estructura se consideran información dada. Si bien las distribuciones de los errores están centradas en puntos dispares, se observa en particular el bajo error (en relación con las demás experiencias) que se obtiene al eliminar conexiones en la fila 8. Una vez más parecen existir sectores más sensibles a cambios locales. Globalmente en cambio[85], se observa una distribución bastante extendida entre 0 y 0,25.

Finalmente se resume en la siguiente figura la distribución del error a permutaciones. Nótese que esta experiencia en parte consiste en combinar los 2 efectos anteriores.

[85] Eliminando conexiones en cualquier lado.

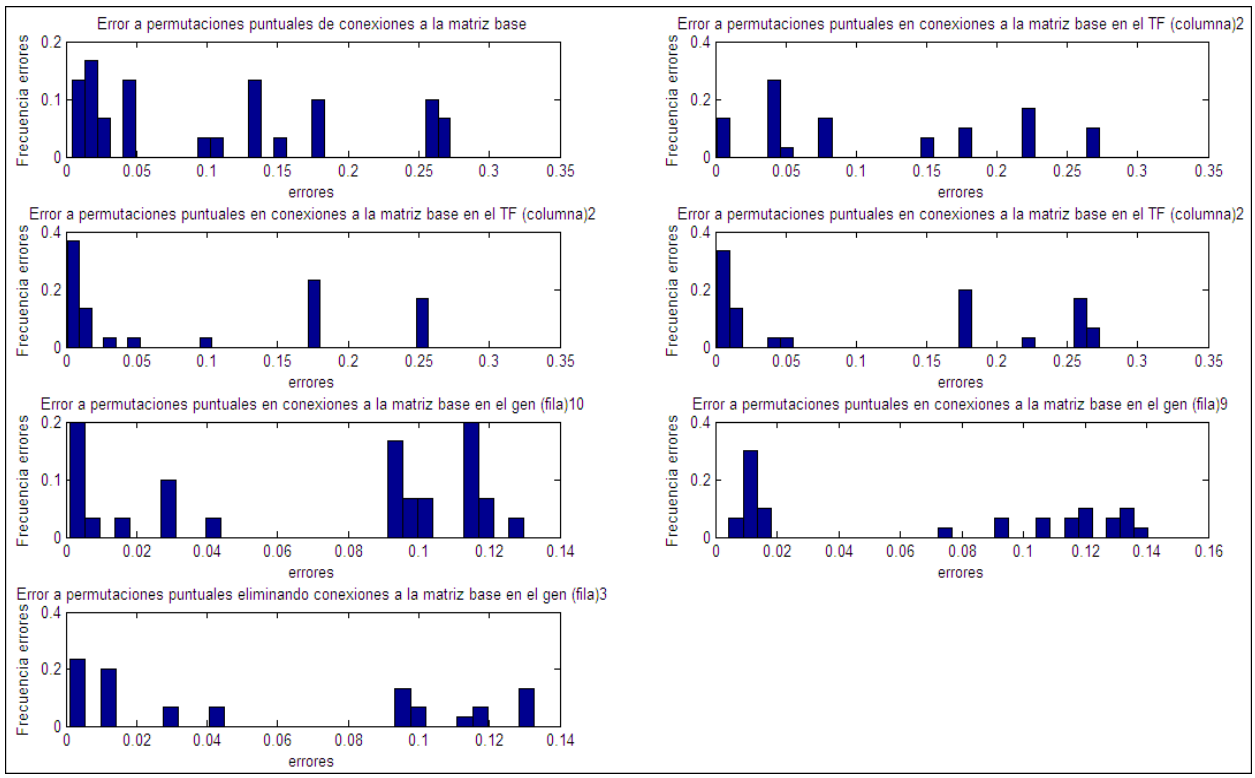


Figura 47: Distribución de los errores a cambios puntuales en la estructura. Permutar conexiones.
Fuente: Elaboración propia.

Se aprecia, que al igual que en el caso de la Figura 46, el error se encuentra bastante distribuido y es mayor (y lejano) al obtenido sólo agregando conexiones. Sin embargo, el error es menor relativo al obtenido con la red original, lo que en parte es producto del efecto de agregar conexiones que tiene la permutación.

Es posible repetir el análisis anterior considerando errores en la matriz de datos (30% en este caso), con el fin de verificar si el efecto antes descrito varía en relación a ese factor. En este caso, la reconstrucción con la red original entrega un error del orden de 10^{-3}. Los resultados se resumen en las siguientes figuras.

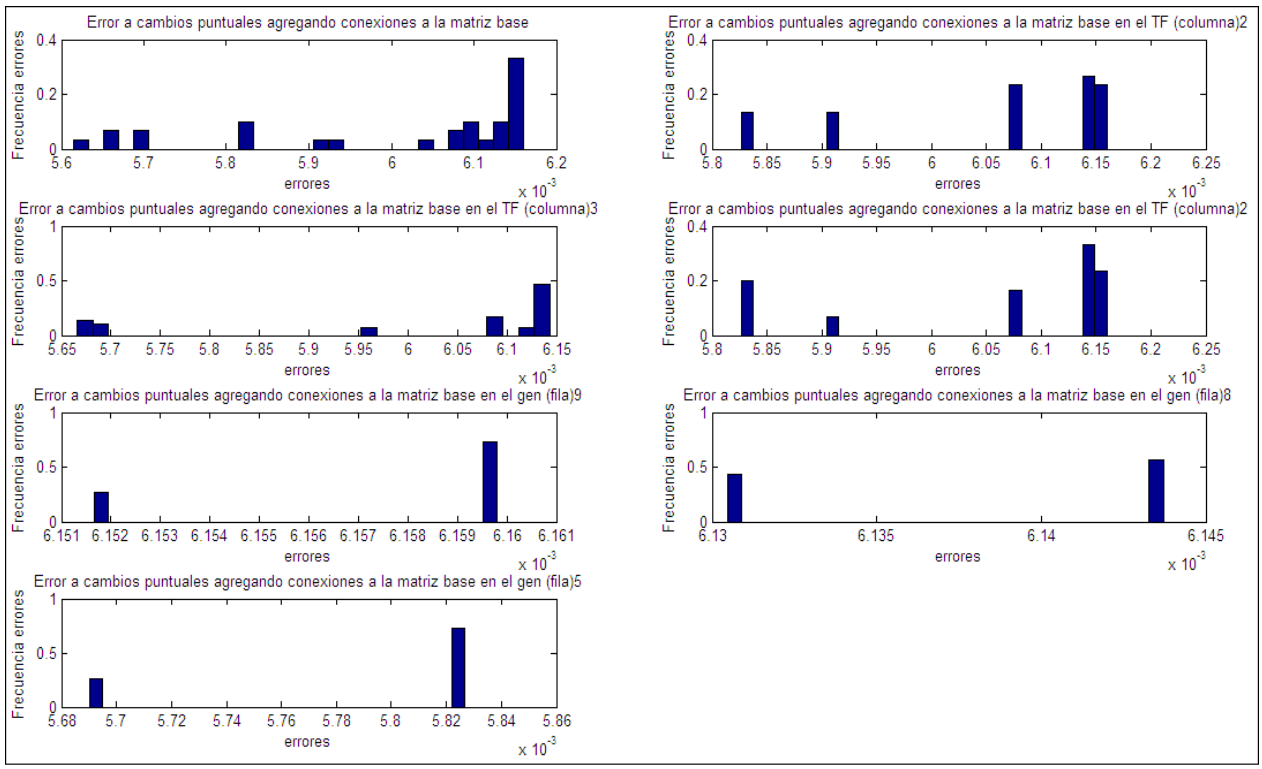


Figura 48: Distribución de los errores a cambios puntuales en la estructura con error en los datos. Agregar conexiones. Fuente: Elaboración propia.

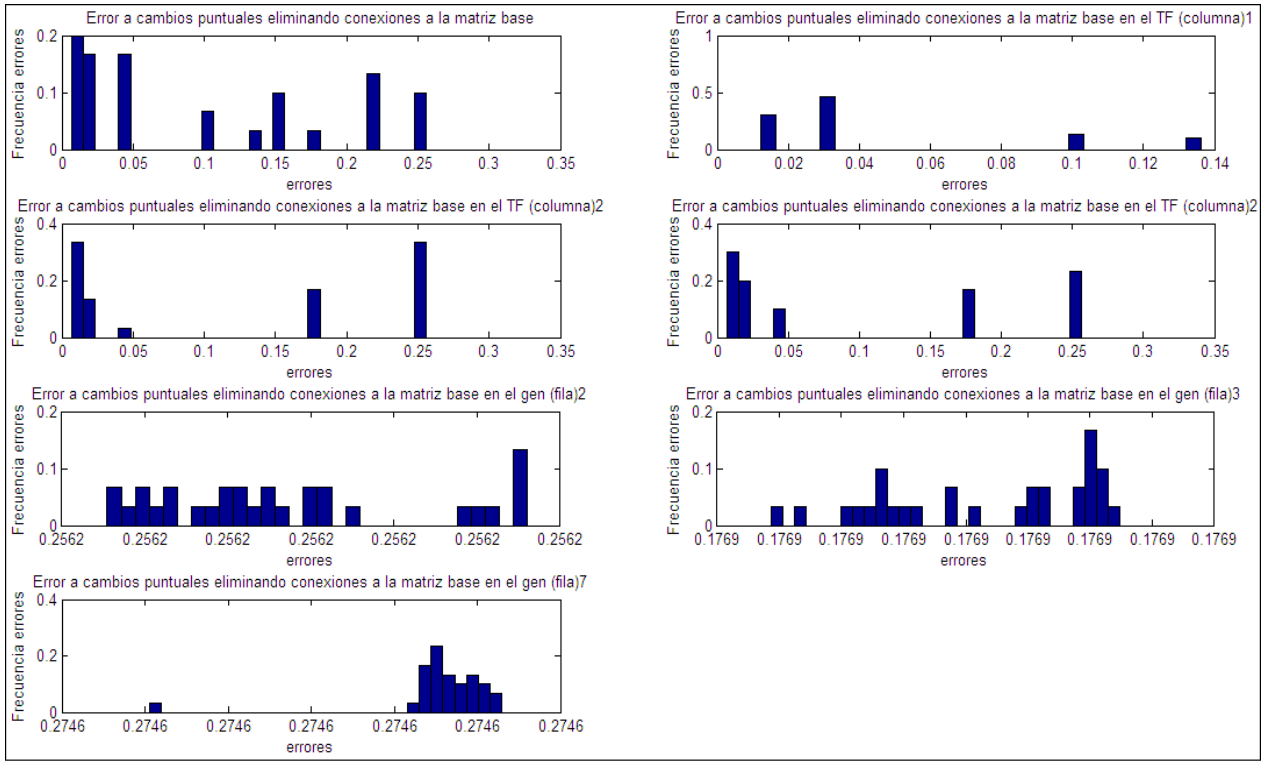


Figura 49: Distribución de los errores a cambios puntuales en la estructura con error en los datos. Eliminar conexiones. Fuente: Elaboración propia.

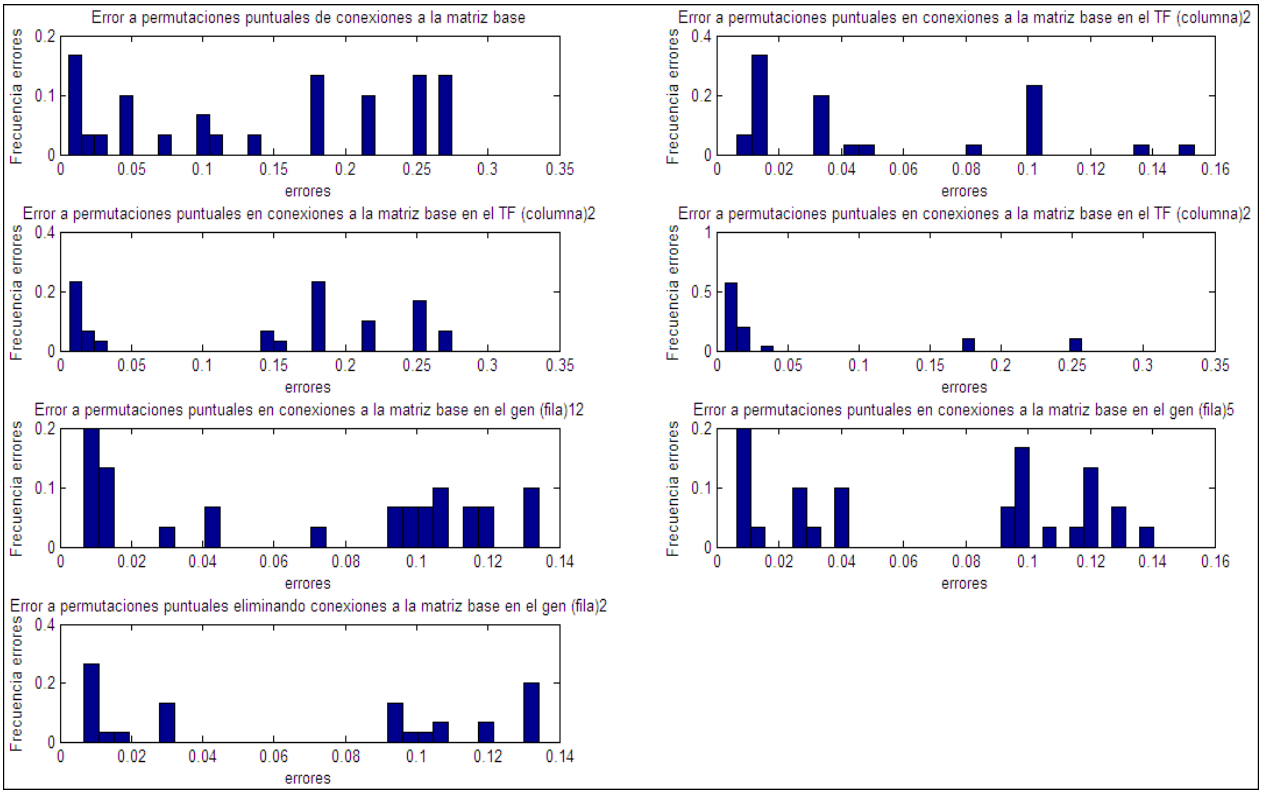


Figura 50: Distribución de los errores a cambios puntuales en la estructura con error en los datos. Permutar conexiones. Fuente: Elaboración propia.

Se observa cualitativamente un efecto similar al caso sin errores. La única diferencia radica en que los errores son mayores, producto de errores de los datos.

Finalmente se repite el procedimiento anterior para una red base distinta a la que originalmente genera los datos. Esto es, partir de una configuración de red errónea y analizar la distribución del error de reconstrucción a cambios puntuales de diferentes tipos. Los gráficos son resumidos a continuación:

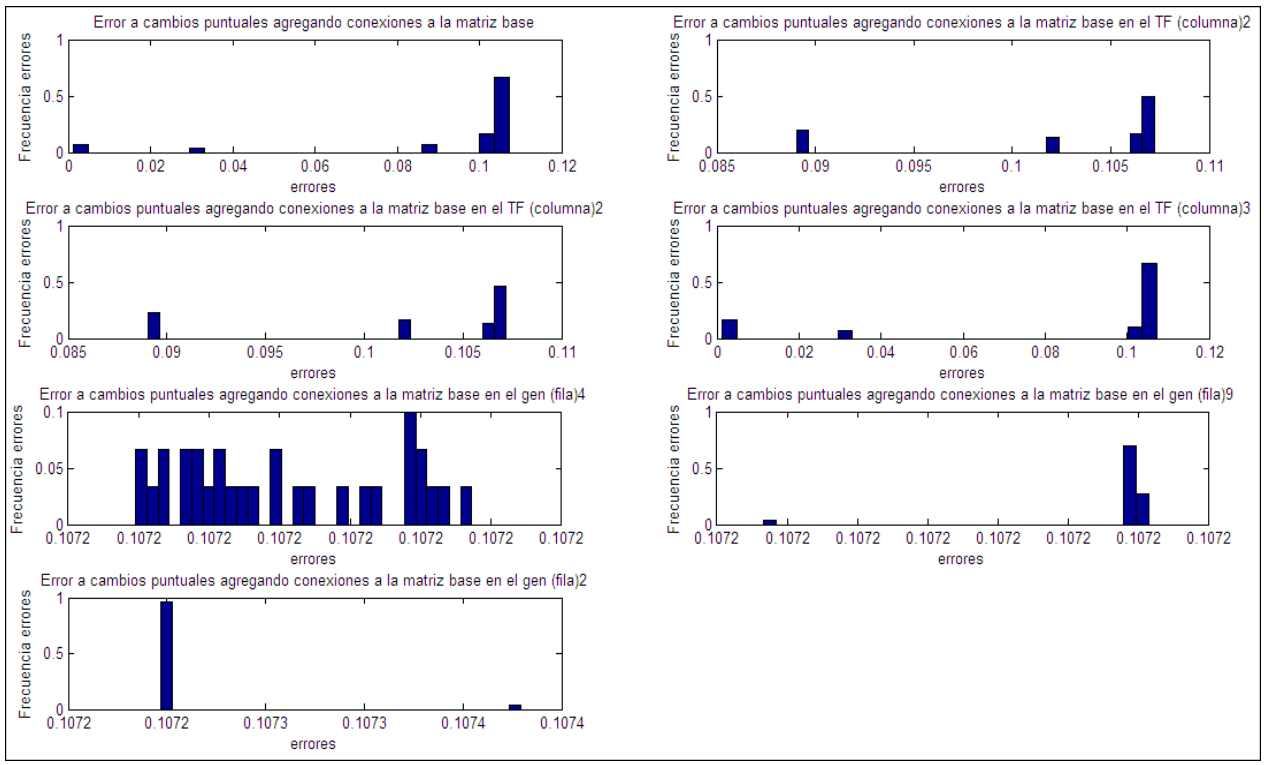


Figura 51: Distribución de los errores a cambios puntuales en la estructura con otra estructura base. Agregar conexiones. Fuente: Elaboración propia.

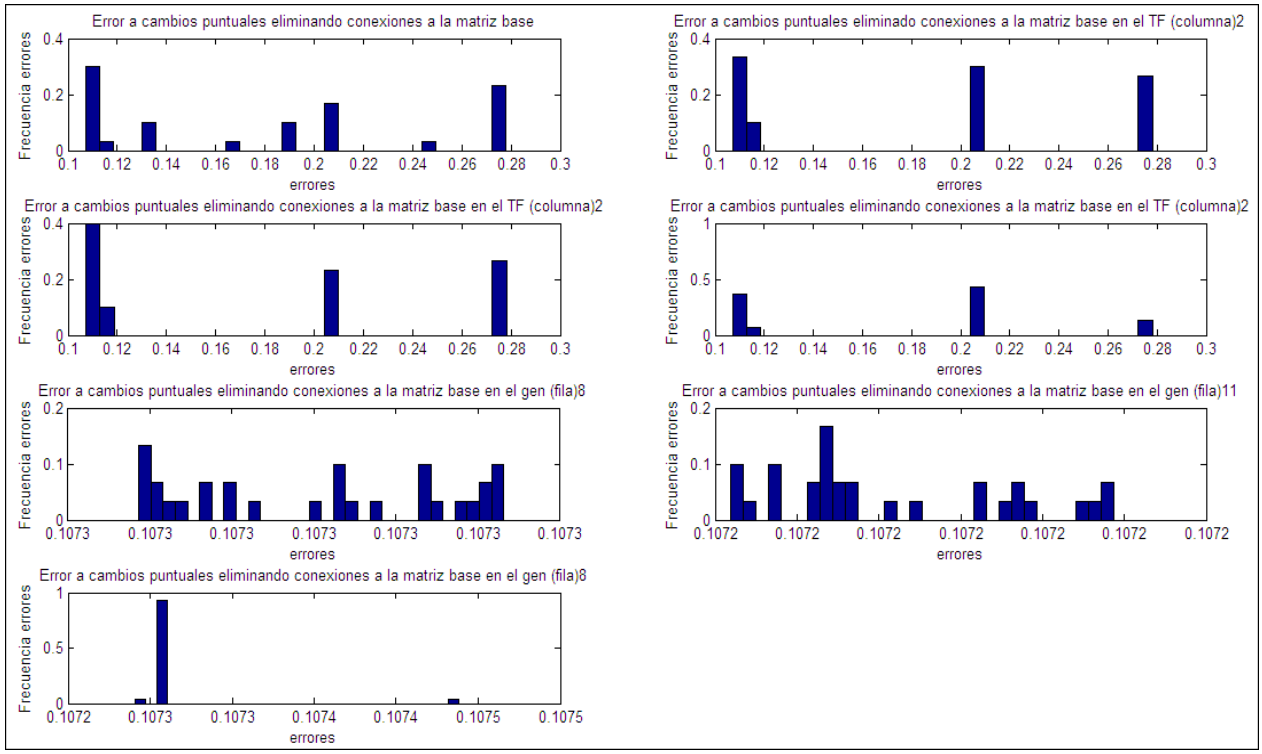


Figura 52: Distribución de los errores a cambios puntuales en la estructura con otra estructura base. Eliminar conexiones. Fuente: Elaboración propia.

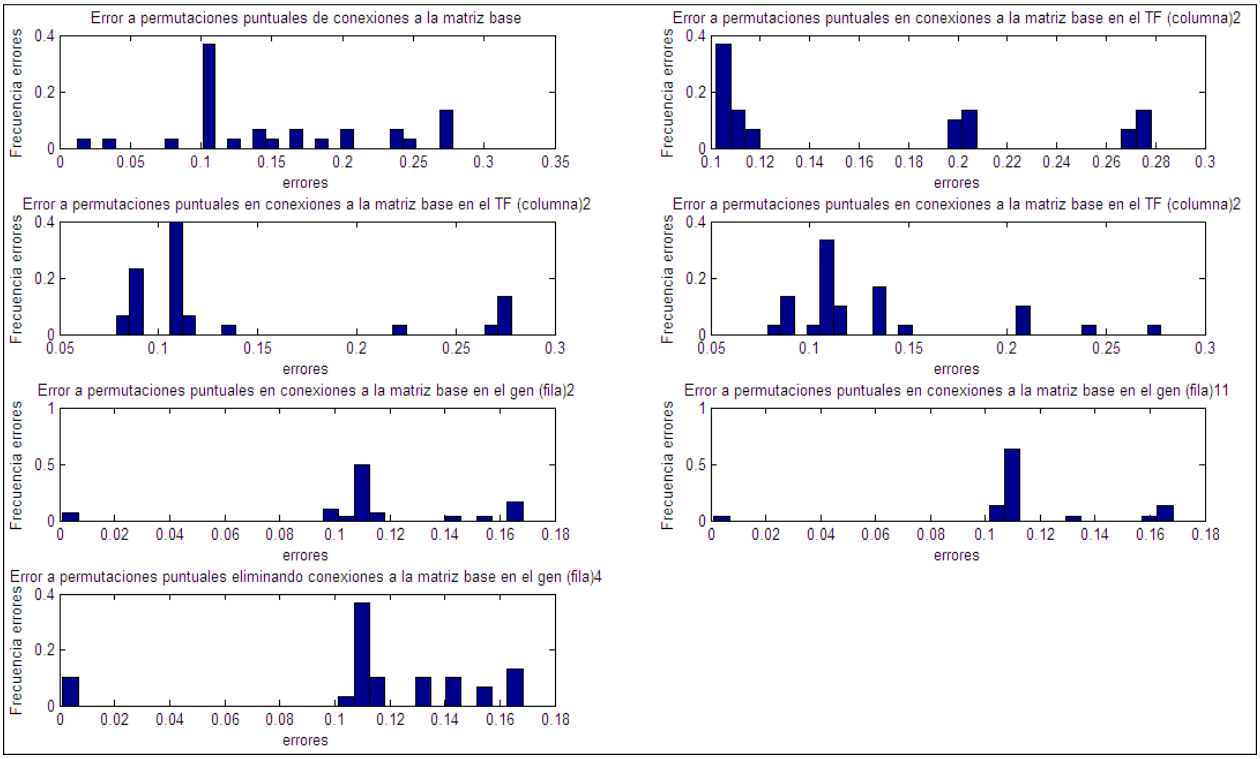


Figura 53: Distribución de los errores a cambios puntuales en la estructura con otra estructura base. Permutar conexiones. Fuente: Elaboración propia.

Como se puede apreciar, en especial en la distribución al agregar conexiones[86], se observa que parte de la distribución se ubica en niveles menores al original (alrededor de 0,1). Esto se explica por supuesto debido a que parte de las conexiones agregadas *"reparan"* conexiones no identificadas en la estructura base. Sin embargo, el efecto no se observa tan claro al momento de eliminar conexiones, y la explicación es simple, dado lo ya comentado y el conocimiento de las capacidades de NCA. La estructura que se perturba mediante modificaciones puntuales en la experiencia anterior posee principalmente 2 tipos de errores: conexiones de sobra, y conexiones inexistentes que no fueron consideradas. El primero de estos problemas no es del todo grave, ya que si estas no dominan la estructura considerada, NCA es capaz de llevar a cerca de cero el peso asignado en la reconstrucción y por ende corregir el problema. El segundo problema en cambio es bastante grave, y debe ser considerado. NCA no es capaz de hacer nada con conexiones no declaradas, ya que asume inequívocamente que dicha conexión no existe. Luego, eliminar conexiones con el fin de corregir conexiones mal consideradas como tal es redundante, en el sentido que se corre un gran riesgo: es posible eliminar conexiones correctas, incurriendo en un gran error por lo mismo, cuando en realidad la ganancia al eliminar conexiones mal consideradas no es contundente, en el sentido de que NCA puede manejar solo este tipo de conflictos. Esto es en efecto observado en la distribución del error al eliminar conexiones de la Figura 52, en donde no se observa una importante

[86] Dada las características de la red base modificada.

frecuencia en niveles de errores más bajos que el base. Agregar conexiones en cambio presenta un riesgo menor: si se asume una conexión falsa, NCA puede manejar la situación, mientras que la ganancia de asumir una conexión que antes no había sido considerada es considerable. La discusión anterior es válida también en el caso de las permutaciones.

La enseñanza parece ir por el lado de tener en extremo precaución a la hora de eliminar conexiones con el fin de modificar la configuración considerada.

3.18.5 Medición de errores entre estructuras

Continuando con el desarrollo de la teoría, es necesario definir forma de medir el error o diferencia que puede existir entre 2 estructuras propuestas, o al comparar una con la estructura original, que es lo que se utilizará en las pruebas sintéticas siguientes.

Sea $A_{o\_}$ la estructura[87] real de una red determinada y sea $A\_$ otra estructura propuesta para dicha red.

Se definen entonces las siguientes medidas de error y distancia:

1. Diferencia de conexiones. Matriz ΔA: Corresponde simplemente a la matriz diferencia de la estructura real y la propuesta (o 2 propuestas).

$$\Delta A = A_{o\_} - A\_$$

Ecuación 90

En la expresión anterior un valor 1 indicará una conexión ausente en $A\_$ relativa a $A_{o\_}$, mientras que un valor -1 indicará una conexión extra.

2. Distancia entre 2 estructuras. Δe: Corresponde a una suerte de métrica de las diferencias puntuales de 2 estructuras, entregada por la suma del valor absoluto de las entradas de ΔA. Esto es:

$$\Delta e = \Sigma[|\Delta A|]$$

Ecuación 91

, donde Σ representa al operador que suma las entradas de la matriz respectiva. Mientras mayor sea Δe mayor serán las diferencias entre 2 estructuras consideradas.

Otro punto interesante de recalcar está motivado en la discusión de la sección anterior. En ocasiones 2 estructuras pueden diferir solo de manera "*ficticia*". Esto es, tener alguna conexión de más, pero que NCA de reparar. Si no hay demasiadas conexiones no consideradas, el método de reconstrucción es capaz de llevar al peso respectivo cerca de cero, por lo que es posible generar una surte de distancia

[87] Una matriz binaria de unos y ceros que indique conectividad.

entre 2 estructuras corregida, que tome en cuenta lo anterior. En lo siguiente, se definirá $\widetilde{\Delta e}$ como la métrica de diferencia definida en la Ecuación 91 corregida por el efecto anterior.

3.18.6 Recocido Simulado NCA

Desarrollada la teoría anterior, es posible aplicar NCA a redes sintéticas como ya se ha realizado anteriormente. Con este objetivo se crea el método recocido_NCA_monot.m, que implementa en NCA el método de recocido simulado monótono descrito con anterioridad. Algunas salvedades respecto al método creado son las que se comentan. De existir algún cambio se aclarará en la experiencia respectiva.

- El proceso que actualiza la temperatura una vez alcanzado el equilibrio térmico corresponde al descrito por la Ecuación 86.

- Con el fin de actualizar el parámetro K una vez alcanzado el equilibrio térmico[88] se utiliza un proceso similar al propuesto para la temperatura. Esto es:

$$K_k = \beta \cdot K_{k-1}$$

Ecuación 92

  , donde k identifica las diferentes etapas de equilibrio térmico[89] y $\beta > 1$.

- Se tiene un número fijo de saltos objetivos en cada etapa de iteración.

- Se utiliza también un número fijo de iteraciones totales.

- Con el fin de permitir un análisis más profundo, el proceso guardará las 10 estructuras de red que produzcan menor error en el proceso de recocido simulado.

3.19 Pruebas sintéticas: Grupo 3

En este set de pruebas finales se pretende analizar el enfoque propuesto para NCA basado en la técnica heurística de recocido simulado. Como se ha comentado, la motivación detrás de esta nueva perspectiva tiene que ver con la gran cantidad de información, y por supuesto, la implicancia que tiene asumir inequívocamente esta información preliminar como correcta. El método pretende disminuir la cantidad de información necesaria, y explorar la posibilidad de generar nuevas redes que se ajusten de mejor manera a los datos, y sean por supuesto, interpretables en un sentido y plano biológico.

[88] Aumentarlo con el fin de permitir mayor número de iteraciones hasta alcanzar el próximo equilibrio.
[89] No el total de iteraciones, que puede ser mayor. En cada etapa existirán como máximo K iteraciones.

Un punto no mencionado al respecto es la suerte de "arte" que es requerido para calibrar la gran cantidad de parámetros que utiliza el método de recocido simulado adaptado a NCA. Es por eso que los resultados pueden variar de acuerdo a la combinación de situaciones exploradas, por lo que los resultados siguientes deben ser interpretados con esto en consideración. En cada caso se indica los parámetros particulares utilizados.

Por último, se aclara que lo siguiente no pretende ser exhaustivo. Simplemente se desea demostrar la técnica de recocido simulado básico, comentar algunos aspectos del mismo, y mencionar las principales ventajas, limitaciones y posibles aplicaciones del enfoque. Se deja abierta la posibilidad de explorar en más detalle los efectos en diferentes tipos de redes, bajo diferentes calibraciones, e incluso, utilizando diferentes métodos NCA.

3.19.1 Prueba simple con una red pequeña.

Experiencia 1: Matriz de datos sin error.

En esta caso se utiliza una red pequeña (12x3x4x80), similar a la que genera los gráficos de frecuencia de errores de la sección anterior. Se utilizó para eso, una estructura base con 2 errores puntuales respecto a la estructura real: una conexión extra, y una no considerada. El resumen de los parámetros utilizados es el siguiente.

Parámetro	Valor
$Iteraciones$	1000
T_o	2
$\propto$	0.99
K	20
β	1.1
A	10

Tabla 72: Detalle parámetros utilizados. Pruebas G3N1E1.
Fuente: Elaboración propia.

En la siguiente figura se representa el desarrollo del algoritmo, donde es posible apreciar esquemáticamente el proceso de recocido simulado adaptado a NCA. En rojo se aprecia la línea que marca el error de ajuste utilizando la estructura correcta, mientras que en verde se aprecia el error de ajuste para la estructura base utilizada. La curva representa el error de ajuste que se obtiene moviéndose a través de diferentes configuraciones en el algoritmo.

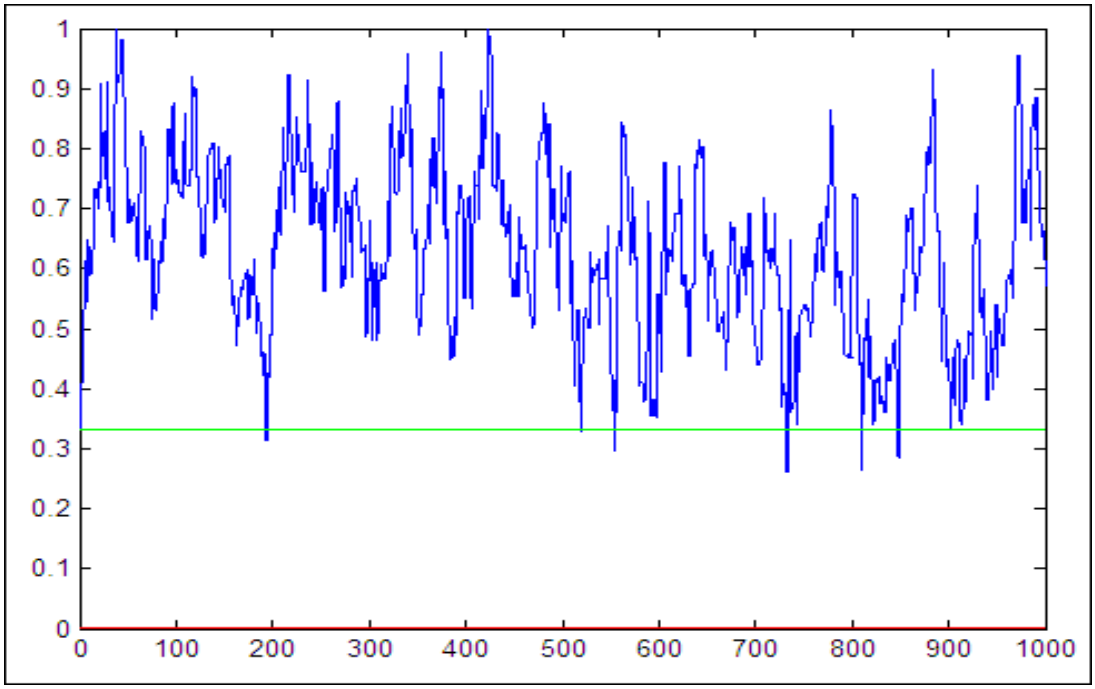


Figura 54: Visualización gráfica del algoritmo de recocido simulado 1. Pruebas G3N1E1.
Fuente: Elaboración propia.

Se observa, al menos en una primera impresión, que los resultados no son muy favorables. El algoritmo comenzó su viaje en la línea de ajuste verde, y luego de eso aumento considerablemente, manteniéndose en promedio sobre ese límite. Localmente se observan que algunas configuraciones alcanzaron un mejor ajuste, con errores de 0,262. Sin embargo, analizando dicha estructura, se observa que $\widetilde{\Delta e} = 18$, por lo que existen 18 diferencias puntuales entre dicha estructura y la real (en contraste a las 2 de la estructura inicial). El algoritmo es repetido una vez más utilizando los mismos parámetros, obteniéndose los resultados de la Figura 55, donde se puede observar cualitativamente el mismo grafico anterior con una ausencia de convergencia[90].

Una primera implicancia del desarrollo anterior tiene relación con la magnitud de $\widetilde{\Delta e}$ y el error de ajuste obtenido. No necesariamente es cierto, al menos al considerar errores de ajuste considerables, que un menor error de ajuste se relacione con menos diferencias puntuales entre 2 estructuras, por lo que habrá que considerar dicho fenómeno en la interpretación de los resultados. Se recalca de todas formas que las estructuras comparadas presentan ambas un gran error de ajuste, y lo anterior no necesariamente es válido para estructuras con errores de ajuste cercanos al real.

[90] Lógicamente la forma no es la misma dado que se trata de un método probabilístico.

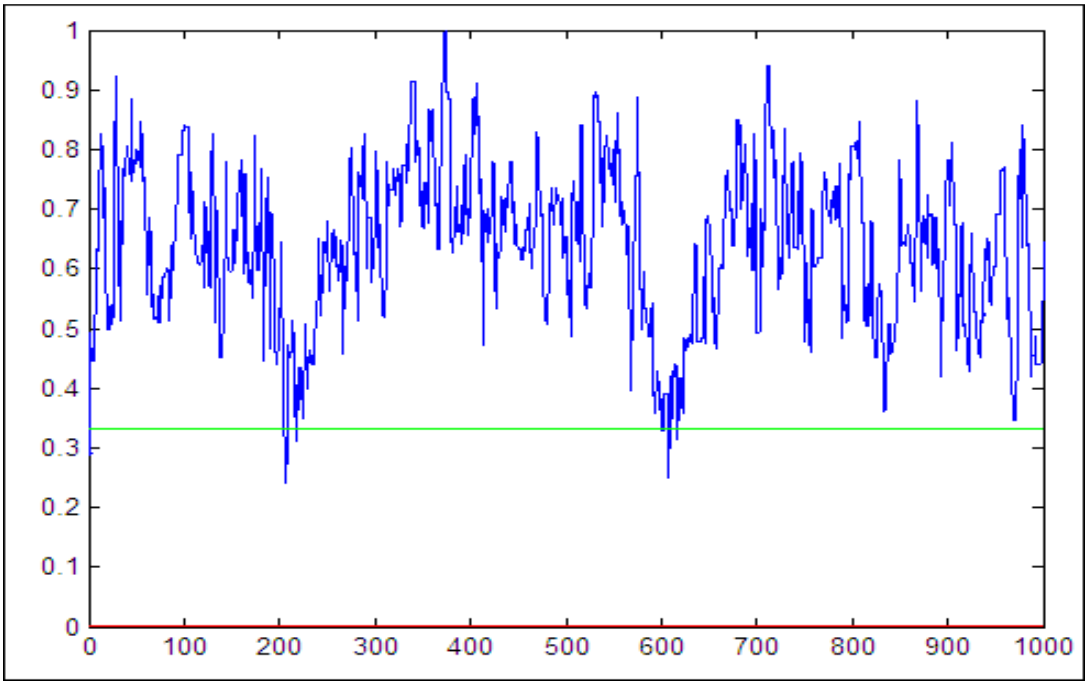


Figura 55: Visualización gráfica del algoritmo de recocido simulado 2. Pruebas G3N1E1.
Fuente: Elaboración propia.

Experiencia 2: Modificación de parámetros.

Como se ha mencionado, la técnica desarrollada tiene mucho de "*arte*", en especial lo que se refiere a ajustar y combinar de buena forma los parámetros para obtener una convergencia. Como una forma de desarrollar intuición al respecto se comentan los principales indicios que llevan a modificar los valores de los paráemtros. De la experiancia anterior se puede notar que el parámetro T_o es muy elevado, dado que la probabilidad de salto de un estado a otro demoestro ser alta incluso en etapas finales del algoritmo. De la misma manera el parámetro K aumenta con demasiada rapidez, por lo que se corrigen los puntos descritos.

Parámetro	Valor
$Iteraciones$	1000
T_o	0.8
$\propto$	0.99
K	10
β	1.01
A	7

Tabla 73: Detalle parámetros utilizados. Pruebas G3N1E2.
Fuente: Elaboración propia.

Los resultados del algoritmo se resumen en la figura.

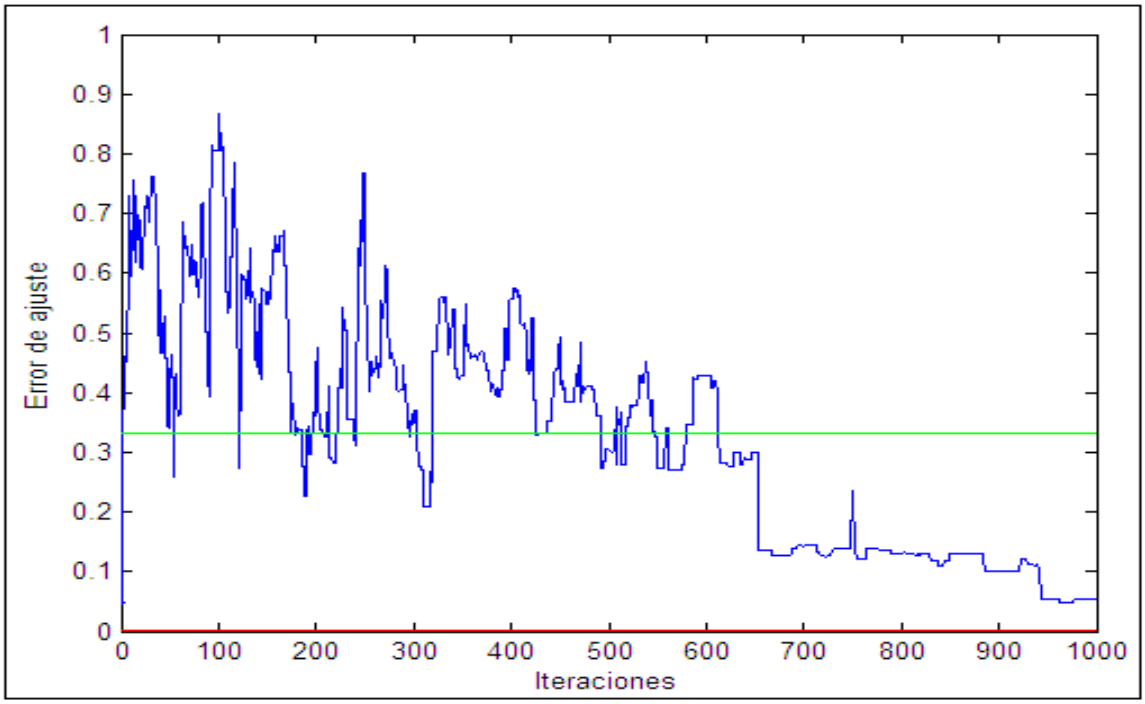


Figura 56: Visualización gráfica del algoritmo de recocido simulado 1. Pruebas G3N1E2.
Fuente: Elaboración propia.

En este caso, a diferencia del anterior, se aprecia claramente una convergencia, en donde dada las características del algoritmo, este se queda "estancado" en dicho nivel de error. Sin embargo, el nivel obtenido no alcanza el umbral de error de la estructura original. Analizando las estructuras que presentaron el menor error, es posible apreciar que se obtienen 5 diferencias puntales corregidas, de las cuales 2 corresponden a conexiones omitidas. Esto es una clara mejora respecto al punto anterior, pero de la misma manera se destaca que la estructura que genera el error en verde presenta tan solo 2 diferencias puntuales.

El siguiente gráfico muestra os resultados al variar levemente los parámetros de la heurística.

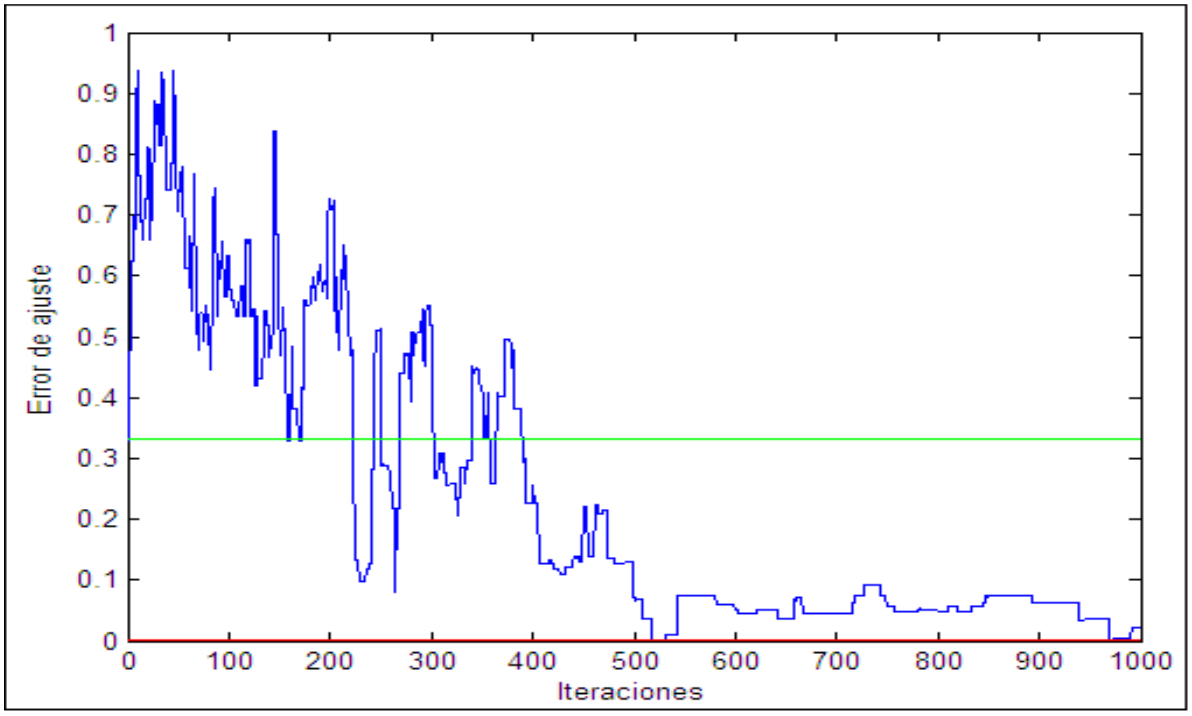


Figura 57: Visualización gráfica del algoritmo de recocido simulado 2. Pruebas G3N1E2.
Fuente: Elaboración propia.

Como se observa, la convergencia una vez más fue alcanzada en un punto bastante inferior al error de la estructura base considerada. Sin embargo, y tal cual demuestra un análisis más profundo de las matrices que conducen a los errores más bajos obtenidos (cercanos a 10^{-5}), la métrica ajustada es del orden de 15. Luego, y dada la cantidad de entradas de la matriz de conectividad, se podría estar hablando de redes completamente diferentes. Lo primero que hay que tener en cuenta es que pese a tratarse de experiencias sintéticas, lo que se está buscando es la posibilidad de explorar nuevas configuraciones posibles, que de alguna forma se ajusten bien a los datos. No se debe olvidar que el método planteado es una herramienta matemática, y como tal, se debe tener la precaución de interpretar los resultados mediante un enfoque adecuado. Es interesante de analizar sin en redes reales las configuraciones alternativas generadas (estén o no cerca de la propuesta) tienen algún sentido físico o biológico, o más aun, tienen más o menos sentido que las que se estaba considerando. En segundo lugar, y dado el bajo error obtenido en el ajuste de las redes alternativas, se sospecha que efectivamente la red encontrada podría tratarse de una transformación de la original, tal cual plantea el teorema 2 descrito. Según la teoría NCA, es posible hablar de redes indistinguibles cuando 2 configuraciones diferentes entregan el mismo residuo dada una matriz de datos, o lo que es equivalente, cuando existe una matriz invertible que las relaciona. En este caso, efectivamente ambas redes resultan ser indistinguibles, lo que explica el gran ajuste de ambas pese a sus considerables diferencias.

Finalmente es necesario recalcar que los grados de libertad del problema son aun elevados, por lo que es esperable resultados como los anteriores, e incluso, una no convergencia al utilizar alguna combinación de parámetros.

Experiencia 3: Error en los datos.

La experiencia anterior es repetida considera los mismos parámetros y un 20% de error porcentual medio en los datos. Los resultados son los siguientes.

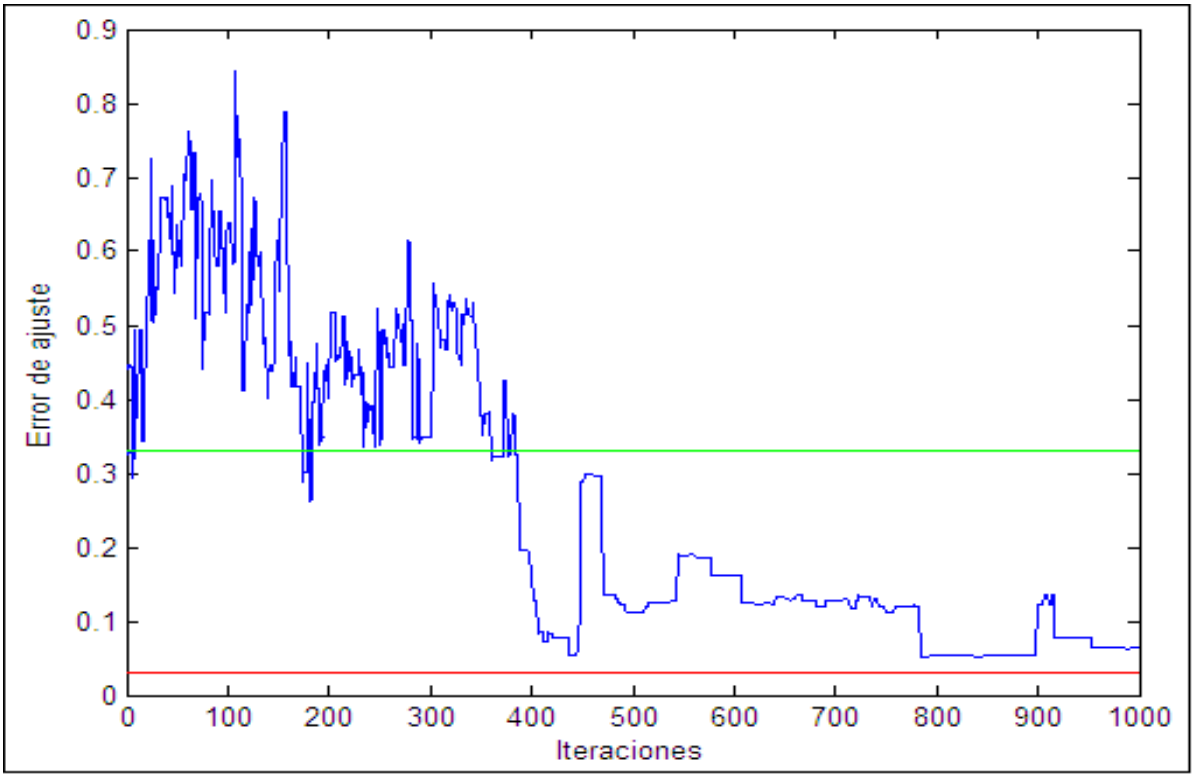


Figura 58: Visualización gráfica del algoritmo de recocido simulado 2. Pruebas G3N1E2.
Fuente: Elaboración propia.

Previo a la discusión de los resultados, es interesante destacar que el error de ajuste es más elevado que en el caso anterior, por ende es difícil pretender que se llegue a un mínimo de error cercano a cero. En teoría, el piso de dicho ajuste debería corresponder a la línea roja mostrada (que en las experiencias anteriores coincidía casi con el cero). Una pregunta directa que surge al respecto es preguntarse si efectivamente esto es así. Si se analiza con calma, y se piensa en los resultados que entrega NCA en las experiencias previas (principalmente lo referido al buen ajuste de los datos a la matriz real, pese a los errores de medición), lo anterior sería válido, y la técnica de recocido simulado sería aplicable aun en presencia de errores en los datos. Luego, se piensa es difícil encontrar una red alternativa que produzca los resultados con errores, y por ende se alcance un residuo cercano a cero. Si bien lo anterior no se

comprueba con certeza en el presente estudio, diferentes repeticiones del mismo experimento parecen avalar dicho comportamiento.

En la simple experiencia anterior se obtiene una matriz de conexión con 12 diferencias puntuales corregidas ($\widetilde{\Delta e} = 12$), lo que es de esperarse ya que incluso gráficamente se obtiene un error menor al umbral. Es posible comprobar también que al variar los parámetros de las iteraciones, es posible converger a un error muy cercano al umbral, y con menores diferencias puntuales.

La discusión de las experiencias anteriores parece entrever que no necesariamente se convergerá a la red que originalmente generó los datos. Si bien para las experiencias puntuales esto puede ser cierto, no es la regla. En las experiencias realizadas se lograron iteraciones que efectivamente convergieron a la red original, por lo que el resultado dependerá mucho de los parámetros del método utilizados y del tipo de red estudiada. Además es necesario recalcar la gran cantidad de grados de libertad que se le da a las configuraciones permitidas de las redes anteriores. Si bien no se desarrolla en este documento, tal cual se mencionó en apartados previos, es posible combinar las experiencias anteriores con otras fuentes de información. La posibilidad de vetar algunas entradas se torna en una herramienta de gran potencial, en cuanto permitiría dirigir de mejor manera el método hacia una convergencia exitosa. Por supuesto dichas entradas estarían asociadas a mecanismos ampliamente conocidos e identificados, por lo que una vez más la presencia y utilización de información a priori se torna relevante. Es incluso posible vetar solo parcialmente algunas entradas (análogo a la confiabilidad de las suposiciones anteriormente utilizadas), lo que abre incluso un nuevo enfoque de estudio. Por otro lado, la posibilidad de obtener redes alternativas que representen a los datos medidas se torna una herramienta de gran interés analítico, en especial cuando el organismo o el mecanismo estudiado son escasamente conocidos.

Conclusiones

El análisis de sistemas biológicos mediante un enfoque de biología de sistemas, y el consecuente uso de herramientas matemáticas y computacionales para la modelación y simulación de los mismos, es fundamental si se desea tener mayor información y conocimiento de los procesos biológicos. La posibilidad de entender de manera global el fenómeno de regulación transcripcional en organismos vía el modelo propuesto es de suma importancia. Pero el modelamiento por si solo de redes de tal envergadura no es suficiente, por lo que el uso de NCA como herramienta de análisis resulta fundamental. NCA destaca no sólo por presentarse como un trasfondo mediante el cual es posible modelar este fenómeno, si no por su capacidad para ayudar a entender mejor los mecanismos de regulación involucrados en un organismo, pudiendo en cierta forma reconstruir los procesos que en éste se producen. La capacidad de NCA permite desentrañar parcialmente un proceso en función de las observaciones de los resultados que éste produce.

En el presente documento se han resumido los esfuerzos por comprender las bases de esta técnica, y en base a la experiencia del autor y a recomendaciones de expertos de diversas áreas se han creado métodos que buscan en parte suplir deficiencias funcionales identificadas en NCA y en otras herramientas de análisis. En la investigación aplicada, especialmente en el campo biológico, es fundamental contar con herramientas versátiles, que permitan una aplicación flexible y robusta y que tengan fundamentos biológicos. Esto último es en extremo importante, y quizás una de las principales falencias identificadas en aplicaciones desarrolladas por no-biólogos. Si bien los enfoques pueden ser interesantes, en muchas ocasiones su aplicabilidad en la biología se ve seriamente limitada debido a las restricciones ajenas a dicho ámbito (por ende las conclusiones obtenidas pueden ser solamente interpretadas considerando dichos errores de planteamiento).

Lo que se intenta en el presente documento es extender una técnica de alta potencia en el análisis biológico haciendo plausible nuevas interrogantes y necesidades. Es así como se propone, por ejemplo, hacer uso de toda la información estadística existente en los datos, incluso la varianza de las mediciones, como una forma de identificar posibles errores de medición (muy comunes por cierto en esta área) y mejorar así los resultados del método. De la misma manera, la posibilidad de incluir suposiciones en los parámetros buscados extiende en gran parte el potencial de reconstrucción, en el sentido que se combina de buena forma la información procedente de los datos con una fuente de información diferente. La experiencia y bagaje del investigador pueden ser plasmadas en los datos, a fin de reconstruir una red que combine ambos aspectos de la problemática. En el cuerpo del documento se presentan además diferentes pruebas que buscan validar el funcionamiento de los métodos propuestos, a fin de guiar las investigaciones futuras.

Un punto importante a recalcar es respecto a la interpretación de los resultados. Como modelo, corresponden a aproximaciones de la realidad y los resultados deben ser analizados con dicho prisma. De la misma manera, se puede notar que los resultados de la reconstrucción son más bien cualitativos, relegando el aspecto numérico a una suerte de relación relativa (nunca absoluta) de las variables. De todas maneras, el modelo utilizado para una red particular puede ser refinado tanto como se quiera, en

la medida que el investigador sea capaz de obtener información de calidad respecto a los reguladores existentes en un organismo, a sus genes, y a una relación al menos parcial de estos elementos.

¿Pero cuál es el objetivo de reconstruir una red de regulación? El objetivo final de analizar la red mediante el enfoque propuesto es utilizar dicha información. Una vez reconstruida una red es posible realizar predicciones para dicho organismos, de forma tal de entrever el comportamiento de los reguladores frente a diferentes condiciones. El potencial es por cierto, como ya se ha comentado, preponderante, sobre todo en aplicaciones biológicas y de ingeniería genética.

Respecto al uso de los métodos planteados y la teoría creada en investigaciones formales, se recomienda una revisión más acabada con experiencias sintéticas, y la validación con una red biológica y datos reales. El primer punto es dirigido, en primer lugar, a repetir las experiencias ya realizadas e identificar posibles errores cometidos, así como a validar las conclusiones obtenidas bajo otra mirada y experiencia. Lo segundo es en extremo necesario y la prueba final y necesaria de lo desarrollado. Es necesario recalcar además la gran cantidad de información necesaria para dicha tarea, en el sentido de conseguir en conjunto a los datos, información respecto a su variabilidad. Si es posible se aconseja la realización de experimentos propios de microarrays, que puedan ser replicados en tríos, controlando las condiciones para obtener los datos precisos. Como ya se ha mencionado, es indispensable además el uso de súper procesadores para llevar a cabo esta tarea.

Un punto transversal a lo anterior y que es desarrollado en la última sección del documento es el enfoque de recocido simulado propuesto. Motivado por la principal falencia de NCA, correspondiente a la gran cantidad de información a priori necesaria, y el gran compromiso de validez inequívoca que es necesario darle a dicha información, el algoritmo busca explorar nuevas topologías que se ajusten de mejor manera a los datos. Si bien es necesario continuar con pruebas en esta línea, testeando nuevos enfoques como los comentados en las discusiones correspondientes, se espera un gran potencia a dicha aplicación. La capacidad de partir sólo de aproximaciones de la red real y obtener redes alternativas que puedan ser interpretadas a la luz de la experiencia e información adicional es de gran relevancia, por lo que se recomienda de sobremanera continuar la investigación y maduración del método.

Finalmente, es necesario destacar que el enfoque NCA no está sólo limitado al campo biológico. La modelación vía una relación log-lineal realizada para el proceso de interés es una de las más básicas y extendidas (pero no por eso inexacta), que combina la simpleza y la sofisticación en forma bastante equilibrada. Luego, NCA puede ser aplicado a problemas de otras áreas de las ciencias, siempre y cuando los fenómenos en estudio puedan ser modelados de forma similar a la presentada: Una red de dos partes con una relación log-lineal entre las mismas. El enfoque de “atractores” muy comentado en la literatura, en donde ciertas piezas de proceso parecen dirigir y controlar la dinámica de un proceso global, es una buena aproximación que es posible encontrar en muchos problemas.

Referencias

1. Gonzalez MA. Estructura y dinámica de redes genéticas. *Gacetas Biomédicas.* UNAM; 2007.

2. Mundo B. Los Secretos del olfato. http://news.bbc.co.uk/hi/spanish/science/newsid_3715000/3715024.stm.[En línea] [Consulta: 5 Marzo 2010].

3. Gámez LA. EL SER HUMANO TIENE MENOS DE 40.000 GENES, EL DOBLE QUE UNA MOSCA O UN GUSANO *El Escéptico Digital*. http://digital.elesceptico.org/leer.php?autor=155&id=1415&tema=30.[En línea] [Consulta: 15 Marzo 2010].

4. Lewis B. *Genes IX*, 9 edn. Mcgraw-hil, 2008, 892pp.

5. Kitano H. *Foundations of Systems Biology*. The MIT Press, 2001.

6. Complejidad genética. http://www.biotech.bioetica.org/ap6.htm.[En línea] [Consulta: 22 Noviembre 2009].

7. Creces. La creciente complejidad del programa genético. http://www.creces.cl/new/index.asp?imat=%20%20%3E%20%207&tc=3&nc=5&art=1827.[En línea]. [Consulta: 18 Marzo 2010].

8. Eucariotas: ¿Cómo controlan la expresión génica? http://www.fcv.unlp.edu.ar/sitioscatedras/87/material/Control%20de%20la%20Expresion%20Genica%20en%20Eucariotas.pdf. [En línea]. [Consulta: Octubre 2009].

9. Lee TI, Rinaldi NJ. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002; **298:** 799-804.

10. Pollack JR, Iyer VR. Characterizing the physical genome. *Nat Genet* 2002; **32:** 515-521.

11. Albert B, Bray D, Lewis J. Biología Molecular de la Célula. In: Omega E (ed). 2002.

12. Control de la expresión genética. http://themedicalbiochemistrypage.org/spanish/gene-regulation-sp.html. [En línea]. [Consulta: Marzo 2010].

13. Regulación de la expresión genética en procariontes. Available at: http://www.ucm.es/info/genetica/grupod/Operon/Operon.htm.

14. Regulación Genética. http://www.ugr.es/~eianez/Microbiologia/15regulacion.htm. [En línea]. [Consulta: Abril 2010].

15. Knudsen S. *A biologist's guide to Analysis of DNA microarrays data*. Wiley-Liss, 2002.

16. Lockhart DJ, Dong H, Byerne MC. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996; **14:** 1675-1680.

17. Knudsen S. *A Biologist's guide to Analysis of DNA microarray data*, 2002 edn., 2002.

18. Sanjuan MA. Más sibre Biología de Sistemas. http://www.madrimasd.org/blogs/complejidad/2006/06/09/29137. [En línea]. [Consulta: Marzo 2010].

19. Lopez M, Romero GR. Biología de Sistemas. Informe de vigilancia tecnológica. http://www.gen-es.org/12_publicaciones/docs/pub_76_d.pdf. [En línea]. [Consulta: Abril 2010]

20. M. A, Balleza E, Kauffman S. Robustness and evolvability in genetic regulatory networks. *Science* 2006; **245**(433-448).

21. Palsson B. *Systems Biology: Properties of Reconstructed Networks*. 2006, 334pp.

22. M. E, Elf J, Aurell E. Systems Biology is Taking Off. *Genome Res* 2003; **13**(2377-2380).

23. Problema Inverso. http://www.worldlingo.com/ma/enwiki/es/Inverse_problem. [En línea]. [Consulta: Marzo 2010].

24. Tarantola A. *Inverse Problem Theory*. SIAM, 2005.

25. De Jong H. Modeling and simulation of genetic regulatory systems: a literature reviews. *J Comput Biol* 2002; **9:** 67.

26. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A* 2003; **100:** 15522-15527.

27. Liebermeister W. Linear modes of gene expression determined by independent component analysis. *Bioinformatics* 2002; **18:** 51-60.

28. Raychaudhuri S, Stuart JM. Principal component cnalysis to summarize microarrays experiments: aplication to sporulation time series. *Pac Symp Biocomput* 2000**:** 455-466.

29. Poyatos J. Introducción a la Biología de Sistemas. http://bioinfo.cnio.es/jpoyatos/ [En línea]. [Consulta: Febrero 2010].

30. Yang YL, Suen J. Inferring yeast cell cycle regulators and interactions using transcription factor activities. *BMC Genomics* 2005; **6:** 6-90.

31. Almeida JS, Voit EO. Neural-network-based parameter estimation in s-system models of biological networks. *Genome Inform* 2003; **14:** 114-123.

32. Bird RB, Stewart WE. *Transport Phenomena.* Wiley: New York, 2002.

33. Tran LM, Brynildsen MP, Kao KC, Suen JK, Liao JC. gNCA: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. *Metab Eng* 2005; **7**(2)**:** 128-141.

34. Kao KC, Yang YL, Boscolo R, Sabatti C, Roychowdhury V, Liao JC. Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis. *Proc Natl Acad Sci U S A* 2004; **101**(2)**:** 641-646.

35. Greene W. *Análisis Econométrico*, 3 edn. Pearson, 2004.

36. Babu MM. Computational approaches to study transcriptional regulation. *Biochemical Society Transactions* 2008; **36:** 758-765.

37. Lee TI. Transcriptional Regulatory Networks in Saccharomyces cerevisiae. *Science* 2002; **298:** 799.

38. Spies RD, Temperinini KG. Un método de molificación para resolver problemas inversos mal condicionados: aplicaciones. *MECOM 2005 - VII Congreso*. 2005.

39. Benavente JM, Otero A, Vasquez J. *Econometría I*. 2004.

40. MathWorks - MATLAB ans Simulink for Technical Computing. http://www.mathworks.com/. [En línea]. [Consulta: Abril 2010]

41. Wolfram Research: Mathematica, software científico y técnico. 2010.

42. Perez C. *MatLab y sus aplicaciones en las Ciencias y la Ingeniería*. Pearson, 2003.

43. Pratap R. *Getting Started with MatLab: A quick introduction for Scientists and Engineers*, 6 edn. Oxford University Press: New York, 2002.

44. Yang WY, Cao W, Chung T, Morris J. *Applied Numerical Methods using MATLAB*. WILEY, 2005.

45. Tikhonov AN, Arsenin VY. *Solutions of ill-Posed Problems.* Winstons & Son: Washington, DC., 1977.

46. Hansen PC, O'Leary DP. The use of L-curve in the regularization of discrete ill-posed problems. *SIAM J Sci Comput* 1993; **14:** 1487-1503.

47. Chip on CHIP. Available at: http://www.chiponchip.org/. [En línea]. [Consulta: Mayo 2010].

48. Yeast Proteome Database. Available at: www.proteome.com/databases/.[En línea]. [Consulta: Mayo 2010].

49. Costanzo MC. *Nucleic Acids Res* 2000; **28:** 73.

Anexos

Anexo 1: Demostración de teoremas

Teorema 1: Solución esencialmente única en gNCA

Extraído de Liao J.C. (2005)

Dado el siguiente problema.

$$\mathbb{P} = \min_{A,P} \; J(E - A \cdot P) = \; J(\Gamma)$$

$$s.t. \;\; A \in Z_A \;\; y \;\; P \in Z_P$$

Dada la matriz E, y ciertas restricciones a la estructura de la red resumidas por los conjuntos Z_A y Z_P, las condiciones necesarias y suficientes para una descomposición esencialmente única de acuerdo a $\mathbb{P}$ son:

1. $A \in Z_A$ tiene rango columna completo.
2. Cada matriz reducida de G, G_{rj} $(j = 1, \dots, L)$ posee rango $L - 1$.
3. $P \in Z_P$ tiene rango fila completo.

Demostración

Dado A, P, $\bar{A}$ y $\bar{P}$ tal que:

$$\mathrm{E} - \Gamma = \mathrm{A} \cdot \mathrm{P} = \overline{A} \cdot \overline{P}$$

Se quiere demostrar que existe una matriz invertible X tal que se cumple lo siguiente:

$$\bar{A} = A \cdot X^{-1}$$

$$\bar{P} = X \cdot P$$

Y más aún, es necesario demostrar que dicha matriz invertible es una matriz diagonal.

Debido a la condición 1, es posible escribir la siguiente expresión:

$$\bar{A}^t A \; P = \bar{A}^t \bar{A} \; \bar{P}$$

$$\bar{A}^t A \; P = \bar{A}^t \bar{A} \; \bar{P}$$

$$\bar{P} = (\bar{A}^t \bar{A})^{-1} \bar{A}^t A \; P \equiv XP$$

, donde:

$$X = (\bar{A}^t \bar{A})^{-1} \bar{A}^t A$$

O equivalentemente:

$$\bar{P}^t = P^t X^t$$

Ecuación 93

Utilizando lo anterior:

$$AP = \bar{A}\, XP$$

O lo que es lo mismo:

$$(A - \bar{A}\, X)P = 0$$

Debido al rango completo de P (condición 3), lo anterior implica que:

$$A = \bar{A}\, X$$

Ecuación 94

Luego, es necesario demostrar que X puede ser sólo diagonal si es que se cumple la condición 2. La ecuación anterior puede ser escrita convenientemente en la siguiente forma, en la que se han agrupado los diferentes sistemas para cada columna de A:

$$\begin{bmatrix} A_{c1} \\ A_{c2} \\ \vdots \\ A_{cL} \end{bmatrix} = \begin{bmatrix} \bar{A} & 0 & & 0 \\ & \bar{A} & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \bar{A} \end{bmatrix} \cdot \begin{bmatrix} X_{c1} \\ X_{c2} \\ \vdots \\ X_{cL} \end{bmatrix}$$

Ecuación 95

, donde M_{cj} es la columna j de la matriz M. De forma similar, es posible escribir la Ecuación 93 como:

$$\begin{bmatrix} \bar{P}^t{}_{c1} \\ \bar{P}^t{}_{c2} \\ \vdots \\ \bar{P}^t{}_{cL} \end{bmatrix} = \begin{bmatrix} P^t & 0 & & 0 \\ & P^t & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & P^t \end{bmatrix} \cdot \begin{bmatrix} X^t{}_{r1} \\ X^t{}_{r2} \\ \vdots \\ X^t{}_{rL} \end{bmatrix}$$

Ecuación 96

Donde $\bar{P}^t{}_{cj}$ es la columna j de la matriz $\bar{P}^t$, y $X^t{}_{rj}$ es la fila j de X. Con el fin de combinar las 2 expresiones anteriores, es posible definir una matriz de permutación $K \in M^{L^2 x L^2}$, tal que:

$$\begin{bmatrix} X_{c1} \\ X_{c2} \\ \vdots \\ X_{cL} \end{bmatrix} = K \cdot \begin{bmatrix} X^t{}_{r1} \\ X^t{}_{r2} \\ \vdots \\ X^t{}_{rL} \end{bmatrix}$$

, donde:

$$K = \begin{bmatrix} K_{11} & 0 & & K_{1L} \\ K_{21} & K_{22} & & K_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ K_{L1} & L_{L2} & \dots & K_{LL} \end{bmatrix}$$

En lo anterior, las matrices K_{ij} son matrices de tamaño LxL. Notar que K es una matriz de permutación simétrica y unitario, por lo que: $I = K^{-1}K = K^tK = K\,K$.

Con lo anterior es posible escribir la Ecuación 96 como:

$$\begin{bmatrix} \bar{P}^t{}_{c1} \\ \bar{P}^t{}_{c2} \\ \vdots \\ \bar{P}^t{}_{cL} \end{bmatrix} = \begin{bmatrix} P^t & 0 & & 0 \\ & P^t & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & P^t \end{bmatrix} K\,K \begin{bmatrix} X^t{}_{r1} \\ X^t{}_{r2} \\ \vdots \\ X^t{}_{rL} \end{bmatrix}$$

$$\begin{bmatrix} \bar{P}^t{}_{c1} \\ \bar{P}^t{}_{c2} \\ \vdots \\ \bar{P}^t{}_{cL} \end{bmatrix} = \begin{bmatrix} P^t & 0 & & 0 \\ & P^t & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & P^t \end{bmatrix} K \begin{bmatrix} X_{c1} \\ X_{c2} \\ \vdots \\ X_{cL} \end{bmatrix}$$

Notar que utilizando la matriz K y la notación $Q(v)$ definida como en el documento, la ecuación anterior se reduce a:

$$\begin{bmatrix} \bar{P}^t{}_{c1} \\ \bar{P}^t{}_{c2} \\ \vdots \\ \bar{P}^t{}_{cL} \end{bmatrix} = \begin{bmatrix} Q(P^t{}_{c1}) & Q(P^t{}_{c1}) & \dots & Q(P^t{}_{c1}) \end{bmatrix} \cdot \begin{bmatrix} X_{c1} \\ X_{c2} \\ \vdots \\ X_{cL} \end{bmatrix}$$

Combinando esto con la Ecuación 95:

$$\begin{bmatrix} \bar{P}^t{}_{c1} \\ \bar{P}^t{}_{c2} \\ \vdots \\ \bar{P}^t{}_{cL} \end{bmatrix} = \begin{bmatrix} \bar{A} & 0 & \dots & 0 \\ 0 & \bar{A} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & 0 & \dots & \bar{A} \\ Q(P^t{}_{c1}) & Q(P^t{}_{c1}) & \dots & Q(P^t{}_{c1}) \end{bmatrix} \cdot \begin{bmatrix} X_{c1} \\ X_{c2} \\ \vdots \\ X_{cL} \end{bmatrix}$$

O equivalentemente, y de acuerdo a la definición de G:

$$\begin{bmatrix} A_c \\ \bar{P}^t{}_c \end{bmatrix} = G \cdot X_c$$

Ecuación 97

Es necesario notar que los elementos no nulos del lado izquierdo de la Ecuación 97 no imponen restricciones para X, mientras que el patrón especifico de ceros en A y$\bar{P}^t$ sí lo hace. Luego, es posible eliminar las filas en G correspondientes a elementos no nulos en el lado derecho de la expresión. Así se obtiene el siguiente sistema reducido:

$$0 = G_r \cdot X_c$$

Ecuación 98

, donde G_r puede ser particionada como sigue:

$$G_r = [G_{r1} \quad G_{r2} \quad \dots \quad G_{rL}]$$

, donde cada G_{rj} contiene L columnas de G_r. Notar además que la columna j de G_{rj} es el vector de ceros. Luego, el máximo rango de cada G_{rj} es $L-1$.

Escribiendo la Ecuación 98 en forma expandida:

$$0 = [G_{r1} \quad G_{r2} \quad \dots \quad G_{rL}] \begin{bmatrix} X_{c1} \\ X_{c2} \\ \\ \vdots \\ \\ X_{cL} \end{bmatrix}$$

Ecuación 99

Ahora, mirando la Ecuación 94 y considerando los elementos ceros en el lado izquierdo de dicha ecuación, es posible escribir:

$$A_{ij} = \sum_{k=1}^{L} \bar{A}_{ik} X_{kj} = 0$$

Y de acuerdo a la misma expresión para P:

$$\bar{P}_{ij} = \sum_{k=1}^{L} X_{ik} P_{kj} = 0$$

Luego, debido a que solo se están considerando los elementos iguale a cero en A y P (que tienen el mismo patrón de ceros que $\bar{A}$ y $\bar{P}$), los términos diagonales en X están siempre acoplados con un multiplicador cero. De esta manera, en la Ecuación 99 existirán $L(L-1)$ incógnitas en X. Si G_r tiene rango $L(L-1)$, entonces todos los términos no diagonales en X desaparecen. Así, el máximo rango permitido para cada sub-matriz de G_r, G_{ri} es $L-1$. En base a esto, la condición necesaria y suficiente para que X sea diagonal y se cumpla lo pedido, es que cada matriz G_{ri} tenga rango $L-1$.

Anexo 2: Particularidades matemáticas

Diferenciación de matrices

Las reglas de derivación de una variable son claras y conocidas extensamente, pero al derivar vectores, y trabajar en planos multidimensionales la intuición no acompaña mucho el desarrollo. En lo siguiente se resumen las reglas de derivación de vectores, extensión de cálculo de una variable, y donde se utiliza la misma notación que en el resto del documento.

Sea y una función de x, en donde x será un vector de dimensión L. Específicamente:

$$y = f(x) = f(x_1, x_2, \dots, x_L)$$

De esta manera se defina la derivada de y respecto a x cómo el gradiente del mismo, esto es:

$$\frac{dy}{dx} = \frac{df(x)}{dx} = x' = \begin{pmatrix} \frac{df(x)}{dx_1} \\ \frac{df(x)}{dx_2} \\ \vdots \\ \frac{df(x)}{dx_L} \end{pmatrix}$$

De esta manera, y de acuerdo a la forma de $f(x)$, se tendrá:

a) Si $f(x) = c$, una constante que no depende de x.

$$\frac{df(x)}{dx} = 0$$

O el vector de ceros para ser más exacto.

b) Si $f(x)$ es una forma lineal, donde β es un vector de constantes de la dimensión adecuada.

$$f(x) = x^t \beta$$

$$\frac{df(x)}{dx} = \beta$$

c) Si $f(x)$ es una forma cuadrática, donde A es una matriz de constantes de constantes de la dimensión adecuada.

$$f(x) = x^t A x$$

$$\frac{df(x)}{dx} = 2Ax$$

, si es que A es simétrica. Si no lo es, la expresión resulta:

$$\frac{df(x)}{dx} = 2(A^t + A)x$$

Anexo 3: Funciones MatLab

Programas y funciones creadas

En la tabla siguiente se resumen las principales funciones creadas, así como una breve descripción de las mismas.

Nombre	Descripción
a_gnca_reg	Técnica gNCA que permite incluir suposiciones de los datos a reconstruir. Programa principal que reproduce el algoritmo de resolución (Ver lanzador respectivo).
ac_gnca_reg	Técnica gNCA que permite incluir suposiciones de los datos a reconstruir y la confiabilidad de las mimas. Programa principal que reproduce el algoritmo de resolución. (Ver lanzador respectivo).
acGNCAreg_	Lanzador técnica acgNCAreg que permite incluir suposiciones de los datos a reconstruir y la confiabilidad de las mimas. Permite inicializar la reconstrucción, y generar un escaneo previo de la red para asegurar la convergencia al mínimo global.
agNCAreg_	Lanzador técnica agNCAreg que permite incluir suposiciones de los datos a reconstruir. Permite inicializar la reconstrucción, y generar un escaneo previo de la red para asegurar la convergencia al mínimo global.
Ar	Permite obtener las matrices reducidas de la matriz A.
block	Retorno una matriz block diagonal del argumento entregado.
c_gnca	Técnica gNCA que permite incluir la confiabilidad de los datos medidos. Programa principal que reproduce el algoritmo de resolución (Ver lanzador respectivo).
c_gnca_reg	Técnica gNCA que permite incluir la confiabilidad de los datos medidos y un término de regularización en P. Programa principal que reproduce el algoritmo de resolución (Ver lanzador respectivo).
cGNCA_	Lanzador técnica cgNCA que permite incluir la confiabilidad de los datos medidos. Permite inicializar la reconstrucción, y generar un escaneo previo de la red para asegurar la convergencia al mínimo global.
change_conex_test	Simulación que permite testear el efecto en el error de ajuste de diferentes redes al realizar cambios puntuales en su estructura. Utiliza NCAbasic_ y entrega gráficos de frecuencia relativa.
comp_NCA	Simulación de reconstrucciones utilizando diferentes métodos. Permite analizar una red particular usando diferentes métodos NCA, y entregando de forma ordenada los resultados más relevantes y matrices reconstruidas para cada una.
coord._f	
count_matrix_value	Cuenta en una matriz de errores porcentuales las entradas menores o iguales a un argumento.
create_A	Crea de manera aleatoria una red de regulación, definiendo la matriz A y sus valores.
create_G	Genera la matriz G del criterio gNCA.
create_Tnetwork	Crea una red aleatoria que cumpla con el criterio gNCA, de una dimensión y densidad dada.
desv_asump_A	Calcula la matriz de varianza de las suposiciones para A.
desv_asump_P	Calcula la matriz de varianza de las suposiciones para P.
disterror_NCA	Genera gráficos de distribución de errores para diferentes métodos NCA.
error_conc	Calcula la matriz de concentraciones de errores para una matriz de errores porcentuales dada.
error_E	Calcula el MSE y la matriz de errores porcentuales entre 2 matrices de datos.
error_resume	Muestra en pantalla el resumen de una reconstrucción.
error_X	Calcula el MSE, la matriz de errores porcentuales y el error porcentual promedio para 2 matrices comparadas.
g_gnca_reg	Técnica gNCA que permite en una función unir todas las funcionalidades NCA creadas y analizadas. Programa principal que reproduce el algoritmo de resolución (Ver lanzador respectivo).
gen_comp	Comprueba si en una red generada, cada gen es controlado por al menos un regulador.
GgNCAreg_	Lanzador técnica GgNCAreg que permite en una función unir todas las funcionalidades NCA creadas y analizadas. Permite inicializar la reconstrucción, y generar un escaneo previo de la red para

	asegurar la convergencia al mínimo global.
gnca_basic	Técnica gNCA original que permite incluir restricciones en la matriz P. Programa principal que reproduce el algoritmo de resolución (Ver lanzador respectivo).
gnca_comp	Comprueba si se cumplen los criterios gNCA.
gnca_n	Técnica gNCA original.
gnca_reg	Técnica gNCA que incluye un término de regularización en la matriz P. Programa principal que reproduce el algoritmo de resolución (Ver lanzador respectivo).
gnca_reg_n	Técnica gNCA original con regularización en P.
gNCAbasic_	Lanzador técnica gNCAbasic. Permite inicializar la reconstrucción, y generar un escaneo previo de la red para asegurar la convergencia al mínimo global.
gNCAreg_	Lanzador técnica gNCAreg. Utiliza un término de regularización en P. Permite inicializar la reconstrucción, y generar un escaneo previo de la red para asegurar la convergencia al mínimo global.
Ident_test	Comprueba la distinguibilidad entre dos redes.
medicion_E	Genera una medición con errores de la matriz de datos.
nca_basic	Técnica NCA. Programa principal que reproduce el algoritmo de resolución (Ver lanzador respectivo).
nca_comp	Comprueba si se cumplen los criterios NCA.
nca_n	Técnica NCA original.
NCAbasic_	Lanzador técnica NCAbasic. Permite inicializar la reconstrucción, y generar un escaneo previo de la red para asegurar la convergencia al mínimo global.
noise	Adiciona error blanco a una matriz de datos.
norm_NCA	Normaliza las matrices reconstruidas de acuerdo a un criterio establecido.
plot_CS	Muestra el ajuste gráfico para la reconstrucción de la matriz A.
plot_TFA	Muestra el ajuste gráfico para la reconstrucción de la matriz P.
poc_one	Retorna una entrada al azar que tenga una conexión real en la matriz A.
poc_zero	Retorna una entrada al azar que no posea una conexión real en la matriz A.
problema_NCA	Recibe una red y analiza en extenso los problemas con el criterio NCA de esta. Retorna detalladamente los nodos con problema.
prom_var_E	Calcula la matriz de datos promedio y la varianza de varias mediciones de los datos.
rand_	Retorno un número al azar entre un intervalo dado.
recocido_NCA_monot	Técnica recocido simulado NCA. Versión monótona.
Var_A_P	Calcula las matrices de varianza de las reconstrucciones en base a los datos del modelo.

Anexo 4: Programas en MatLab

En el siguiente CD se pone a disposición las funciones y métodos utilizados programados en MatLab.