Universidad Carlos III de Madrid

Tesis Doctoral

Búsqueda de relaciones causales para aplicaciones en cardiología y psiquiatría

Autor:
Mario de Prado Cumplido

Director:
Antonio Artés Rodríguez

Departamento de Teoría de la Señal y Comunicaciones

Leganés, Enero de 2012

Tesis Doctoral:

Búsqueda de relaciones causales para aplicaciones en cardiología y psiquiatría.

Autor:

Mario de Prado Cumplido

Director:

Dr. Antonio Artés Rodríguez

El tribunal nombrado para juzgar la tesis doctoral arriba citada, compuesto por los doctores

Presidente:

Vocales:

Secretario:

acuerda otorgarle la calificación de

Leganés, a

A mis padres,
Alfonso y Marina.

A Marta.

Dos cosas llenan el ánimo de admiración y respeto, siempre nuevos y crecientes: el cielo estrellado sobre mí y la ley moral en mí.

I. K.

En vano los hombres, amontonados por centenares y miles sobre una estrecha extensión, procuraban mutilar la tierra sobre la cual se apretujan; en vano la cubrían de piedras a fin de que nada pudiese germinar en ella; [...] la primavera era la primavera incluso en la ciudad. El sol calentaba, brotaba la hierba y verdeaba entre las grietas del pavimento.

L. T.
Resurrección

Resumen

En esta tesis se presenta un estudio sobre la teoría de la causalidad y se desarrollan nuevas técnicas de inferencia causal en datos discretos y en señales continuas. Estos métodos se han utilizado para avanzar en el conocimiento de dos problemas médicos: uno en el campo de la psiquiatría y otro en cardiología.

Los métodos clásicos para inferir relaciones causales a partir de muestras se basan en múltiples análisis de independencia estadística condicional. Estos métodos no dan buenos resultados cuando el número de variables aleatorias es elevado. Por el contrario, las técnicas de aprendizaje máquina son razonablemente robustas frente a problemas de alta dimensionalidad. Se han sustituido los test de independencia por una batería de clasificadores en el algoritmo de búsqueda causal propuesto. En la tesis se muestra cómo es posible identificar las relaciones causales en función del número de variables relevantes en una clasificación. Se ha ensayado con dos clasificadores, un k vecinos más próximos y una máquina de vectores soporte (SVM). Para aligerar la carga computacional impuesta por el gran número de clasificadores a entrenar, se ha recurrido a una permutación aleatoria para obtener la salida de problemas de menor dimensión a partir de máquinas entrenadas en dimensión superior. Se ha empleado la técnica de remuestreo "bootstrap" para etapas intermedias de estos algoritmos.

Para señales en tiempo continuo, el criterio de inferencia causal de Granger afirma que si una señal ayuda a la predicción de otra, la primera tiene influencia causal en la segunda. El criterio se implementa comparando los residuos de modelos autorregresivos ARMA de las señales. La inferencia causal tipo Granger obtiene mejores resultados forzando dispersidad en los coeficientes del modelo. Con este objetivo, en la tesis se ha empleado una aproximación basada en SVM para el modelado autorregresivo (cSVARMA); adicionalmente, se ha desarrollado una versión multivariable del algoritmo cSVARMA, que supera las prestaciones del modelo unidimensional. La función de coste robusta de los algoritmos propuestos hace que se obtengan mejores resultados, especialmente en escenarios con ruido tipo impulsivo.

Se presentan simulaciones con datos sintéticos comparando los métodos anteriores con los más relevantes en la literatura; los resultados validan el buen comportamiento de los algoritmos propuestos para identificar causalidad en datos discretos y en series temporales.

La algoritmia desarrollada se ha aplicado a dos problemas médicos. En el problema psiquiátrico, se han estudiado los factores más relevantes relacionados con la repetición de intentos de suicidio. No existen modelos etiológicos definidos para esta problemática. A partir de una base de datos de pacientes psiquiátricos, con información clínica y sociológica, se han seleccionado las variables fuertemente relevantes y se ha generado su árbol causal. La interpretación del modelo permite generar nuevas hipótesis de trabajo. También se ha entrenado un clasificador con las variables causantes de la repetición de intentos de suicidio, que puede emplearse para valorar el riesgo de nuevos pacientes y prevenir otros intentos.

En cardiología, se ha desarrollado una herramienta de representación de relaciones causales que permite identificar los mecanismos de generación y mantenimiento de fibrilaciones auriculares. Existen distintas hipótesis sobre esta patología, y cada una de ellas conlleva un tratamiento terapéutico distinto. Los métodos clásicos de análisis de estas señales, basados en las frecuencias dominantes de las ondas auriculares, presentan una importante limitación, pues usando únicamente la frecuencia no es posible identificar los focos y la manera de propagarse de los frentes de onda por el miocardio. Se ha recogido una base de datos de electrogramas de fibrilación auricular y se han aplicado a este problema los métodos de inferencia causal desarrollados. Finalmente, se ha diseñado una herramienta de visualización de las interacciones causales, que puede servir de apoyo al médico en el estudio de estas arritmias.

Abstract

In this PhD thesis, an exposition of the theory of causation is presented, and new techniques of causal inference for discrete and continuous data are developed. These methods are used to advance in the knowledge of two medical problems within the fields of psychiatry and cardiology.

The classical methods to infer causal relations from sample sets rest on multiple statistical independence tests. These methods operate bad in coping with high dimensional random variable spaces. On the contrary, machine learning techniques have shown good performance and robustness in these kind of problems. The proposal of the thesis is to substitute the independence tests with an ensemble of classifiers. In the thesis it is shown how the number of relevant features for a classification task is useful to identify causal relationships. Two classifiers have been used: a k nearest neighbour and a support vector machine. In order to improve the effectiveness of the method, in terms of processing time, a random permutation have been used to reduce the number of trainings. Some phases of the causal classifiers algorithm have been resolved with bootstrap techniques.

In the field of continuous signals, the Granger criterion states that if a signal helps in the prediction of another, the former has a causal influence on the latter. The implementation of Granger causality rely on the comparison of the residuals of auto-regressive ARMA models of the signals. These causal relations can be improved if sparsity in the coefficients of the ARMA model is imposed. To this end, an ARMA model based in support vector machines (SVM) has been implemented and applied to causal inference (cSVARMA). A multivariate version of the algorithm has been derived, which improves the performance of the unidimensional model. The robust cost function of SVM based ARMA models makes them soundness, specially in contexts with impulsive noise.

Some toy problems with synthetic data are provided to prove the good performance of the algorithms compared with the state-of-the-art methods; the results show the validity of the proposed algorithms, both in discrete and continuous data.

The causal inference methods presented in the thesis have been applied to a pair of medical problems. In psychiatry there are few etiological models for the factors that impels for the repetition of suicide attempts. A data base of sociodemographic and clinical variables has been used to select the strongly relevant features, and their causal graph has been generated. The interpretation of the findings in the causal models permits to provide new hypothesis to be proved. A classifier trained with the identified causes of suicidal behaviour is provided, that can be used to prevent new attempts and evaluate their risk.

In cardiology, the physiological mechanisms responsible for the onset and maintenance of the atrial fibrillation are not completely known. There are several hypothesis of this phenomenon, which lead to different therapeutic strategies. The typical analysis procedures, related with the detection of the fibrillation dominant frequency, present a serious limitation, because the information of the frequency is not enough to allow the identification of the triggering area of the arrhythmia. A data base of fibrillation electrograms has been recollected. Causal methodology for signals in continuous time have been applied to this problem, and the causal relationships found in the electrograms have been shown thanks to the design of a visualization tool. The methodology presented can be used to identify the propagation patterns of signals in the heart atria.

Agradecimientos

Hace ya años que comenzó la travesía que ha sido esta tesis, que ahora llega a buen puerto. Desde luego, no hubiera finalizado este trabajo sin la ayuda de muchas personas, a las que escribir estas pocas palabras de agradecimiento quedará, sin duda, escaso e incompleto. Vaya a todos, de corazón, mis gracias más auténticas.

Como es imperativo, agradecer en primer lugar la dirección del Dr. Antonio Artés Rodríguez y su apoyo para que finalice esta tesis. Mucho del trabajo que sigue se debe a sus intuiciones y aportaciones. También quiero agradecerle la libertad otorgada para trabajar en aquello en lo que creo, y para no hacerlo en lo que no creo. Al Dr. Artés y al Dr. Aníbal Figueiras Vidal agradezco la oportunidad de aprender, enseñar e investigar en un grupo de investigación del nivel que tiene el departamento de Teoría de la Señal y Comunicaciones. A los doctores Ángel Arenal Maíz y Enrique Baca García les debo la oportunidad de haberme podido asomar a problemas médico reales y aprender de su experiencia y buen hacer.

Son tantas las personas a las que recordar que alguna se quedará en el tintero; aunque no estén en estas líneas, sí lo están en la memoria de mi trayectoria vital. Mi experiencia en la universidad hubiera sido muy distinta sin los varios grupos humanos con los que he vivido. Quiero expresar mi más sentido agradecimiento a Sancho, José Miguel, Ricardo, José Emilio, Ricardo. Han sido para mí ejemplo de muchas cosas. De ilusión y esfuerzo. De nobleza, fe y fina ironía. De valentía (¿imprudencia tal vez?), generosidad y comprensión. De sencillez y sinceridad. De convicciones (aunque no sean compartidas) y buen humor. A Jero quiero agradecer su apoyo inquebrantable. A los que me acogieron desde el principio: José Luis, Mati, Fernando, Emilio. A Manuel, Sergio y Rubén, por los últimos tiempos vividos juntos. A todo el, muy numeroso, GTS, por lo mucho que se puede aprender de vosotros. A los trabajadores de la Salud, Jorge, Pablo, Antonio, Leonardo. Finalmente, agradecer los ratos pasados con aquellos que he compartido mesa, conversación y polémica.

A quien por mucho tiempo me ha acompañado en la aventura de la do-

cencia, Harold, gracias por tu actitud de servicio y buen humor, canastos.

Muchas gracias a los varios correctores (hasta donde se puede corregir) de la tesis, Montse, Marta, David, Marce. Algún día acertaré a poner las comas, en su sitio.

Si la inversión en tiempo y en renuncias ha merecido la pena, ha sido principalmente por los nombres aquí recogidos.

Tengo que recordar también a aquellos con los que he compartido el resto de mis afanes, trabajo nada remunerado, nocturno (pues crece como la hierba) y preocupado en arribar a puertos de tan difícil acceso que tal vez nunca se alcancen. Mis más preciadas gracias a los que han comprometido su vida a educar más allá del colegio, el instituto o la universidad, y de los que he aprendido, lección nada fácil, que merece la pena jugar de manera diferente, aunque sea para perder.

A las varias víctimas colaterales, Bortis y Víctor, gracias por este último año irrepetible. Al final, pude girar la llave. Aunque quien sabe donde nos volveremos a encontrar, la vida da muchas vueltas. A Netis y Jaime, por su sacrificio, paciencia y por la oportunidad que regalan a los demás.

Querría finalizar los agradecimientos citando las personas a las que dedico el libro. A mis padres y a mi hermana Montse, porque lo que soy y porque lo más preciado de lo que he aprendido se lo debo a ellos; efectivamente, mucha de la sabiduría es de los sencillos. Y esta vez sí, he dicho que acababa y he cumplido. A Marta, por la paciencia, el apoyo y la exigencia. Ha costado, pero ahora se podrán rellenar las lagunas que tenemos pendientes, codo a codo, en nuestro proyecto.

Índice general

Capítulo 1

Introducción

1.1. Motivación y objetivos

La teoría estadística ha sido amplia y exitosamente utilizada a la hora de resolver problemas en diversos campos como por ejemplo en ingeniería (comunicaciones digitales, procesado de señal...) o en medicina (psiquiatría, cardiología...). Muchos fenómenos presentes en el mundo real son demasiado complejos para ser descritos de manera determinista. También puede darse por caso el que existan excepciones, tal vez escasas en número, al caso general, que lo invaliden para una utilización universal. En estas situaciones, un tratamiento estadístico presenta evidentes ventajas. Por ejemplo, resulta de mayor utilidad describir globalmente el efecto del ruido térmico en un conductor, antes que especificar las ecuaciones que rigen el comportamiento de cada átomo y electrón.

Un ejemplo paradigmático de esta situación, dentro del mundo de la ingeniería, se encuentra en un sencillo modelo de un canal de comunicaciones.

$$y(t) = x(t) + n(t) \tag{1.1}$$

En él, la información de fuente $x(t)$ se ve contaminada por un ruido aditivo $n(t)$ independiente; un modelo de este sistema se muestra en la Figura 1.1. Una vez observada la salida del canal, $y(t)$, y asumiendo cierto conocimiento de las propiedades estadísticas del ruido, es posible reconstruir la información enviada.

El modelo (1.1) puede llevar a interpretaciones erróneas, aunque correctas desde el punto de vista matemático. Por ejemplo, podría interpretarse que la fuente de información genera datos siguiendo la señal de salida $y(t)$ menos un cierto ruido:

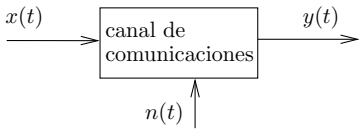


Figura 1.1: Modelo de un sistema de comunicaciones.

$$x(t) = y(t) - n(t) \tag{1.2}$$

La aparente paradoja surge por la polisemia implícita al signo de igualdad, que no es capaz de recoger de manera unívoca la dirección en la que fluye la información. Por ejemplo, en algunos lenguajes de programación se emplea un símbolo específico para la operación de asignación ($a := b$), que resalta el hecho de que el valor del término de la derecha se asigna a la variable de la izquierda del símbolo de asignación. Aunque esta operación iguala el contenido de ambas variables a y b, el valor de a no puede pasar a b. Otro símbolo, ==, se utiliza para expresar no asignación, sino equivalencia o igualdad.

Continuando con el ejemplo del canal de comunicaciones, es inmediato apreciar que esta aproximación estadística tiene una limitación intrínseca, a saber, la correlación entre dos señales o vectores de datos indica que existe una relación entre ellas. Sin embargo, las medidas estadísticas habituales no son capaces de distinguir la *dirección* en la que fluye la información. El novedoso paradigma *causal* (Pearl, 2000; Spirtes et al., 2000) se ha demostrado útil en los últimos años para distinguir entre causas y efectos.

Si el modelo anterior se representa con un modelo gráfico, como en la Figura 1.2, la ambigüedad queda resuelta.

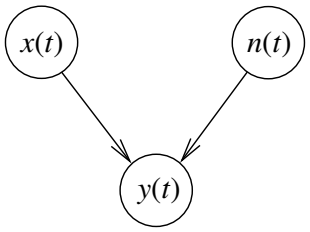


Figura 1.2: Ejemplo de grafo para el sistema de comunicaciones del ejemplo $y(t) = x(t) + n(t)$.

El modelo gráfico de la Figura 1.2 aporta más información que la Ecuación 1.1. Por ejemplo, la ausencia de enlace entre las variables $x(t)$ y $n(t)$ indica

que son independientes. Sin embargo, una vez que se tiene conocimiento de la variable $y(t)$ dicha independencia pierde vigencia. Dado el modelo gráfico, la paradoja antes expresada queda resuelta. Las limitaciones implícitas al signo de igualdad y a las ecuaciones tradicionales pueden ser superadas mediante el uso de modelos gráficos con significado causal.

Los conceptos habituales en estadística como, por ejemplo, correlación o independencia, son complementados en el paradigma causal con nuevos conceptos como *manipulación* o *consecuencia*. En el ejemplo anterior las variables de entrada y de salida, $x(t)$ e $y(t)$ respectivamente, están correlacionadas. Sin embargo, realizar algún tipo de *manipulación* sobre $y(t)$ no repercutirá en la señal de entrada. La operación contraria, manipular $x(t)$, sí tendría *consecuencias* en la salida $y(t)$.

El nuevo problema que se plantea consiste en la necesidad de modelar matemáticamente las relaciones de causalidad. En el pensamiento de la filosofía el concepto de causalidad ha sido el centro de amplios debates. Desde el Renacimiento y hasta el racionalismo de René Descartes e Immanuel Kant se ha tenido la creencia de que la naturaleza y el hombre están regidos por leyes fijas e inmutables, que el pensamiento humano puede hacer inteligibles.

Sin embargo, corrientes escépticas con esta filosofía han tenido una perspectiva opuesta, a saber, que no es posible colegir mediante observaciones de la naturaleza reglas deterministas que predigan el comportamiento de los acontecimientos y de las personas. El filósofo inglés David Hume lo resumiría en la metáfora del juego de billar, en el que él no podía observar "nada que saliera de una bola para entrar en la otra" cuando éstas chocaban. Esta corriente escéptica tan sólo confía en la conjunción temporal de eventos, no en la relación causal. Estas posiciones, unidas al éxito de las herramientas basadas en probabilidad y estadística, han relegado la utilización de técnicas de búsqueda de causalidad. No es hasta hace poco, y basado principalmente en los trabajos pioneros de Judea Pearl (Pearl, 2000) y Peter Spirtes (Spirtes et al., 2000) en las dos últimas décadas del siglo XX, que ha aparecido un nuevo paradigma matemático acerca de la *causalidad*.

Estudios controlados aleatorios

Antes de entrar a debatir acerca del paradigma introducido en dichos trabajos, se presentará otro método científico empleado tradicionalmente para encontrar relaciones causales. Se trata de los *estudios controlados aleatorios*. En estos experimentos los casos a estudiar se dividen en varios grupos, típicamente en dos: grupo de prueba y grupo de control. Al primero de ellos se le somete al experimento, mientras que al de control se le expone a unas

circunstancias parecidas. Por ejemplo, si se quisiera demostrar la utilidad de cierto medicamento para curar una enfermedad, habría que administrarlo al grupo de prueba, mientras que el grupo de control recibiría un placebo. Tras un cierto período, la observación comparada en la evolución del estado de salud de ambos grupos puede dar una idea de la influencia causal del medicamento sobre la enfermedad.

Este procedimiento busca relaciones de causalidad mediante una configuración que promueve el que no haya ninguna variable común entre los grupos de control y prueba. Más adelante, cuando se presente el modelo gráfico causal, se observará que ésta es una asunción habitual para asegurar el funcionamiento de los métodos de inferencia.

Los estudios controlados aleatorios no siempre son la mejor opción, debido principalmente a tres problemas:

- No son siempre éticamente realizables. Por ejemplo, no sería éticamente aceptable forzar a parte de la población a tomar alguna droga para comprobar sus efectos en el organismo.
- Suelen ser costosos.
- Pueden ser imposibles de realizar. Por ejemplo, no es posible establecer un estudio aleatorio para conocer el efecto de una ley o de una determinada política económica.

En este contexto, la elaboración de un paradigma causal basado en muestras, sin exigir mayores asunciones, se convierte en una herramienta muy útil.

Influencia del tiempo en la inferencia causal

Aunque se abordará en mayor detalle en las siguientes secciones, resulta apropiado aclarar en este punto el papel que desempeña el *tiempo* en los análisis de causalidad. Es evidente que en cualquier sistema del mundo real las causas preceden a los efectos; en sistemas lineales se identifica como *causal* a aquel sistema que cumple que su respuesta al impulso $h(t)$ verifica que $h(t) = 0, \forall t < 0$, que viene a ser otra manera de expresar la misma idea. Esta condición se reflejará de manera distinta a la hora de aplicar métodos en el campo de los datos discretos y continuos. En el primero, es habitual asumir que la recogida de datos se ha realizado en un intervalo temporal (de duración dependiente del problema concreto) que permite asignar un tiempo común a todas las variables. Así, la variable tiempo queda fuera de las herramientas para hacer inferencia causal. En series temporales no es posible,

ni interesante, realizar una operación similar; en estos casos es precisamente esa componente de dependencia temporal intrínseca a la estructura de las señales la que se emplea, mayoritariamente, para identificar las relaciones de causa y efecto. El método clásico para cuantificar y orientar la causalidad, conocido como *causalidad de Granger* (Granger, 1969; Marinazzo et al., 2006), emplea predictores entre series temporales que se basan precisamente en el ordenamiento temporal de las señales.

Sin embargo, más allá de esta diferencia, el paradigma causal que se presenta a continuación hace abstracción del método de inferencia particular y de las propiedades en las que se basa, y engloba ambos dominios, tanto discreto como continuo.

Objetivo de la tesis

La presente tesis aborda el problema de la extracción de información causal a partir de observaciones. Los métodos estadísticos clásicos son capaces de establecer medidas del parecido entre variables. Sin embargo, no pueden determinar la dirección en la que fluye la información. Existen numerosos métodos de inferencia causal, que recuperan el árbol de relaciones entre variables o entre señales. Muchos de ellos son computacionalmente costosos, y sufren una rápida degradación según se incrementa la dimensionalidad del problema.

El desarrollo de este capítulo introductorio será el siguiente: en primer lugar se presentarán los problemas médicos que se afrontarán con las herramientas de búsqueda causal; después se describirán brevemente los modelos gráficos y redes bayesianas que servirán de base para el posterior desarrollo del paradigma causal; una vez adquirida esta base teórica se estudiarán diversos algoritmos de inferencia causal; finalmente, se presentará el esquema del resto de la tesis.

1.2. Problemas clínicos abordados en la tesis

El paradigma causal se está aplicando en un número creciente de problemas de bioingeniería, donde prima no sólo la capacidad predictiva o discriminante sino especialmente la comprensión de los mecanismos subyacentes. Algunos ejemplos de esto se encuadran en el ámbito de la electroencefalografía, para comprender la comunicación entre subsistemas cerebrales (Marinazzo et al., 2011; Liao et al., 2009); de la genética, para identificar caminos en datos de expresión génica (Yoo y Cooper, 2004; Mukhopadhyay y Chatterjee, 2007); o en el análisis de sintomatologías cardiorespiratorias (Rosenblum

et al., 2002).

Mediante técnicas estadísticas y de tratamiento de señal, se han podido obtener interesantes resultados al analizar señales bioeléctricas, como las producidas en el ser humano, que principalmente son las generadas por el corazón (ECG), el cerebro (EEG) o los músculos (EMG[1]). Otras muchas señales del cuerpo humano proveen de información acerca del estado de la persona, como por ejemplo, sin ser exhaustivos: el ritmo respiratorio, la presión sanguínea, el nivel de sudoración de la piel, la temperatura corporal o los movimientos musculares.

La captación cada vez más eficiente de estas señales, unido a las técnicas avanzadas de tratamiento, ha permitido obtener información relevante sobre diversas enfermedades y patologías, así como mejorar las técnicas de diagnóstico de las mismas (Sörnmo y Laguna, 2005). Una aplicación prometedora es la telemedicina o el tratamiento médico a distancia. Estos servicios suelen basarse en la monitorización remota de señales bioeléctricas, por lo general no invasivas. Con esta metodología es posible dar una atención médica más rápida, y disminuir los costes del sistema sanitario.

Otra aplicación que está teniendo una creciente atención consiste en la monitorización constante del paciente mediante sensores o redes de sensores ubicados en distintas prendas y partes del cuerpo (Fletcher et al., 2010). En psiquiatría, estas herramientas son de interés para caracterizar el estado fisiopsicológico del paciente; en muchos casos los test de preguntas que le realizan no obtienen respuestas ajustadas a la verdad, pues el paciente puede mentir o sencillamente desconocer su estado real (Patel et al., 2009). Los dispositivos implantables, como los desfibriladores automáticos o DAI[2], monitorizan y adoptan medidas terapéuticas ante arritmias cardíacas; los algoritmos que toman esas decisiones son tema de intenso debate, pues deben combinar altas tasas de sensibilidad y especificidad manteniendo la sencillez y la baja carga computacional (Aliot et al., 2004; de-Prado-Cumplido et al., 2005; Arenal-Maíz et al., 2007).

La presente tesis es un intento de avanzar en el conocimiento del nuevo paradigma causal, con el desarrollo de nuevos métodos y la profundización en su teoría y aplicación, y ha tenido una inspiración en dos problemas del campo de la bioingeniería: uno de ellos es un estudio de factores causantes de la repetición en intentos de suicidio; el otro consiste en el análisis de señales cardíacas durante episodios de fibrilación auricular. Estos problemas clínicos se enmarcan, respectivamente, en métodos de resolución en el campo discreto y en el campo de las series temporales. Se procede a introducir estos

[1]Electromiogramas, estímulos eléctricos que atraviesan los músculos no cardiacos.
[2]Desfibrilador Automático Implantable, (IACD por sus siglas inglesas).

problemas médicos a continuación.

1.2.1. Intentos repetidos de suicidio

El número de enfermedades mentales ha experimentado un significativo incremento en las últimas décadas, especialmente en los países desarrollados. A diferencia de otras especialidades médicas, la psiquiatría debe enfrentarse a las enfermedades mentales basándose muchas veces en decisiones subjetivas sobre los síntomas del paciente. En numerosas ocasiones el tratamiento psiquiátrico se fundamenta en cuestionarios respondidos por los pacientes, que pueden ocultar o tergiversar la información o, simplemente, no conocer la respuesta auténtica. Por ejemplo, en trastornos de ansiedad es capital conocer las horas de sueño del paciente, pero su opinión suele distanciarse del número real de horas.

La categorización de las enfermedades o desórdenes mentales se recogen en los documentos DSM ("Diagnostic and Statistical Manual of Mental Disorders") y el ICD-10 ("International Classification of Diseases"). Esta taxología, aun teniendo base científica, consiste principalmente en un consenso entre la comunidad clínica psiquiátrica. Además de esto, los métodos de análisis empleados suelen ser sencillos descriptores estadísticos, cocientes de verosimilitud y simples regresores como la regresión logística.

En los últimos años tanto los avances tecnológicos como la aplicación de métodos de análisis novedosos están cambiando el panorama de la investigación en psiquiatría. Por una parte se estima que la herencia genética determina al 50 % los trastornos mentales. La posibilidad de registrar esa información a un precio cada vez más asequible, unido a las mejores posibilidades de tratamiento informático han abierto una prometedora vía de investigación. Una medida que puede ser de gran utilidad es el electroencefalograma; si bien el EEG no registra la localización de la actividad eléctrica, técnicas como la magnetoencefalografía combinan los EEGs con técnicas de imagen y localización anatómica. Cada vez mejores sistemas de sensado y algoritmos más complejos y potentes van a aportar mucha luz sobre la relación entre actividad neurológica y enfermedad mental. Por otro lado, está generalizándose la recopilación y posterior tratamiento de señales psicofisiológicas, como por ejemplo el ritmo cardiaco o respiratorio, la sudoración de la piel, la temperatura corporal o esquemas de movimiento. Estas medidas objetivas son un reflejo de la actividad neuronal simpática y parasimpática, que pueden aportar valiosa información acerca del estado mental del paciente, sin que deba intermediar su opinión, potencialmente sesgada (Fletcher et al., 2010).

El otro camino de innovación en psiquiatría viene de la mano de no-

vedosas técnicas de análisis, principalmente de la aplicación de técnicas de aprendizaje máquina (Baca-García et al., 2006; Oquendo et al., 2012). La relativa sencillez de los métodos que se han venido empleando en la literatura permitía a los clínicos extraer conclusiones de una manera directa. Los resultados de métodos más avanzados, aunque tengan una mejor capacidad predictiva o discriminativa, son menos manejables cuando lo que se desea es comprender los mecanismos que gobiernan el problema. Por ejemplo, en un clasificador lineal de tipo máquina de vectores soporte (Schölkopf y Smola, 2001), los pesos de la máquina pueden dar una idea de cómo de relevante es cada variable en comparación al resto, y en qué sentido ejerce su influencia en el resultado final. Sin embargo, esa misma máquina de vectores soporte con un núcleo gausiano, convierte el problema en no lineal, y la proyección de los datos sobre un espacio de dimensión infinita hace que los pesos del clasificador resulten crípticos para extraer conocimiento acerca de cómo interactúan las variables individuales. En este contexto, la inferencia causal proporciona, gracias a métodos de gran capacidad de análisis, resultados fáciles de comprender y sobre los que es posible realizar nuevas hipótesis y pruebas.

En el problema psiquiátrico con el que se ha trabajado, se han analizado las variables relevantes y su interacción a la hora de caracterizar las causas que provocan la repetición de intentos de suicidio. Los estudios realizados por la Organización Mundial de la Salud revelan que anualmente hay en torno a 10 ó 20 millones de intentos de suicidio, de los cuales un millón se llegan a consumar. Si el estudio se limita únicamente a Europa los datos son similares, estimándose que los intentos de suicidio multiplican por 10 ó 40 el número de suicidios consumados. Las autolesiones y los intentos de suicidio están muy correlacionados con la repetición de estos comportamientos. Se calcula que el 16 % de las personas que cuentan en su haber con un intento de suicidio o de autolesión vuelven a repetir en el primer año. Si se amplia la ventana temporal a cuatro años el porcentaje sube al 23 % y alcanza el 40 % cuando el seguimiento se alarga hasta los ocho años tras la primera tentativa. Centrándose en los suicidios consumados, el 7 % de estas personas mueren en los nueve años posteriores al inicio de estos comportamientos. El riesgo de suicidio entre las personas que se autolesionan está dos órdenes de magnitud por encima del riesgo de la población en general (Owens et al., 2002). También es preciso reseñar el impacto económico de este problema social. Los costes de tratamiento de estos pacientes son elevados, además del sobrecoste adicional debido al tiempo de baja laboral.

En el Capítulo 3 se presenta la aplicación de técnicas de aprendizaje causal a un problema psiquiátrico. Esta metodología es posible aplicarla en otros muchos escenarios, con una gran potencialidad a la hora de mejorar tanto el diagnóstico como el tratamiento de enfermedades mentales.

1.2.2. Identificación de focos en fibrilaciones auriculares

A continuación se presenta de manera sucinta el otro problema médico que se ha abordado en esta tesis. Dentro del campo de la electrofisiología cardíaca, uno de los retos planteados es el estudio, comprensión y tratamiento de las *fibrilaciones auriculares*. Se han empleado las herramientas de análisis causal para el estudio de esta patología cardiaca, con el fin de proveer al personal médico de una metodología que permita validar las hipótesis actualmente vigentes sobre el origen y mantenimiento de estas taquiarritmias.

Cuando el corazón funciona con normalidad, el flujo de sangre sigue el camino marcado por las flechas blancas en la Figura 1.3 (Dubin, 2000). La sangre entra al corazón por la vena cava inferior hacia la aurícula derecha. De aquí pasa al ventrículo derecho que la impele hacia los pulmones. Una vez que la corriente sanguínea ha sido oxigenada, vuelve a la aurícula izquierda a través de las venas pulmonares. Desde la aurícula pasa al ventrículo izquierdo. Éste se encarga de eyectar la sangre al resto del sistema circulatorio a través de la arteria aorta.

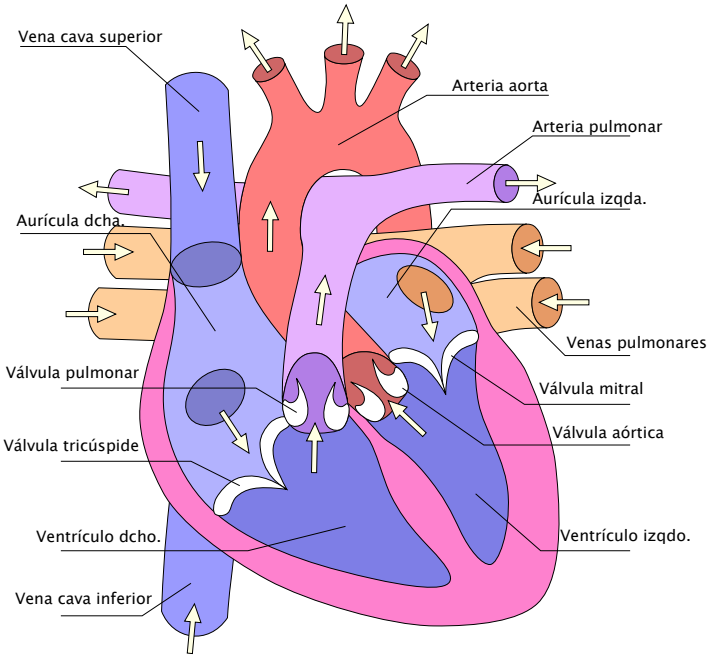


Figura 1.3: Esquema frontal de un corazón humano, con sus partes principales; la circulación sanguínea sigue las flechas blancas. Figura con licencia GPL.

Son dos los movimientos periódicos que realiza el corazón para forzar este recorrido: diástole y sístole. Durante la diástole auricular, estas cavidades se relajan, dejando pasar la sangre a su interior. En la diástole ventricular, la relajación de los ventrículos permite que pase parte de la sangre de las aurículas. El movimiento de sístole es una contracción rápida, que impulsa la sangre a proseguir en el circuito descrito. El ventrículo izquierdo es la cavidad que debe realizar un mayor esfuerzo, pues debe asegurar el riego al resto del cuerpo. Este movimiento es la parte más crítica del corazón.

Este movimiento sanguíneo y cardiaco se regula mediante impulsos eléctricos, controlados desde los sistemas simpático y parasimpático (Jalife et al., 1999). En funcionamiento normal cardiaco se dice que el corazón late en ritmo sinusal. El nodo sinusal situado en la aurícula derecha es el encargado de generar la estimulación cardíaca que se propaga por el miocardio, a través del sistema de haces que se puede ver esquemáticamente en la Figura 1.4.

Esta estimulación eléctrica provoca las contracciones de las distintas cavidades, y consecuentemente las sístoles auriculares y ventriculares. En el nodo auriculo-ventricular (A-V), situado en la confluencia de aurículas y ventrículos, este impulso se ve retenido debido a la distinta configuración de las células que lo componen. De esta forma, la sangre dispone del tiempo suficiente para abandonar las aurículas e ingresar en los ventrículos. Una vez superado

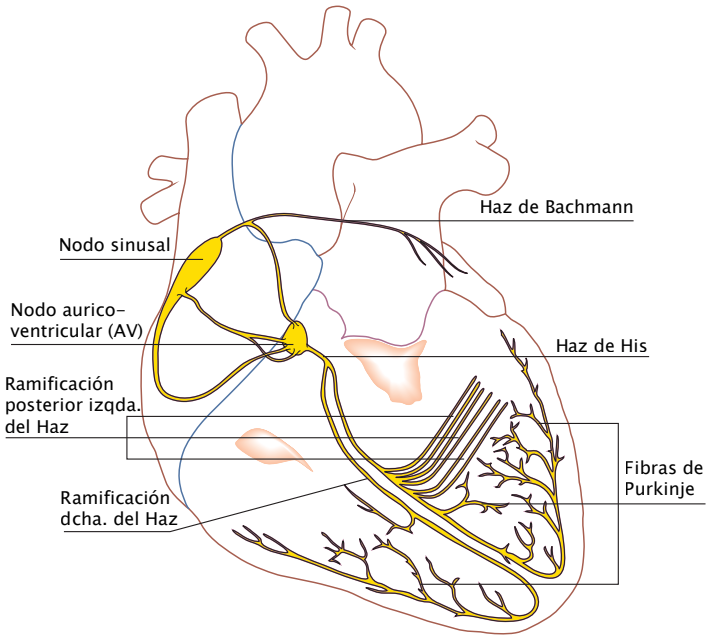


Figura 1.4: Vías de conducción eléctrica del corazón.

el nodo A-V, el Haz de His distribuye el impulso eléctrico por un camino de fibras de conducción eléctrica rápida, conocido como fibras de Purkinje, que permite que los impulsos eléctricos provoquen una contracción rápida y coordinada del tejido ventricular.

Si se registra la actividad eléctrica del corazón de manera no invasiva, se obtienen trenes de ondas similares al electrocardiograma (ECG) de la Figura 1.5. La onda P representa la despolarización auricular (contracción). El complejo QRS es la despolarización ventricular. El tiempo entre P y QRS es el necesario por la estimulación para atravesar el nodo A-V, de conducción más lenta. Durante el QRS se produce la sístole ventricular. Finalmente, la onda T marca el final de la repolarización ventricular.

El que se acaba de describir es el funcionamiento normal de las corrientes de estimulación cardíaca. Los ritmos anormales que pueden comprometer el funcionamiento del corazón son varios, y de diverso origen. Sin embargo, la redundancia del sistema de propagación hace difícil, si bien no imposible, un fallo general de este órgano. Por ejemplo: si el nodo sinusal cesa su estimulación, o ésta es bloqueada, otros puntos de la pared muscular auricular comenzarían automáticamente a liderar una nueva estimulación cardíaca, a un ritmo entre 60 y 80 latidos/min. Incluso si toda la aurícula estuviera imposibilitada, el nodo A-V es capaz de provocar la sístole ventricular (a una tasa entre 40 y 60 lat/min). Finalmente, el ventrículo puede tomar el relevo si se produce un fallo de los sistemas anteriores (entre 20 y 40 lat/min). En el caso de las fibrilaciones auriculares, la estimulación auricular es de naturaleza caótica e inhibe un correcto movimiento de sístole y diástole auricular. Sin embargo el mero movimiento de relajación de los ventrículos es capaz de atraer la sangre contenida en las aurículas, permitiendo que la persona con

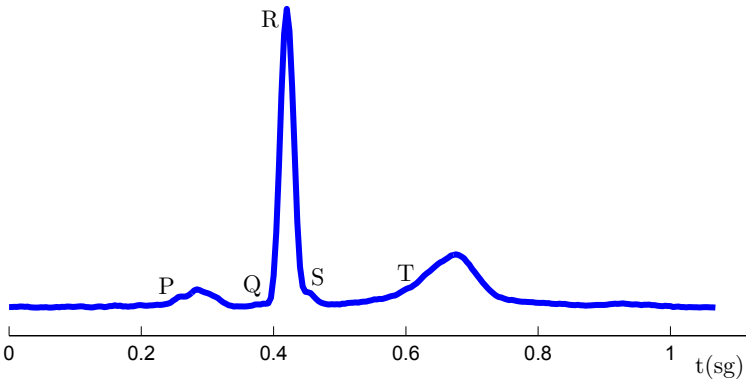


Figura 1.5: Ejemplo de un eletrocardiograma. Las componentes principales son la onda P, el complejo QRS y la onda T.

esta dolencia mantenga un estado vital estable.

En general los ritmos eléctricos anómalos pueden provocar dolencias cardiacas que se dividen en ritmos lentos, rápidos, o de alto grado de descoordinación, llamados, respectivamente, bradicardias, taquicardias o fibrilaciones. Pueden afectar a las cavidades superiores o inferiores. Se realiza una taxología general atendiendo al mecanismo que origina la arritmia: éstas se pueden clasificar en ritmos irregulares, latidos y ritmos ectópicos, latidos prematuros, taquiarritmias y bloqueos. Se procede a definirlos brevemente.

Ritmos irregulares

Son ritmos generados por varios puntos de estimulación anómala, con un patrón de latido mayor que la variabilidad normal del ritmo sinusal. En un corazón sano, la variabilidad cardíaca es necesaria, y suele ser del 10 % del ritmo. Valores mayores entran en la categoría de irregulares. Se pueden dividir en:

- Marcapasos migratorio. En ocasiones el foco de estimulación del nodo sinusal se desplaza de ubicación, migra, y produce variaciones en el ciclo cardiaco, y por lo tanto en la longitud de la onda P (despolarización auricular) del ECG, aún cuando el ritmo general se mantiene en valores normales.

- Taquicardia auricular multifoco. A diferencia del caso anterior, en esta patología son varios los focos activos de la aurícula, que además laten a un ritmo superior a 100 lat/min. Es propio de pacientes con enfermedad de obstrucción pulmonar crónica.

- Fibrilación auricular. En este caso el ritmo se dispara a valores superiores a 150 lat/min; los mecanismos de generación provocan controversia entre la comunidad científica. La apariencia de los electrogramas[3] son ondas caóticas, con un ritmo muy irregular, que no son capaces de producir una sístole auricular eficaz.

Latidos y ritmos ectópicos

Un latido ectópico ocurre cuando alguno de los tres sistemas de respaldo del nodo sinusal genera erróneamente un impulso. Los tres puntos de generación ectópica están en las aurículas, el nodo A-V y los ventrículos. Los sistemas de respaldo comienzan su actividad en caso de que el nodo sinusal tarde más de lo habitual, dejando así de inhibir al resto de puntos activos.

[3]Variaciones eléctricas sensadas localmente.

Cuando la pausa en el nodo sinusal no es temporal sino permanente, los puntos activos de la aurícula, el nodo A-V, o los puntos activos del ventrículo, de manera jerárquica, toman el relevo para marcar el ritmo cardiaco.

Latidos prematuros

Son latidos generados antes de lo habitual en algún punto excitable del miocardio. Aparecen por diversos motivos, entre los que se encuentran la presencia de estimulantes como la adrenalina, la cafeína o las anfetaminas, la presencia de algunas toxinas o problemas de hipertiroidismo.

Taquiarritmias

En esta categoría se encuentran los ritmos mayores de 150 lat/min, ya sean generados por uno o más puntos activos. Entre 150 y 250 se denominan taquicardias paroxísticas (de aparición repentina y reaparición frecuente). Por encima de 250 lat/min y por debajo de 350 se denomina "flutter" (aleteo) a estos ritmos rápidos. Por encima de este valor se encontrarían las fibrilaciones. El punto activo que causa la taquiarritmia puede ser auricular, ventricular o en la unión de estos. Las arritmias supraventriculares (auriculares o en el nodo A-V) no son directamente peligrosas para la vida del paciente, pero pueden desembocar en otros ritmos anómalos más dañinos. El "flutter" auricular se produce cuando un automatismo auricular, distinto del nodo sinusal, genera rápidas sucesiones de impulsos idénticos. Si el "flutter" sucede en los ventrículos, el aspecto de la forma de onda es parecido a una señal sinusoidal. Cuando el ritmo cardiaco supera los 350 lat/min se dice que se ha entrado en fibrilación. Múltiples puntos de las aurículas o de los ventrículos generan impulsos eléctricos, que se propagan de manera desorganizada por el miocardio. En estos casos la eficacia mecánica del corazón se anula, lo cual es crítico si la fibrilación tiene lugar en los ventrículos, pues es preciso realizar una desfibrilación inmediata.

Bloqueos cardiacos

Los bloqueos cardiacos son zonas del miocardio que impiden o retardan el paso de la estimulación eléctrica. En función de si son persistentes u ocasionales, y del grado de atenuación del impulso, se catalogan en diversos estadios de gravedad. La supresión del automatismo de mayor jerarquía puede provocar que otros focos activos del corazón tomen el relevo de marcar el ritmo.

Los posibles tratamientos de las afecciones cardíacas incluyen medicación, el uso de dispositivos como marcapasos o desfibriladores, medidas más inva-

sivas y drásticas como la ablación o una combinación de los anteriores. En el caso concreto de las fibrilaciones auriculares, son dos las medidas terapéuticas que se toman en general: terminar la fibrilación auricular para volver a ritmo sinusal normal, o bien tolerar la fibrilación en la aurícula, a la vez que se monitorizan los ventrículos para asegurar que están en ritmo normal. En los últimos tiempos la ablación de la fibrilación auricular se ha convertido en la medida habitual en muchos servicios de electrofisiología cardíaca. En cualquiera de los casos, existe la dificultad de que el mecanismo de generación y mantenimiento de las fibrilaciones no es del todo conocido. Las dos posibles hipótesis más ampliamente aceptadas actualmente teorizan, una de ellas, que la causa está en ondas de propagación caótica que circulan por toda la superficie de la aurícula, o bien, la segunda teoría, defiende que hay ciertos puntos de la aurícula, de frecuencia de estimulación eléctrica mayor que el resto del músculo, que gobiernan la fibrilación. Por supuesto, el tratamiento a aplicar con mayor posibilidad de éxito será el que se corresponda con el auténtico mecanismo de generación de fibrilaciones.

El trabajo en esta tesis se ha centrado en el análisis de señales de electrogramas, es decir, registros de actividad eléctrica tomados de manera invasiva. Hay que distinguir los electrogramas así descritos del más común electrocardiograma, que se registra de manera no invasiva al cuerpo del paciente, y que muestra la actividad agregada del corazón como un todo. El instrumento de medida es un lazo circular con varios pares de sensores, que miden actividad eléctrica local en ciertos puntos de la aurícula izquierda. Esta configuración de la información registrada es idónea para aplicar métodos causales, con el objeto de identificar en qué forma y dirección se propagan las ondas eléctricas. Esta información resulta muy valiosa de cara a resolver el debate clínico acerca de la naturaleza de las fibrilaciones y a su posible tratamiento terapéutico.

1.3. Modelos gráficos para causalidad

Los modelos gráficos son una manera de visualizar un conjunto de variables y las relaciones probabilísticas establecidas entre ellas (Bishop, 2007). Si bien no es una técnica novedosa, ni revolucionaria por sí misma, sí permite una representación sencilla y pedagógica de conceptos más complejos. Además, es una herramienta básica no sólo para la representación de relaciones causales, sino para la extracción de conclusiones *causales* a partir del grafo.

A continuación se presenta la teoría básica de grafos y redes bayesianas que servirá de base para introducir los conceptos de causalidad.

1.3.1. Elementos de grafos dirigidos acíclicos

Un *grafo* $\mathcal{G}$ es un conjunto de nodos o vértices (variables o señales) y de enlaces (relaciones), que codifican las relaciones existentes entre dichas variables. Estos enlaces pueden ser dirigidos. Un grafo o red estará formada por un conjunto de nodos o vértices V y un conjunto de enlaces E, que serán pares de nodos. Dos nodos unidos por un enlace se dice que son *adyacentes*. Si a estos pares de nodos se les confiere orientación, el primero de cada par será el origen. En este caso el grafo recibe el nombre de *grafo dirigido*. Se procede a presentar un ejemplo. El grafo $\mathcal{G}$ está formado por los conjuntos $V = \{X, Y, Z, W\}$ y $E = \{(X, Z), (Y, Z), (Z, W)\}$. En la Figura 1.6 se muestra gráficamente la misma información.

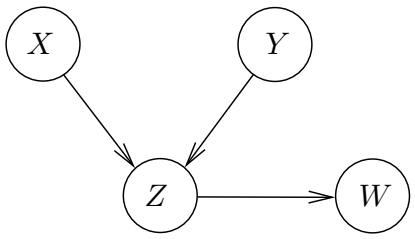


Figura 1.6: Ejemplo de grafo $\mathcal{G}$ formado por $V = \{X, Y, Z, W\}$.

En un grafo dirigido se pueden encontrar los siguientes elementos:

- Camino: secuencia de n nodos $\{X_1, X_2, \ldots, X_n\}$ que cumple que los enlaces (X_{i-1}, X_i), con $i \in \{2, \ldots, n\}$ pertenecen a E.
- Cadena: secuencia de nodos $\{X_1, X_2, \ldots, X_n\}$ que cumplen que o bien $(X_{i-1}, X_i) \in E$ o bien $(X_i, X_{i-1}) \in E$, con $i \in \{2, \ldots, n\}$. La diferencia con respecto al camino consiste en que las cadenas siempre recorren nodos adyacentes, pero no tienen por qué hacerlo en el sentido de las flechas que los unen.
- Factor de confusión ("confounder" en inglés): dos nodos adyacentes unidos por un enlace bidireccional. La aparición de estos enlaces bidireccionales suele significar la existencia de una variable adicional con enlaces dirigidos hacia las variables del factor de confusión.
- Lazo o bucle: se dice que un grafo contiene un lazo si existe un camino que parta de un nodo y vuelva al mismo.
- Esqueleto: el esqueleto de un grafo consiste en la eliminación de la orientación de los enlaces. Por consiguiente, distintos grafos pueden tener idéntico esqueleto si comparten nodos y adyacencias.

- Ancestros o predecesores de un nodo: es el conjunto de nodos desde los que se puede llegar al nodo objetivo mediante un camino. En los grafos dirigidos se emplean términos de relación familiar (padres, hijos, descendientes) para designar a conjuntos de nodos en función de su posición relativa dentro del grafo.

Si todos los enlaces están dirigidos y no hay bucles el grafo se denomina *Dirigido y Acíclico* (DAG). Los grafos tipo DAG son un componente esencial de la teoría causal. El número de posibles DAGs en función de los nodos crece de la forma que se muestra en la Figura 1.7, donde el eje de ordenadas tiene escalado logarítmico. Hay $\mathcal{O}(d!2^{\binom{d}{2}})$ posible grafos de d nodos. El número exacto se puede determinar mediante la siguiente recursión, según (Robinson, 1977):

$$a_d = \sum_{i=1}^{d} (-1)^{i-1} \binom{d}{i} 2^{i(d-i)} a_{d-i} \tag{1.3}$$

donde se asume que $a_0 = 1$.

Como se puede comprobar, el espacio de búsqueda para realizar inferencia es muy grande. Para cuatro nodos, el número de posibles DAGs es de 543, mientras que para diez nodos las posibilidades ascienden a un orden de 10^{20}.

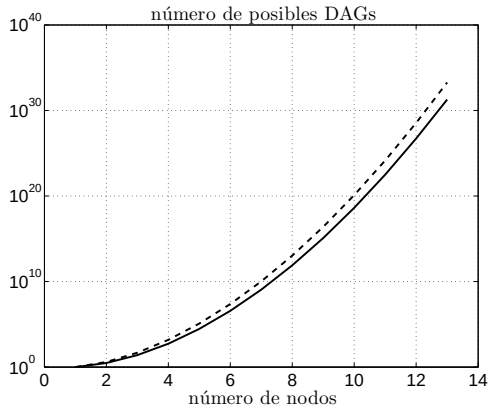


Figura 1.7: Número posible de grafos dirigidos sin bucles (DAG) en función de los nodos que los componen, en escala logarítmica. En línea punteada se representa la aproximación $(d!2^{\binom{d}{2}})$.

1.3.2. Redes bayesianas

En aquellos casos en los que los nodos representan variables aleatorias, el grafo recibe el nombre de *Red Bayesiana*. Por lo tanto una red bayesiana consiste en un grafo $\mathcal{G}$, formado por d nodos, y una distribución de probabilidad $P(x_1, \ldots, x_d) = P(\mathbf{x})$; de esta manera los enlaces del grafo indican las relaciones de dependencia entre las variables. Como se verá, la caracterización de la probabilidad empleando el grafo resulta mucho más compacta que si se diera la distribución conjunta al completo.

Cualquier distribución P se puede expresar como:

$$P(x_1, \ldots, x_d) = \prod_{i=1}^{d} P(x_i | x_1, \ldots, x_{i-1}) \tag{1.4}$$

en virtud de la regla de la cadena de probabilidades.

Ahora bien, suponiendo que x_i es independiente de todos sus predecesores, excepto de un subconjunto pa_i, entonces la distribución $P(\mathbf{x})$ se puede expresar como un producto de los llamados *núcleos de Markov*:

$$P(x_1, \ldots, x_d) = \prod_{i=1}^{d} P(x_i | pa_i) \tag{1.5}$$

donde pa_i es el conjunto de predecesores de x_i que no son estadísticamente independientes de dicha variable. Por lo tanto $P(x_i|x_1, \ldots, x_{i-1}) = P(x_i|pa_i)$. Al conjunto pa_i se le conoce por el nombre de *padres markovianos* o simplemente *padres*. Más formalmente:

Definición 1.3.1 (Padres markovianos)**.** Sea un conjunto ordenado de variables $V = \{X_1, \ldots, X_d\}$ y $P(\mathbf{x})$ su distribución de probabilidad asociada. Se dice que el conjunto $pa_i \in V$ son los *padres markovianos* si es el mínimo conjunto de predecesores de x_i que lo hace estadísticamente independiente de todos los demás predecesores. Por tanto pa_i es el mínimo conjunto de nodos pertenecientes a V que verifican:

$$P(x_i|pa_i) = P(x_i|x_i, \ldots, x_{i-1}) \tag{1.6}$$

El grafo de la Figura 1.6, asumiendo que se cumple la Definición 1.3.1, sería un sencillo ejemplo de una red bayesiana. En virtud de dicha propiedad se verificaría que $W \perp\!\!\!\perp X, Y|Z$, de forma que Z apantalla a la variable W del resto del grafo. La expresión $X \perp\!\!\!\perp Y$ se empleará en este documento como abreviatura de la relación de independencia estadística $P(X, Y) = P(X) \cdot P(Y)$. De forma análoga, $X \perp\!\!\!\perp Y|Z$ hará referencia a

la independencia condicional $P(X,Y|Z) = P(X|Z) \cdot P(Y|Z)$. La expresión $X \not\perp\!\!\!\perp Y$ significará dependencia estadística.

De la Definición 1.3.1 se puede derivar un método para elaborar una red bayesiana mediante la aplicación recursiva de test de independencia. Este método es aplicable debido a la particularidad de que es un conjunto ordenado, por lo que *a priori* se conoce dicha información del problema. El método sería como se describe a continuación. Se comprueba la independencia entre X_1 y X_2; en caso de dependencia se unen mediante un enlace. Si se cumple que (X_1, X_2) son independientes de X_3, no se establece ningún nuevo enlace. Si se cumple que $X_3 \not\perp\!\!\!\perp X_1$ y que $X_3 \perp\!\!\!\perp X_1|X_2$ entonces se orientará $X_2 \rightarrow X_3$, puesto que al ser un conjunto ordenado, la flecha no puede ir hacia X_2, y que el condicionamiento sobre X_2 independiza las variables. En caso contrario, es decir, si $X_3 \perp\!\!\!\perp X_2|X_1$ entonces el enlace dirigido será $X_1 \rightarrow X_3$.

En la i-ésima iteración de este sencillo algoritmo habría que encontrar el conjunto markoviano pa_i que independiza a x_i del resto de sus predecesores. No es un método práctico para construir una red bayesiana debido al elevado número de test de independencia a realizar. Sin embargo, este algoritmo de fuerza bruta basado directamente en la implementación de la Definición 1.3.1 servirá de inspiración para los métodos de inferencia que se presentarán más adelante.

El conjunto de nodos pa_i es único si $P(\mathbf{x})$ es estrictamente positivo, i.e., $\nexists P(\mathbf{x}_i) = 0$ (Pearl 1998b, (Spirtes et al., 2000)).

Si una distribución P puede descomponerse como en (1.5) respecto del grafo $\mathcal{G}$, entonces se dice que $\mathcal{G}$ representa a P o que $\mathcal{G}$ y P *son compatibles*.

1.3.3. Redes bayesianas causales

Una *red bayesiana causal* es una red bayesiana que cumple el criterio de intervención o manipulación. Antes de definir estos conceptos, se introduce y discute la *d-separación*. Las variables X, Y, o Z pueden referirse tanto a variables individuales como a subconjuntos de variables del grafo; la diferencia queda clara en función del contexto.

El concepto de *d-separación* es un criterio gráfico para buscar independencias condicionales. Se empleará la siguiente notación:

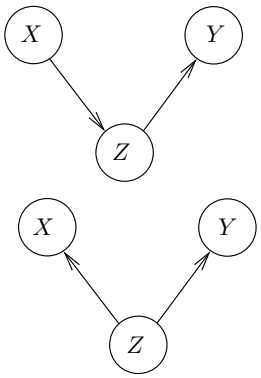
$$\begin{cases} (X \perp\!\!\!\perp Y|Z)_P & \text{Independencia condicional estadística} \\ (X \perp\!\!\!\perp Y|Z)_{\mathcal{G}} & \text{D-separación} \end{cases} \tag{1.7}$$

Se dice que Z d-separa X de Y si y sólo si Z bloquea cada camino entre un nodo de X y otro de Y. Un camino está d-separado (o bloqueado) por Z si y sólo si se cumplen estas dos condiciones:

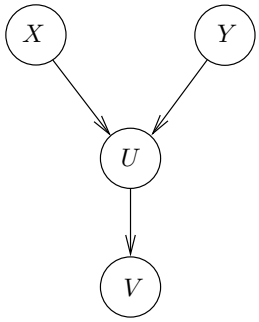
1. el camino es una cadena $X_i \to X_m \to X_j$ o una división ("fork") $X_i \leftarrow X_m \to X_j$ y además $X_m \in Z$. Ver Figura 1.8(a).

2. el camino es una división invertida ("collider") $X_i \to X_m \leftarrow X_j$ y además $X_m \notin Z$ ni tampoco ninguno de los hijos de X_m. Ver Figura 1.8(b). Este tipo de relación también se denomina como *colisión* o *estructura en "V"*.

Se expone un ejemplo de una estructura en "V" o colisión. Se supondrá que X indica nota alta, Y talento musical y Z la aceptación en un colegio. Entonces, si el criterio de aceptación en el colegio es nota alta o bien talento musical, los alumnos tendrán una correlación negativa entre X e Y, aún cuando en la población en general sean independientes. Por tanto, en este ejemplo conocer Z hace dependientes a X e Y.



(a) Ejemplo de cadena (arriba) y división (abajo).



(b) Ejemplo de colisión o estructura en "V".

Figura 1.8: En (a), Z d-separa X e Y. En (b), cualquier conjunto Z que no contenga ni U ni ningún descendiente de U d-separa X e Y; en el ejemplo (b), Z debe ser el conjunto vacío para d-separarlos.

Existe una relación entre independencia estadística y d-separación. Para los subconjuntos disjuntos de nodos X, Y y Z del DAG $\mathcal{G}$ y para todas las distribuciones $P(\mathbf{x})$:

1. D-separación implicará independencia estadística, i.e., $(X \perp\!\!\!\perp Y|Z)_{\mathcal{G}} \Rightarrow (X \perp\!\!\!\perp Y|Z)_P$ siempre que $\mathcal{G}$ y P sean compatibles. El grafo y la probabilidad son compatibles si ésta última puede descomponerse en producto de núcleos de Markov, como en (1.5), siguiendo la estructura del grafo.

2. Si X e Y son estadísticamente independientes dado Z, $(X \perp\!\!\!\perp Y|Z)_P$, en todas las distribuciones compatibles con $\mathcal{G}$, entonces Z d-separa X e Y, $(X \perp\!\!\!\perp Y|Z)_{\mathcal{G}}$.

Una propiedad importante, que impone un límite a los algoritmos de inferencia, derivado de la d-separación, se conoce como el principio de equivalencia en observación. Se dice que dos DAGs son equivalentes en observación si y sólo si tienen el mismo esqueleto y las mismas estructuras en "V".

Una red bayesiana puede ser el soporte de relaciones causales. Se dice que X_i es una *causa directa* de X_j cuando existan dos manipulaciones sobre $V \backslash X_j$ (todos los nodos excepto el efecto X_j) que difieran sólo en los valores de la causa X_i y que produzcan diferentes distribuciones de X_j. En la notación de (Pearl, 2009), las manipulaciones o intervenciones se indican como:

$$P(X_j|do(X_i)) \tag{1.8}$$

que no debe confundirse con la marginalización $P(X_j|X_i)$. La Ecuación 1.8 hace referencia a la asignación de un valor determinado a X_i. Si efectivamente existe una relación causal directa $X_i \rightarrow X_j$, la distribución en (1.8) variará para distintos valores de la causa X_i.

Definición 1.3.2 (Red Bayesiana Causal)**.** Una red bayesiana $(\mathcal{G}, P(\mathbf{x}))$ es un DAG causal si y sólo para cada enlace $X_i \rightarrow X_j$ del grafo, X_i es una causa directa de X_j.

Mediante la herramienta de d-separación, es posible detectar a partir de un grafo causal si existe independencia estadística entre dos variables X e Y, y consecuentemente identificar relaciones causales.

Para poder inferir modelos causales a partir de muestras, se deben asumir dos condiciones adicionales a la de Markov: de minimalidad y de estabilidad.

- Condición de *minimalidad*: se dice que un grafo $\mathcal{G}$ cumple la condición de minimalidad si cada enlace de $\mathcal{G}$ previene, si no estuviera presente, una relación de independencia condicional existente en $P(\mathbf{x})$.

- Condición de *estabilidad o fidelidad*: si todas y únicamente las relaciones de independencia condicional ciertas en P se muestran en la condición de Markov aplicada al grafo $\mathcal{G}$, entonces P y $\mathcal{G}$ son *estables* una con respecto al otro. Más aún, se dice que P es estable si hay un DAG para el cual es estable.

  Intuitivamente, estabilidad significa que las distribuciones P compatibles con el grafo $\mathcal{G}$ no destruyen ninguna independencia. Si P es "estable" entonces X e Y son dependientes sólo si hay un camino de X a Y.

1.3.4. Propiedades relevantes de la causalidad

El paradigma de búsqueda de relaciones causales persigue encontrar un modelo que explique las relaciones entre las variables de un problema. Generalmente se asume una relación probabilística entre estas variables, de forma que las relaciones son simétricas. Las relaciones causa efecto son claramente asimétricas, y se sustentan en el proceso real de generación de los datos. Sin embargo, es de utilidad utilizar un análisis probabilístico de la causalidad. De manera intuitiva, es claro que causa implica mayor verosimilitud del efecto, pero no siempre certeza. En las relaciones causa-efecto puede haber excepciones. Por ejemplo, en el ejemplo de la Figura 1.1 del canal de comunicaciones, aunque el modelo es fijo tal que $y(t) = x(t) + n(t)$, no siempre será posible recuperar el valor de la señal de entrada $x(t)$ conocida la salida $y(t)$, pues dependerá de la estadística del ruido $n(t)$.

Los modelos causales permiten realizar las tareas de

1. Predicción, basado en observación ($P(X_j|X_i)$).
2. Intervención: $P(X_j|do(X_i))$.
3. Estudio de casos *contrafácticos* ("counterfactuals" en inglés), cuando se desea analizar qué hubiera pasado en caso de que una variable hubiera tenido cierto valor.

La primera tarea es predominantemente probabilística, mientras que las dos últimas son operaciones causales. Se enumeran a continuación alguna de las propiedades de la relación de independencia estadística condicional $X \perp\!\!\!\perp Y|Z$.

1. Simetría: $(X \perp\!\!\!\perp Y|Z) \Rightarrow (Y \perp\!\!\!\perp X|Z)$. Si en el contexto de Z, X no dice nada de Y, entonces Y no dice nada de X.
2. Descomposición: $(X \perp\!\!\!\perp YW|Z) \Rightarrow (X \perp\!\!\!\perp Y|Z)$. Si la combinación de dos variables es irrelevante a X también serán irrelevantes por separado.
3. Unión débil: $(X \perp\!\!\!\perp YW|Z) \Rightarrow (X \perp\!\!\!\perp Y|ZW)$. Aprender información irrelevante W no puede ayudar a hacerse relevante a la información irrelevante Y.
4. Contracción: $(X \perp\!\!\!\perp Y|Z) \wedge (X \perp\!\!\!\perp W|ZY) \Rightarrow (X \perp\!\!\!\perp YW|Z)$. Si W es irrelevante después de aprender otra información independiente Y, entonces W ya era irrelevante antes de conocer Y. Las dos últimas propiedades aseguran que la información irrelevante no altera el estado de relevancia de cualquier combinación: lo irrelevante lo seguirá siendo, y lo relevante también.

5. Intersección: $(X \perp\!\!\!\perp W|ZY) \wedge (X \perp\!\!\!\perp Y|ZW) \Rightarrow (X \perp\!\!\!\perp YW|Z)$. Si Y es irrelevante conocido W, y W es irrelevante conocido Y, entonces ni W ni Y ni su combinación serán relevantes.

1.4. Causalidad y aprendizaje máquina

Los objetivos de estas dos disciplinas son coincidentes en gran medida, sin embargo presentan particularidades (Guyon et al., 2007). Ambas buscan comprender la relación entre variables, y generar modelos que permitan hacer predicción. Sin embargo, los métodos causales tratan de desvelar el mecanismo de generación de los datos. Los algoritmos de aprendizaje máquina han evolucionado en las últimas décadas hasta un punto en el que las tareas de clasificación o de regresión, por citar las más habituales, han alcanzado una alta eficacia, y son capaces de resolver problemas de alta dimensionalidad y bajo número de muestras. Los algoritmos de búsqueda causal, sin embargo, suelen adolecer de mayores problemas si el número de dimensiones es elevado o el número de muestras limitado. Por tanto, transferir ideas del ámbito del aprendizaje máquina al incipiente campo de investigación en causalidad se presenta prometedor. Además de esto, un enfoque desde el paradigma causal puede dar un sustento teórico mayor a los distintos problemas del aprendizaje máquina, ofreciendo visiones alternativas a la Teoría del Aprendizaje Estadístico (Vapnik, 1995).

1.4.1. Máquinas de vectores soporte en clasificación

Dentro de la disciplina del aprendizaje máquina, las máquinas de vectores soporte (SVM[4]) pueden considerarse el estado del arte en discriminación, especialmente binaria (Burges, 1998). Además, su desarrollo tiene un sólido fundamento teórico en la Teoría del Aprendizaje Estadístico (Vapnik, 1995, 1998). En este apartado se introducen los conceptos más importantes que serán empleados en capítulos posteriores.

Se supondrá que se dispone de un conjunto muestral formado por N muestras de dimensión d: $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$, donde $\mathbf{x}_i \in \mathbb{R}^d$ e $y_i \in \{+1, -1\}$. El hiperplano $\mathbf{w}$ que separa las muestras, suponiéndolas linealmente separables, con un mayor margen viene dado por la minimización del funcional:

[4] "Support Vector Machine".

$$\min_{\mathbf{w},b} \quad \frac{1}{2}||\mathbf{w}||^2 \tag{1.9}$$

$$\text{restringido a} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \tag{1.10}$$

con $i = 1, \ldots, N$, que es solucionable mediante multiplicadores de Lagrange y programación cuadrática (Boyd y Vandenberghe, 2004).

Se van a introducir dos modificaciones al funcional en (1.9) y (1.10). Por un lado se va a permitir que los conjuntos de las clases $+1$ y -1 sean no separables de manera lineal; por otro lado, se va aplicar una transformación no lineal $\phi(\mathbf{x}_i)$, a un espacio de dimensión superior, llamado espacio de características: $\phi(\cdot), \mathbb{R}^d \xrightarrow{\phi(\cdot)} \mathbb{R}^H$. El nuevo funcional es el siguiente:

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i \tag{1.11}$$

$$\text{restringido a} \quad y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \tag{1.12}$$

$$\xi_i \geq 0 \tag{1.13}$$

Las variables ξ_i permiten que se cometan violaciones del margen, haciendo resolubles los problemas no separables. Una muestra con $\xi_i > 1$ se ubicará en la zona del espacio correspondiente a la otra clase. Para controlar estos errores, se incluye en el funcional a minimizar el término $\sum_i \xi_i$. El parámetro C, o coste de la SVM, controla el balance entre la suavidad de la solución y el número de errores cometidos.

La solución de (1.11)-(1.13) se puede obtener mediante el método de los multiplicadores de Lagrange, introduciendo las restricciones, multiplicadas por α_i y μ_i, en el funcional:

$$L_P = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i - \sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^T\phi(\mathbf{x}) + b) - 1 + \xi_i) \tag{1.14}$$

La solución que minimiza (1.14) respecto a $\mathbf{w}$, b, ξ_i, $-\alpha_i$ y $-\mu_i$ está determinada por las condiciones KKT, de Karush-Kuhn-Tucker (Burges, 1998; Schölkopf y Smola, 2001). Se deben cumplir las restricciones (1.12) y (1.13), los multiplicadores (α_i y μ_i) por sus restricciones deben anularse, deben ser estrictamente positivos, y, finalmente, las derivadas parciales deben ser nulas:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \phi(\mathbf{x}_i) = 0 \tag{1.15}$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{1.16}$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \mu_i = 0 \tag{1.17}$$

La Ecuación 1.15 determina que la solución a la SVM es una combinación lineal de las muestras en el espacio de características: $\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \phi(\mathbf{x}_i)$. Siendo esto así, la SVM plantea soluciones de la forma:

$$\begin{aligned} f(\mathbf{x}) &= \text{signo}\left(\mathbf{w}^T \phi(\mathbf{x}) + b\right) = \\ &= \text{signo}\left(\sum_{i=1}^{N} \alpha_i y_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x})\right) = \qquad (1.18) \\ &= \text{signo}\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x})\right) \qquad (1.19) \end{aligned}$$

En esta última ecuación, el producto interno de funciones $\phi(\cdot)$ se puede sustituir por una función $K(\mathbf{x}_i, \mathbf{x}) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x})$, llamada *núcleo*. De esta manera, no es necesario conocer explícitamente la función de transformación, tan sólo la función núcleo. Las funciones núcleo deben cumplir el Teorema de Mercer (Vapnik, 1998) para representar un producto escalar en un espacio de Hilbert. Las más habituales son el núcleo lineal (1.20), polinómico (1.21) y gausiano (1.22).

$$K(\mathbf{x}_i, \mathbf{x}) = \mathbf{x}_i^T \mathbf{x} \tag{1.20}$$

$$K(\mathbf{x}_i, \mathbf{x}) = (\mathbf{x}_i^T \mathbf{x} + 1)^p \tag{1.21}$$

$$K(\mathbf{x}_i, \mathbf{x}) = e^{-||\mathbf{x}_i - \mathbf{x}||^2/(2\sigma^2)} \tag{1.22}$$

En los núcleos polinómico y gausiano se han introducido unos nuevos hiperparámetros, p, el orden del polinomio, y σ, el ancho de la función de base radial. Junto con el coste C de (1.11), serán los valores que controlarán la capacidad de generalización de la máquina, balanceando la regularización con la comisión de errores. El procedimiento habitual de optimización de

los hiperparámetros de la SVM se lleva a cabo mediante validación cruzada (Schölkopf y Smola, 2001).

El papel de los multiplicadores de Lagrange α_i en (1.19) es importante; el número de α_i distintos de cero es el número de vectores soporte, siendo estos vectores aquellas muestras que participan en la solución. La inclusión de las variables ξ_i hace que las α_i estén acotadas superiormente por el coste: $0 \leq \alpha_i \leq C$. El hecho de que sólo los valores cercanos al hiperplano de separación intervengan en la solución hace que la SVM sea dispersa.

El desarrollo realizado, y la solución en (1.19), definen una SVM para clasificación binaria. Es posible adaptar la máquina para resolver problemas de clasificación multiclase, donde las etiquetas y_i pueden adoptar cualquier valor en un conjunto finito. También existe una versión de la SVM para resolver problemas de regresión (Schölkopf y Smola, 2001).

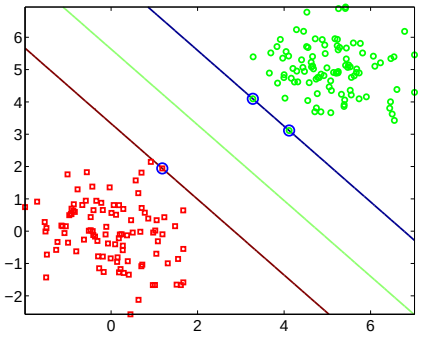
En la Figura 1.9 se muestran varios casos de clasificación con SVM. En (a) se puede ver una clasificación con núcleo lineal, de un conjunto de datos separable linealmente. La frontera de clasificación se dibuja en trazo verde, y los planos $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$ en azul y marrón. Los vectores soporte se han resaltado con círculo exterior. Como se ve en la Figura 1.9(a), sólo tres muestras son vectores soporte.

Las gráficas de la Figura 1.9(b)-(d) representan la solución a un problema no separable linealmente. El núcleo empleado en la SVM es gausiano, y se ha variado el parámetro σ de la Ecuación 1.22 para valores $\sigma = \{0.4, 1.5, 7\}$. Como puede verse en la Figura 1.9(b), un valor pequeño del ancho de la gausiana conduce a una solución sobreajustada. La probabilidad de acierto de la SVM en un conjunto de muestras no empleado para entrenar el clasificador es del 92.6 %. La máquina tiene una frontera de clasificación muy compleja, ajustada en exceso al conjunto de entrenamiento. Este hecho también se comprueba en el número de vectores soporte, que asciende al 79 % de las muestras de entrenamiento. Un valor muy alto del parámetro σ, como en la Figura 1.9(d), genera una solución demasiado regularizada, que generaliza mal (un acierto del 86.4 %). El valor óptimo del ancho de la gausiana se sitúa en torno a $\sigma = 1.5$, figura (c), que con un número bajo de vectores soporte (28), alcanza en generalización un acierto del 96.4 %. En esta gráfica se comprueba cómo los vectores soporte están en el margen entre las clases.

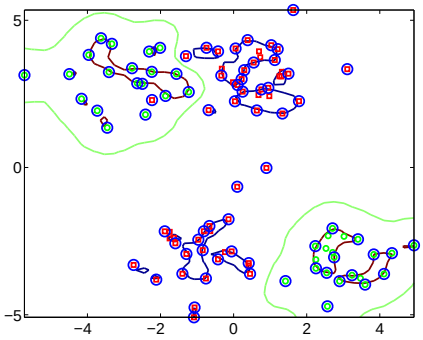
La SVM servirá de base para los algoritmos de inferencia causal que se desarrollarán.

1.4.2. Causalidad y selección de variables

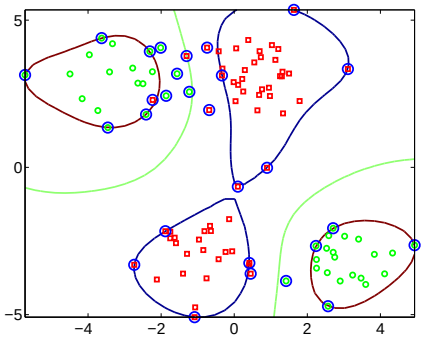
Una posible aplicación del paradigma causal es la selección de características (FS, "Feature Selection", por sus siglas inglesas) (Guyon et al., 2007).



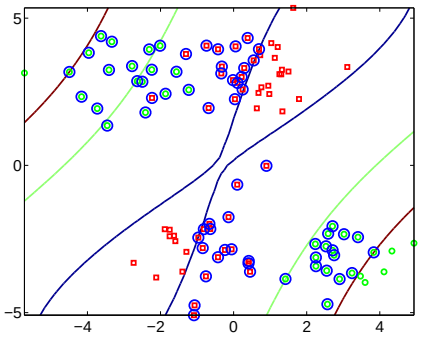
(a) SVM lineal.



(b) SVM gausiana. $\sigma = 0.4$. Ac= 92.6 %. #SV= 79.



(c) SVM gausiana. $\sigma = 1.5$. Ac= 96.4 %. #SV= 28.



(d) SVM gausiana. $\sigma = 7$. Ac= 86.4 %. #SV= 68.

Figura 1.9: Ejemplo de clasificadores binarios SVM. Los círculos verdes representan la clase $y_i = 1$, y los cuadrados rojos la clase -1. La frontera de clasificación se dibuja en verde y los márgenes $y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) = 1$ en azul y marrón. Los vectores soporte se circunscriben para resaltarlos.

Los métodos de FS actuales son capaces de extraer información relevante de conjuntos con pocas muestras y muchas variables, si bien los resultados son poco interpretables y por lo general carecen de fundamentación teórica, más allá de que son heurísticos para acelerar la búsqueda (Leiva-Murillo, 2007). La selección de características causal (CFS) hasta ahora sólo se ha podido aplicar a bases de datos amplias y con pocas dimensiones. La gran ventaja es la propia de los métodos causales: interpretabilidad y justificación, posibilidad de intervención y buena fundamentación teórica.

Además, la asunción habitual en FS es disponer de muestras i.i.d[5]. Con el paradigma causal se podrá relajar esta condición y permitir cambios en la distribución de las causas. La asunción que debe mantenerse es la de estacionariedad en la distribución condicional $P(Y|X)$. Esto es así porque si el modelo causal responde a la realidad, el mecanismo que relaciona el efecto con su causa no cambia, aún cuando ésta última sí lo haga. Finalmente, intervenir en una causa modificará su efecto, cosa que no sucede en el sentido inverso. Una selección de variables causal permite realizar intervenciones en el problema asegurando una repercusión en la variable bajo estudio.

Se presenta ahora el esquema de relevancia de variables desarrollado en (Kohavi y John, 1998). Posteriormente se relaciona con el paradigma causal, siguiendo las ideas de (Guyon et al., 2007). Se amplía la notación anterior, donde las muestras eran $\mathbf{x} = \{X_1, \ldots, X_i, \ldots, X_d\}$ con los nuevos conjuntos de muestras $\mathbf{x}^{\setminus i} = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d\}$, en el que se ha eliminado la componente i. También se usa el subconjunto $\mathbf{v}^{\setminus i} \subseteq \mathbf{x}^{\setminus i}$. Es posible definir los siguientes tres conceptos.

- *Variables irrelevantes*. Una variable X_i es irrelevante para otra variable objetivo Y si y sólo si para todo subconjunto $\mathbf{v}^{\setminus i}$ se cumple:

$$P(X_i, Y|\mathbf{v}^{\setminus i}) = P(X_i|\mathbf{v}^{\setminus i})P(Y|\mathbf{v}^{\setminus i}) \tag{1.23}$$

- *Relevancia fuerte*. Una variable X_i es fuertemente relevante para otra variable Y si y sólo si existen algunos valores $x, y, \mathbf{v}$ con $P(X_i = x, \mathbf{x}^{\setminus i} = \mathbf{v}) > 0$ tal que:

$$P(Y = y|X_i = x, \mathbf{x}^{\setminus i} = \mathbf{v}) \neq P(Y = y|\mathbf{x}^{\setminus i} = \mathbf{v}) \tag{1.24}$$

- *Relevancia débil*. Una variable X_i es débilmente relevante para otra variable Y si y sólo si no es fuertemente relevante y existe un subconjunto de variables $\mathbf{v}^{\setminus i}$ para el que existen algunos valores $x, y, \mathbf{v}$ con $P(X_i = x, \mathbf{v}^{\setminus i} = \mathbf{v}) > 0$ tal que:

$$P(Y = y|X_i = x, \mathbf{v}^{\setminus i} = \mathbf{v}) \neq P(Y = y|\mathbf{v}^{\setminus i} = \mathbf{v}) \tag{1.25}$$

Estas definiciones están basadas en la independencia estadística entre las variables X_i e Y, condicionadas sobre otro subconjunto de variables. En (Kohavi y John, 1998) también se distingue entre los conceptos de *relevancia*, según las definiciones previas, y el de *utilidad*. Este último está relacionado

[5]Independientes e idénticamente distribuidas.

con la capacidad predictiva de las variables; por lo tanto una variable relevante (fuerte o débil) puede no ser útil, si, por ejemplo, la información que aporta es redundante a la de otras variables.

Si se analizan estos conceptos desde el punto de vista del paradigma causal, se puede dar una interpretación más robusta al problema de la selección de variables.

- *"Markov Blanket"*. El subconjunto de nodos MB es el "Markov Blanket"[6] de Y si y sólo si para cualquier subconjunto V disjunto de MB se cumple $Y \perp\!\!\!\perp V|\text{MB}$.

  El MB comprende los padres de Y, los hijos y los padres de los hijos.

  Es posible realizar una asignación entre los componentes del MB y los grados de relevancia desde un punto de vista de FS.

  1. Irrelevancia: nodos no conectados con Y.
  2. Relevancia débil: nodos conectados con Y pero $\notin$ MB.
  3. Relevancia fuerte: directamente el conjunto MB.

Un ejemplo de algoritmo que emplea el "Markov Blanket" de manera eficiente para hacer selección de características es el método "HITON" (Aliferis et al., 2003).

1.4.3. Remuestreo "bootstrap" y test de hipótesis

Los algoritmos que se se van a presentar más adelante requieren de la estimación de distribuciones de probabilidad. También será necesario emplear test de hipótesis con el objetivo de discriminar entre varias posibilidades. Se exponen someramente los fundamentos del remuestreo "bootstrap", así como del test de hipótesis.

Remuestreo "bootstrap" para estimación de distribuciones

Se presenta a continuación una herramienta que será necesario emplear en el algoritmo de búsqueda de causalidad propuesto. Se trata de una técnica introducida por Bradley Efron (Efron y Tibshirani, 1993), que tiene aplicación en la determinación de la calidad de estimaciones estadísticas. En general, es una técnica computacional para extraer conocimiento de bases de datos.

[6] "Markov Blanket" o cobertura markoviana. Se empleará el nombre en inglés en adelante.

La idea en la que se basa esta técnica de remuestreo es muy sencilla, y se estructura en dos pasos. Se parte únicamente de un conjunto de muestras, sin asumir ningún otro conocimiento a priori. A partir de esos datos se estima su distribución según la Función de Distribución Empírica (EDF por sus siglas en inglés). Esta estima posee un buen comportamiento asintótico. En una segunda etapa, se realiza un remuestreo de Monte Carlo siguiendo la EDF.

El resultado es, desde un punto de vista computacional, equivalente a un remuestreo con repetición de la base de datos original. Finalmente, para cada uno de esos nuevos conjuntos de datos, o *réplicas bootstrap*, se aplica el estadístico que se desea analizar, que puede ir desde una sencilla media muestral hasta una elaborada estimación de la densidad espectral de potencia (Zoubir y Iskander, 2004).

No es posible, sin embargo, hacer un análisis teórico de la convergencia del algoritmo en un caso general; éste deberá realizarse para cada estadístico concreto.

En el libro de Efron (Efron y Tibshirani, 1993) se presenta una abstracción del método, que se replica en la Figura 1.10. En la parte de la izquierda de la figura se representa el esquema habitual. Se ha observado un conjunto de datos $\mathbf{X} = \{X_1, \dots X_n\}$, generados por una función de distribución desconocida, F. A estos datos se les aplica una función $s(\mathbf{X})$ para estimar cierto estadístico de interés $\hat{\theta}$.

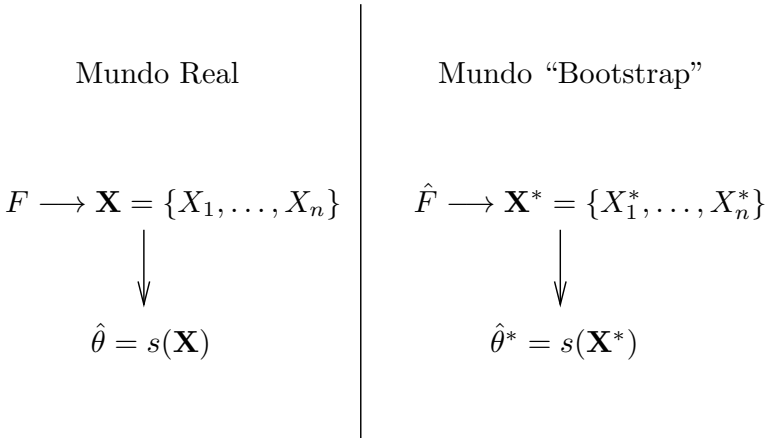


Figura 1.10: Aproximación "bootstrap" (Efron y Tibshirani, 1993).

En la aproximación "bootstrap", el primer paso consiste en obtener una aproximación a la función de distribución $F(x)$. Este paso se realiza mediante la *Función de Distribución Empírica*. La estimación mencionada asigna un

peso de probabilidad de $1/n$ a cada muestra; la ecuación queda como:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} u(x - x_i) \tag{1.26}$$

donde $u(x)$ es la función escalón,

$$u(x) = \begin{cases} 1 & \text{si } x > 0. \\ 0 & \text{resto} \end{cases}$$

En los casos en los que la variable x sea multidimensional, la función escalón será no nula cuando todas las componentes de x sean mayores que cero. Un ejemplo de esta estima se puede ver en la Figura 1.11, donde se han generado 10, 50 y 1000 muestras independientes e idénticamente distribuidas de una misma distribución desconocida F. Se aprecia cómo la suavidad de la curva $\hat{F}_n(x)$ aumenta sensiblemente con el número de muestras.

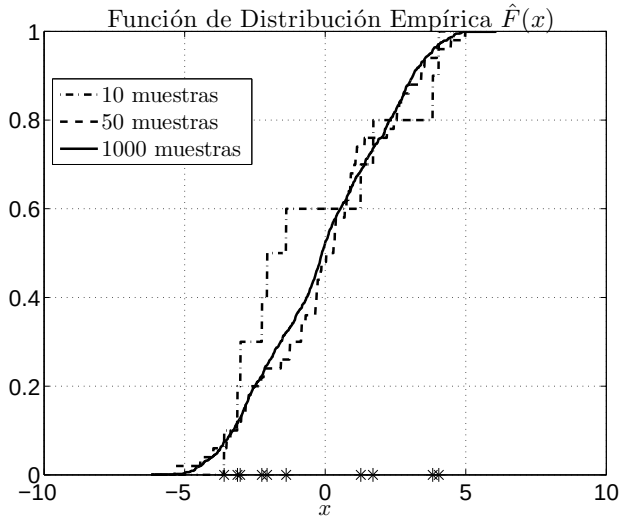


Figura 1.11: Comparación de la estima empírica de la Función de Distribución para tres conjuntos de muestras, de tamaño creciente. En el eje de abscisas están representadas las muestras del conjunto de datos más reducido.

Esta aproximación, según el teorema de Glivenko-Cantelli, (Vapnik, 1998), converge en probabilidad a la distribución real en el límite de muestras infinitas. Con más precisión, esta propiedad queda enunciada según el Teorema 1:

Teorema 1. *La siguiente convergencia en probabilidad es cierta:*

$$\sup_x |F(x) - \hat{F}_n(x)| \xrightarrow[n\to\infty]{P} 0$$

Este resultado provee la base teórica que justifica el buen comportamiento de la Función de Distribución Empírica en el remuestreo "bootstrap".

La estima de la EDF es la herramienta que permite el paso del "Mundo Real" al "Mundo bootstrap", según se muestra en la Figura 1.10.

La segunda etapa del algoritmo "bootstrap" consiste en la generación de múltiples copias siguiendo el método de Monte Carlo. Este método es de utilidad cuando se desea realizar un cálculo que implica una distribución de probabilidad, y cuando el desarrollo matemático del cálculo es demasiado complejo o costoso. El cálculo exacto se reemplaza por una colección de conjuntos de muestras $x_1, \ldots, x_n \overset{\text{iid}}{\sim} F(x)$. Una de las ventajas de este método numérico es la baja sensibilidad al número de dimensiones, por lo que resulta útil su empleo en problemas de alta dimensionalidad como por ejemplo el filtrado de partículas (Gordon et al., 1993).

El procedimiento para implementar el método "bootstrap" se recoge en los siguientes pasos, donde se aprecia la aplicación sucesiva de la estima EDF y el método Monte Carlo.

1. Recoger las muestras del experimento aleatorio: $X = \{x_1, \ldots, x_n\}$.

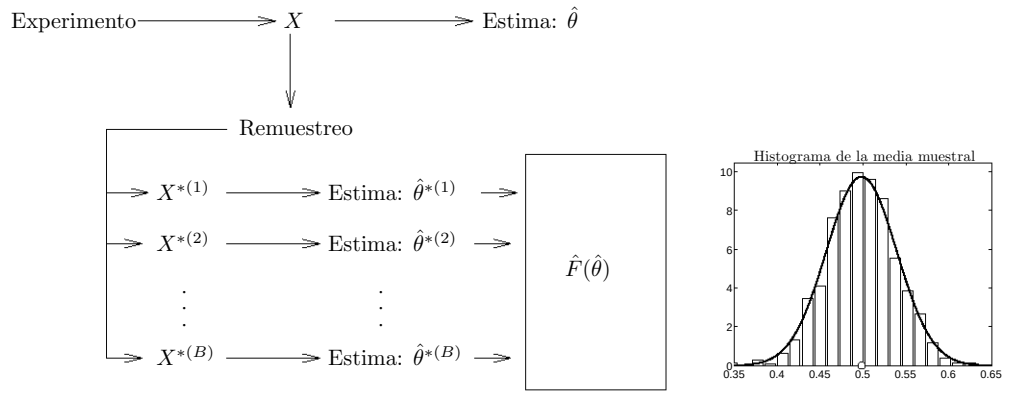
2. Calcular la Función de Distribución Empírica $\hat{F}_n(x)$ a partir de las muestras X.

3. Obtener nuevas réplicas "bootstrap" a partir de $\hat{F}_n(x)$, $X^* = \{x_1^*, \ldots\ x_n^*\}$.

4. Aproximar la distribución $\hat{F}(\hat{\theta})$ a partir de la distribución de $\hat{\theta}^*$ obtenida de las réplicas "bootstrap".

En la práctica no es necesario realizar el paso intermedio de estimar la EDF. Un remuestreo con repetición genera las mismas réplicas X^*, por lo que se simplifica notablemente el procedimiento. Únicamente hay que extraer con una probabilidad uniforme muestras del conjunto original X, hasta completar réplicas "bootstrap" de tamaño n. Cada réplica, por tanto, contiene n muestras del conjunto original, de forma que las muestras pueden aparecer varias veces o no aparecer. Dos réplicas "bootstrap" válidas, para un conjunto original de dimensión 4, $X = \{x_1, x_2, x_3, x_4\}$, pueden ser:

$$\begin{aligned} X^{*(1)} &= \{x_1, x_1, x_4, x_3\} \\ X^{*(2)} &= \{x_2, x_2, x_2, x_2\} \end{aligned}$$



(a) Procedimiento "bootstrap". Se remuestrea el conjunto original con reinserción y con distribución uniforme. A cada muestra "bootstrap" se le aplica el estadístico de interés.

(b) Histograma de la media muestral. La curva continua es una normal $N(\hat{\theta}, \sigma_X/\sqrt{n})$. En el eje de abscisas se indica la media muestral $\hat{\theta}$.

Figura 1.12: Procedimiento "bootstrap" y ejemplo de obtención de la distribución de un estadístico.

En la primera réplica no se encuentra presente la muestra x_2, y la réplica $X^{*(2)}$ consiste en una repetición de únicamente la muestra x_2.

El procedimiento se describe de manera gráfica en la Figura 1.12(a).

A modo de ejemplo, se ha calculado la media muestral de un conjunto $X = \{x_1, \ldots, x_n\}$ con $n = 100$ muestras generadas a partir de una distribución uniforme entre 0 y 1. Este conjunto X se ha remuestreado $B = 1000$ veces. El estadístico de interés en este caso es la media, y se aproximará por la media muestral:

$$\hat{\theta} = s(X) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.27}$$

y su aproximación en el "mundo bootstrap" es:

$$\hat{\theta}^* = s(X^*) = \frac{1}{n}\sum_{i=1}^{n} x_i^* \tag{1.28}$$

En caso que $E[X^2] < \infty$, entonces, como se muestra en (Shao y Tu, 1995), se puede demostrar que:

$$\sup_x \left| P_* \left(\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \leq x \right) - P\left(\sqrt{n}(\hat{\theta} - \theta) \leq x \right) \right| \to 0 \quad \text{a.s.} \tag{1.29}$$

Como es sabido, la distribución de la media muestral es una gausiana de media la media de la distribución, en este caso la de $U(0,1)$, y de varianza $\frac{\sigma_X^2}{n}$. Esto se representa en la Figura 1.12(b).

Las técnicas "bootstrap" se pueden emplear para otras aplicaciones, entre las que se encuentran la estima de errores estándar y de intervalos de confianza, test de permutación y de hipótesis, regresión o modelado autorregresivo. También es posible realizar un "bootstrap" paramétrico. A continuación se desarrolla el cálculo de un test de hipótesis con "bootstrap", pues será necesario para uno de los algoritmos que se presentan posteriormente.

Tests de hipótesis con "bootstrap"

Un test de hipótesis binario (Scharf, 1991) sirve para comprobar si una determinada hipótesis es o no válida. Si la hipótesis es cierta, se dirá que no es posible rechazar H_0, llamada también hipótesis nula. Por el contrario, si se tiene suficiente certeza de que H_0 no es correcta, el test se decantará por la alternativa H_1.

Para realizar el modelado matemático de este esquema de decisión, se asigna una función de densidad de probabilidad a los datos del experimento, que se suponen parte de la familia paramétrica $\{f_{\boldsymbol{\theta}}(\mathbf{x}) : \boldsymbol{\theta} \in \Theta\}$. Los posibles valores de $\boldsymbol{\theta}$ se dividen en dos conjuntos disjuntos, tal que $\Theta = \Theta_0 \cup \Theta_1$, y $\Theta_0 \cap \Theta_1 = \emptyset$. Por consiguiente, otra manera de expresar el test es:

$$\begin{cases} H_0 : & \boldsymbol{\theta} \in \Theta_0 \\ H_1 : & \boldsymbol{\theta} \in \Theta_1 \end{cases} \tag{1.30}$$

para el caso de test binarios. Si las hipótesis a verificar son múltiples, se dice que es un test M-ario. Si la cardinalidad de los conjuntos Θ_i es la unidad, la hipótesis es simple; de lo contrario, es compuesta.

Un test binario toma la forma:

$$\phi(\mathbf{x}) = \begin{cases} 1 \sim H_1, & \mathbf{x} \in R \\ 0 \sim H_0, & \mathbf{x} \in A \end{cases} \tag{1.31}$$

Es decir, el espacio de $\mathbf{x}$ se particiona en dos áreas, A y R, que indican respectivamente las zonas en las que se acepta y rechaza la hipótesis nula H_0. La probabilidad de falsa alarma se define como

$$\alpha = P_{\boldsymbol{\theta}_0}\left(\phi(\mathbf{x}) = 1\right) = \int_{\mathbf{x} \in R} \phi(\mathbf{x}) f_{\boldsymbol{\theta}_0}(\mathbf{x}) d\mathbf{x} \tag{1.32}$$

es decir, la probabilidad de que, siendo cierta la hipótesis nula H_0, se elija H_1. En la Figura 1.13, ejemplo de un caso unidimensional y unimodal, se

correspondería con la región sombreada en horizontal. La probabilidad de detección por el contrario, es:

$$\beta = P_{\theta_1}\left(\phi(\mathbf{x}) = 1\right) = \int_{\mathbf{x}\in R} \phi(\mathbf{x}) f_{\theta_1}(\mathbf{x}) d\mathbf{x} \tag{1.33}$$

es decir, la probabilidad de rechazar correctamente la hipótesis nula y decidir consecuentemente H_1. El umbral entre las regiones R y A se ha situado en el ejemplo de la Figura 1.13 en el punto donde $f_{\theta_0}(x) = f_{\theta_1}(x)$.

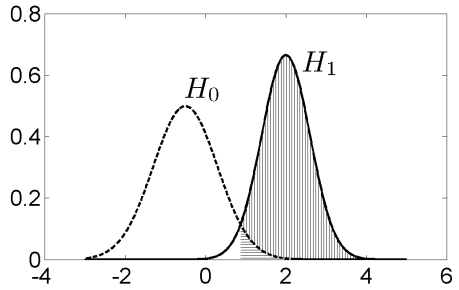


Figura 1.13: Funciones de densidad de probabilidad de un test de hipótesis. Se muestran la probabilidad de falsa alarma, α (área rallada en horizontal) y de la probabilidad de detección, β (área rallada) cuando el umbral es u : $f_{\theta_0}(u) = f_{\theta_1}(u)$.

El test de Neyman-Pearson define el test de mayor capacidad de detección (UMP por las siglas inglesas de *Universal Most Powerfull*). El mejor test según este criterio es aquel que a igualdad de probabilidad de falsa alarma, tiene una mayor probabilidad de detección. El lema de Neyman-Pearson aplica a test simples, e identifica las regiones de decisión óptimas. Si el conjunto de parámetros es $\Theta = \{\theta_0, \theta_1\}$, y la densidad de probabilidad de los datos es f_{θ_i}, y llamando $l(\mathbf{x})$ al cociente de verosimilitud $l(\mathbf{x}) = \frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}$, el lema queda así:

$$\phi(\mathbf{x}) = \begin{cases} 1, & l(\mathbf{x}) > k \\ 0, & l(\mathbf{x}) < k \end{cases} \tag{1.34}$$

para aquellos casos en los que sea nula la probabilidad de que $l(\mathbf{x}) = k$. El umbral k se localiza fijando la probabilidad de falsa alarma α:

$$\alpha = \int_k^\infty f_{\theta_0}(l) dl = P_{\theta_0}(l(\mathbf{x}) > k) \tag{1.35}$$

Cuando la distribución no sea conocida, o no haya suficientes muestras para que no apliquen los métodos asintóticos, el test de hipótesis "bootstrap"

puede resultar interesante (Zoubir y Iskander, 2004). Se va a modificar ligeramente el test simple presentado anteriormente, pues será más útil de cara a los algoritmos de inferencia causal. En este caso la hipótesis a probar es $H_0 \sim \theta_0 = 0$ contra $H_1 \sim \theta_1 > 0$. Se supondrá que $\hat{\theta}$ y $\hat{\sigma}^2$ son, respectivamente, el estimador de θ y de su varianza. Se define el test estadístico

$$T_n = \frac{\hat{\theta} - 0}{\hat{\sigma}} \tag{1.36}$$

donde la normalización por la desviación típica se incluye para asegurar que el estadístico sea fundamental ("pivotal" en inglés), es decir, que sea independiente de la varianza de los datos. Si la distribución de T_n fuera conocida, se podría emplear un test de hipótesis clásico, fijando la probabilidad de falsa alarma α y usando la Ecuación 1.35.

Siguiendo la aproximación "bootstrap", se procede a remuestrear los datos para aplicar sobre las réplicas $X^{*(b)}$ el estadístico $T_n^{*(b)}$. Se obtiene así el conjunto $T_n^{*(1)}, \ldots, T_n^{*(B)}$. Una vez ordenadas las salidas $T_n^{*[1]} \leq \ldots \leq T_n^{*[B]}$, el umbral se calcula como: $T_\alpha = T_n^{*[q]}$, donde $q = \lceil (1-\alpha)(B+1) \rceil$.

1.5. Métodos de inferencia causal basados en muestras

Si bien uno de los principales investigadores en causalidad, Judea Pearl, niega la posibilidad de realizar inferencia causal a partir de muestras para el caso general, bajo las asunciones de minimalidad y estabilidad pueden generarse verosímilmente inferencias causales. De hecho, son muchos los métodos que se han ensayado para recuperar la red bayesiana causal a partir de un conjunto de muestras. Este autor argumenta en (Pearl, 2009, Sec. 2.9) que estadística y causalidad han de diferenciarse con claridad. No reconoce como posible ninguna hibridación, más allá de la necesaria para la teoría de redes bayesianas.

El Cuadro 1.1 recoge conceptos pertenecientes al campo de la estadística y de la causalidad, según dicho autor.

Judea Pearl entiende que el árbol causal debe obtenerse como una hipótesis de trabajo, en función del conocimiento que se tiene del problema. Una vez dado aquél, se pueden llevar a cabo operaciones causales sobre él, que pueden validar o no la hipótesis de partida. De esta forma, conceptos económicos como los que sustentan el trabajo de Granger (Granger, 1969), Pearl los encuadra en el ámbito estadístico. Este mismo razonamiento llevaría a afirmar que los modelos generativos, que se basan en ecuaciones como $P(y_t|y_{t-1}, x_t)$

Estadística	Causalidad
Inferencia	Manipulación
Causalidad de Granger	Aleatorización
Independencia condicional	Efecto
Correlación	"Markov Blanket"
Regresión	Contrafáctico
Verosimilitud	"Confounding"

Cuadro 1.1: Conceptos enfrentados y excluyentes de estadística y causalidad. Basado en (Pearl, 2000, Sec. 1.5).

son meramente estadísticos, a menos que se asuman ciertas propiedades causales del modelo de generación de datos. Los filtros de Kalman, por ejemplo, entrarían en esta categoría.

A pesar de todo lo anterior, se han desarrollado numerosos métodos y se ha comprobado su eficacia a la hora de recuperar una estructura causal que replique los mecanismos que gobiernan un cierto conjunto de variables.

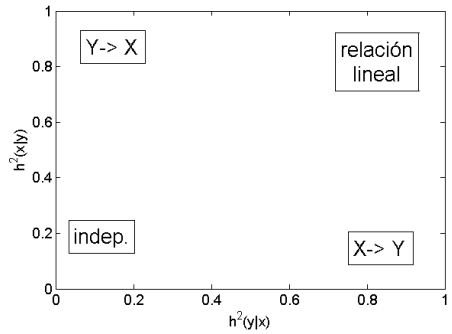
Ejemplo de criterio causal

A modo de ejemplo introductorio a esta sección de inferencia causal, se presenta un método muy sencillo para orientar la relación entre dos series temporales. En los apartados posteriores se realiza una taxonomía general de los métodos, pero sirva este ejemplo de ilustración. En este caso, siguiendo el método expuesto en (Kalitzin et al., 2007), se calcula el llamado *índice* h^2, que intenta cuantificar el grado de asociación no lineal entre las señales $x(t)$ e $y(t)$ definido como:

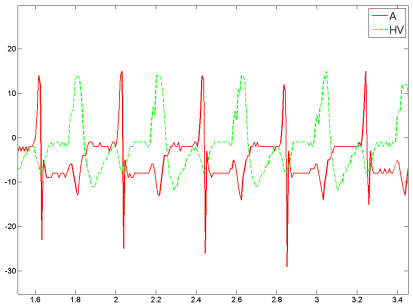
$$\begin{aligned} h^2(y|x) &= 1 - \frac{1}{N\sigma_y^2} \sum_{a=1}^{M} \sum_{i, x_i \in B_a} (y_i - \langle y \rangle_a)^2 \\ \langle y \rangle_a &= \frac{1}{N a_a} \sum_{i, x_i \in B_a} y_i, \\ \sum_{a=1}^{M} N_a &= N \end{aligned} \tag{1.37}$$

La interpretación de esta fórmula es la siguiente: el dominio de x se divide en una cobertura de M regiones, llamadas B_a. Para cada una de ellas se calcula la varianza de $y(t) \quad \forall t : x(t) \in B_a$.

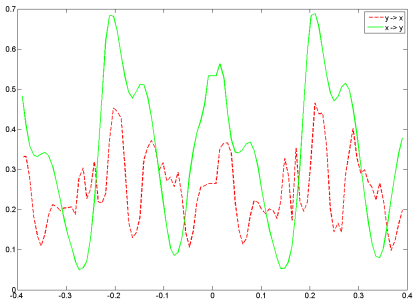
Es preciso hacer notar los siguientes puntos, que se muestran de manera gráfica en la Figura 1.14(a):



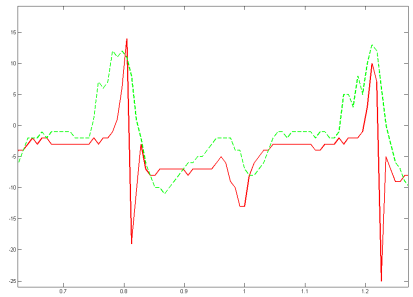
(a) Interpretación de $h^2(x|y)$ y $h^2(y|x)$.



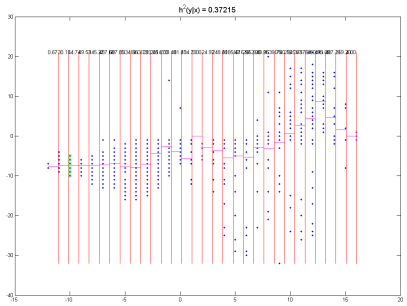
(b) Señales Atip-Aring y HVA-HVB.



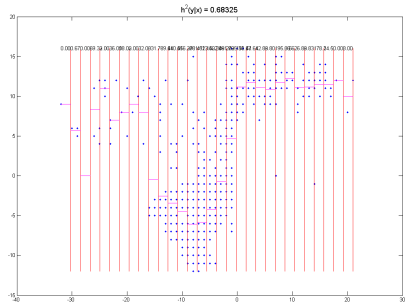
(c) Índice h^2 retardado para los dos sentidos.



(d) Señales alineadas según el máximo h^2.



(e) HV→A.



(f) A→HV.

Figura 1.14: Ejemplo de algoritmo para orientar el flujo de información entre dos series temporales empleando el índice h^2.

- El índice está acotado: $0 \leq h^2 \leq 1$.

- Si la varianza de $y(t)$ dentro de cada subregión B_a es pequeña, el valor de $h^2(y|x)$ será alto, significando que existe una transformación no lineal de x en y.
- Si existe una relación lineal entre x e y entonces tanto $h^2(y|x)$ como $h^2(x|y)$ tienen que ser próximos a 1.
- Si $x(t)$ e $y(t)$ son estadísticamente independientes h^2 tiende a 0.

Basado en (Kalitzin et al., 2007) se ha buscado una ordenación causa/efecto entre dos electrogramas medidos simultáneamente por un DAI (Desfibrilador Automático Implantable) bicameral (Hesselson, 2004). Un ejemplo de las señales se muestra en la Figura 1.14(b). La señal dibujada con línea continua Atip-Aring mide el electrograma auricular; la señal dibujada en trazo discontinuo, HVA-HVB, mide la tensión entre la caja del DAI y la sonda de desfibrilación en el ventrículo izquierdo. Por consiguiente, como la estimulación cardíaca se propaga desde las aurículas hacia los ventrículos, existe una relación causal desde Atip-Aring hacia HVA-HVB.

En la Figura 1.14(c) se han dibujado los índices $h^2(y|x)$ (verde continuo) y $h^2(x|y)$ (rojo discontinuo). El máximo del índice h^2 se obtiene con un retardo de 200 msg.; empleando dicha información, se ha procedido a alinear las señales, como se ve en 1.14(d). Los valores promedios de $h^2(y|x)$ (es decir, Atip-Aring es causa de HVA-HVB) y de $h^2(x|y)$ (HVA-HVB es causa de Atip-Aring) son $h^2(y|x) = 0.68$ y $h^2(x|y) = 0.37$. Con estas cifras, el algoritmo se decantaría, acertadamente, porque la señal auricular es causa de la ventricular.

Sin embargo, el valor del índice en el sentido contrario (señal ventricular es causa de la señal auricular) es relativamente alto con $h^2(x|y) = 0.37$. Como se ha indicado, este método sólo detecta causalidad si las relaciones son no lineales. Si bien es cierto que el proceso físico de propagación por el miocardio es no lineal (nodo A-V, fibras de Purkinje), las señales Atip-Aring y HVA-HVB son relativamente parecidas. Además, el sensado en la aurícula es local, y el ventricular se realiza con sensores más lejanos, que pueden captar información de las aurículas.

Este algoritmo ha intentado buscar asimetrías en la relación entre dos series temporales, basándose en la *intuición* de que la representación cartesiana de las señales $x(t)$ (en abscisas) e $y(t)$ (en ordenadas) y su contraria, como se muestra en las Figuras 1.14(e-f), pueden hacer significativa la dirección en la que se propaga la información.

En los siguientes apartados se sistematizará este proceso para diversos criterios. En primer lugar se abordarán técnicas para el caso de variables discretas y posteriormente, en la Sección 1.7 para series temporales.

1.6. Métodos de búsqueda causal en el dominio discreto

A continuación se listan los diversos algoritmos presentes en la literatura para realizar inferencia causal en el dominio discreto; también se agrupan en función de la propiedad que explotan para localizar la influencia causal.

1.6.1. Métodos basados en restricciones

El primero de los algoritmos que se diseñó para la obtención de relaciones de causalidad es conocido como algoritmo "Inductive Causation" (IC). También se le conoce por PC, debido a una modificación del mismo presentada por Peter Spirtes y Clark Glymour (Spirtes et al., 2000). El número de grafos válidos en función del número de nodos d crece superexponencialmente, como se vio en la Ecuación 1.3. No es viable por lo tanto una búsqueda exhaustiva.

El método IC empieza con un grafo DAG completamente desconectado. El algoritmo se desarrolla de manera iterativa siguiendo los pasos indicados en el Algoritmo 1. Dado el grafo $\mathcal{G}$, formado por el conjunto de nodos $V = \{X_1, X_2, X_3, \ldots, X_d\}$, se asumirá de partida que está totalmente desconectado, por lo que el conjunto de enlaces será $E = \emptyset$. La distribución de probabilidad asociada a $\mathcal{G}$ se asume que es $P(\mathbf{x})$.

Algoritmo 1 Algoritmo de Inducción Causal (IC) estándar

1: Para cada par de variables X_i, X_j, buscar el conjunto $S_{XY} \in V$ tal que $X \perp\!\!\!\perp Y | S_{XY}$, esto es, que X e Y sean independientes condicionadas al conjunto S_{XY}. Establecer una conexión entre $X - Y$ si y sólo si no existe ningún conjunto de variables S_{XY} que cumpla dicha condición.

2: En las estructuras tipo $X - Z - Y$ (con X e Y no adyacentes), orientar las conexiones así: $X \rightarrow Z \leftarrow Y$, en caso que $Z \notin S_{XY}$. Esta estructura se conoce como estructura en "V".

3: Orientar todas las conexiones que sea posible respetando las condiciones de no crear nuevas estructuras en "V" ni generar lazos dirigidos.

La manera de implementar los pasos 1 y 3 del Algoritmo 1 no es única, y es susceptible de diversos esquemas de optimización. El más habitual es la implementación de Peter Spirtes y Clark Glymour (Spirtes et al., 2000), conocido como algoritmo PC (por los nombres de los autores). En cuanto a la implementación de los heurísticos del paso 3 del Algoritmo 1, las siguientes

reglas aplicadas de manera reiterativa conducen a un grafo máximamente orientado:

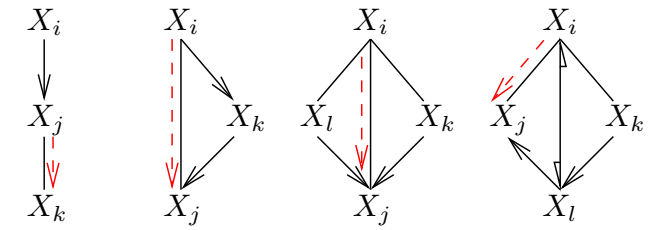
R_1 Orientar $X_j - X_k$ como $X_j \to X_k$ en caso de que exista $X_i \to X_j$ y no sean adjuntos X_i y X_k.

R_2 Orientar $X_i - X_j$ como $X_i \to X_j$ si hay una cadena $X_i \to X_k \to X_j$.

R_3 Orientar $X_i - X_j$ como $X_i \to X_j$ cuando hay dos cadenas $X_i - X_k \to X_j$ y $X_i - X_l \to X_j$ tal que no sean adjuntos X_k y X_l.

R_4 Orientar $X_i - X_j$ como $X_i \to X_j$ cuando hay dos cadenas $X_i - X_k \to X_l$ y $X_k \to X_l \to X_j$ tal que no sean adjuntos X_k y X_j pero sí lo sean X_i y X_l.

Estas reglas se han representado de manera gráfica en la Figura 1.15. La línea punteada muestra el enlace que es posible orientar siguiendo estas reglas. En la regla R_1 la orientación contraria ($X_k \to X_j$) del enlace generaría una nueva estructura en "V". En el caso siguiente, orientar el enlace como $X_j \to X_i$ produciría un ciclo en el grafo. La regla R_3 se aplica para evitar una violación de las dos condiciones (evitar nuevas estructuras en "V" y evitar caminos cerrados). Si el enlace en el caso de la Figura 1.15(c) tuviera orientación $X_j \to X_i$, habría que forzar $X_k \to X_i$ y $X_l \to X_i$ para evitar crear ciclos. Pero en este caso se habría introducido una nueva colisión en $X_l \to X_i \leftarrow X_k$. Finalmente, la regla R_4 también se puede demostrar por reducción al absurdo. Si la orientación en la Figura 1.15(d) fuera $X_j \to X_i$, habría que forzar, para evitar bucles, los enlaces $X_l \to X_i$ y $X_k \to X_i$.



(a) R_1 (b) R_2 (c) R_3 (d) R_4

Figura 1.15: Heurísticos del paso 3 del Algoritmo IC. El enlace punteado indica cuál es el enlace que se orienta para cada regla.

El test de independencia se puede realizar mediante varios algoritmos. No se abordan en este trabajo las muy diversas maneras de establecer la dependencia o independencia estadística. Sin embargo sí se detallan los métodos

empleados habitualmente en la literatura causal. Se citan dos de ellos, el estadístico G^2 y un método núcleo, el "Hilbert Space Independence Criterium" (HSIC).

- Test estadístico G^2. Se desea determina si es cierto el test $X_i \perp\!\!\!\perp X_j | \mathbf{X}_k$. Sea $S_{ijk}^{ab\mathbf{c}}$ el número de veces que $X_i = a$, $X_j = b$ y $\mathbf{X}_k = \mathbf{c}$.

$$G^2 = 2 \sum_{ab\mathbf{c}} S_{ijk}^{ab\mathbf{c}} \ln \frac{S_{ijk}^{ab\mathbf{c}}}{\frac{S_{ik}^{a\mathbf{c}} \; S_{jk}^{b\mathbf{c}}}{S_k^{\mathbf{c}}}} \tag{1.38}$$

  que es equivalente a:

$$\sum (\text{valor\_observado}) \ln \frac{(\text{valor\_observado})}{(\text{valor\_esperado\_si\_indep})}$$

  El test tiene una distribución asintótica χ^2, con un grado de libertad dado por la fórmula: $df = (|X_i| - 1)\,(|X_j| - 1) \prod_{m=1}^{d} |X_{km}|$.

  Un ejemplo de algoritmos que hacen uso de este estadístico son el PC clásico, HITON y MMHC (Tsamardinos et al., 2006).

- Métodos núcleo (Sun et al., 2007b; Sun y Janzing, 2007), donde la dependencia estadística se mide con *normas de Hilbert-Schmidt* (Gretton et al., 2005). El "Hilbert-Schmidt Independence Criterion" (HSIC) es la suma de los autovalores al cuadrado de la covarianza cruzada (evaluada en el espacio transformado RKHS), y se emplea como una medida de la dependencia de las variables en el espacio original. Un posible estimador de esta medida es:

$$\text{HSIC} = (m-1)^{-2} \text{traza}\,(KHLH) \tag{1.39}$$

  donde $K_{ij} = k(x_i, y_j)$; $L_{ij} = l(y_i, y_j)$ y finalmente, $H_{ij} = \delta_{ij} - m^{-1}$.

1.6.2. Métodos basados en puntuación de DAGs

Estos algoritmos hacen una búsqueda para maximizar una cierta función de coste, por ejemplo la probabilidad a posteriori de una red bayesiana dado un conjunto de muestras. En esta categoría entran algoritmos de búsqueda local y algoritmos que emplean heurísticos, como por ejemplo técnicas tipo "greedy".

En su mayor parte son los mismos métodos que se vienen empleando para recuperar la estructura de una red bayesiana. La diferencia radica en la interpretación causal que se hace de las relaciones entre variables de la

red. A partir de un conjunto de muestras, se puede definir un problema de optimización, aunque su resolución es de tipo NP-completo, es decir, que sus soluciones no son verificables en tiempo polinómico.

Los elementos a tener en cuenta son la estructura candidata en cada momento de la red bayesiana, el conjunto de muestras disponibles y una función de coste o de puntuación que los relaciona. El objetivo de estos algoritmos es buscar eficazmente la estructura que maximiza la función de puntuación para el conjunto de muestras dado.

Ejemplos son el "Optimal Reinsertion" (OR) (Moore y keen Wong, 2003) o el "Greedy Hill-Climbing Search" (GS). Ambos usan como función de coste el estadístico "BDEu", una métrica bayesiana similar a la verosimilitud (Heckerman et al., 1995). Esta métrica se define como:

$$P(\mathcal{G}, D) = P(\mathcal{G}) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma\left(\alpha_{ij} + N_{ij}\right)} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{1.40}$$

donde,

- $P(\mathcal{G})$ es la probabilidad a priori del grafo.
- n es el número de variables del grafo.
- Γ es la función gamma.
- q_i es la cardinalidad del conjunto de padres del nodo i en los datos D. Si el nodo i no tiene padres, $q_i = 1$.
- r_i es la cardinalidad del nodo i.
- N_{ijk} es el número de muestras en los que la variable i tiene el valor indexado por k cuando los padres están en el estado j.
- N_{ij} es el número de muestras en el que los padres de la variable i están en el estado j.
- α_{ijk} es la probabilidad a priori de que la variable i tenga el valor indexado por k cuando los padres estén en el estado j.
- α_{ij} es la probabilidad a priori de que los padres de la variable i estén en el estado j.

Hay que notar que, por definición, $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ y $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Para las probabilidades a priori, la variante BDEu usa una distribución uniforme, de forma que se asignan $\alpha_{ij} = 1/q_i$ y $\alpha_{ijk} = 1/q_i r_i$.

En el caso del algoritmo OR de Andrew Moore, se busca maximizar la función de puntuación dividiendo el problema en suma de puntuaciones de los padres markovianos, como puede verse en (1.41). La puntuación de la red al completo, $\mathcal{G}$, se divide en la puntuación de cada nodo dados sus padres, pa_i: $\text{NodeScore}(\text{pa}_i \to i)$ (ejemplo para el nodo i).

$$\begin{gathered} \arg\max_{\mathcal{G}} \text{DagScore}(\mathcal{G}) \\ \text{DagScore}(\mathcal{G}) = \sum_{i=1}^{n} \text{NodeScore}(\text{pa}_i \to i) \end{gathered} \tag{1.41}$$

La *optimización* de (1.41) se suele llevar a cabo con un algoritmo de ascenso de colina ("hill climbing"), por el que se aplica una serie de operaciones que modifican los enlaces del grafo, como "añadir", "eliminar" o "invertir sentido". Por supuesto, en cada paso de la optimización se debe asegurar que no se producen ciclos en el DAG $\mathcal{G}$. El ascenso de colina aplica las modificaciones locales que inducen un mayor incremento en la puntuación $\text{NodeScore}(\cdot)$. Una vez modificada la estructura del grafo en la iteración, se procede a la búsqueda de una nueva modificación. El algoritmo continúa hasta que no es posible encontrar mejoras al grafo.

Otro algoritmo relevante que se basa en el mismo principio de búsqueda y puntuación es el "Max-Min Hill-Climbing" (MMHC), descrito en (Tsamardinos et al., 2006).

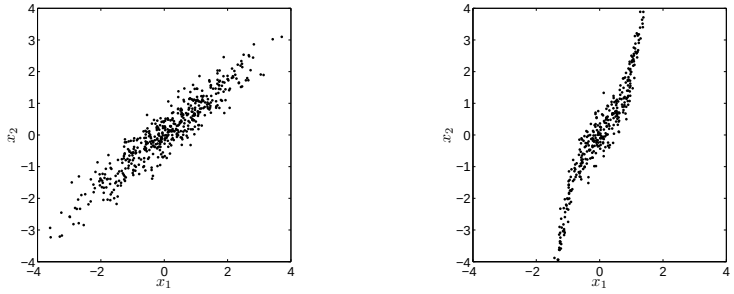
Si bien el método MMHC comparte el principio de funcionamiento basado en búsqueda y puntuación, incluye otros mecanismos que lo convierten en un algoritmo híbrido. El algoritmo se puede dividir en dos etapas. En la primera de ellas, el esqueleto del árbol causal se recupera mediante una búsqueda local de padres e hijos. Esta etapa se lleva a cabo de una manera similar a cómo funciona el PC (Spirtes et al., 2000), tratando de identificar los subconjuntos de variables que independizan una variable del resto. La etapa de identificación del esqueleto converge al real asintóticamente con el número de muestras. La segunda etapa tiene por objeto orientar los enlaces entre variables. Para ello se lleva a cabo una búsqueda heurística de "ascenso de colina".

Se aplican de manera inteligente los operadores de *añadir-enlace*, *eliminar-enlace* y *revertir-enlace*, para maximizar la verosimilitud BDEu (Heckerman et al., 1995). El operador de *añadir-enlace* sólo se emplea entre dos variables X e Y en caso de que ese enlace haya sido identificado como perteneciente al esqueleto en la etapa previa.

1.6.3. Métodos que aprovechan no linealidades

Cuando las relaciones entre variables son lineales y están contaminadas por ruido aditivo gausiano, se hace más difícil recuperar la orientación causal. Sin embargo, cuando el problema es no lineal es posible aprovechar esta peculiaridad para determinar la relación causal.

Es posible recuperar la estructura causal en los casos en los que la relación entre variables es una función no lineal (Hoyer et al., 2009). El desarrollo de esta idea se describirá para el caso de dos variables, por simplicidad, aunque el resultado es válido para un número mayor de variables. Se asume que el modelo es $x_i = f_i(pa_i) + \epsilon_i$. En la Figura 1.16 se muestran datos de una función lineal y de una no lineal. En ambos casos el ruido es gausiano, por lo que si se dibujara el histograma para distintos cortes de la variable x_1 y de x_2 de la función lineal de la Figura 1.16(a), éstos serían una gausiana de igual varianza. Por el contrario, en el escenario no lineal, el histograma correspondiente a los cortes de la variable x_2 no presenta una distribución gausiana. Aún cuando $p(x_2|x_1)$ es gausiano, no lo es $p(x_1|x_2)$.



(a) $x_2 = 0.8x_1 + \epsilon$. (b) $x_2 = x_1 + x_1^3 + \epsilon$.

Figura 1.16: Distribución conjunta de las variables x_1 y x_2 con relación lineal y no lineal.

El principio de estimación queda reflejado en el Algoritmo 2. El procedimiento puede dar lugar a cualquiera de los cinco posibles escenarios: $x_1 \quad x_2$, $x_1 \rightarrow x_2$, $x_1 \leftarrow x_2$, $x_1 \leftrightarrow x_2$, $x_1 - x_2$. El primer caso se dará si las variables son independientes. En caso contrario, el algoritmo puede decidir que hay evidencia de relación causal en uno de los dos sentidos, en los dos, o en ninguno de ellos. En esta última situación, sólo se podrá afirmar que hay dependencia, pero que el modelo no dispone de la capacidad expresiva suficiente como para identificar el flujo de relación causal.

El desarrollo previo es válido para dos variables. La generalización a problemas multidimensionales es directa, aunque no tiene buenas propiedades

Algoritmo 2 Determinación de orientación causal aprovechando las propiedades de la relación no lineal entre variables.

si Comprobar la independencia estadística de x_1 y x_2 **entonces**
  Terminar algoritmo.
si no
  Estimar regresión no lineal $x_2 := f(x_1) + \epsilon$.
  si $\hat{\epsilon} = x_2 - \hat{f}(x_1)$ es independiente de x_1 **entonces**
    aceptar modelo $x_2 \leftarrow x_1$
  si no
    rechazar modelo $x_2 \leftarrow x_1$
  fin si
  Estimar regresión lineal $x_1 := g(x_2) + \epsilon'$.
  si $\hat{\epsilon}' = x_1 - \hat{g}(x_2)$ es independiente de x_2 **entonces**
    aceptar modelo $x_1 \leftarrow x_2$
  si no
    rechazar modelo $x_1 \leftarrow x_2$
  fin si
fin si

de escalado. Para cada uno de los posibles grafos $\mathcal{G}_i$ sería preciso estimar un regresor no lineal por variable, empleando los padres que indique el grafo. La independencia estadística de los residuos con el conjunto de padres validaría esos enlaces. En el citado trabajo de (Hoyer et al., 2009), se menciona un límite de siete variables como máximo para que el tiempo de cálculo sea razonable.

1.6.4. Métodos que explotan la complejidad

Las relaciones de causalidad imponen una serie de propiedades a las distribuciones que pueden ser usadas para inferir relaciones de causa-efecto.

Suponiendo que X sea causa de Y, entonces la complejidad de $P(X|Y)$ debe ser mayor que la de $P(Y|X)$. Este es el razonamiento detrás de artículos como (Sun et al., 2006, 2007a), donde la complejidad se define como la seminorma del logaritmo de la función de distribución:

$$C\left(P(Y|X)\right) = \text{mín}\left\{||\phi||^2 \text{ tal que } P(y|x) = \exp(\phi(x,y) - \ln z_\phi(x))\right\}$$
$$\phi(x,y) = \sum_{i=1}^{n} c_i k((x_i,y_i),(x,y)) \tag{1.42}$$

Finalmente, se usa un estimador de máxima verosimilitud para ajustar el modelo exponencial a los datos.

En (Sun et al., 2005) la dirección más plausible viene dada por la distribución condicional que maximiza la entropía condicional, con las restricciones de ajustarse a los momentos de los datos. El problema de optimización es:

$$\max_{P(Y|X)} -\sum_x \sum_y P(x)P(y|x)\ln(P(y|x))$$

sujeto a una serie de condiciones para ajustarse a los momentos de los datos.

Las ideas anteriores se pueden visualizar fácilmente mediante el siguiente ejemplo. En el caso de que una de las variables sea binaria, por ejemplo, suponiendo X un vector $\in \mathbb{R}^d$ e $Y \in \{-1, 1\}$, entonces pueden darse dos posibilidades:

- Si $Y \to X$, entonces habrá dos agrupamientos diferenciados.
- Si $X \to Y$, entonces los datos (sin diferenciar por la etiqueta Y) estarán agrupados en un único "cluster" y habrá una frontera de decisión razonablemente bien definida.

La regla de la navaja de Occam aboga por apostar por el más sencillo de los dos modelos. En el ejemplo anterior, parecería más lógico que la relación fuera $X \to Y$. Por supuesto, es preciso evaluar la complejidad de las distribuciones de probabilidad en cada caso.

Se han presentado numerosos trabajos que intentan explotar estos conceptos. Sin embargo carecen de una sustentación teórica más allá de la intuición inicial de la idea. Si tiene algún futuro esta aproximación será para problemas muy específicos.

1.7. Métodos de búsqueda causal para series temporales

Como se ha comentado previamente, en el dominio continuo es viable explotar la estructura temporal de las señales para encontrar la dirección causal. El método de referencia en este campo es el propuesto por Granger (Granger, 1969) desde el área de la economía, si bien tiene aplicaciones en campos tan diversos como el estudio climático, la biomedicina, sociología o la ingeniería (Chu et al., 2005; Nolte et al., 2008; Marinazzo et al., 2006).

Anterior a los desarrollos basados en (Spirtes et al., 2000; Pearl, 2000) se había considerado una aproximación a la búsqueda de relaciones causales

inspirada en las ideas de Granger y Wiener (Granger, 1969; Hlavackova-Schindler y Verdes, 2007), (Ancona et al., 2004; Angelini et al., 2007; Marinazzo et al., 2006).

Según el criterio de causalidad de Granger entre dos series temporales, *"si el error de predicción de la primera serie se reduce al usar la segunda serie temporal en una regresión lineal, entonces la segunda serie tiene una influencia causal en la primera"*.

Para estimar la relación causal entre un par de series temporales X e Y habría que resolver estos problemas:

$$\begin{aligned} x &= W_{11}X + W_{12}Y, \text{varianza del error: } e_{xy} \\ y &= W_{21}X + W_{22}Y, \text{varianza del error: } e_{yx} \end{aligned} \tag{1.43}$$

$$\begin{aligned} x &= V_1X, \text{varianza del error: } e_x \\ y &= V_2Y, \text{varianza del error: } e_y \end{aligned} \tag{1.44}$$

Se dan las siguientes relaciones: X tiene influencia causal en Y si $\mathrm{Var}(e_{yx}) < \mathrm{Var}(e_y)$. Igualmente, Y tiene influencia causal en X si $\mathrm{Var}(e_{xy}) < \mathrm{Var}(e_x)$. Es posible combinar toda esta información haciendo que $c_1 = \mathrm{Var}(e_x) - \mathrm{Var}(e_{xy})$ y $c_2 = \mathrm{Var}(e_y) - \mathrm{Var}(e_{yx})$. El cociente:

$$D = \frac{c_2 - c_1}{c_1 + c_2} \tag{1.45}$$

cuantifica la influencia causal entre las series. En los casos extremos, $D = 1$ implicará una relación causal exclusiva $X \to Y$, y $D = -1$ significará que $Y \to X$.

Esta aproximación es fácilmente extensible al caso no lineal, empleando otros regresores en (1.43) y (1.44), por ejemplo un regresor RBF (Funciones de Base Radial) (Ancona et al., 2004; Marinazzo et al., 2006). Otra posibilidad para la extensión no lineal está basada en una formulación de *Teoría de la Información* (Hlavackova-Schindler y Verdes, 2007).

Este método no está exento de críticas, algunas de las cuales se recogen en (Hu et al., 2011), donde también se esbozan posibles soluciones. En dicho trabajo se expone que la causalidad de Granger no siempre determina la intensidad de la relación causal, ya sea en el método directo o en modificaciones en el dominio de la frecuencia, y se presentan algunos contraejemplos. Si bien es cierto que en algunos casos concretos la causalidad de Granger puede dar resultados equívocos, en el caso general recupera exitosamente la estructura de los problemas. Las modificaciones al método estándar presentadas en (Hu et al., 2011), para el caso multivariable, consisten en introducir la idea de *proporción*, de manera que el nuevo índice causal entre las series X y Y tiene

en cuenta qué "porcentaje" o parte de Y influye en X. En el fondo es una modificación ligera al concepto de Granger de causalidad y se apoya en la misma intuición.

1.7.1. Métodos dispersos

Las líneas de investigación recientes buscan métodos dispersos a la hora de modelar la predicción en las series. La dispersidad ayuda a la estima de la medida de causalidad, pues los coeficientes del modelado ARMA tienden a ser nulos para las relaciones de causalidad no existentes. Se van a emplear tres algoritmos dispersos para estudiar su repercusión en la inferencia causal, que son: LASSO, "Group LASSO" y RVM.

Dada la naturaleza del problema es beneficioso que los coeficientes del modelo autorregresivo sean dispersos, esto es, que sean idénticamente nulos en las direcciones en las que no hay causalidad.

LASSO y "Group LASSO"

El método de modelado LASSO[7] (Tibshirani, 1994; Nardi y Rinaldo, 2011) minimiza el error cuadrático, a la vez que impone una penalización a los coeficientes del modelo mediante una norma l_1:

$$\hat{A}^{lasso} = \arg\min_{A} ||Y - XA||^2 + \lambda||A||_{l_1} \tag{1.46}$$

La norma $||\mathbf{a}||_{l_1} = \sum_{i=1}^{d} |a_i|$ controla la complejidad del modelo, forzando un número pequeño de coeficientes A_{ij}^{p} distintos de cero.

En (Haufe et al., 2008) se ha publicado un intento de calcular el índice causal de Granger empleando un método disperso, tipo LASSO, haciendo agrupación de coeficientes. En dicha publicación se intenta resolver el siguiente problema conocido como *"Group LASSO"*:

$$\mathbf{z}(t) = \sum_{i=1}^{P} \mathbf{A}\mathbf{z}(t-i) \tag{1.47}$$

$$\begin{pmatrix}\mathbf{z}_1(t)\\ \mathbf{z}_2(t)\end{pmatrix} = \begin{pmatrix}A_{11}^1 & A_{12}^1\\ A_{21}^1 & A_{22}^1\end{pmatrix}\begin{pmatrix}\mathbf{z}_1(t-1)\\ \mathbf{z}_2(t-1)\end{pmatrix} + \cdots + \begin{pmatrix}A_{11}^P & A_{12}^P\\ A_{21}^P & A_{22}^P\end{pmatrix}\begin{pmatrix}\mathbf{z}_1(t-P)\\ \mathbf{z}_2(t-P)\end{pmatrix}$$

El funcional a minimizar intenta que los coeficientes sean simultáneamente cero en los P retardos:

[7] "Least Absolute Selection and Shrinkage Operator"

$$\min_{\mathbf{A}} \|\mathbf{Y}^* - \mathbf{X}^*\mathbf{A}\|^2 \tag{1.48}$$
$$\text{restringido a } \| \begin{pmatrix} A_{11}^{1,\ldots,P} & A_{22}^{1,\ldots,P} \end{pmatrix} \| + \|A_{12}^{1,\ldots,P}\| + \|A_{21}^{1,\ldots,P}\| \leq k$$

En la Figura 1.17 puede verse cómo la constante k, que limita la suma l_1 de las normas l_2 de los coeficientes permite obtener una solución dispersa en los coeficientes de la relación no causal.

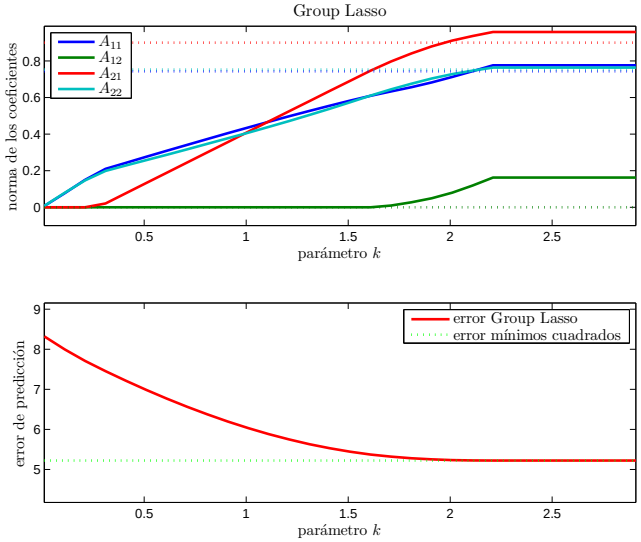


Figura 1.17: Norma de los coeficientes del modelo ARMA y error de predicción, para "Group LASSO". Se compara con la solución de mínimos cuadrados.

"Relevance Vector Machine"

Las Máquinas de Vectores Soporte (SVM) han demostrado ser el estado del arte en cuanto a clasificación se refiere, especialmente en clasificación binaria. Sin embargo presentan ciertas limitaciones, que otros enfoques han intentado solventar.

Un ejemplo de esto son las Máquinas de Vectores Relevantes (RVM) (Tipping, 2001). El algoritmo está basado en el paradigma bayesiano, y destaca porque obtiene soluciones dispersas en los parámetros de la máquina. Más adelante se aplicará este algoritmo como regresor en el modelado ARMA para la obtención de relaciones causales.

Se supondrá un conjunto de datos de la forma $\{\mathbf{x}_n, t_n\}_{n=1}^N$, en el que las etiquetas están contaminadas con ruido gausiano del tipo $t_n = y(\mathbf{x}_n; \mathbf{w}) + \epsilon_n$. El proceso ϵ_n será blanco, de media nula y varianza σ^2.

La verosimilitud de la base de datos completa es

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}||\mathbf{t} - \mathbf{\Phi}\mathbf{w}||^2\right\}$$

Con el fin de mantener la solución de la ecuación anterior libre del efecto de sobreajuste, se habrá de imponer una "regularización" sobre los parámetros de la máquina. En concreto, esto se puede llevar a cabo de la siguiente manera:

$$p(\mathbf{w}|\alpha) = \prod_{i=0}^{N} \mathcal{N}(w_i|0, \alpha_i^{-1})$$

Este a priori gausiano de media nula en los pesos de la máquina hace más probable pesos pequeños o idénticamente nulos. Finalmente, es necesario definir una distribución a los hiperparámetros α_i y σ.

$$p(\alpha) = \prod_{i=0}^{N} \text{Gamma}(\alpha_i|a, b) \tag{1.49}$$

$$p(\beta) = \text{Gamma}(\beta|c, d) \tag{1.50}$$

donde $\beta = \sigma^{-2}$ y la función en la que se apoyan las distribuciones es $\text{Gamma}(\alpha|a, b) = \Gamma(a)^{-1} b^a \alpha^{a-1} e^{-b\alpha}$.

1.7.2. Orientación de la flecha temporal

Otro enfoque, distinto a la aproximación de Granger, basado en la no gausianidad de los datos se ha presentado en (Peters et al., 2009). Si una serie temporal es generada con un ruido no gausiano, las propiedades de la misma en el sentido directo e inverso del tiempo son distintas. En dicho trabajo se emplea un algoritmo de búsqueda de la flecha temporal en el que se ajusta un modelo ARMA (autorregresivos de media móvil) en la serie temporal y en la misma pero con la referencia temporal invertida. En caso de que el residuo de la regresión sea independiente de la señal en un sentido, y dependiente en el otro, entonces el primer sentido es el auténtico sentido causal. En otros trabajos se emplean medidas de distancia entre la distribución de los residuos y la distribución gausiana (Hernández-Lobato et al., 2011).

La aplicación más interesante de este método es la de encontrar relaciones causales entre dos o más series temporales; sin embargo es un trabajo en fase de investigación, y aún no se han presentado resultados satisfactorios.

En concreto, dada la serie temporal $x(t),\ 1 \leq t \leq N$, puede interesar preguntarse cuál es el sentido temporal en el que la serie ha sido generada, es decir, si $x(1), \ldots, x(N)$ es la secuencia correcta o bien lo es $x(N), \ldots, x(1)$. El algoritmo ajusta un modelo ARMA, siendo $\epsilon(t)$ ruido i.i.d. de media nula:

$$x(t) = \sum_{i=1}^{P} a_i x(t-i) + \sum_{j=1}^{Q} \epsilon(t-j) + \epsilon(t) \tag{1.51}$$

Se dice que un proceso ARMA es causal si $\epsilon(t)$ es independiente de valores pasados de la señal $x(\tau)$, con $\tau < t$. Una señal se denomina *reversible en el tiempo* si existe el proceso i.i.d. $\tilde{\epsilon}(t)$ tal que:

$$x(t) = \sum_{i=1}^{\tilde{P}} \tilde{a}_i x(t+i) + \sum_{j=1}^{\tilde{Q}} \tilde{\epsilon}(t+j) + \tilde{\epsilon}(t) \tag{1.52}$$

donde $\tilde{\epsilon}(t)$ es independiente de $x(\tau)$ para $\tau > t$. Es posible demostrar que si y sólo si el ruido del proceso estocástico es gausiano la serie es reversible en tiempo (Peters et al., 2009).

Por lo tanto, en caso de que los residuos $\epsilon(t)$ tengan una distribución gausiana no habrá información suficiente para tomar ninguna decisión. En caso contrario, pueden darse tres situaciones:

$$\begin{cases} \text{Si } (\epsilon(t) \perp\!\!\!\perp x(\tau),\ \tau < t) \wedge (\tilde{\epsilon}(t) \not\perp\!\!\!\perp x(\tau),\ \tau > t), & \text{entonces, } x(1), \ldots, x(N) \\ \text{Si } (\epsilon(t) \not\perp\!\!\!\perp x(\tau),\ \tau < t) \wedge (\tilde{\epsilon}(t) \perp\!\!\!\perp x(\tau),\ \tau > t), & \text{entonces, } x(N), \ldots, x(1) \\ \text{Si } (\epsilon(t) \not\perp\!\!\!\perp x(\tau),\ \tau < t) \wedge (\tilde{\epsilon}(t) \not\perp\!\!\!\perp x(\tau),\ \tau > t), & \text{orden no identificable} \end{cases}$$

Sólo es posible identificar el sentido de la flecha temporal si el residuo es independiente de la señal en una ordenación, pero dependiente en la otra.

Existen numerosos test de gausianidad, para comprobar la hipótesis de que los residuos presenten dicha distribución. Una posibilidad es emplear el conocido como test de Jarque-Bera, basado en momentos de orden 3 y 4 de los datos. El siguiente paso del algoritmo verifica la independencia estadística entre la serie temporal y su inversa y los respectivos residuos de los modelos ARMA. Si los resultados presentan dependencia en la serie temporal directa y en su inversa, tampoco es posible tomar ninguna decisión. Sin embargo, si en un sentido hay dependencia entre residuos y los datos, pero en el otro sentido son independientes, entonces éste último será el correcto sentido causal.

La aplicación de esta idea a un caso más general, que involucre más de una serie temporal es una aproximación interesante, que puede ser de interés para la identificación de relaciones causales.

1.7.3. Causalidad mediante ICA

Suponiendo que el proceso de generación de los datos es lineal y que la perturbación en las relaciones no es gausiana, en (Shimizu et al., 2006) se ha desarrollado un algoritmo basado en el análisis de componentes independientes, (ICA, "Independent Component Analysis") para encontrar relaciones causales. A diferencia de otros muchos métodos, cuando las relaciones entre variables están contaminadas con ruido no gausiano, es posible recuperar una estructura única, no una familia de soluciones equivalentes desde el punto de vista estadístico.

Dado un orden causal $k(i)$, tal que los efectos siempre están ordenados después de las causas, y un conjunto de muestras sobre el que aplica, $x_i, i \in \{1, \ldots, m\}$, el problema se reduce a encontrar los coeficientes b_{ij} tal que:

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i + c_i \tag{1.53}$$

que es un problema canónico de ICA. La interpretación de (1.53) es directa: cada variable es una combinación lineal de las variables que le preceden en la ordenación causal, más una constante c_i y un ruido no gausiano e_i, de varianza distinta de cero e independientes entre sí, de forma que $P(\mathbf{e}) = \prod_{i=1}^{m} P(e_i)$.

Las restricciones anteriores implican que existe suficiencia causal, es decir, que no hay variables latentes, aunque no tiene por qué cumplir con el criterio de estabilidad.

Por tanto el problema se puede expresar como

$$\mathbf{x} = B\mathbf{x} + \mathbf{e} \tag{1.54}$$

donde la matriz B es triangular inferior, si se conociera el orden causal y se realizaran las permutaciones correspondientes. La solución al problema enunciado en (1.54) viene dado por

$$\mathbf{x} = A\mathbf{e} \tag{1.55}$$

Para tener en cuenta las particularidades de la inferencia causal mediante componentes independientes es preciso tener en consideración dos particularidades del problema: el orden y el escalado. Las columnas de A en (1.55) se pueden reordenar sin afectar a la distribución de los datos $\mathbf{x}$. En un problema estándar de ICA la ordenación no es relevante, sin embargo a la hora de buscar causalidad es un elemento clave, según se ha expuesto en el desarrollo anterior; por lo tanto uno de los pasos del algoritmo es encontrar la permutación que se corresponda con el árbol causal. El escalado de la solución es la otra característica a tener en cuenta. En ICA estándar se ajusta la matriz

A para que las componentes de la solución tengan varianza unidad. En la versión causal del ICA, se acepta que el ruido tenga una varianza distinta de cero (y posiblemente distinta de la unidad); adicionalmente, se fuerza para que A^{-1} tenga en la diagonal valores unitarios, para dar el mismo peso a la componente de ruido y de la variable observada.

El método LiNGAM expuesto en (Shimizu et al., 2006) se presenta en el Algoritmo 3, que en esencia es un ICA estándar más una etapa de permutación y normalización de la matriz de mezclas A.

Algoritmo 3 Algoritmo LiNGAM (Shimizu et al., 2006) de inferencia causal para problemas con relación lineal y ruido no gausiano.

Requerir: Dado un conjunto de m muestras d-dimensionales X_{dxm}, con $d << m$.

1: Obtener la solución ICA $X = AS$, donde S tiene las componentes independientes. Por comodidad, se usará $W = A^{-1}$.
2: Encontrar la permutación de columnas de W, $\widetilde{W}$ que minimiza $\sum_i 1/|\widetilde{W}_{ii}|$, es decir, la permutación que hace prácticamente nula la diagonal principal.
3: Dividir cada columna de $\widetilde{W}$ por el elemento de la diagonal correspondiente, para obtener la versión normalizada $\widetilde{W}'$.
4: Estimar $\widehat{B} = I - \widetilde{W}$.
5: Para establecer el orden causal, encontrar la matriz de permutación P que aplicada a $\widehat{B}$ genere la matriz $\widetilde{B} = P\widehat{B}P^T$ que sea lo más próxima posible a una matriz triangular inferior, para lo cual se puede minimizar el funcional $\sum_{i \leq j} \widetilde{B}_{ij}^2$.

El algoritmo se ha ampliado también al caso de variables latentes (no presentes en los datos) (Hoyer et al., 2006), aunque como contrapartida no es posible encontrar una solución única, sino un conjunto finito de soluciones equivalentes.

1.8. Aportaciones y estructura de la tesis

En esta tesis se presentan dos técnicas novedosas de inferencia causal. En función de si los datos son discretos o continuos, se ensayan distintas aproximaciones. En ambos casos se emplean técnicas de aprendizaje máquina para aprovechar sus ventajas en el tratamiento de datos, a saber: buen comportamiento frente a altos ratios del cociente variables entre muestras y tiempos de procesado razonables. En el campo discreto, se relacionan los conceptos de independencia estadística y de relevancia de variables para un clasificador k

vecinos más próximos y para una Máquina de Vectores Soporte. En el campo de las señales continuas, se emplea una Máquina de Vectores Soporte para entrenar un modelo ARMA. La agrupación de coeficientes que realiza este método hace que de manera natural agrupe los coeficientes correspondientes a las señales causa y efecto y los haga tender a cero.

Estos métodos se aplican a dos problemas médicos: para establecer un árbol causal entre las variables significativas de intentos repetidos de suicidio y para detectar patrones de propagación en fibrilaciones auriculares.

El Capítulo 2 desarrolla en su totalidad las aportaciones aquí descritas brevemente. La aplicación al problema de psiquiatría se muestra en el Capítulo 3; en este caso se dispone de una base de datos clínica, con datos discretos. El problema de cardiología, para estudiar las fibrilaciones en la aurícula izquierda, se presenta en el Capítulo 4. Finalmente, el Capítulo 5 recoge las conclusiones y trabajo futuro.

Capítulo 2

Métodos de búsqueda de relaciones causales

Una vez realizada la revisión acerca del paradigma causal y de los métodos de inferencia, en este capítulo se presentan las aportaciones de la tesis, en concreto en el desarrollo de algoritmia de inferencia causal, dejando las aplicaciones clínicas para su desarrollo en los capítulos posteriores.

La posibilidad de tener acceso a la variable *tiempo*, tal y como se ha explicado en el capítulo introductorio, va a condicionar sobremanera el tipo de algoritmo que se puede emplear, así como los criterios en los que se basa. En el campo de los datos discretos se presenta un método novedoso de inferencia causal, que emplea una batería de clasificadores sencillos en baja dimensión para orientar los enlaces entre las variables (de-Prado-Cumplido y Artés-Rodríguez, 2008). Si en una cadena de tres variables $X - Y - T$ tanto X como Y son relevantes para clasificar T, entonces sólo hay una orientación causal posible. Este método se aplicará a un problema clínico en psiquiatría, en el Capítulo 3, y se comparará con otros resultados presentes en la literatura (López-Castromán et al., 2011).

En el campo de las series temporales, y debido a la estructura de dependencia temporal de estos datos, se propone aplicar un método de modelado ARMA basado en máquinas de vectores soporte (Rojo-Álvarez et al., 2004) para estimar las relaciones causales. En el presente capítulo se evalúa y compara el comportamiento de este método para estimar relaciones causales. En el Capítulo 4 de la tesis se aplicará este método para avanzar en la comprensión de cuestiones abiertas acerca del mecanismo de generación de fibrilaciones auriculares (de-Prado-Cumplido y Artés-Rodríguez, 2010).

2.1. Inferencia causal para datos discretos

Se procede en primer lugar a presentar los métodos desarrollados para realizar inferencia causal. A lo largo de esta sección de la tesis se asume que no se dispone de la información temporal, o bien, expresado de otra manera, se asume que los datos se han recogido en un intervalo de tiempo diferencial.

La notación que se emplea en esta sección es la siguiente. Las variables se llaman X_i, organizadas en grafos de d nodos. El conjunto de todos los nodos de un grafo se denomina $\mathbf{X} = (X_1, \ldots, X_d)$. Las bases de datos estarán compuestas por N muestras, del tipo $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^T$, con $i = 1, \ldots, N$. Un ejemplo de base de datos se recoje en el Cuadro 2.1, y un grafo que establece la jerarquía entre las variables en la Figura 2.1.

$\mathbf{X}$		$\mathbf{x}_1$	$\mathbf{x}_2$	$\cdots$	$\mathbf{x}_N$
X_1	Educación	4	2	$\cdots$	3
X_2	Edad	2	3	$\cdots$	1
X_3	Edad 1$^{\text{er}}$ intento	2	2	$\cdots$	3
X_4	Ansiedad	1	0	$\cdots$	0
X_5	Uso de drogas	1	1	$\cdots$	0
X_6	Tratamiento	0	0	$\cdots$	1
X_7	Intentos repetidos	3	2	$\cdots$	0

Cuadro 2.1: Ejemplo de base de datos discretos. El espacio muestral tiene siete variables de tipo clínico, y consta de N pacientes (muestras). Cada valor de las variables codifica un estado. Por ejemplo, un valor 2 en *educación* indica graduado, y un 4, estudios de bachillerato.

A partir de un conjunto de datos, $\mathbf{X}$, se pretende desarrollar un método que infiera una jerarquización entre las variables, de forma que se obtenga un grafo o árbol causal. Se va a desarrollar un algoritmo basado en clasificadores para realizar inferencia causal. Si la relación entre variables es $X \to Y \to T$, toda la información útil para clasificar T se encuentra en Y. Sin embargo, si la relación es $X \to Y \leftarrow T$, tanto X como Y son relevantes de cara a clasificar T. De esta forma será posible identificar las estructuras en "V" presentes en el grafo. Se ha ensayado el uso de dos tipos de clasificador: un k vecinos más próximos y la salida blanda de una máquina de vectores soporte (SVM) multiclase. Estos algoritmos recibirán el nombre, respectivamente, de ccKnn y ccMSVM. En la próxima sección se presenta el método ccKnn y su relación con otros métodos de inferencia causal. En la Sección 2.1.3 se presenta el desarrollo del ccMSVM. Finalmente se realizan experimentos con datos sintéticos para comprobar la validez de los métodos.

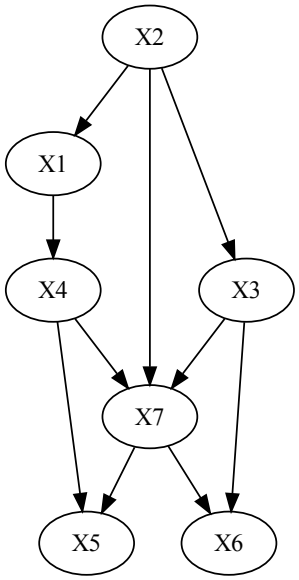


Figura 2.1: Ejemplo de grafo para la base de datos presentada en el ejemplo introductorio del Cuadro 2.1.

2.1.1. Método basado en clasificadores k-NN

El primero de los algoritmos propuestos parte de un esquema de funcionamiento similar al método PC (ver Algoritmo 1 en la Sección 1.6.1). Sin embargo, el funcionamiento del PC depende de una batería de test de independencia estadística, que en situaciones con pocas muestras, o alta dimensionalidad, puede dar lugar a resultados inexactos. Esto genera árboles causales con muchos enlaces sin dirigir. Es beneficiosa la sustitución de los test por otras medidas de relación entre variables. Basándonos en los trabajos de Kohavi (Kohavi y John, 1998) sobre variables relevantes, se va a emplear un criterio basado en la probabilidad de error de clasificación.

Métodos de inferencia

Como se ha visto en el capítulo introductorio, el mecanismo principal para orientar las relaciones causales consiste en buscar estructuras en "V", es decir, conjuntos de tres variables o nodos que cumplan las siguientes condiciones:

$$X \perp\!\!\!\perp T \tag{2.1}$$

$$X \not\perp\!\!\!\perp T|Y \tag{2.2}$$

El criterio para realizar los cálculos de la ecuación anterior se basa tradicionalmente en medidas estadísticas. Sin embargo, dicha relación también establece la relevancia o irrelevancia de las variables cuando se usan para clasificar una de ellas. Por ejemplo, si la relación causal entre las variables fuera $X \rightarrow Y \rightarrow T$, una máquina que tratara de clasificar T tendría toda la información relevante contenida en la variable Y. Sin embargo, en el supuesto que se cumpliera lo establecido en la Ecuación 2.1, tanto X como Y aportan información no redundante útil en la clasificación de T. Este criterio es el que se va a emplear como fundamento de los algoritmos en tiempo discreto que se proponen en esta tesis.

Clasificadores por vecinos más próximos

El algoritmo va a sustituir los test estadísticos necesarios para desvelar relaciones tipo (2.1) por dos clasificadores de la variable T: $\hat{T} = f_1(Y)$ y $\hat{T} = f_2(Y, X)$, que estimarán la variable T basados en, respectivamente, Y y X e Y.

Debido a que la dimensionalidad de este problema es muy reducida, se puede experimentar con esquemas de clasificación sencillos. Un clasificador que cumple este criterio y ofrece un buen comportamiento es el conocido como k vecinos más próximos.

Se supondrá que el conjunto de N datos viene dado por $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, donde cada $\mathbf{x}$ posee una cardinalidad finita, que se recoge en $\{c_1, \ldots, c_N\}$. El vecino más próximo a una muestra $\mathbf{x}$, que se llamará $\mathbf{x}'$, vendrá dado por:

$$\mathbf{x}' = \min_i(\|\mathbf{x} - \mathbf{x}_i\|^2) \tag{2.3}$$

La etiqueta asociada a cada muestra tendrá una probabilidad $P(\omega_i|\mathbf{x}')$. Cuando el número de muestras es elevado, es intuitivo apreciar que $P(\omega_i|\mathbf{x}') \simeq P(\omega_i|\mathbf{x})$. Siguiendo el razonamiento, y si se define $\omega_m(\mathbf{x})$ como

$$P(\omega_m|\mathbf{x}) = \max_i P(\omega_i|\mathbf{x}) \tag{2.4}$$

entonces la regla de decisión bayesiana siempre elige ω_m. Esta configuración produce una división del espacio muestral en el conocido como teselado de Voronoi.

La tasa de error del algoritmo *vecino más próximo* siempre se encuentra acotada entre el error de Bayes y el doble del error de Bayes, aunque en la práctica la cota superior es más ajustada. En la medida en la que aumenta el número de muestras N, el rendimiento del algoritmo mejora.

La ampliación de esta idea para tener en cuenta no sólo al vecino más próximo, sino a un conjunto de vecinos más próximos, es inmediata. En este

caso, la regla de los k vecinos más próximos asigna a una muestra el estado o etiqueta ω_m si la mayoría simple de muestras presentan dicha etiqueta, y lo hacen con una probabilidad:

$$\sum_{i=(k+1)/2}^{k} \binom{k}{i} P(\omega_m|\mathbf{x})^i \left[1 - P(\omega_m|\mathbf{x})\right]^{k-i} \tag{2.5}$$

La elección del parámetro k de número de vecinos supondrá un compromiso entre un valor alto que dé una estimación más fiable y un valor reducido que mantenga $P(\omega_m|\mathbf{x}')$ similar a $P(\omega_m|\mathbf{x})$. Un procedimiento habitual es realizar la selección mediante validación cruzada.

Post-tratamiento

Debido a la forma en la que se exploran todos los caminos de tres nodos, se hace necesaria una purga posterior a la ejecución de los clasificadores. Es posible que dos caminos de 3 nodos (nodos adjuntos en el grafo original) sean etiquetados como colisiones, y que alguno de esos enlaces se orienten en sentidos opuestos. En la Figura 2.2 se muestra un ejemplo. Los caminos formados por los tríos de nodos $2-4-6$ y $4-6-5$ han sido etiquetados por el algoritmo como colisiones o estructuras en "V".

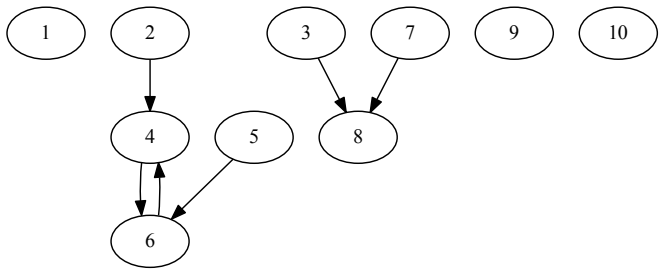


Figura 2.2: Ejemplo de determinación contradictoria de un enlace. Se han identificado dos colisiones, que indican sentidos opuestos entre los nodos $4-6$.

Se exploran dos posibilidades para eliminar los enlaces contradictorios: pueden premiarse aquellos asociados a una menor probabilidad de error en los clasificadores, o bien los que presenten mayor credibilidad según el test de Wilcoxon (Conover, 1999), o bien una combinación de los anteriores.

Si se emplea la probabilidad de error de los clasificadores como criterio de elección, habrá que tener en cuenta dos probabilidades. Si la cadena es $X-Y-T$, se implementan dos clasificadores, que dan las probabilidades de error: $P_{e,T}(Y)$ y $P_{e,T}(Y,X)$.

La otra opción, como se ha indicado, consiste en emplear el test de verosimilitud, que es una medida derivada de las probabilidades de error. Resultaría más eficiente si se emplea la información sin procesar contenida en el resultado de los clasificadores.

Sin embargo, se ha optado por la solución más sencilla, consistente en recurrir al resultado del test de Wilcoxon. A mayor valor de dicho test, más verosímil es el enlace.

Heurísticos de completitud del grafo

Finalmente, una vez realizado el proceso anterior, se aplica lo visto en el capítulo introductorio, en el Apartado 1.6.1, consistente en un conjunto de reglas heurísticas para completar los enlaces, con las restricciones de no generar ciclos ni nuevas estructuras en "V". Aunque las colisiones han sido determinadas mediante un conjunto de clasificadores, en lugar de los test estadísticos habituales, las reglas heurísticas para terminar de orientar los enlaces siguen siendo aplicables.

Estimación de densidad de probabilidad y test de hipótesis

En el Apartado 1.4.3 se ha introducido la técnica de remuestreo "bootstrap". Los algoritmos de inferencia causal van a requerir en uno de sus pasos la estima de la función de densidad de probabilidad del error de clasificación. Este cometido se lleva a cabo mediante remuestreo de la función de distribución empírica de la salida de los clasificadores. Para el segundo de los algoritmos de búsqueda causal propuesto, es preciso emplear un test de hipótesis. Su implementación también se apoya en una técnica derivada del remuestreo "bootstrap".

Algoritmo de clasificadores causales ccKnn

Una vez presentadas las herramientas principales necesarias para el algoritmo de inferencia causal, se detalla su desarrollo.

El algoritmo debe conocer el esqueleto de la red bayesiana. En esto es similar a otros como el MMHC (Tsamardinos et al., 2006), que emplean dos aproximaciones distintas para obtener en primer lugar el esqueleto del grafo y posteriormente orientar los enlaces. En el presente caso, se supondrá conocido el esqueleto auténtico, aunque su estimación se puede realizar con cualquier otro método de inferencia. En el desarrollo posterior se emplea el esqueleto generado por el mencionado método MMHC.

El esqueleto define las relaciones entre variables, eliminando las conexiones que son redundantes en el sentido de que la información que aporta una

variable sobre la otra se puede identificar a través de una tercera. Por consiguiente, el esqueleto de un grafo se sitúa al nivel de las relaciones estadísticas clásicas, en las que se conoce la relación entre variables, e incluso mediante la operación de "condicionado a" se pueden establecer caminos entre variables. Sin embargo, para conocer la verdadera relación de causa y efecto, el mecanismo que gobierna el flujo de información, es preciso orientar los enlaces entre variables. Esta tarea es la que afrontan los algoritmos de inferencia causal que se proponen en la presente tesis.

Supuesto conocido el esqueleto del grafo, se buscan todas las cadenas de tres variables

$$X - Y - T \tag{2.6}$$

y se construyen dos clasificadores: $\hat{T} = f_1(Y)$ y $\hat{T} = f_2(Y, X)$. A la hora de orientar los enlaces de la terna de variables (2.6) se pueden dar los cuatro casos recogidos en el Cuadro 2.2.

X | Y | T	mín P_e	Indep. condicional
$\circ \rightarrow \circ \rightarrow \bullet$	sólo Y es relevante	$X \perp\!\!\!\perp T \mid Y$
$\circ \leftarrow \circ \leftarrow \bullet$	sólo Y es relevante	$X \perp\!\!\!\perp T \mid Y$
$\circ \leftarrow \circ \rightarrow \bullet$	sólo Y es relevante	$X \perp\!\!\!\perp T \mid Y$
$\circ \rightarrow \circ \leftarrow \bullet$	X e Y son relevantes	$X \not\perp\!\!\!\perp T \mid Y$

Cuadro 2.2: Variables relevantes para clasificar la variable T, en función de la relación causal entre ellas. Sólo cuando la configuración es una estructura en "V", las variables X e Y son necesarias para obtener mínima probabilidad de error del clasificador.

Si la condición de Markov se cumple (ver Apartado 1.3.3), sólo hay una configuración del Cuadro 2.2 en el que X contiene información relevante de T no contenida en Y y, por lo tanto, en la que X es útil para la clasificación: el caso de colisión o estructura en "V": $X \rightarrow Y \leftarrow T$. Como muestra la tercera columna del Cuadro 2.2, para las tres primeras relaciones, una vez conocida la variable Y, X y T son estadísticamente independientes. En la categorización de (Kohavi y John, 1998) se estaría hablando de variables de *relevancia débil*, variables conectadas con T, pero a través del "Markov Blanket". Sin embargo, con la configuración de la cuarta entrada del Cuadro, la variable X es estadísticamente dependiente aún conociendo la consecuencia común, que es T. Ahora la variable X sí es miembro del "Markov Blanket", y entra en el conjunto de variables *fuertemente relevantes* para la variable T. Consecuentemente, es de esperar que en caso de colisión, el uso de la variable X reduzca el error de clasificación.

Esta propiedad se puede aprovechar para identificar las colisiones, y de esta manera orientar el árbol causal. La sustitución de los test estadísticos, que se emplean en los métodos habituales como el PC, por sencillos clasificadores abre la puerta a introducir métodos de aprendizaje máquina, que mejoren las deficiencias de los test estadísticos.

En este trabajo se ha ensayado con dos clasificadores, conocidos por sus buenas propiedades. La primera versión del método de inferencia causal con clasificadores emplea el k vecinos más próximos, k-NN (por sus siglas en inglés "Nearest Neighbor"). El método se resume a continuación, en el Algoritmo 4.

Algoritmo 4 ccKnn: inferencia causal usando clasificadores k-NN.

1: Identificar todas las cadenas de tres variables $X - Y - T$.
2: **para** cada cadena **hacer**
3: Entrenar los clasificadores $\hat{T} = f_1(Y)$ y $\hat{T} = f_2(Y, X)$ mediante k vecinos más próximos.
4: Calcular $p_Y = P_{e,T}(Y)$ y $p_{Y,X} = P_{e,T}(Y, X)$, y su distribución, mediante remuestreo "bootstrap".
5: **fin para**
6: Mediante un test de hipótesis, comprobar si $P_{e,T}(Y, X) < P_{e,T}(Y)$. Si la hipótesis nula (es mayor o igual) se rechaza, entonces orientar los enlaces como: $X \to Y \leftarrow T$.

La distribución del error de clasificación se estima empleando un remuestreo "bootstrap" de $B = 250$ réplicas. El test de hipótesis es un test de Wilcoxon (Conover, 1999), pues los clasificadores se entrenan con el mismo conjunto de entrenamiento en cada caso.

Se desea saber si las probabilidades de error p_Y y $p_{Y,X}$ de los clasificadores de T tienen la misma media, donde p_Y es la probabilidad de error del clasificador que usa la variable $f_1(Y)$, y $p_{Y,X}$ el error del clasificador $f_2(Y, X)$. Un posible test a emplear fue presentado por Wilcoxon (Conover, 1999). En este test, los datos son B observaciones $(p_Y^1, p_{Y,X}^1), \ldots, (p_Y^B, p_{Y,X}^B)$ que provienen de una variable aleatoria bidimensional. Se calculan las diferencias $|D_i| = |p_Y^i - p_{Y,X}^i|$, y se eliminan aquellos pares que presentan diferencia nula ($|D_i| = 0$). Se asignan rangos o intervalos con el siguiente protocolo: el primer intervalo se asigna al par $(p_Y^i - p_{Y,X}^i)$ con la menor diferencia en valor absoluto; los sucesivos intervalos se asignan por diferencias absolutas crecientes. En caso de que haya varios pares con igual diferencia, se les asigna a todos ellos un rango igual al promedio de los rangos que se les hubiera asignado.

Las asunciones que se han de respetar son las siguientes:

1. La simetría en la distribución de cada diferencia D_i

2. La independencia de cada D_i

3. La media de los D_i es igual

4. El dominio de los D_i es finito

Los datos usados en el presente trabajo para el test de Wilcoxon consisten en las B réplicas "bootstrap". Puesto que los conjuntos de entrenamiento de $\hat{T} = f_1(Y)$ y $\hat{T} = f_2(Y, X)$ son los mismos, el uso de este test es apropiado. Se quiere saber si la diferencia

$$D_i = p^i_{Y,X} - p^i_Y$$

es negativa.

El test de Wilcoxon calcula el rango con signo de la siguiente manera:

- R_i es el rango asignado a $(p^i_Y, p^i_{Y,X})$ si D_i es positivo.

- R_i es el negativo del rango asignado a $(p^i_Y, p^i_{Y,X})$ si D_i es negativo.

Y el estadístico es: $T = \frac{\sum_{i=1}^{B} R_i}{\sqrt{\sum_{i=1}^{B} R_i^2}}$, cuya distribución asintótica es, aproximadamente, gausiana.

Una vez que se ha procedido a identificar todas las colisiones o estructuras en "V", se procede a explorar el árbol causal con el objeto de orientar todas las posibles conexiones, sin generar nuevas colisiones ni caminos cerrados o ciclos, con las cuatro reglas o heurísticos que se mencionaron en la Sección 1.6.1.

R_1 Orientar $X_j - X_k$ como $X_j \to X_k$ en caso de que exista $X_i \to X_j$ y no sean adjuntos X_i y X_k.

R_2 Orientar $X_i - X_j$ como $X_i \to X_j$ si hay una cadena $X_i \to X_k \to X_j$.

R_3 Orientar $X_i - X_j$ como $X_i \to X_j$ cuando hay dos cadenas $X_i - X_k \to X_j$ y $X_i - X_l \to X_j$ tal que no sean adjuntos X_k y X_l.

R_4 Orientar $X_i - X_j$ como $X_i \to X_j$ cuando hay dos cadenas $X_i - X_k \to X_l$ y $X_k \to X_l \to X_j$ tal que no sean adjuntos X_k y X_j pero sí lo sean X_i y X_l.

2.1.2. Experimentos

El método ccKnn se ha probado en un conjunto de datos sintéticos, para comprobar su capacidad en diversas situaciones. Para ello se han generado grafos de manera aleatoria, siguiendo (Spirtes et al., 2000), con estas configuraciones:

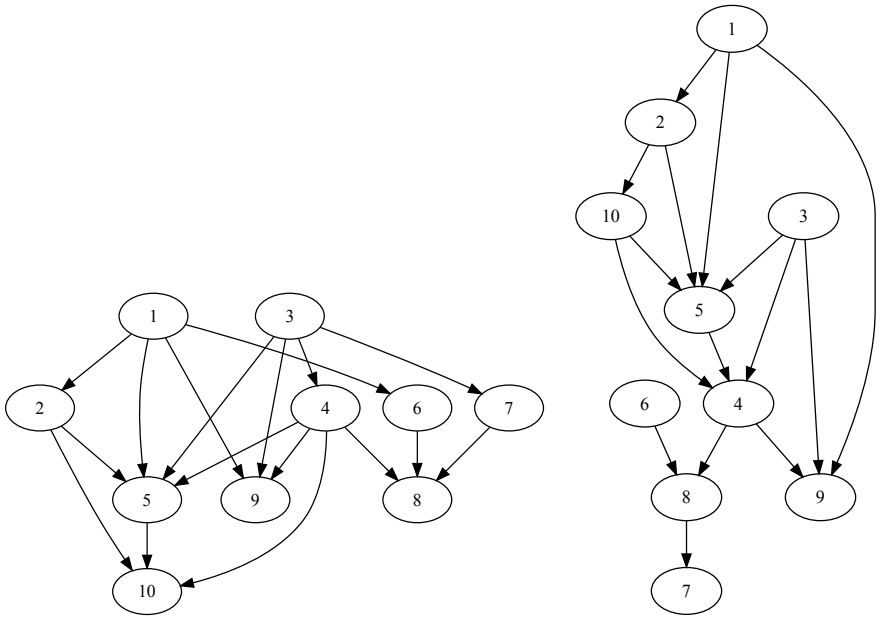
- Número de variables: 10 ó 50.
- Grado o "fan in": número medio de padres de cada variable, $\{2, \ldots, 5\}$.

Se han generado 20 grafos por cada combinación de parámetros, y de cada grafo se han extraído 5.000 muestras. En la Figura 2.3 se muestra un ejemplo de grafo de 10 nodos y grado 2, junto con su estima con el método propuesto. Puede apreciarse cómo las estructuras en "V" del nodo X_9 se recuperan correctamente: $X_1 \rightarrow X_9 \leftarrow X_3$ y $X_1 \rightarrow X_9 \leftarrow X_4$. Sin embargo, en la variable X_5 se ha cometido un error en la estimación del sentido del enlace $X_4 - X_5$, que ha provocado la identificación errónea de una colisión en $X_5 \rightarrow X_4 \leftarrow X_3$.

En este ejemplo, la probabilidad de pérdida en la orientación de los enlaces asciende a un valor de $P_e = 26,7\,\%$. Es preciso destacar que la interpretación de las probabilidades de error es diferente a la que se realiza en otras aplicaciones de aprendizaje máquina como, por ejemplo, en clasificación. Esta interpretación debe tener en cuenta un factor cuantitativo y otro de valoración cualitativa. Por un lado, el grafo de la Figura 2.3 tiene únicamente 17 enlaces. Si un algoritmo fuera capaz de localizar sin error dicho número de enlaces, y sólo se cometiera una equivocación orientando un enlace, se estaría cometiendo un error de pérdida del 6 %. En la estimación realizada para este ejemplo, sólo ha sido posible orientar 15 enlaces de los 17 totales; de esos 15 se han cometido 4 errores, que arrojan el error mencionado del $26,7\,\%$. Por supuesto, es posible mejorar la estima aumentando el número de enlaces, de nodos, o de ambos a la vez. Sin embargo, la complejidad de estos grafos limitan sobremanera la capacidad de inferencia de muchos algoritmos, que para redes poco densas pueden ser eficaces y pierden sus buenas propiedades en problemas con un alto número de enlaces. Por otro lado, y de manera cualitativa, el significado de las variables puede hacer que el error a la hora de orientar un enlace se convierta en muy grave o en insignificante. Dicho de otra manera, no deberían tener el mismo peso todos los errores en la orientación de los enlaces, sino que habría que ponderarlos en función del valor de la variable. Esta ponderación dependerá del problema concreto, y debería realizarla un experto según su conocimiento a priori.

Esta reflexión conduce a otra pregunta. ¿Es posible una medida de la confianza o verosimilitud de los enlaces causales? A pesar de que sería una medida muy interesante para valorar los resultados de una inferencia causal, no existen en la actualidad métodos que de manera directa ofrezcan ese resultado. Queda como tema de investigación abierto de interesante uso potencial.

A modo de ilustración, también se muestran en la Figura 2.4 dos ejemplos de grafos de 50 nodos, para valores extremos de la variable grado. Como se ve, el número de enlaces crece hasta ser de 125 para la red de la Figura 2.4(b).



(a) Grafo generado de grado 2. (b) Estimación.

Figura 2.3: Ejemplo de grafo generado sintéticamente y estimación mediante el método ccKnn. Las conexiones cuya orientación no ha sido posible determinar, se han eliminado.

La siguiente Figura 2.5 muestra los resultados de aprender los grafos mediante el método de *clasificadores causales*, ccKnn (de-Prado-Cumplido y Artés-Rodríguez, 2008), el PC (Spirtes et al., 2000) y el MMHC (Tsamardinos et al., 2006). Las probabilidades de error se pueden catalogar de la siguiente forma:

1. Pérdida en enlaces: probabilidad de que un enlace esté presente en el grafo real pero no en el estimado.

2. Falsa alarma en enlaces: probabilidad de que un enlace no esté presente en el grafo real pero sí en el grafo estimado.

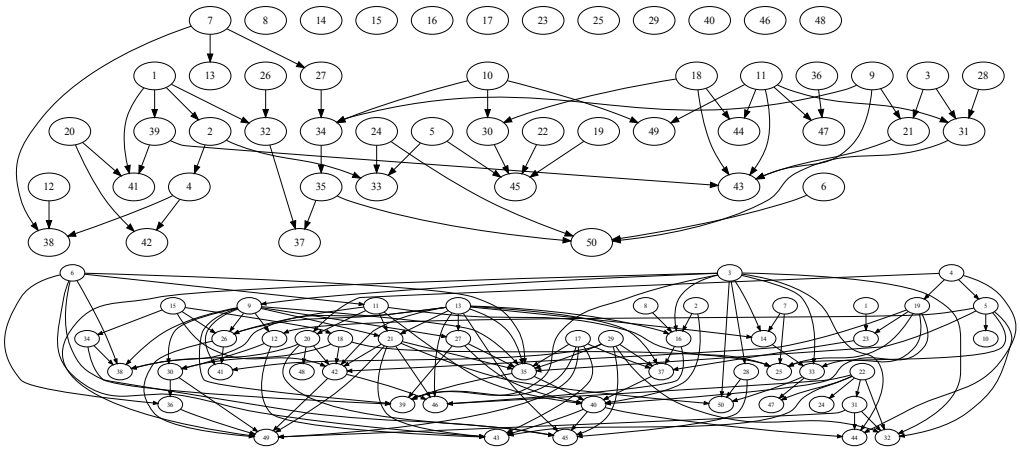


Figura 2.4: Ejemplo de grafo de 50 nodos generado sintéticamente, para dos valores de la variable grado: en la figura superior de grado 2 y en la figura inferior de grado 5.

Lo cual se corresponde con los errores tipo I y II habituales. En nuestro caso esta medida no aporta información del rendimiento del algoritmo, puesto que se da por conocido el esqueleto auténtico del grafo. Sin embargo, se ha de añadir un tipo de error adicional, aquel que se comete en la orientación de los enlaces correctamente detectados.

1. Pérdida en orientación: para aquellos enlaces correctamente detectados, se comete un error de pérdida en la orientación si en el grafo real hay una flecha que no se encuentra en el grafo estimado.

2. Falsa alarma en orientación: para enlaces correctamente detectados, se comete este error al estimar la existencia de una flecha en un enlace del grafo estimado que no se corresponde con la realidad.

Este segundo conjunto de errores es el que se aplica para cuantificar el rendimiento del algoritmo ccKnn; sin embargo, debido a la configuración del problema, un error de pérdida de una orientación implica directamente una falsa alarma en el extremo contrario del enlace, y viceversa. Ambas medidas van a ser iguales una vez se condiciona al conjunto de enlaces correctamente detectados. Por este motivo únicamente se muestran los datos para las pérdidas en orientación de enlaces.

De las gráficas en la Figura 2.5 se pueden extraer algunas conclusiones interesantes. Por ejemplo, el método MMHC -también el método ccKnn propuesto- emplea un algoritmo en dos pasos, estimando por separado esqueleto y orientación de los enlaces. Si se compara el método de referencia

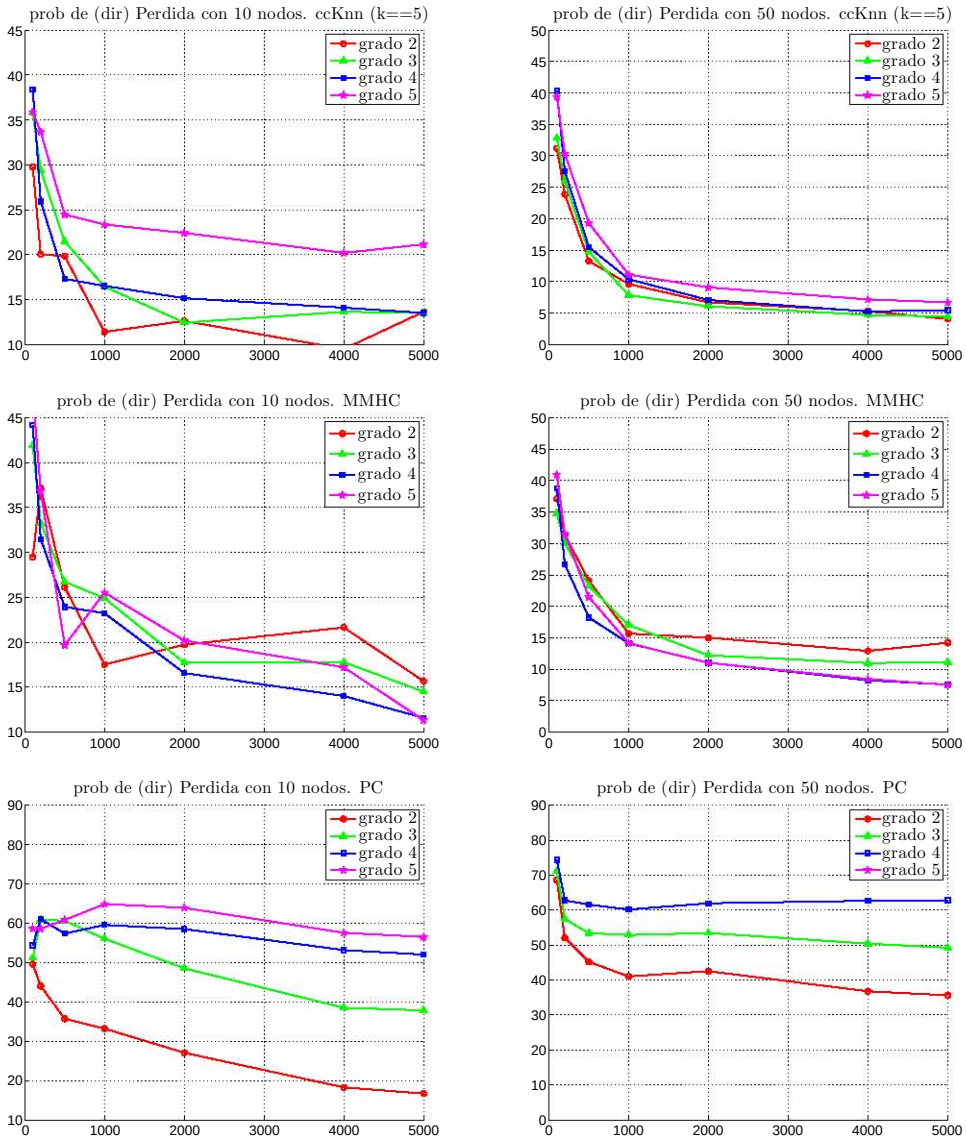


Figura 2.5: Probabilidad de Pérdida en la orientación de los enlaces. Resultados para grafos de 10 nodos (columna izquierda) y de 50 nodos (columna derecha). Se comparan los métodos PC, MMHC y ccKnn.

PC con el MMHC, se observa que éste último presenta un mejor comportamiento. Se puede concluir que la división de tareas no supone una mayor probabilidad de error, sino todo lo contrario. Hay que notar que la escala de las gráficas correspondientes al método PC es mayor que las otras. Los

resultados, para un subconjunto de parámetros, se presentan en el Cuadro 2.3.

	10 nodos					
	grado 2			grado 4		
	200	1000	5000	200	1000	5000
PC	44.0	33.2	16.7	61.1	59.5	52.0
MMHC	37.1	17.5	15.7	31.5	23.2	**11.6**
ccKnn	**20.1**	**11.4**	**13.6**	**25.9**	**16.5**	13.5
	50 nodos					
	grado 2			grado 4		
	200	1000	5000	200	1000	5000
PC	52.0	41.0	35.6	62.8	60.2	62.8
MMHC	31.4	15.7	14.2	**26.7**	14.1	7.6
ccKnn	**23.9**	**9.6**	**4.1**	27.5	**10.4**	**5.5**

Cuadro 2.3: Probabilidad de acierto ponderado a la hora de orientar enlaces causales para los métodos PC, MMHC y ccKnn. Datos en tanto por ciento.

En general, el método ccKnn obtiene unos mejores resultados que el MMHC. Para el caso de problemas menos densos, con sólo 10 variables, y grado (número medio de padres) alto, los resultados son ligeramente peores. Esto puede deberse al elevado número de enlaces respecto al número de nodos, por lo que el algoritmo debe tomar decisiones sobre muchos enlaces con la información redundante de unos pocos nodos. Cuando el número de nodos aumenta de 10 a 50, los resultados mejoran sensiblemente.

2.1.3. Método basado en salida blanda de la SVM

Una mejora inmediata del método expuesto consiste en sustituir el clasificador de k vecinos más próximos por una máquina más potente, como por ejemplo una Máquina de Vectores Soporte (SVM por sus siglas en inglés). Sin embargo, esto añadiría mayor complejidad al método, al ser por lo general más costosas de entrenar las SVM. Por otra parte, una de las ventajas de emplear estas máquinas es que se puede interpretar la salida de la SVM, con cierta modificación, como una distribución probabilística de las etiquetas dadas las muestras, es decir, que la SVM nos proporciona una aproximación a $P(c|\mathbf{x}_i)$. Gracias a esta propiedad, es posible eludir el uso de un remuestreo tipo "bootstrap" para modelar las distribuciones de salida, aligerando considerablemente la carga de la batería de clasificadores.

Salida probabilística de la SVM

Para aligerar la carga computacional del algoritmo previo introducida por la etapa "bootstrap", se puede usar la salida probabilística de la SVM.

La salida de la SVM, como se vio en la Sección 1.4.1, es:

$$f(\mathbf{x}) = \sum_{i=1}^{nSV} \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b$$

Gracias al empleo de una función sigmoide, esta salida puede emplearse en un modelo paramétrico e interpretarse como una probabilidad a posteriori (Platt, 1999) $P(c|\mathbf{x})$.

$$P(c|\mathbf{x}) = \frac{1}{1 + \exp(Af(\mathbf{x}) + B)} \tag{2.7}$$

donde c identifica la clase o etiqueta de la muestra, y A y B son parámetros que deberán asegurar que la operación representa una auténtica probabilidad.

Los parámetros A y B se determinan por máxima verosimilitud, para lo que se optimiza una función de error de entropía cruzada:

$$\text{mín} \quad -\sum_i t_i \log(p_i) + (1 - t_i) \log(1 - p_i) \tag{2.8}$$

donde t_i es una transformación de escala de las etiquetas: $t_i = (y_i + 1)/2$ para que pertenezcan al conjunto $[0, 1]$, y donde $p_i = 1/\left(1 + \exp(Af_i + B)\right)$.

Es necesario tener en consideración dos aspectos de la minimización de (2.8): la elección del conjunto de entrenamiento (p_i, t_i) y el posible sobreajuste de las probabilidades obtenidas. La solución más directa para la elección del conjunto de entrenamiento es emplear los mismos datos que se usaron para obtener la solución de la SVM. Por lo tanto, $f_i = f(\mathbf{x}_i)$. Para evitar el sobreajuste se toman dos medidas. En primer lugar se realiza una validación cruzada, dividiendo el conjunto de entrenamiento en bloques, en el que de manera sucesiva, uno se emplea para validar los parámetros obtenidos al entrenar con el resto de bloques. Esto se repite hasta que todos los bloques se han usado en la evaluación. La validación cruzada no es suficiente por sí misma, por lo que se necesita un regularizador adicional. Se podría imponer un a priori sobre los parámetros A y B, o bien sobre la distribución de los datos, que es la opción más simple. Puesto que $y_i = \text{signo}(f_i)$, el valor de t_i para $f_i > 0$ debería ser 1. Pero para aceptar cierto grado de incertidumbre y suavizar así los datos, se puede asignar el valor $t_i = 1 - \epsilon_+$. De manera análoga, para valores negativos de la salida blanda de la SVM, se asignaría $t_i = \epsilon_-$. En concreto, estos valores calculados por el criterio MAP (máximo

a posteriori) son $t_+ = \frac{N_+ +1}{N_+ +2}$ y $t_- = \frac{1}{N_- +2}$, donde N_+ y N_- son el número de muestras de la clases $+1$ y -1, respectivamente.

Con el fin de mejorar el rendimiento de esta estimación, es posible realizar algunas ligeras modificaciones, siguiendo (Lin et al., 2007). En concreto, la fórmula (2.8) se puede reemplazar por

$$- (t_i \log p_i + (1 - t_i) \log(1 - p_i)) = \tag{2.9}$$
$$= (t_i - 1)(Af_i + B) + \log(1 + \exp(Af_i + B)) = \tag{2.10}$$
$$= t_i(Af_i + B) + \log(1 + \exp(-Af_i + B)) \tag{2.11}$$

Así, en caso que $Af_i + B \geq 0$, se usaría (2.9), y en caso contrario (2.10).

Análogamente, la evaluación de (2.7), en los casos que $Af_i + B \geq 0$, se haría con $(\exp(-Af - B)/(1 + \exp(-Af - B))$.

Las probabilidades de las clases $P_1(c|Y)$ y $P_2(c|\{Y, X\})$ se pueden calcular usando el método anterior. En la Figura 2.6 se visualiza un ejemplo de la configuración del problema que a resolver. Por lo tanto, una medida de relevancia de la variable X respecto de la clasificación de T se cuantifica mediante esta integral:

$$D = \int |P_1(c|Y) - P_2(c|\{Y, X\})| P(\mathbf{x}) d\mathbf{x} \tag{2.12}$$

La salida probabilística de la SVM aproxima las probabilidades $\hat{P}(c|\mathbf{x})$, por lo que el criterio es ahora:

$$\hat{D} = \frac{1}{N} \sum_{i=1}^{N} |\hat{P}_1(c|Y) - \hat{P}_2(c|\{Y, X\})| \tag{2.13}$$

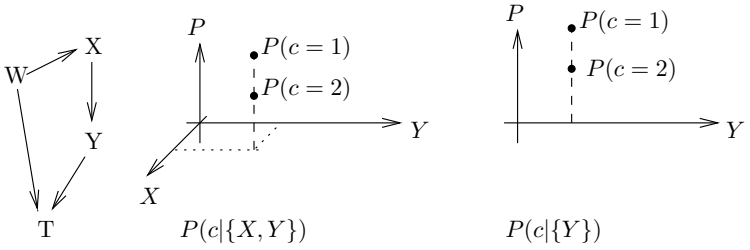


Figura 2.6: Comparativa de las distribuciones de los clasificadores individuales.

En los casos en los que la etiqueta c no sea binaria sino multiclase, con $\mathcal{C}$ clases, se hará una ligera modificación, quedando como:

$$\hat{D} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{\mathcal{C}-1} |\hat{P}_1(c|Y) - \hat{P}_2(c|\{Y, X\})| \tag{2.14}$$

Finalmente, se determina si la medida $\hat{D}$ es significativamente mayor que 0. En dicho caso, la variable X se puede considerar relevante para clasificar T, verificando por tanto el hallazgo de una nueva estructura en "V". Con este fin, se emplea el siguiente test:

$$\begin{cases} H_0 & P_1(c|Y) = P_2(c|\{Y, X\}) \quad \text{si } \hat{D} = 0 \\ H_1 & P_1(c|Y) \neq P_2(c|\{Y, X\}) \quad \text{si } \hat{D} \neq 0 \end{cases} \tag{2.15}$$

Se ha llevado a cabo mediante un test de hipótesis con "bootstrap". Se define el estadístico del test como

$$T_n = \frac{|\hat{D} - 0|}{\hat{\sigma}} \tag{2.16}$$

donde $\hat{\sigma}$ es una estimación de la desviación típica de $\hat{D}$.

Ambas estimaciones en la Ecuación 2.16 se pueden realizar mediante la media y varianza muestral. El método "bootstrap" remuestrea D y para cada una de las B réplicas calcula el estadístico del test. El α-percentil de las réplicas es el umbral con el que comparar T_n. El procedimiento se recoge en el Algoritmo 5.

Algoritmo 5 Test de Hipótesis mediante "bootstrap" para determinar si $\hat{D}$ es nulo.

1: Obtener estimaciones $\hat{D}_i$ para cada cadena de tres nodos.
2: Calcular el estadístico $T_n(\hat{\mathbf{D}})$.
3: **para** $j = 1 \rightarrow B$ **hacer**
4: Remuestrear $\hat{\mathbf{D}}^{*j}$.
5: Calcular $T_n^{*j}(\hat{\mathbf{D}}^{*j})$.
6: **fin para**
7: Ordenar los estadísticos de las réplicas: $T_n^{*(1)}, \ldots, T_n^{*(B)}$.
8: Fijar $T_\alpha = T_n^{*(q)}$, donde $q = \lceil (1-\alpha) \rceil (B+1)$.
9: Comparar T_n con T_α. Si $T_n < T_\alpha$, no se puede rechazar la hipótesis nula, por lo que se asume que $\hat{D} = 0$.

Clasificadores multiclase y salida probabilística

El procedimiento para obtener una salida probabilística cuando el problema de clasificación presenta más de dos clases no es tan directo como

el que se ha expuesto en la sección anterior. A continuación se presentan las particularidades en la obtención de una salida blanda en clasificadores multiclase.

Cuando las etiquetas del conjunto de datos tienen un número de clases $\mathcal{C} > 2$ son varias las estrategias que se pueden ensayar. Una de ellas es la de "uno-contra-el resto", en la que se entrenan $\mathcal{C}$ clasificadores $f^1, \ldots, f^{\mathcal{C}}$ binarios, para separar cada clase de todas las demás combinadas. De esta forma, el clasificador f^i se entrenaría con datos binarios formados por la clase i y por todas las demás exceptuando dicha etiqueta. El último paso consistiría en combinar las salidas de cada clasificador seleccionando aquella con mayor margen:

$$\arg \max_{j=1,\ldots,\mathcal{C}} g^j(\mathbf{x}) \tag{2.17}$$

$$g^j(\mathbf{x}) = \sum_{i=1}^{m} y_i \alpha_i^j k(\mathbf{x}, \mathbf{x}_i) + b^j \tag{2.18}$$

donde los clasificadores f^i son el signo de las salidas blandas: $f^i(\mathbf{x}) = \text{signo}(g^i(\mathbf{x}))$.

La aproximación "uno-contra-el resto" tiene varias desventajas. Por un lado, resulta cuestionable la comparación directa de la salida blanda de distintos clasificadores, pues pueden tener distintas escalas. Por otra parte, los problemas a resolver tienden a ser muy asimétricos en número de muestras, al agrupar todas las clases menos una.

Otra estrategia que no adolece de estos inconvenientes es la conocida como "uno-contra-uno". En este caso, se emplean $\mathcal{C}(\mathcal{C}-1)/2$ clasificadores entrenados con pares de etiquetas. Aunque el número de máquinas a entrenar crece de manera cuadrática con el número de etiquetas, la sencillez de cada una de estas hace que resulte ventajoso en muchas ocasiones. Una vez obtenidas las decisiones de cada uno de los $\mathcal{C}(\mathcal{C}-1)/2$ clasificadores, éstas se combinan en una decisión global para cada muestra por mayoría simple.

La tercera aproximación al problema que se expone aquí consiste en un punto intermedio entre las ideas previas. En lugar de clasificar las clases i y la j, como en el caso "uno-contra-uno", o de clasificar i contra $\{\mathcal{C} \setminus i\}$, se pueden hacer distintas agrupaciones, como por ejemplo, clases pares contra clases impares, o el primer tercio contra el último tercio de las clases. Si las divisiones se realizan de manera inteligente, y se interpretan las decisiones parciales como bits de un *código protección frente a errores*, entonces, aunque alguna de las máquinas cometiera una equivocación al clasificar, se podría detectar e incluso corregir la salida errónea (Dietterich y Bakiri, 1995; Allwein et al., 2001).

Se va a emplear el método "uno-contra-uno" por sus buenas propiedades y sencillez. Una vez que el problema de multiclasificación ha sido resuelto, las probabilidades se obtienen siguiendo las ideas de (Platt, 1999) para el caso binario expuestas en el apartado anterior. Para obtener las probabilidades $p_c = P(c|\mathbf{x})$, $c = 1, \ldots, \mathcal{C}$, en primer lugar se determinan las probabilidades por pares de clases: $r_{ij} = P(y = i | y = i \text{ o } j, \mathbf{x})$. Se asume que $r_{ij} = \frac{1}{1+\exp(Af(\mathbf{x})+B)}$. Finalmente, las probabilidades p_c se calculan como se expone en (Chang y Lin, 2011), resolviendo el siguiente problema de optimización:

$$\begin{aligned} \min_{\mathbf{p}} \ & \frac{1}{2} \sum_{i=1}^{\mathcal{C}} \sum_{j:j\neq i} (r_{ij} p_i - r_{ij} p_j)^2 \\ \text{restringido a } & p_i \geq 0, \forall i, \quad \sum_{i=1}^{\mathcal{C}} p_i = 1 \end{aligned} \tag{2.19}$$

Permutación aleatoria en clasificadores multiclase

Es posible demostrar que $P(c|Y) = P(c|\{Y, X^{perm}\})$, donde X^{perm} es una permutación aleatoria de dicha variable, como se desarrolla en (Shen et al., 2008) basándose en las ideas de Leo Breiman (Breiman, 2001). Gracias a esta propiedad sólo el clasificador para $\{Y, X\}$ debe ser entrenado. Por consiguiente, la misma máquina $f(\mathbf{x})$ con la variable X permutada se puede sustituir en la Ecuación 2.7 para determinar la estima de $P(c|Y)$.

La demostración de esta propiedad es sencilla. En primer lugar se define la permutación aleatoria de la variable i del vector multidimensional $\mathbf{x}_{(i)}$ de la siguiente forma. Se generan N números aleatorios uniformemente entre 0 y 1: $\xi_1, \ldots, \xi_N$. A continuación se realiza el intercambio entre los valores $\mathbf{x}_k^i$ y $\mathbf{x}_j^i$, donde $k = 1, \ldots, N-1$ y $j = \lfloor N \cdot \xi_k \rfloor + 1$. Por otro lado, se define $\mathbf{x}_{\setminus i}$ como la variable en la que se ha eliminado la componente i-ésima.

Teorema 2. *En los supuestos enunciados, las siguientes probabilidades son equivalentes:* $P(c|\mathbf{x}_{(i)}) = P(c|\mathbf{x}_{\setminus i})$.

Demostración. Dado que se ha empleado una distribución uniforme ξ para la permutación de $\mathbf{x}_{(i)}$, la distribución de esa variable permanece inalterada: $P(\mathbf{x}_{(i)}^i) = P(\mathbf{x}^i)$.

Por tanto, $P(\mathbf{x}_{(i)}) = P(\mathbf{x}_{(i)}^i, \mathbf{x}_{\setminus i}) = P(\mathbf{x}_{(i)}^i) \cdot P(\mathbf{x}_{\setminus i}) = P(\mathbf{x}^i) \cdot P(\mathbf{x}_{\setminus i})$, que se verifica pues las distribuciones de $\mathbf{x}_{(i)}^i$ y $\mathbf{x}^i$ son independientes debido a la permutación. Igualmente, se puede afirmar que

$$P(\mathbf{x}_{(i)}, c) = P(\mathbf{x}^i_{(i)}) \cdot P(\mathbf{x}_{\setminus i}, c) = P(\mathbf{x}^i) \cdot P(\mathbf{x}_{\setminus i}, c)$$

Finalmente,

$$P(c|\mathbf{x}_{(i)}) = \frac{P(c, \mathbf{x}_{(i)})}{P(\mathbf{x}_{(i)})} = \frac{P(\mathbf{x}^i) \cdot P(\mathbf{x}_{\setminus i}, c)}{P(\mathbf{x}^i) \cdot P(\mathbf{x}_{\setminus i})} = P(c|\mathbf{x}_{\setminus i})$$

□

Un corolario directo del Teorema 2 concluye que la información mutua entre etiqueta y datos es la misma si se omite la variable que si se permuta la misma, es decir, $I(c, \mathbf{x}_{(i)}) = I(c, \mathbf{x}_{\setminus i})$, puesto que:

$$\begin{aligned} I(c, \mathbf{x}_{(i)}) &= \sum_c \int_{\mathbf{x}_{(i)}} P(c, \mathbf{x}_{(i)}) \log \frac{P(c, \mathbf{x}_{(i)})}{P(c) \cdot P(\mathbf{x}_{(i)})} d\mathbf{x}_{(i)} = \\ &= \sum_c \int_{\mathbf{x}_{\setminus i}} \int_{\mathbf{x}^i_{(i)}} P(\mathbf{x}^i_{(i)}) P(c, \mathbf{x}_{\setminus i}) \log \frac{P(c, \mathbf{x}_{\setminus i})}{P(c) \cdot P(\mathbf{x}_{\setminus i})} d\mathbf{x}^i_{(i)} d\mathbf{x}_{\setminus i} = \\ &= \sum_c \int_{\mathbf{x}_{\setminus i}} P(c, \mathbf{x}_{\setminus i}) \log \frac{P(c, \mathbf{x}_{\setminus i})}{P(c) \cdot P(\mathbf{x}_{\setminus i})} d\mathbf{x}_{\setminus i} \int_{\mathbf{x}^i_{(i)}} P(\mathbf{x}^i_{(i)}) d\mathbf{x}^i_{(i)} = \\ &= I(c, \mathbf{x}_{\setminus i}) \end{aligned} \tag{2.20}$$

Continuando con el razonamiento, la medida $\hat{D}$ de la Ecuación 2.13 se puede determinar entrenando un sólo clasificador y su salida probabilística asociada, $\hat{P}_2(c|Y, X)$, mientras que se asigna $\hat{P}_1(c|Y) = \hat{P}_2(c|Y, X^{perm})$ realizando la permutación aleatoria de la variable X y usando ese vector en la máquina entrenada para $\hat{P}_2(\cdot)$.

Algoritmo de clasificadores causales ccMSVM

Recogiendo todo lo anterior, el algoritmo de inferencia causal con clasificadores SVM quedaría definitivamente como se expone en el Algoritmo 6.

Es preciso hacer constar cómo, a diferencia del método de clasificación causal ccKnn, en este caso la comparación entre máquinas en una y dos dimensiones se realiza a partir de las salidas de las SVM. Así, no es preciso dar el paso intermedio de calcular la probabilidad de error de clasificación antes de proceder a la comparación, como sí ocurre en el algoritmo con clasificadores k-NN. El Cuadro 2.4 recoge estas diferencias.

Si bien el método ccMSVM también requiere de una etapa de remuestreo, ésta no es tan costosa computacionalmente como en el caso del ccKnn. Más

Algoritmo 6 ccMSVM: inferencia causal usando la salida blanda de la M-SVM.

1: Identificar todas las cadenas de tres variables $X - Y - T$.
2: **para** cada cadena **hacer**
3: Entrenar los clasificadores $\hat{T} = f_1(Y)$ y $\hat{T} = f_2(Y, X)$ mediante M-SVM.
4: Calcular $P_{e,T}(Y, X)$ y $P_e(Y)$, y su distribución, mediante la salida blanda de la SVM y la permutación aleatoria de X.
5: **fin para**
6: Mediante un test de hipótesis "bootstrap", comprobar si $\hat{D} = 0$ (Ver Algoritmo 5). Si la hipótesis nula (el estadístico $\hat{D}$ es cero) se rechaza, entonces orientar los enlaces como: $X \rightarrow Y \leftarrow T$.

adelante se ofrecen resultados de rendimiento en tiempo de cómputo, donde se aprecia este efecto.

Método	$X - Y - T$	$Y - T$	Medida	Test
ccKnn	$P^*_{e,T}(X, Y)$	$P^*_{e,T}(Y)$	$P_{e,T}(X, Y) - P_{e,T}(Y)$	Wilcoxon
ccMSVM	$P^\dagger(c\|X, Y)$	$P^\dagger(c\|Y)$	$\sum \|P(c\|Y) - P(c\|Y, X)\|$	"Bootstrap"

Cuadro 2.4: Comparativa de métodos ccKnn y ccMSVM. *Obtenido mediante remuestreo "bootstrap". †Método de J. Platt para salida probabilística.

2.1.4. Experimentos

Se ha aplicado el algoritmo ccMSVM a la base de datos descrita en el Apartado 2.1.2.

Los resultados en cuanto a probabilidad de error son peores que para ccKnn, pues el clasificador SVM con núcleo gausiano no es tan discriminante como el k-NN para este tipo de datos. Sin embargo ofrece una mayor flexibilidad, y es posible explorar el uso de distintos núcleos, adaptándose a diversos tipos de datos. Los resultados son levemente mejores que para el método PC clásico, además de suponer un ahorro significativo en tiempo respecto al ccKnn.

Las curvas de probabilidad de pérdida en la orientación de los enlaces se muestran en la Figura 2.8. Se incluyen los dos algoritmos propuestos, ccKnn y ccMSVM, y los algoritmos PC y MMHC, tanto para la configuración de redes de 10 y 50 nodos. La escala del algoritmo PC es distinta al resto de curvas para poderla visualizar completamente. También se muestra en el Cuadro

2.5 un resumen del comportamiento de los distintos algoritmos, para puntos localizados de las gráficas.

La elección de los hiperparámetros C y σ de la máquina de vectores soporte con núcleo gausiano, se ha realizado mediante una validación cruzada. Para cada una de las cadenas de tres nodos, se han dividido los datos en sendos conjuntos de entrenamiento y validación. La salida probabilística del conjunto que no ha sido empleado en la fase de entrenamiento, es la que se utiliza para determinar la diferencia entre los clasificadores de una y dos variables.

Puede verse cómo la inclusión de los clasificadores M-SVM empeora los resultados respecto a la versión con k vecinos más próximos. La M-SVM con núcleos lineal o polinómico ofrecen peores resultados.

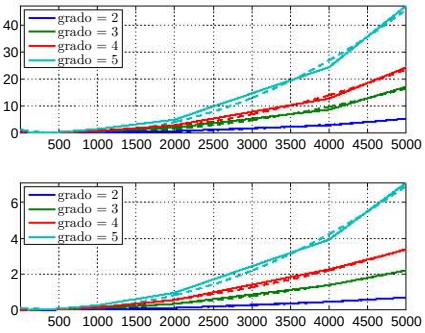
	10 nodos					
	grado 2			grado 4		
	200	1000	5000	200	1000	5000
PC	44.0	33.2	16.7	61.1	59.5	52.0
MMHC	37.1	17.5	15.7	31.5	23.2	**11.6**
ccKnn	**20.1**	**11.4**	**13.6**	**25.9**	**16.5**	13.5
ccMSVM	30.3	22.6	16.2	31.2	29.1	25.3
	50 nodos					
	grado 2			grado 4		
	200	1000	5000	200	1000	5000
PC	52.0	41.0	35.6	62.8	60.2	62.8
MMHC	31.4	15.7	14.2	**26.7**	14.1	7.6
ccKnn	**23.9**	**9.6**	**4.1**	27.5	**10.4**	**5.5**
ccMSVM	35.7	21.5	17.1	36.2	24.0	19.8

Cuadro 2.5: Probabilidad de acierto ponderado a la hora de orientar enlaces causales. Datos en tanto por ciento.

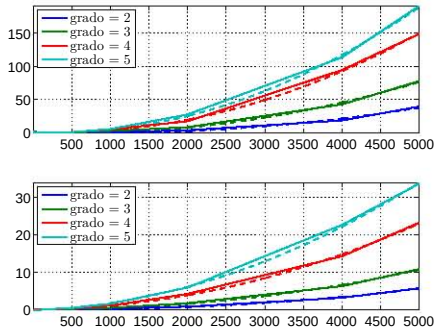
Sin embargo, ccMSVM no necesita de un remuestreo para obtener la distribución del error de los clasificadores, por lo que su tiempo de cálculo se reduce considerablemente. Este aumento de la velocidad se ve reforzado por el uso de permutación aleatoria de los datos bidimensionales para evitar entrenar una nueva M-SVM en cada cadena de tres nodos. El Teorema 2 asegura la validez teórica de este proceso. Además, las relaciones entre tasa de error de los clasificadores $f_1(Y)$ y $f_2(Y, X)$, son similares con k vecinos más próximos y con M-SVM.

Los tiempos de cálculo promedio de las simulaciones se muestran en la Figura 2.7. El clasificador k vecinos más próximos tiene una complejidad

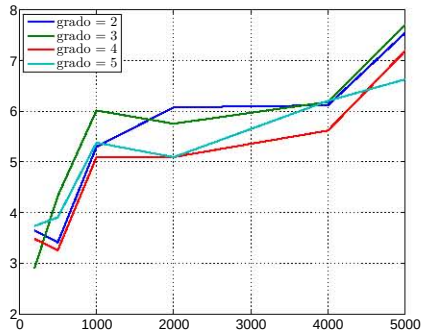
$\mathcal{O}(dN^2)$. Se ha ajustado una curva polinómica de orden 2, en línea discontinua en las figuras, que confirma el comportamiento cuadrático del tiempo empleado por el ccKnn.



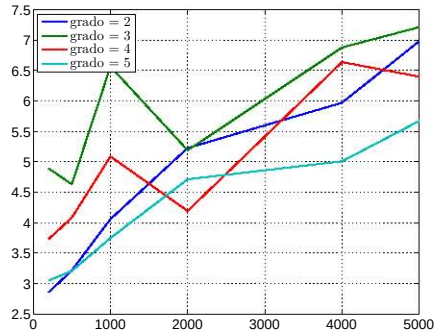
(a) Grafos de 10 nodos. ccKnn (panel superior) y ccMSVM (panel inferior).



(b) Grafos de 50 nodos. ccKnn (panel superior) y ccMSVM (panel inferior).



(c) Cociente tiempos, 10 nodos.



(d) Cociente tiempos, 50 nodos.

Figura 2.7: Tiempo (en horas) frente al número de muestras invertidos por ccKnn y ccMSVM. Éste último acelera el proceso en un factor entre 5 y 7.

En los paneles de la Figura 2.7(a-b) se comprueba la reducción significativa de tiempo del ccMSVM, representado en la parte inferior de las gráficas. El eje de ordenadas tiene una escala distinta en cada figura. La complejidad de la implementación de la M-SVM empleada (Chang y Lin, 2011) está situada entre $\mathcal{O}(N^2)$ y $\mathcal{O}(dN^2)$. Las simulaciones también contrastan este comportamiento cuadrático.

En función del número de muestras, la aceleración del ccMSVM puede llegar casi a un orden de magnitud. Hay que notar que las simulaciones se han llevado a cabo con una granja de computación, compuesto por máquinas similares en potencia de cálculo, aunque no completamente homogéneas. Sin

embargo, debido a que las curvas de tiempo están promediadas para las 20 realizaciones por cada conjunto de parámetros, el comportamiento detallado previamente es válido.

La otra ventaja del ccMSVM consiste en la posibilidad de emplear núcleos adaptados a los datos con los que se trabaja. Por ejemplo, si se quisiera explorar relaciones causales en documentos de texto, se podría emplear un núcleo de cadenas de texto ("string kernel") (Lodhi et al., 2002). Otra aplicación donde este mismo núcleo ha sido empleado con buenos resultados es en problemas de genética. La inferencia causal en estos problemas es una aplicación muy interesante.

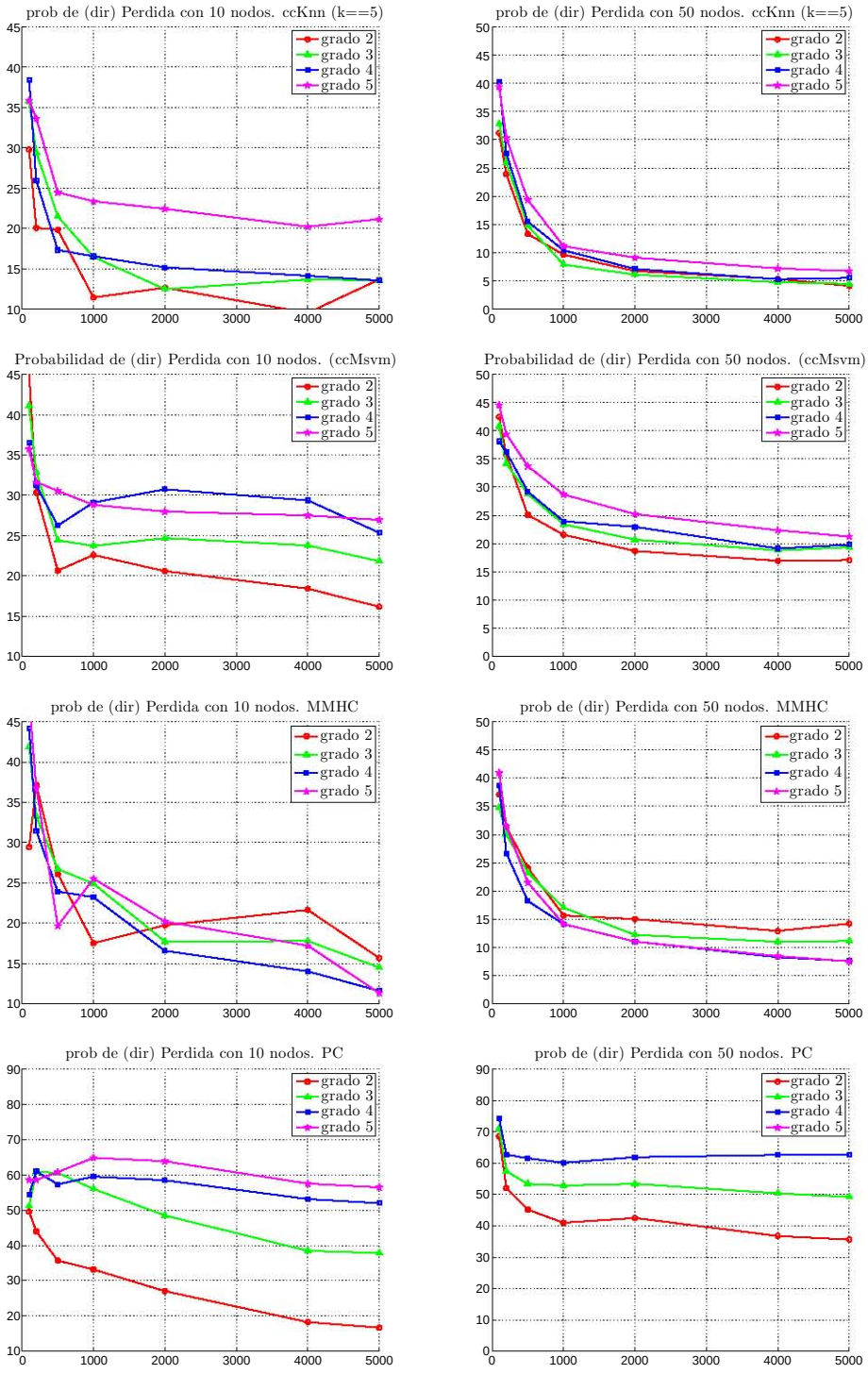


Figura 2.8: Probabilidad de pérdida de orientación causal. Comparativa de los métodos ccKnn, ccMSVM, PC y MMHC, para grafos de 10 y 50 nodos. Los resultados de ccMSVM mejoran ligeramente los del algoritmo de referencia PC.

2.2. Inferencia causal para series temporales

Hasta este punto se ha desestimado el uso de la información temporal, tal y como se explicaba en el Apartado 1.1; únicamente se asumía como válida la condición de Markov. En esta sección se dejará de lado este marco teórico para explorar las posibilidades que ofrece otro criterio de búsqueda causal, la causalidad tipo Granger. Se procede ahora a presentar el método clásico de Granger; en esta tesis se presenta como aportación la aplicación del SVARMA (Rojo-Álvarez et al., 2004), un método de modelado autorregresivo con vectores soporte, a la inferencia causal; también se aplica una modificación del SVARMA para espacios multidimensionales, inspirado en (Pérez-Cruz et al., 2002; Sánchez-Férnandez et al., 2004). Finalmente, se compara con otros métodos que son el estado del arte en búsqueda causal para series temporales.

En la introducción, en el Apartado 1.7, se han presentado las ideas principales que dan sustento a la causalidad tipo Granger. El concepto fue esbozado por Norbert Wiener, y más tarde concretado por el economista Clive Granger (Granger, 1969). Dadas dos series temporales, si se desea conocer la posible influencia causal de una de ellas sobre la otra, se deben ajustar dos modelos autorregresivos. Si la varianza del error de predicción de una serie es mayor que la predicción que incluye valores de la otra serie, se puede afirmar que la segunda tiene influencia causal sobre la primera. El procedimiento se puede repetir para determinar la existencia de relación causal en el sentido inverso.

La formulación matemática de estas ideas se implementa mediante las siguientes ecuaciones de predicción ARMA, dadas las series temporales $x_1(t)$ y $x_2(t)$:

$$\begin{aligned}
x_1(t) &= \sum_{p=1}^{P} A_{11}(p)x_1(t-p) + \sum_{p=1}^{P} A_{12}(p)x_2(t-p) + e_{12}(t) \\
x_1(t) &= \sum_{p=1}^{P} B_1(p)x_1(t-p) + e_1(t) \\
x_2(t) &= \sum_{p=1}^{P} A_{21}(p)x_1(t-p) + \sum_{p=1}^{P} A_{22}(p)x_2(t-p) + e_{21}(t) \\
x_2(t) &= \sum_{p=1}^{P} B_1(p)x_2(t-p) + e_2(t) \\
t &= 1, \ldots, N
\end{aligned} \tag{2.21}$$

En forma vectorial las Ecuaciones (2.21) quedarían como:

$$\mathbf{x}(t) = \sum_{i=1}^{P} A(p)\mathbf{x}(t-p) + \mathbf{e_a}(t)$$
$$\mathbf{x}(t) = \sum_{i=1}^{P} B(p)\mathbf{x}(t-p) + \mathbf{e_b}(t)$$
$$t = 1, \ldots, N \tag{2.22}$$

donde se han agrupado las variables:

$$A(p) = \{A_{ij}\}_{i,j=1,\ldots,2}\,;\ \mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix};\ \mathbf{e}_a(t) = \begin{pmatrix} e_{12}(t) \\ e_{21}(t) \end{pmatrix} \tag{2.23}$$

Para el modelo puramente autorregresivo, la agrupación de variables queda como:

$$B = \begin{pmatrix} B_{11} & \mathbf{0} \\ \mathbf{0} & B_{22} \end{pmatrix};\ \mathbf{e}_b(t) = \begin{pmatrix} e_1(t) \\ e_2(t) \end{pmatrix} \tag{2.24}$$

A partir de estos modelos, hay varias formas de calcular el índice causal. En el presente trabajo, siguiendo (Schelter et al., 2006), se empleará la siguiente:

$$c_{j\to i} = \ln \frac{\sigma^2_{e_i}}{\sigma^2_{e_{ij}}} \tag{2.25}$$

donde $\sigma^2_{e_i}$ y $\sigma^2_{e_{ij}}$ son las varianzas de los respectivos errores de predicción. Por ejemplo, $c_{2\to 1}$ mide la influencia causal de $x_2(t)$ sobre $x_1(t)$; intuitivamente, se comprueba que si la segunda serie no tiene influencia en la primera, las varianzas de los ruidos e_1 y e_{12} serán similares, por lo que $c_{2\to 1}$ tenderá a 0.

A modo de ejemplo se muestra en la Figura 2.9 un conjunto de señales registradas en quirófano por un polígrafo durante un episodio de fibrilación auricular. Estas señales se recogen de manera síncrona por un catéter de geometría circular, posicionado en una de las venas pulmonares. La relación causal entre las señales puede dar una idea de cómo se propagan las ondas por la pared auricular.

En la Figura 2.10 se presenta el resultado de aplicar las Ecuaciones 2.22 y 2.25 a unas señales sintéticas que modelan las recogidas por el polígrafo. El índice causal $c_{i\to j}$ se ha calculado para un conjunto de ventanas deslizantes, con una superposición del 50 %. La señal $x_2(t)$ se ha construido como una versión retardada de $x_1(t)$ más ruido gausiano: $x_2(t) = x_1(t-\tau) + \eta(t)$.

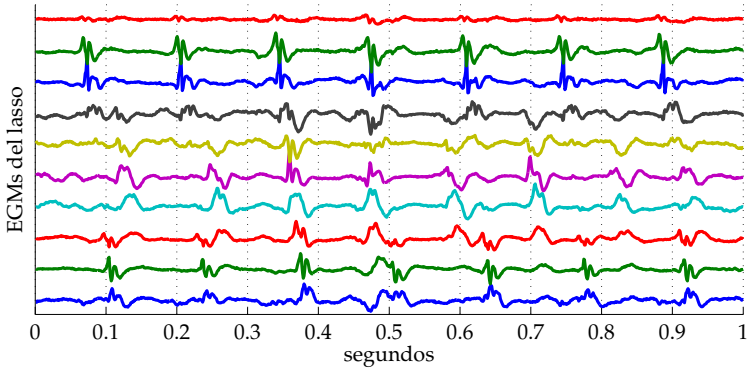


Figura 2.9: Ejemplo de señales síncronas registradas durante un episodio de fibrilación auricular.

Como era de esperar la influencia causal de la señal 2 sobre la 1, $c_{1\to 2}$, es despreciable, mientras que la inversa es mayor que cero durante todo el registro. También se presenta en la figura otro índice causal, éste calculado como $g = \frac{g_2 - g_1}{g_1 + g_2}$, con $g_1 = e_1 - e_{12}$ y $g_2 = e_2 - e_{21}$. Este índice tenderá a la unidad cuando la relación sea puramente $x_1(t) \to x_2(t)$, mientras que tenderá a -1 en caso contrario. La contrapartida del índice g es que no discrimina de manera explícita la posible relación causal cruzada entre los pares de señales, por lo que se emplea la medida definida en la Ecuación 2.25.

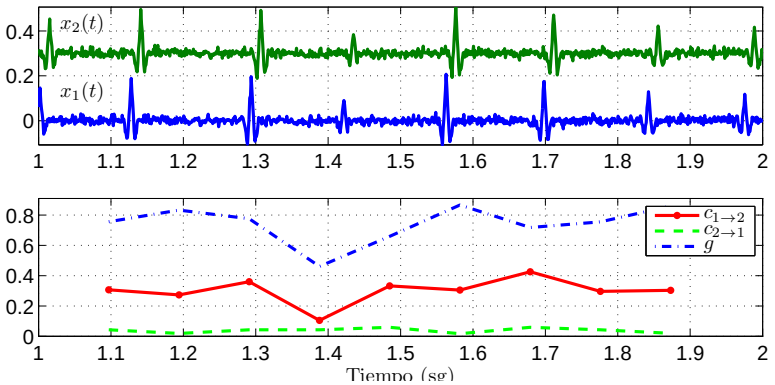


Figura 2.10: Señales de fibrilación sintéticas, generadas por el modelo $x_2(t) = x_1(t - \tau) + \eta(t)$. En la parte inferior se muestran los índices de causalidad Granger calculados según dos medidas distintas, $c_{i\to j}$ y g.

2.2.1. Representación dispersa y causalidad

A la hora de realizar cualquier modelado, es importante no sólo marcar el objetivo de mínimo error (de predicción o clasificación), sino también de controlar la complejidad del modelo. Como guía general, el conocido como principio de la regla de Occam aboga por, a igualdad de prestaciones, elegir siempre la opción o hipótesis más sencilla, que menos asunciones tome. En el campo del aprendizaje estadístico, para datos discretos, la dimensión VC, de Vapnik-Chervonenkis (Vapnik, 1998), es un ejemplo de control de la complejidad basado en el caso peor. A la hora de hacer modelado autorregresivo, el orden de las ecuaciones en diferencias, dado por el parámetro P de la Ecuación 2.22, se suele estimar mediante criterios como el AIC ("Akaike Information Criterion") o el BIC ("Bayesian Information Criterion") (Schelter et al., 2006). A mayor orden P, mayor complejidad del modelo, aunque también es habitual que el error de ajuste, cuantificado por la varianza de los residuos σ_e^2, disminuya. Alguna de las formulaciones que tratan de balancear complejidad y ajuste se muestra en las siguientes ecuaciones.

$$\begin{aligned} AIC(P) &= 2\ln(\sigma_e^2) + \frac{2d^2P}{N} \\ BIC(P) &= 2\ln(\sigma_e^2) + \frac{2d^2P\ln(N)}{N} \end{aligned} \tag{2.26}$$

donde d es la dimensión del espacio de entrada.

En este sentido, los algoritmos que buscan una representación dispersa están emergiendo como una interesante alternativa a los métodos actuales, con mejores prestaciones y un comportamiento más robusto (Zibulevsky y Elad, 2010).

Particularizando para el caso de relaciones causales, lo habitual es que los grafos sean relativamente dispersos. Sin embargo, los métodos tipo Granger que se apoyan en modelado ARMA tienden a generar soluciones de coeficientes densos. En la Ecuación 2.22, esto se traduciría en pocos elementos nulos en las matrices $A(p)$.

Para encontrar la solución del modelado ARMA, se procede a reformular ligeramente el problema. En su versión vectorial, se recuerda de nuevo las ecuaciones $\mathbf{x}(t) = \sum_{p=1}^{P} A(p)\mathbf{x}(t-p) + \mathbf{e}(t)$.

Aunque el desarrollo se hará para el caso bidimensional, las ecuaciones son válidas para problemas de dimensionalidad mayor, en los que $\mathbf{x}(t) \in \mathbb{R}^d$ y $A(p) \in \mathbb{R}^{d\times d}$. En desarrollos posteriores se empleará la siguiente notación:

$$A = (A(1), \ldots, A(P))^T \tag{2.27}$$

$$X = (X_1, \ldots, X_P) \tag{2.28}$$

$$Y = X_0 \tag{2.29}$$

$$X_p = \begin{pmatrix} \mathbf{x}(P+1-p)^T \\ \mathbf{x}(P+2-p)^T \\ \vdots \\ \mathbf{x}(T-p)^T \end{pmatrix} \overset{d=2}{=} \begin{pmatrix} x_1(P+1-p), x_1(P+2-p), \ldots, x_1(T-p) \\ x_2(P+1-p), x_2(P+2-p), \ldots, x_2(T-p) \end{pmatrix} \tag{2.30}$$

La solución de mínimos cuadrados (LS[1]) es la matriz que minimiza el siguiente funcional:

$$\hat{A}^{LS} = \arg\min_A ||Y - XA||^2 \tag{2.31}$$

En este contexto se entenderá que cuando se calcula una norma a una matriz, aquella aplica realmente sobre la versión vectorizada de ésta, de forma que todos los elementos se alinean en una única columna. La solución de este problema tiende a dar soluciones sobreajustadas debido al alto número de coeficientes del modelo, que asciende a d^2P parámetros en el caso multidimensional. Una manera de atenuar esta contrapartida es aplicar una función de coste a los coeficientes del modelo; cuando esta función es cuadrática, al estimador se le conoce como regresor tipo "Ridge":

$$\hat{A}^{ridge} = \arg\min_A ||Y - XA||^2 + \lambda||A||^2 \tag{2.32}$$

El regresor "Ridge" introduce un nuevo parámetro en (2.32), λ, que controla el balance entre la complejidad del modelo, menor a medida que $||A||^2$ disminuye, y la potencia de los errores, cuantificado en $||Y - XA||^2$.

El funcional recogido en (2.32), si bien fuerza coeficientes de norma cuadrática baja, no llegan a ser idénticamente nulos. Para provocar la deseada dispersidad del grafo causal, se podría emplear un test estadístico que mida la relevancia de dichos coeficientes, como se propone en (Haufe et al., 2008).

Sin embargo, resultaría más apropiado una estrategia que englobara las dos etapas, la de ajuste del modelo y la de búsqueda de la dispersidad, en un sólo algoritmo. Además, existe una relación entre la dispersidad en el modelado y la capacidad de generalización (Herbrich, 2002). El siguiente funcional, de norma cero, cumple con dicho objetivo:

[1]LS, siglas en inglés de "Least Squares".

$$\hat{A}^{l_0} = \arg\min_{A} ||A||_{l_0} \tag{2.33}$$

$$\text{restringido a } ||Y - XA|| \leq k \tag{2.34}$$

donde k es un factor de tolerancia o desviación del modelo respecto a la señal real. La norma l_0, aún no siendo propiamente una distancia, calcula la cardinalidad o número de elementos distintos de cero[2]: $||\mathbf{a}||_{l_0} = \sum_{i=1}^{d} |a_i|^0$. La solución de (2.33) es NP-completa, de difícil obtención. Se pueden emplear métodos numéricos para aproximarla, o bien relajar la norma l_0 por la l_1, para la que sí hay algoritmos con coste asumible. Esta sustitución da lugar al algoritmo conocido como LASSO[3] (Tibshirani, 1994; Nardi y Rinaldo, 2011), que se ha introducido en la Sección 1.7.1. A continuación se detalla el funcional del algoritmo:

$$\hat{A}^{lasso} = \arg\min_{A} ||Y - XA||^2 + \lambda ||A||_{l_1} \tag{2.35}$$

La diferencia con el funcional (2.33) es que ahora se aplica una norma $||\mathbf{a}||_{l_1} = \sum_{i=1}^{d} |a_i|$, que permite el ajuste del modelo de manera simultánea con un control de la complejidad, traducido en el número de coeficientes A_{ij}^{p} distintos de cero. Es posible resolver el LASSO mediante algoritmos rápidos. En concreto, en (Friedman et al., 2010) se propone una implementación flexible, que combina la solución "Ridge" y LASSO, que se conoce como "elastic net". El problema que resuelve es $||Y - XA||^2 + \lambda P_\alpha(A)$, donde $P_\alpha(A) = (1-\alpha)||A||_{l_2}^2 + \alpha||A||_{l_1}$. El término de penalización balancea el funcional "Ridge" ($\alpha = 0$) con el LASSO ($\alpha = 1$). Esta implementación, además de la flexibilidad, ofrece un buen rendimiento en carga computacional.

Una manera de refinar el LASSO aplicado a causalidad es agrupar los coeficientes del modelo según las series temporales a las que pertenecen. Esta idea está propuesta en (Yuan y Lin, 2006) y se aplica a causalidad en (Haufe et al., 2008); recibe el nombre de "Group LASSO".

$$\hat{A}^{Glasso} = \arg\min_{A} ||Y - XA||^2 \tag{2.36}$$

$$\text{restringido a } ||A_{11}(p), \ldots, A_{dd}(P)|| + \sum_{i \neq j} || \left(A_{ij}(1), \ldots, A_{ij}(P)\right) || \leq k$$

Esta aproximación asigna los coeficientes puramente autorregresivos a un grupo, y hace una suma de norma l_1 de las normas l_2 del resto de coeficientes.

[2] Se asume que $0^0 = 0$.

[3] "Least Absolute Selection and Shrinkage Operator"

En la Figura 2.11 puede verse cómo la constante k, que limita la suma l_1 de las normas l_2 de los coeficientes, permite obtener una solución dispersa en los coeficientes de la relación causal. En la parte superior, la gráfica muestra la norma de los coeficientes de la solución de mínimos cuadrados (LS) en línea punteada; la línea continua muestra la evolución de la norma de los coeficientes obtenidos con "Group LASSO". En la parte inferior de la Figura 2.11 se muestra la varianza del error de predicción, en línea punteada para mínimos cuadrados y línea continua para "Group LASSO". Puede apreciarse como para $k > 1.5$ los coeficientes que rigen la relación $x_2(t) \leftarrow x_1(t)$ (A_{12}) dejan de ser idénticamente nulos. El precio a pagar por mantener la solución sencilla, es que para este valor el error de predicción todavía no alcanza al de mínimos cuadrados. Sin embargo, cuando el objetivo es orientar la relación causal entre las series, interesa más un modelo disperso y adaptado al mecanismo real de generación que un error pequeño.

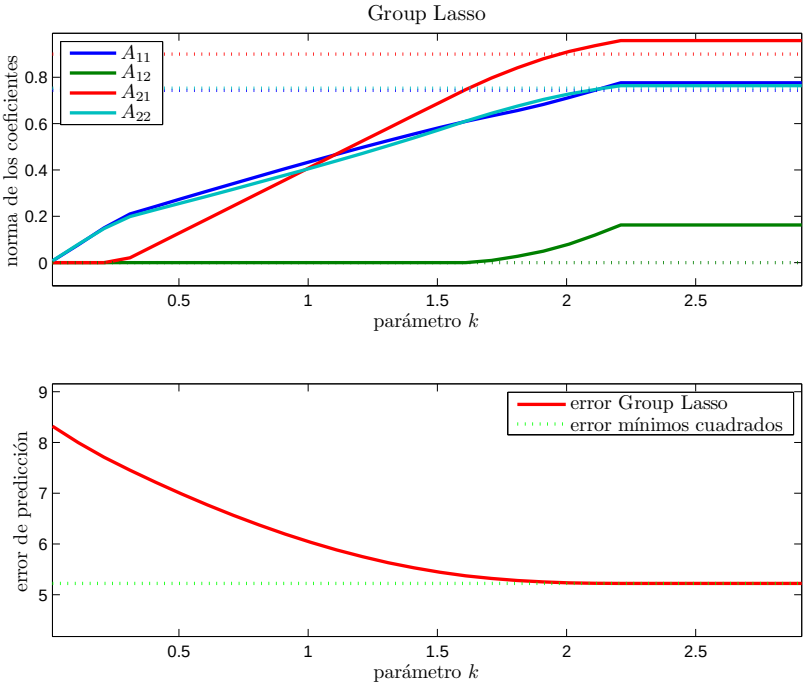


Figura 2.11: Norma de los coeficientes del modelo ARMA y error de predicción. Se compara con la solución de mínimos cuadrados.

Se va a estudiar un último algoritmo para desarrollar la causalidad de Granger: las máquinas de vectores relevantes (RVM por sus siglas inglesas

de "Relevance Vector Machine") (Tipping, 2001). La RVM es un modelo lineal en los parámetros, que obtiene soluciones dispersas en clasificación y regresión, empleando una aproximación bayesiana. De hecho, la RVM es un caso particular de un modelo bayesiano disperso, en el que se formula la solución al estilo de la máquina de vectores soporte. El modelo de una máquina de vectores relevantes es:

$$x_i(t) = \sum_{j=1}^{M} w_j K(\mathbf{x}_i(t), \mathbf{x}_j(t)) + w_0 \tag{2.37}$$

Los pesos $\mathbf{w}$ de la Ecuación 2.37 hacen las veces de los coeficientes A. Para hacer modelado ARMA, las funciones base $K(\cdot)$ de la Ecuación (2.37) son directamente el producto interior de las matrices $X^S \cdot (X^S)^T$, donde

$$X^S = \left(\begin{array}{c|c} X & \mathbf{0} \\ \hline \mathbf{0} & X \end{array}\right)$$

y donde se asume que X viene dado por (2.23). Se define una probabilidad a priori sobre los hiperparámetros $\mathbf{w}$ del modelo:

$$P(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{j=1}^{M} P(w_j|\alpha_j) \propto \prod_{j=1}^{M} \alpha_j^{1/2} \exp(-\frac{1}{2}\alpha_j w_j^2) \tag{2.38}$$

El método RVM, dados el modelo (2.37) y la asunción a priori de la Ecuación 2.38, calcula entonces el máximo de la función de verosimilitud, y devuelve los valores más probables de la distribución a posteriori de los hiperparámetros $\boldsymbol{\alpha}$. Este procedimiento se puede llevar a cabo computacionalmente con el paquete de software (Tipping, 2009).

Optimización convexa en procesado de señal

Aunque el problema "Group LASSO" formulado en la Ecuación 2.36 no es diferenciable, sí es convexo, y se puede optimizar como se describe en el siguiente apartado. En general, los métodos de optimización son ampliamente utilizados en el campo del procesado de señal (Boyd y Vandenberghe, 2004; Luo y Yu, 2006). Cuando esta optimización cae en la categoría de problemas convexos, existen varias formas de resolver estos problemas eficientemente.

Un problema convexo se define por la siguiente propiedad. Dados los puntos $x, y \in S$ del conjunto $S \in \mathbb{R}^d$, se dice que éste es convexo si $\lambda x + (1 - \lambda)y \in S$, $\forall \lambda \in [0, 1]$, es decir, si todos los puntos del segmento que los une pertenecen al conjunto S.

Un cono convexo $\mathcal{C}_t$ es un tipo especial de espacio convexo, que está acotado por un escalado positivo. Un cono de segundo grado se define como $\mathcal{C}_t = \{\mathbf{x}| \quad x_1 \geq ||\mathbf{x}||\}$. Si se aplica una transformación lineal a este espacio, se obtiene un cono rotado, que se llamará Rot$\mathcal{C}_t$ y se define como Rot$\mathcal{C}_t = \{\mathbf{x}| \quad 2x_1x_2 \geq ||\mathbf{x}||^2\}$.

Un ejemplo de conjunto convexo es la bola unitaria $S = \{\mathbf{x}| \ ||\mathbf{x}|| \leq 1\}$. Sin embargo, hay que notar que la hiperesfera de radio unidad, por ejemplo, no es convexo, puesto que los puntos interiores a la hiperesfera no pertenecen al conjunto.

Antes de definir un problema de minimización convexa, es necesario establecer qué son las funciones convexas. Se dice que una función $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^d$ es convexa si para cualesquier dos puntos $\mathbf{x}$, $\mathbf{y}$ en $\mathbb{R}^d$ se cumple que:

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}), \quad \forall\theta \in [0,1] \tag{2.39}$$

es decir, que la función siempre está por debajo del segmento que une $f(\mathbf{x})$ y $f(\mathbf{y})$.

Con estos elementos ya se puede definir en qué consiste un problema de optimización convexo. En concreto, el problema consiste en minimizar el funcional:

$$\begin{aligned} &\min_{\mathbf{x}} f_0(\mathbf{x}) \\ \text{restringido a } &f_i(\mathbf{x}) \leq 0 \\ &h_j(\mathbf{x}) = 0 \\ &\mathbf{x} \in S \end{aligned} \tag{2.40}$$

donde las funciones $f_i(\mathbf{x})$ deben ser convexas, $h_j(\mathbf{x})$ funciones afines[4], y el conjunto S un conjunto convexo. Si la solución $\mathbf{x}^*$ pertenece al conjunto S se dice que la minimización converge a una solución única.

Si un problema se puede formular como un funcional convexo, es posible aplicar técnicas de resolución eficientes. En la siguiente sección se detalla cómo resolver el problema del "Group LASSO" mediante programación convexa cónica. También es necesario recurrir a programación cónica para solucionar la versión multidimensional del SVARMA.

Obtención de la solución del "Group LASSO"

Para encontrar la solución del funcional (2.36) se ha de aplicar una técnica de minimización conocida como Programación Cónica de segundo grado

[4]Función afín es aquella que se expresa de la forma $\mathbf{a}_j^T\mathbf{x} + b_j$, con $\mathbf{a}_j \in \mathbb{R}^d$ y $b_j \in \mathbb{R}$.

(SOCP, por las siglas inglesas de "Second Order Cone Programming") (Roth y Fischer, 2008).

El funcional a minimizar se modifica para que las restricciones sean de tipo cónico. Los detalles de implementación pueden consultarse en el Apéndice A. El problema (2.36) queda como la minimización de:

$$\begin{aligned} \min_{A^S} \quad & ((X^S)^T Y^S) A^S + 2v + c_{fix} \\ \text{restringido a} \quad & \|\mathbf{q}\| \leq y_1,\ \|\mathbf{r}\| \leq y_2,\ \|\mathbf{s}\| \leq y_3 \\ & y_1 + y_2 + y_3 \leq k \\ & X^S A^S - \vec{t} = \mathbf{0} \\ & w = 1,\ v \geq 0 \\ & \|\vec{t}\|^2 \leq 2vw \end{aligned} \tag{2.41}$$

Funciones de coste de los métodos dispersos

Como resumen, en el Cuadro 2.6 se muestran los costes aplicados por cada método de regresión. El algoritmo RVM, al ser bayesiano, fuerza la dispersidad de la solución al imponer un a priori sobre los parámetros e hiperparámetros.

Coste aplicado a error y pesos según método

Método	Errores	Pesos
LS	l_2	-
Regresor "Ridge"	l_2	l_2
Norma cero	l_2	l_0
RVM	modelo gausiano	a priori disperso
LASSO	l_2	l_1
Group LASSO	l_2	l_1 de l_2
c(Multi-)SVARMA	Coste Huber y zona de ϵ-insensibilidad†	l_2 por grupos

Cuadro 2.6: Comparación de la función de error aplicada a los errores y pesos, según métodos. †Ver Figura 2.12.

A continuación se presenta la aplicación de dos técnicas de modelado ARMA, basado en máquinas de vectores soporte, a la orientación causal de relaciones entre señales.

2.2.2. Causalidad de Granger mediante vectores soporte

Para estimar el modelo ARMA a partir de series temporales, la técnica de mínimos cuadrados ofrece una solución sencilla. Si el modelo es:

$$\mathbf{x}(t) = \sum_{p=1}^{P} A(p)\mathbf{x}(t-p) + \mathbf{e}(t) \tag{2.42}$$

y se agrupan las variables tal que $A = \big(A(1), \ldots, A(P)\big)$, $X = \big(X_1, \ldots, X_P\big)$, $Y = X_0$ y $X_p = \big(\mathbf{x}(P+1-p), \ldots, \mathbf{x}(N-p)\big)^T$, la solución que minimiza la norma l_2, $||Y - XA||^2$, es:

$$\hat{A}^{LS} = (X^T X)^{-1} X^T Y \tag{2.43}$$

Sin embargo este método no da buenas soluciones cuando en las series de entrenamiento hay presencia de muestras fuera de rango, amén de que tiende a sobreajustar debido al alto número de parámetros del modelo.

Resulta beneficioso emplear otra aproximación, basada en la minimización de un funcional tipo máquina de vectores soporte (SVM) (Rojo-Álvarez et al., 2004). En este caso la dispersidad se consigue en el espacio de las muestras de entrenamiento, no directamente en los coeficientes. Sin embargo, el funcional minimiza una suma l_1 de normas l_2 de los coeficientes, por lo que, de manera automática, supone una agrupación de los coeficientes al estilo "Group LASSO".

La función de coste que los algoritmos de la familia SVM aplican sobre los residuos del modelado se conoce como función de Vapnik, o de ϵ-insensibilidad:

$$L_\epsilon(e) = \begin{cases} 0 & \text{si } |e| < \epsilon \\ |e| - \epsilon & \text{si } |e| \geq \epsilon \end{cases} \tag{2.44}$$

El modelado ARMA mediante vectores soporte trata cada dimensión del problema (2.42) de manera independiente. Resulta más cómodo separar la formulación vectorial. Para la primera dimensión, el problema queda como:

$$x_1(t) = \sum_{p=1}^{P} a_p x_1(t-p) + \sum_{p=1}^{P} b_p x_2(t-p) + e_1(t) \tag{2.45}$$

y equivalentemente, de forma vectorial, que hará más claro el desarrollo posterior, $x_1(t) = \mathbf{a}^T\mathbf{x}_{1t} + \mathbf{b}^T\mathbf{x}_{2t} + e_1(t)$, donde los vectores columna $\mathbf{x}_{it}$ son las P

muestras anteriores al instante t: $\mathbf{x}_{it} = \big(x_i(t-1), x_i(t-2), \ldots, x_i(t-P)\big)^T$. Adicionalmente, se llamará $\mathbf{x}_i = \big(x_i(P+1), x_i(P+2), \ldots, x_i(N)\big)^T$.

El funcional para el modelado autorregresivo con vectores soporte queda como:

$$\min_{\mathbf{a},\mathbf{b},\mathbf{e}} \ \frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right) + C\sum_{t=P+1}^{N} L_\epsilon(e_1(t)) \tag{2.46}$$

Se demuestra (Vapnik, 1998) que el anterior funcional es equivalente a la minimización de:

$$\min_{\mathbf{a},\mathbf{b},\boldsymbol{\xi}^{(*)}} \ \frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right) + C\sum_{t=P+1}^{N} (\xi_t + \xi_t^*) \tag{2.47}$$

restringido a

$$x_1(t) - (\mathbf{a}^T\mathbf{x}_{1t} + \mathbf{b}^T\mathbf{x}_{2t}) \leq \epsilon + \xi_t \tag{2.48}$$

$$-x_1(t) + (\mathbf{a}^T\mathbf{x}_{1t} + \mathbf{b}^T\mathbf{x}_{2t}) \leq \epsilon + \xi_t^* \tag{2.49}$$

$$\xi_t^{(*)} \geq 0 \tag{2.50}$$

donde ahora, siguiendo la notación de la Ecuación 2.22, la matriz A queda como

$$A(p) = \begin{pmatrix} a_p & b_p \\ c_p & d_p \end{pmatrix} \tag{2.51}$$

Los coeficientes $\mathbf{c}$ y $\mathbf{d}$ resolverían el modelado de $x_2(t)$ dados $x_1(t)$ y $x_2(t)$. En (2.47) se han introducido unas variables auxiliares positivas ξ_t, que, cuando son distintas de cero, permiten incumplir las condiciones de margen impuestas por (2.48) y (2.49).

Es habitual introducir un término adicional para regularizar la minimización. Adicionalmente, el nuevo término resuelve posibles problemas de convergencia. No se suele prestar atención a esto en la literatura, sin embargo, este cambio en realidad induce una nueva función de coste, que queda como:

$$\min_{\mathbf{a},\mathbf{b},\boldsymbol{\xi}^{(*)}} \ \frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right) + C\sum_{t\in I_2} (\xi_t + \xi_t^*) + \frac{1}{2\gamma}\sum_{t\in I_1}\left(\xi_t^2 + \xi_t^{*2}\right) \tag{2.52}$$

restringido a (2.48)-(2.50). Gráficamente esta relación puede verse en la Figura 2.12.

Para justificar el cambio que se ha introducido en la función de coste, se desarrolla a continuación el procedimiento de optimización de (2.47).

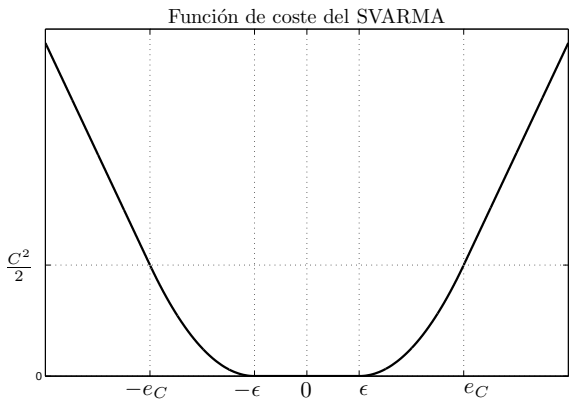


Figura 2.12: Función de Coste inducida por el funcional del SVARMA, donde se aprecian las tres zonas diferencias: ϵ-insensibilidad, zona cuadrática y zona lineal.

El funcional de Lagrange incorpora las restricciones mediante la introducción de unas nuevas variables a optimizar, $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$:

$$\begin{aligned}\min_{\mathbf{a},\mathbf{b},\boldsymbol{\xi}^{(*)},\boldsymbol{\alpha}^{(*)},\boldsymbol{\beta}^{(*)}} \quad & \frac{1}{2}\left(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2\right)+C\sum_{t=P+1}^{N}\left(\xi_t+\xi_t^*\right) \\ & -\left(\boldsymbol{\beta}^T\boldsymbol{\xi}+\boldsymbol{\beta}^{*T}\boldsymbol{\xi}^*\right) \\ & +\boldsymbol{\alpha}^T\left(\mathbf{x}_1-\left(\mathbf{a}^T\mathbf{x}_{1t}+\mathbf{b}^T\mathbf{x}_{2t}\right)-\epsilon-\boldsymbol{\xi}\right) \\ & +\boldsymbol{\alpha}^{*T}\left(-\mathbf{x}_1+\left(\mathbf{a}^T\mathbf{x}_{1t}+\mathbf{b}^T\mathbf{x}_{2t}\right)-\epsilon-\boldsymbol{\xi}^*\right)\end{aligned} \tag{2.53}$$

Las variables auxiliares del lagrangiano deben ser no negativas, $\boldsymbol{\alpha}^{(*)} \geq \mathbf{0}$, $\boldsymbol{\beta}^{(*)} \geq \mathbf{0}$ y las variables $\boldsymbol{\xi}^{(*)} \geq \mathbf{0}$, condición que se mantiene del funcional (2.48).

Para obtener el óptimo de la Ecuación (2.53) se determinan las derivadas parciales respecto a las variables:

$$\frac{\partial L_{PD}}{\partial \mathbf{a}} = 0 \tag{2.54}$$

$$\frac{\partial L_{PD}}{\partial \mathbf{b}} = 0 \tag{2.55}$$

$$\frac{\partial L_{PD}}{\partial \boldsymbol{\xi}^{(*)}} = 0 \tag{2.56}$$

De la Ecuación 2.56 se concluye que $\mathbf{0} \leq \boldsymbol{\alpha}^{(*)} \leq \mathbf{1}C$. De las Ecuaciones 2.54-2.55 se concluye que los coeficientes del modelo ARMA son una

combinación lineal de las series temporales:

$$\mathbf{a} = \left(\mathbf{x}_{1P}, \mathbf{x}_{1(P+1)}, \ldots, \mathbf{x}_{1(N-1)}\right)\left(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\right) \tag{2.57}$$

$$\mathbf{b} = \left(\mathbf{x}_{2P}, \mathbf{x}_{2(P+1)}, \ldots, \mathbf{x}_{2(N-1)}\right)\left(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\right) \tag{2.58}$$

Se realiza una última agrupación de variables, de forma que las funciones de autocorrelación son $R_1(t_1, t_2) = \mathbf{x}_{1t_1}^T \mathbf{x}_{1t_2}$ y $R_2(t_1, t_2) = \mathbf{x}_{2t_1}^T \mathbf{x}_{2t_2}$. Se agrupan estas autocorrelaciones en las matrices $\mathbf{R}_i = \{R_i(t_1, t_2)\}_{t_1, t_2 = P+1}^{N}$. Junto a (2.57), una vez introducidos en (2.53), da un funcional dual de la forma:

$$L_D = -\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T (\mathbf{R}_1 + \mathbf{R}_2)(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T \mathbf{x}_1 - \epsilon \mathbf{1}^T (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \tag{2.59}$$

Finalmente, el problema (2.59) se formula como un problema convexo cuadrático:

$$\min_{\mathbf{z}} L_D = \min_{\mathbf{z}} \mathbf{z}^T K \mathbf{z} + \mathbf{f}^T \mathbf{z} \tag{2.60}$$

donde se han agrupado las variables de forma que $\mathbf{z} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^* \end{pmatrix}$, $\mathbf{f} = \begin{pmatrix} \mathbf{x}_1 - \epsilon \\ -\mathbf{x}_1 - \epsilon \end{pmatrix}$ y

$$K = \begin{pmatrix} \mathbf{R}_2 + \mathbf{R}_1 & -\mathbf{R}_2 - \mathbf{R}_1 \\ -\mathbf{R}_2 - \mathbf{R}_1 & \mathbf{R}_2 + \mathbf{R}_1 \end{pmatrix} \tag{2.61}$$

La matriz K puede generar problemas a la hora de invertirla. Por esto es habitual emplear un truco numérico, que induce una regularización adicional. En concreto, para poder invertir la matriz se suma una pequeña cantidad γ a la diagonal. La matriz queda por tanto como $K' = K + \gamma I$.

Si en lugar de entender el valor γ como una operación necesaria para la convergencia del algoritmo, se asume que es un término nuevo del funcional a minimizar, el funcional crece con un nuevo término cuadrático:

$$\begin{aligned} L'_D = &-\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T (\mathbf{R}_1 + \mathbf{R}_2)(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \\ &+ (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T \mathbf{x}_1 - \epsilon \mathbf{1}^T (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \\ &- \frac{\gamma}{2}(\boldsymbol{\alpha}^T \mathbf{I} \boldsymbol{\alpha} + \boldsymbol{\alpha}^{*T} \mathbf{I} \boldsymbol{\alpha}^*) \end{aligned} \tag{2.62}$$

Ahora es posible volver al inicio del desarrollo, y hacer explícita la verdadera función de coste que se aplica a los residuos del modelado. Es una combinación de tres zonas: una, como el coste de Vapnik, donde no se penaliza los errores, de ancho ϵ. Después una zona cuadrática, que depende del

regularizador γ, óptima cuando el ruido de observación es gausiano. Finalmente, un coste lineal, que suaviza el efecto de muestras fuera de rango en las series de entrenamiento. Analíticamente, esta función queda como se expresa en la siguiente ecuación, y se muestra en la Figura 2.12.

$$L_\epsilon(e) = \begin{cases} 0 & \text{si } |e| < \epsilon \\ \frac{1}{2\gamma}(|e| - \epsilon)^2 & \text{si } \epsilon \leq |e| \leq e_C \quad \text{(zona } I_1\text{)} \\ C(|e| - \epsilon) - \frac{1}{2}\gamma C^2 & \text{si } |e| \geq e_C \quad \text{(zona } I_2\text{)} \end{cases} \tag{2.63}$$

donde $e_C = \epsilon + \gamma C$.

En el espacio primal, la inclusión del valor γ hace que el funcional a minimizar sea, en lugar de (2.47), el siguiente:

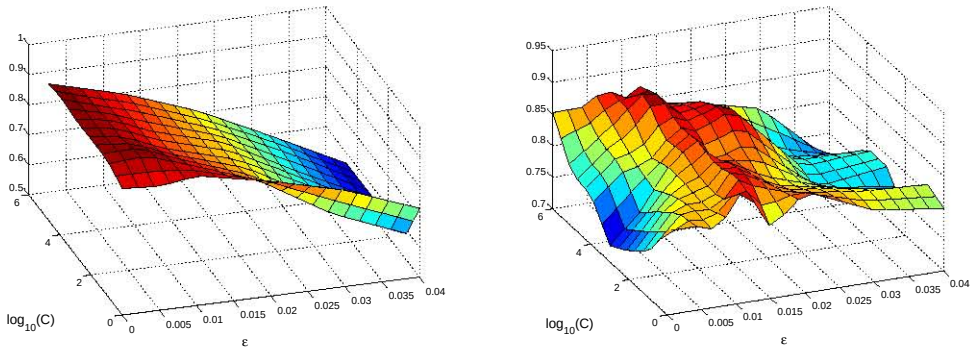
$$\min_{\mathbf{a},\mathbf{b},\boldsymbol{\xi}^{(*)}} \frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right) + C\sum_{t\in I_2}(\xi_t + \xi_t^*) + \frac{1}{2\gamma}\sum_{t\in I_1}\left(\xi_t^2 + \xi_t^{*2}\right) \tag{2.64}$$

restringido a (2.48)-(2.50).

Los hiperparámetros que controlan el funcionamiento de la máquina, y que regulan su capacidad de generalización son el coste C, la zona de insensibilidad ϵ y el factor de regularización γ, que controla el ancho de la zona cuadrática (ver Figura 2.12). Sin embargo, las variables de coste y regularización numérica están vinculadas, pues el ancho de la zona cuadrática I_1 viene dado por γC. Si se asigna un valor pequeño a γ, como suele ser habitual, por ejemplo de $\gamma = 10^{-6}$, habrá que compensarlo con un valor alto de C, para que la zona cuadrática sea significativa.

El efecto de los hiperparámetros puede comprobarse en la Figura 2.13. Las curvas representan el acierto ponderado a la hora de decidir la dirección causal. Dicha probabilidad de acierto se determina como la media de la sensibilidad y la especificidad. Es posible observar cómo el parámetro de insensibilidad ϵ no ayuda con ruido normal (Figura 2.13(a)), mientras que en el caso de ruido impulsivo los mejores resultados se obtienen en torno a la varianza del ruido impulsivo. Éste se ha generado siguiendo un proceso uniforme durante el 3 % de la duración de la señal. La influencia del parámetro C es poco relevante en un amplio rango de valores.

En la Figura 2.14 se muestra el error de predicción tras usar un modelado cSVARMA con $C = 5 \cdot 10^5$, $\gamma = 10^{-6}$ y $\epsilon = 0.5$. De acuerdo con (2.63), las zonas aplicarían en este caso en las zonas cuadrática $I_1 \equiv 0.5 \leq e \leq 1$ y lineal $I_2 \equiv e \geq 1$. El ruido impulsivo, con amplitud en torno a 2, queda en la zona lineal. El algoritmo así minimiza su efecto en la solución.



(a) Ruido normal.

(b) Ruido impulsivo.

Figura 2.13: Curvas de acierto ponderado para los hiperparámetros C y ϵ del cSVARMA, en caso de ruido normal e impulsivo.

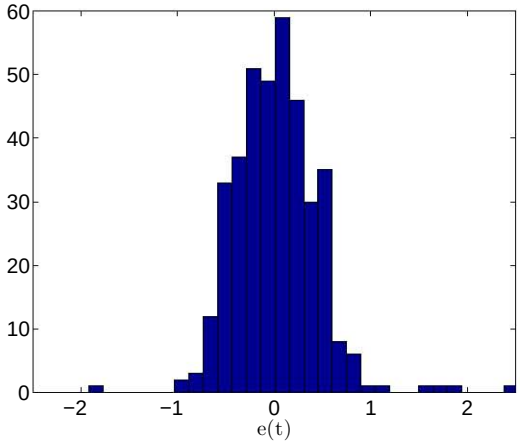


Figura 2.14: Histograma de los residuos de un modelado con cSVARMA. En esta simulación se han usado los parámetros $C = 5 \cdot 10^5$, $\gamma = 10^{-6}$ y $\epsilon = 0.5$.

El algoritmo de modelado cSVARMA, aplicado al problema de inferencia causal, presenta las ventajas de otros métodos que son el estado del arte, a la vez que aporta una flexibilidad adicional a la hora de configurar la función de coste y de definir los parámetros de la máquina. En los experimentos realizados, que se presentan al final del capítulo, se comprueba que el método es competitivo, e incluso mejora cuando las señales presentan ruido de carácter impulsivo.

A continuación se presenta una modificación al método cSVARMA, pa-

ra el caso multidimensional, inspirado en el trabajo de (Sánchez-Férnandez et al., 2004).

2.2.3. Causalidad de Granger multidimensional

Uno de los inconvenientes del cSVARMA es que ajusta de manera independiente los modelos para cada una de las componentes del problema, es decir, para las señales $x_1(t)$ y $x_2(t)$. Una manera de solucionar esta posible limitación es realizar una modificación al coste de los errores, de manera que se penalicen conjuntamente. Esta modificación, inspirada en (Sánchez-Férnandez et al., 2004), recibe el nombre de cMultiSVARMA, y el funcional a minimizar es:

$$\min_{\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d},\boldsymbol{\xi}} \ \frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + \|\mathbf{c}\|^2 + \|\mathbf{d}\|^2\right) + C\sum_{t=P+1}^{N}\xi_n \tag{2.65}$$

restringido a

$$\frac{1}{2}\|\mathbf{x}(t) - \sum_{i=1}^{P} A(p)\mathbf{x}(t-p)\|^2 \leq \epsilon + \xi_t \tag{2.66}$$

$$\boldsymbol{\xi} \geq \mathbf{0} \tag{2.67}$$

A diferencia del caso unidimensional del cSVARMA (2.63), la función de coste que se aplica a los errores de predicción es:

$$L_\epsilon(\mathbf{e}) = \begin{cases} 0 & \text{si } ||\mathbf{e}|| < \epsilon \\ (||\mathbf{e}|| - \epsilon)^2 & \text{si } ||\mathbf{e}|| \geq \epsilon \end{cases} \tag{2.68}$$

Con esta función, se penalizan conjuntamente los errores de predicción de las distintas componentes.

Este funcional se resuelve, similarmente al caso del "Group LASSO", mediante un problema de optimización cónica de segundo grado.

$$\begin{aligned} \text{mín} \quad & \sum_{i=1}^{d^2} v_i + C \sum_n \xi_n \\ \text{restringido a} \quad & ||a_i||^2 \leq 2 v_i w_i, \;\; w_i = 1, \;\; v_i \geq 0 \\ & \mathbf{X_n}^* A^S - \vec{t} = \mathbf{0} \\ & p_n - \mathbf{y}_n^T \mathbf{X}_n^* A^* - \xi_n \leq -\frac{1}{2} \mathbf{y}_n^T \mathbf{y}_n + \epsilon \\ & q_n = 1 \\ & \|\vec{t}\|^2 \leq 2 p_n q_n \\ & \big(i = 1, \ldots, d^2. \quad t = P+1, \ldots, N.\big) \end{aligned} \tag{2.69}$$

que, una vez adaptado a formas canónicas, se resuelve recurriendo al paquete de software "Mosek" (Mosek y ApS, 2010).

En la Figura 2.15 puede verse cómo se comportan los algoritmos cSVARMA y cMultiSVARMA en función del parámetro de insensibilidad ϵ. Aunque las gráficas son similares a las del "Group LASSO", (Figura 2.11), los coeficientes del modelo ARMA no llegan a ser idénticamente cero.

Cuando el parámetro ϵ es cero, se obtiene la solución de mínimos cuadrados. Para cSVARMA, este error no coincide con el de mínimos cuadrados pues cada componente de la señal $\mathbf{x}(t)$ se modela por separado. El acoplamiento de los coeficientes es mayor en cMultiSVARMA, como era de esperar al vincular con una norma l_2 los residuos de la predicción.

Cuando las observaciones se ven contaminadas por ruido gausiano, el "Group LASSO" ya no presenta el buen comportamiento que tiene sin ruido. Como se puede comprobar en la Figura 2.16, no es posible establecer un parámetro k que haga idénticamente cero los coeficientes A_{12}^p, sin afectar también a los otros parámetros del modelo. Rápidamente el algoritmo alcanza el error de mínimos cuadrados, a costa de perder la dispersidad en los coeficientes del modelo. En (Haufe et al., 2008) se sugiere usar la norma de los coeficientes para determinar la dirección causal. Sin embargo, la figura muestra un contraejemplo que invalida ese criterio. Parece más adecuado emplear (2.25) para determinar la dirección de flujo de información.

En el siguiente apartado se comparan los distintos métodos presentados, frente a ruido aditivo gausiano e impulsivo.

2.2.4. Resultados

Para comprobar el comportamiento de las diversas técnicas presentadas en este apartado, se han generado pares de secuencias sintéticas, añadiendo

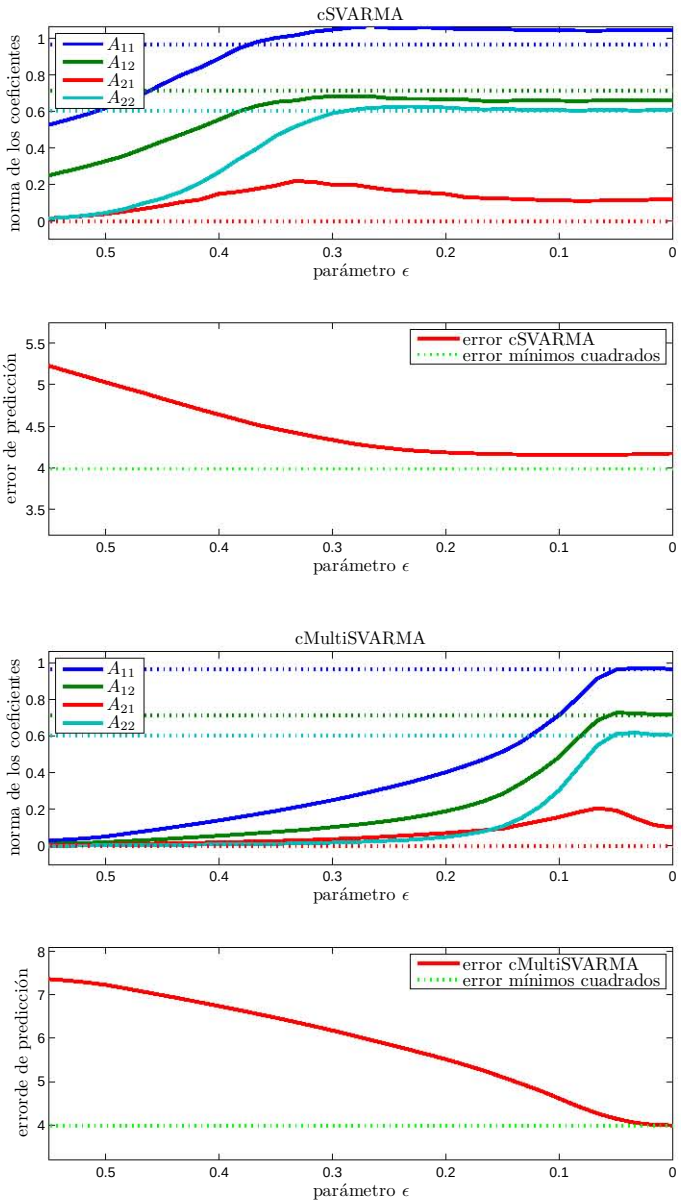


Figura 2.15: Norma de los coeficientes del modelo cMultiSVARMA y error de predicción. Se compara con la solución de mínimos cuadrados.

ruido gausiano.

Adicionalmente a los métodos que calculan un modelo autorregresivo, se ha simulado el método conocido como PSI ("Phase Slope Index") (Nolte

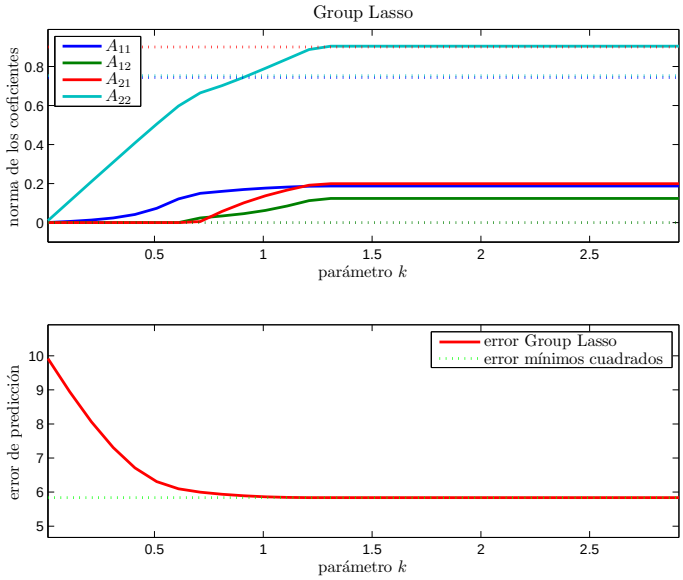


Figura 2.16: Norma de los coeficientes del "Group LASSO" con ruido aditivo.

et al., 2008, 2010). El índice PSI se calcula como la parte imaginaria del sumatorio:

$$\Phi_{ij} = \Im \sum_{f \in F} \left(C_{ij}^*(f) C_{ij}(f + \delta f) \right) \tag{2.70}$$

donde $C_{ij}(f) = \frac{S_{ij}(f)}{\sqrt{S_{ii}(f) S_{jj}(f)}}$ es la coherencia compleja, que se calcula a partir de la densidad espectral cruzada $S_{ij}(f)$[5]. Básicamente, este método calcula cuál de las series se ha producido antes, según se refleja en la fase del espectro.

En primer lugar se han generado un par de series temporales que presentan una influencia causal de la segunda componente hacia la primera. Con este fin, se define la matriz generadora $A(p) = \begin{pmatrix} A_{11}^p & A_{12}^p \\ 0 & A_{22}^p \end{pmatrix}$. Puesto que los coeficientes A_{21}^p son nulos $\forall p = 1, \ldots, P$, la señal $x_1(t)$ no tendrá influencia en la otra señal $x_2(t)$.

El ruido es independiente para cada canal; su matriz generadora es $N(p) = \begin{pmatrix} n_{11} & 0 \\ 0 & n_{22} \end{pmatrix}$. Se supondrá que los procesos generados de ruido y señal tienen el mismo orden P. En las simulaciones se ha fijado este valor a $P = 5$.

[5] $S_{ij}(f) = \sum_t \mathcal{F}(x_i(t)) \mathcal{F}^*(x_j(t))$

$$\boldsymbol{\eta}(t) = \sum_{p=1}^{P} N(p)\boldsymbol{\eta}(t-p) + \mathbf{e}_{\eta}(t) \tag{2.71}$$

$$\mathbf{x}(t) = \sum_{p=1}^{P} A(p)\mathbf{x}(t-p) + \mathbf{e}(t) \tag{2.72}$$

Las señales $\mathbf{e}_{\eta}(t)$ y $\mathbf{e}(t)$ se han generado como procesos estacionarios gaussianos, con matriz de covarianza la matriz identidad.

Finalmente, la secuencia de prueba consiste en un promediado aditivo entre señal y ruido controlado por el factor de ruido ρ:

$$\mathbf{y}(t) = (1-\rho)\mathbf{x}(t) + \rho B\boldsymbol{\eta}(t) \tag{2.73}$$

donde la matriz B se genera aleatoriamente. Las componentes $\mathbf{x}(t)$ y $B\boldsymbol{\eta}(t)$ se han normalizado de forma que tienen potencia unidad. Así las cosas, el nivel de señal a ruido vendrá dado por:

$$SNR_{dB} = 20\log_{10}\frac{1-\rho}{\rho} \tag{2.74}$$

Para comprobar las prestaciones de los métodos frente a ruido impulsivo, se ha generado una secuencia aleatoria uniforme, $U(-3,3)$, con un ciclo de trabajo del 3 %. En la Figura 2.17 se muestra una realización de la componente impulsiva del ruido. En estas pruebas, por lo tanto, la señal de prueba es

$$\mathbf{y}_{imp}(t) = (1-\rho)\mathbf{x}(t) + \rho B\boldsymbol{\eta}(t) + \boldsymbol{\eta}_{imp}(t) \tag{2.75}$$

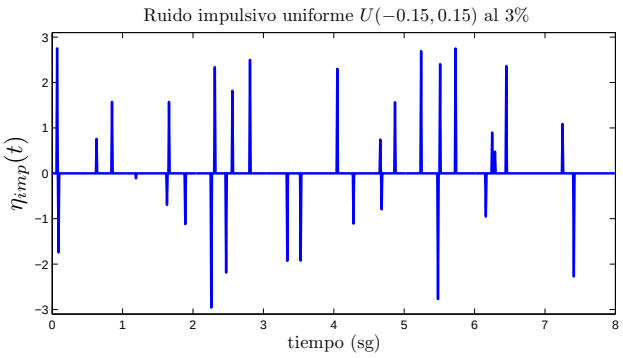


Figura 2.17: Realización de ruido impulsivo $\boldsymbol{\eta}_{imp}(t)$ uniforme, con un ciclo de trabajo del 3 %.

Se han comparado diversos métodos de modelado autorregresivo para calcular la causalidad de Granger. La Figura 2.18 sintetiza los resultados de las pruebas. La primera de las gráficas incluye ruido gausiano, y la situada debajo, el caso con ruido impulsivo.

En todos los modelos autorregresivos se ha empleado un orden de 10, doble que el empleado para la generación de las señales. Se han generado secuencias de 10 sg., muestreadas a $f_s = 100$ Hz. El índice de ruido se ha variado linealmente entre $\rho = 0.1$ y $\rho = 0.9$, dándole 10 valores. Para cada combinación, se han generado 250 réplicas. Se ha implementado un esquema de ventanas deslizantes, de 0.2 sg., solapadas al 50 %, como se mostraba en la Figura 2.10. Los modelos se entrenan con el primer tercio de cada ventana, y se calcula el error sobre la parte del segmento restante.

Para cada ventana se calcula el índice de Granger, y se promedia para la duración completa de la señal. Finalmente, se compara la decisión causal de los métodos con la verdadera orientación de las series. La Figura 2.18 muestra el acierto ponderado, calculado como el promedio entre sensibilidad y especificidad.

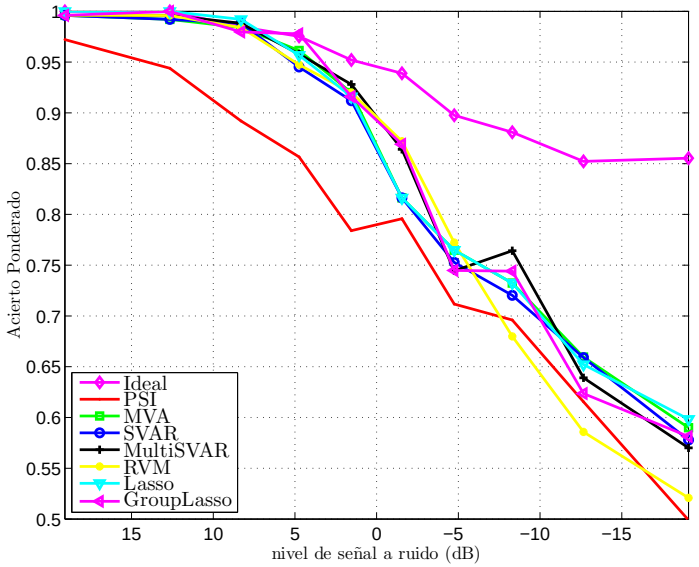
En el Cuadro 2.7 se muestra un subconjunto de medidas de las gráficas. En negrita se resalta el mejor resultado, excluyendo el ideal, para cada nivel de señal a ruido. Para ruido impulsivo, los métodos de vectores soporte son los mejores en todo el rango de señal.

	Ruido Normal (en dBs)				Ruido Impulsivo (en dBs)			
Método	19	8	2	-8	19	8	2	-8
Ideal†	100.0	98.5	95.2	88.1	100.0	99.6	97.6	92.9
PSI	97.2	89.2	78.4	69.6	72.9	68.5	61.6	53.1
MVA	99.6	98.4	92.0	73.2	96.0	91.6	83.2	60.4
SVARMA	99.6	98.8	91.2	72.0	**98.8**	95.2	**86.8**	64.8
MultiSVARMA	99.6	98.8	**92.8**	**76.4**	98.4	**96.4**	85.6	**66.4**
RVM	99.6	98.4	92.0	68.0	**98.8**	96.0	84.8	62.4
LASSO	**100.0**	**99.2**	91.6	73.2	97.2	94.4	83.6	61.6
Group LASSO	99.6	98.0	91.6	74.4	96.0	89.9	80.8	61.6

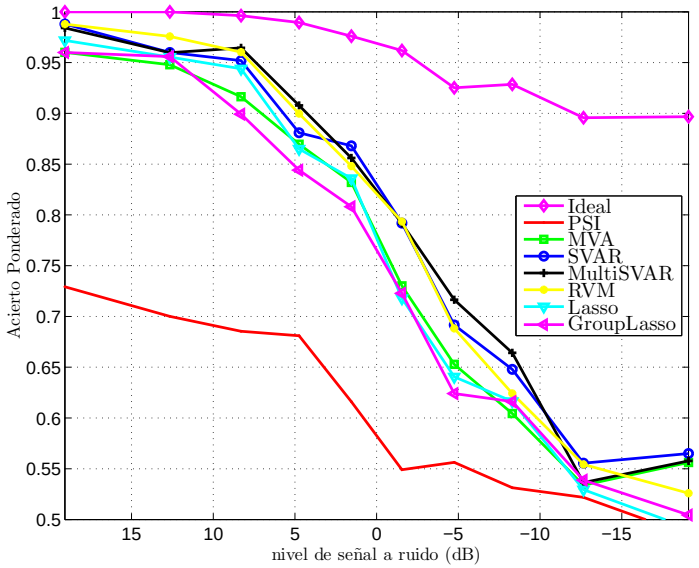
Cuadro 2.7: Acierto ponderado para diversos métodos y niveles de ruido. †Resultado suponiendo conocido el modelo ARMA auténtico.

Además de los métodos ya explicados, se ha incluido en las pruebas un modelado clásico de mínimos cuadrados por el método de Vieria-Morf[6], que aparece como el mejor en (Schlögl, 2006); una implementación computacional se encuentra en (Schlögl, 2005).

[6]En dicho artículo el método se nombra erróneamente como Nutall-Strand.



(a) Ruido normal



(b) Ruido impulsivo

Figura 2.18: Comparación de diversos métodos para valores decrecientes del nivel de señal a ruido (en dBs).

En las gráficas se comprueba que el método PSI es el peor para niveles de SNR altos, aunque cuando ruido y señal comienzan a ser comparables se coloca al nivel de los métodos Granger. Frente a ruido impulsivo, el índice PSI empeora aún más. Esto es normal, puesto que un impulso se traduce en una contaminación de todo el espectro. Se incluye en la comparativa, pues es el método que se usará en un capítulo posterior para el análisis de fibrilaciones cardíacas. A pesar de sus regulares prestaciones, es parecido a un criterio usado en la comunidad médica, basado en frecuencias dominantes.

En las figuras de 2.18 también se muestra el resultado cuando se suponen conocidos los coeficientes auténticos que han generado el modelo. Esta curva, indicada como "ideal" en las figuras, supone el límite teórico alcanzable. Se puede comprobar cómo, cuando el nivel de ruido comienza a ser comparable al de señal, queda mucho espacio para la mejora.

Si bien las curvas son relativamente similares, se observa un grupo de algoritmos que ofrecen mejores prestaciones. En concreto, al emplear RVM, cSVARMA y cMultiSVARMA frente a ruido impulsivo, se consiguen mejores resultados que con el resto.

Con el fin de mostrar cuantitativamente la mejora que ofrece emplear los métodos basados en vectores soporte, se detallan estos datos en el Cuadro 2.8. El cuadro recoge los datos en el caso de ruido impulsivo.

	cSVARMA	cMultiSVARMA
MVA	4.1 %	4.5 %
cSVARMA	-	0.4 %
cMultiSVARMA	-0.3 %	-
RVM	1.0 %	1.3 %
LASSO	5.2 %	5.6 %
"Group LASSO"	6.2 %	6.6 %

Cuadro 2.8: Porcentaje de mejora media, en acierto ponderado, de los métodos cSVARMA y cMultiSVARMA, con ruido impulsivo.

Se advierte que el método RVM, que comparte unos fundamentos similares a los métodos de vectores soporte, ofrece cifras bastante parecidas, en torno a un 1 % peor en probabilidad de acierto. Con respecto al método clásico de Granger, se mejora en más del 4 %. También hay que resaltar que la propiedad multidimensional del cMultiSVARMA permite mejorar, si bien ligeramente, el resultado del cSVARMA. No se incorporan al cuadro los valores del método PSI por ser muy grande su diferencia.

Capítulo 3

Causas de repetición de intentos de suicidio

En este capítulo de la tesis se aborda la utilización de técnicas de inferencia causal en un problema de psiquiatría, en concreto, para la obtención y jerarquización de los factores más relevantes a la hora de predecir la comisión de intentos de suicidio repetidos. Para el estudio se emplean varios métodos con el fin de identificar las variables más significativas. El resultado principal se ha obtenido mediante la aplicación de los métodos de inferencia causal presentados en esta tesis. El estudio clínico se encuentra publicado en (López-Castromán et al., 2011), en el que se han empleado métodos clásicos de búsqueda de relaciones causales. En el presente capítulo se exponen los principales resultados de dicho trabajo, así como se aplica al mismo y se discuten los resultados del método ccKnn, desarrollado en la Sección 2.1.1.

Los métodos de aprendizaje máquina se llevan estudiando desde hace varias décadas, y se han aplicado con éxito en diferentes ámbitos. Sin embargo, son varias las áreas científicas donde aún contemplan con recelo el empleo de métodos diferentes a los habituales: estadística clásica, indicadores del riesgo como los "odd ratio", etcétera. La ventaja de esta aproximación es su sencillez y la facilidad con la que se interpretan los resultados obtenidos. Por el contrario, el aprendizaje máquina ofrece resultados superiores, en muchos casos a costa de dificultar la interpretabilidad de los mismos. En diferentes comunidades, estas nuevas técnicas reciben diferentes nombres: minería de datos, inteligencia artificial o búsqueda de conocimiento en bases de datos. Pero todos ellos describen una realidad común: el tratamiento automático de amplias colecciones de datos para buscar conocimiento, realizando tareas como la clasificación, predicción o segmentado. En el campo de la psiquiatría son cada vez mayores y mejores las bases de datos de las que se dispone: más homogéneas, cubriendo periodos de tiempo mayores, más relacionadas con

otras bases de datos.

El aprendizaje máquina permite, a partir de una base de datos y sin apenas asunciones, aprender y extraer conocimiento de la misma; en concreto, una ventaja es la posibilidad de extraer nuevas hipótesis, que posteriormente pueden dar lugar a experimentos y trabajos que las contrasten. En áreas como la genética, donde las bases de datos son de una magnitud considerable, este tipo de herramientas se hace incluso más importante (Oquendo et al., 2012). Sin embargo, cualquier disciplina puede beneficiarse de herramientas como el aprendizaje máquina o la inferencia causal.

Los métodos máquina mejoran sustancialmente a los algoritmos clásicos. Por ejemplo, en (Baca-García et al., 2006) se mejora una regresión logística que da un 81 % de sensibilidad y un 91 % de especifidad a la hora de decidir la hospitalización de un paciente que ha cometido un intento de suicidio (IS). Empleando una máquina de vectores soporte se alcanza un acierto de generalización del 99 % y 100 % en sensibilidad y especificidad, respectivamente. En otro estudio, (Baca-García et al., 2007a), se emplean "Random Forest"[1] para identificar las variables asociadas con IS en la historia familiar. En (Baca-García et al., 2010) se presenta un estudio genético, también con "Random Forest" de SNPs[2] indicadores asociados con IS; con tres SNPs de tres genes se obtuvo una probabilidad de acierto del 67 % a la hora de clasificar personas con intentos de suicidio. Finalmente, también se han aplicado técnicas de corte más clásico, como son los modelos de Markov, para estudiar la estabilidad en el diagnóstico del trastorno bipolar (Baca-García et al., 2007b).

Aunque el aprendizaje máquina resulta prometedor en este tipo de estudios, especialmente en cuanto a la capacidad predictiva de los modelos que genera, puede adolecer de falta de transparencia o interpretabilidad. Los métodos que proveen de un modelo causal superan este obstáculo.

En este capítulo se presenta el problema médico de los intentos de suicidio (IS) repetidos, y la base de datos con la que se ha trabajado. Se procede a obtener un mapa causal con las variables de la base de datos, especialmente con aquellas que contienen más información relevante acerca del número de intentos de suicidio. Para ello se van a aplicar las herramientas descritas en las secciones previas, de aprendizaje máquina e inferencia causal. Para la obtención del mapa causal sólo se tendrán en cuenta los datos, sin más consideraciones a priori.

[1] Una técnica de clasificación y selección de variables basada en el remuestreo de árboles de clasificadores.

[2] "Single Nucleotide Polymorphisms", variación en la secuencia de ADN que afecta a una sola base o nucleótido, y que puede estar presente en un pequeño porcentaje de la población.

La búsqueda directa del mapa causal no resulta satisfactoria debido a lo reducido de la relación número de muestras / número de variables. Se afronta el problema en dos etapas; primero se reduce el conjunto de variables a aquellas que son fuertemente relevantes, las cuales incluirán el "Markov Blanket" de la variable de interés. Posteriormente se buscará el árbol causal de dichas variables. Se han utilizado dos métodos, el MMHC (Tsamardinos et al., 2006) y el ccKnn (de-Prado-Cumplido y Artés-Rodríguez, 2008). El capítulo concluye con la interpretación de los resultados, tanto clínicos como de los métodos empleados.

3.1. Intentos de suicidio repetidos

El número de muertes por suicidio a nivel internacional asciende a casi un millón al año, según datos de la Organización Mundial de la Salud (World Health Organization, 2009). Sin embargo el número de intentos de suicidio se encuentra en un margen de entre 10 y 20 veces el número de intentos consumados. En un porcentaje muy elevado de los casos, un intento fallido de suicidio conduce a un nuevo intento.

En general, los afectados por enfermedades mentales suponen un grupo poblacional relativamente numeroso. Los costes asociados a los tratamientos suponen una carga económica muy importante para los Estados. Desde un punto de vista sanitario, pero también económico, profundizar en la comprensión de estos problemas y articular medidas para la prevención y tratamiento de estas enfermedades resulta prioritario.

Si bien es conocido el hecho de que los intentos de suicidio (IS) frustrados conducen a otros nuevos, no existen unas guías aceptadas por la comunidad de cómo prevenir estos casos. Se conocen algunas de las variables sociodemográficas (edad, estado civil, nivel educativo) y clínicas (salud física, letalidad de intentos previos de suicidio, comorbilidad psiquiátrica), pero no la relación entre ellas y con la repetición de intentos.

3.1.1. Base de datos y características

Para tratar de aclarar las preguntas abiertas acerca de los motivos y las variables más influyentes en la repetición de IS, se cuenta con una base de datos que recoge 3347 pacientes (muestras) con 140 variables de tipo clínico y genético.

Los pacientes provienen del servicio de urgencias de dos hospitales, el Hospital Universitario Ramón y Cajal, en Madrid, y el Hospital Universitario Lapeyronie, en Montpellier. La recogida de datos abarca el periodo entre los

años 1994 y 2006. Las zonas de cobertura de los hospitales abarcan una población de medio millón y de 400 mil personas, en Madrid y Montpellier respectivamente. Se contó con el consentimiento de los pacientes y con la autorización por parte de sendos comités de ética.

El listado completo de variables se recoge en el Apéndice B. Sin embargo, se han suprimido ciertas variables en el estudio, hasta reducir la base de datos a únicamente 44. En el Cuadro B.1 del apéndice se han resaltado en cursiva las variables que han entrado en el estudio. Todas las variables de genética se han descartado, pues los registros de los pacientes franceses son escasos. La base de datos incluía dos grupos de control, uno de pacientes sin IS y otro de donantes de sangre. Estos registros serían útiles en otro tipo de estudio, pero no resultan informativos para la generación de un modelo causal de IS. Finalmente, se han eliminado aquellas variables que son propias de un solo sexo.

Se supondrá que los datos se almacenan en las $N = 1349$ muestras $(\mathbf{x}_i, y_i)$, con $i = 1, \ldots, N$, $\mathbf{x}_i \in \mathbb{R}^{43}$ e $y_i \in \{0, 1\}$. La etiqueta de las muestras, el número de intentos de suicidio, se ha dividido en dos grupos, de forma que la clase 0 hace referencia a pacientes con 2 o menos IS y la clase 1 a los que han cometido más de 2 intentos. El problema queda balanceado, pues las clases tienen un 54.3 % y un 45.7 % de pacientes para los grupos $y_i = 0$ e $y_i = 1$, respectivamente.

El histograma del número de IS se recoge en la Figura 3.1. Aproximadamente un 90 % de los pacientes ha cometido menos de 6 IS. La media se sitúa en 3.3 intentos por paciente, aunque algunos casos atípicos llegan a 50.

El conjunto de variables que se han empleado se muestra en el Cuadro 3.1, junto a una breve descripción de su significado. El estudio engloba factores socio-demográficos: edad, género, profesión, estado laboral y civil, hijos y nivel educativo, edad en el primer intento, historia familiar de IS y violencia en el intento (variables 28 a 34).

También se recogen diagnósticos psiquiátricos: trastornos de la personalidad, trastornos de depresión bipolar, ansiedad, trastorno obsesivo-compulsivo, consumo de drogas, psicosis y trastornos de la conducta alimentaria (variables 37 a 43).

Finalmente, un grupo de variables caracterizan los intentos de suicidio: edad en el primer intento, si el método empleado fue o no violento y, como se ha indicado, la variable principal de intentos de suicidio, que separa a los pacientes con 2 o menos intentos (#IS= 0) de los reincidentes múltiples (#IS> 2).

Nº var.	Etiqueta	Descripción de la variable
1	his_fam_s	Historial familiar de comportamiento suicida

Nº var.	Etiqueta	Descripción de la variable
2	age	Edad
3	ri_agent	Agente empleado†
4	ri_conci	Estado disminuido de consciencia†
5	ri_lesio	Heridas y toxicidad†
6	ri_rever	Reversibilidad†
7	ri_tto	Tratamiento necesario†
8	re_ubica	Lugar del IS‡
9	re_perso	Persona que avisa del rescate‡
10	re_descu	Probabilidad de ser detenido‡
11	re_acces	Accesibilidad al rescate‡
12	re_retra	Tiempo hasta ser descubierto‡
13	aislamie	Aislamiento⋆
14	tiempo	Tiempo⋆
15	precauci	Precauciones frente descubrimiento o intervención⋆
16	ayuda	Actos para solicitar ayuda durante o tras el intento⋆
17	acto_fin	Últimos actos anticipando la muerte (seguros, regalos, ...)⋆
18	prepa_in	Preparación activa del intento⋆
19	nota	Nota de suicidio⋆
20	comunic	Puesta en conocimiento de la intención antes del intento⋆
21	propo_ii	Propósito del intento⋆
22	expect	Expectativa de la letalidad del intento⋆
23	letal	Conocimiento de la letalidad del método⋆
24	seriedad	Gravedad del intento⋆
25	actit_mu	Actitud sobre la vida y la muerte⋆
26	inter_me	Opiniones acerca de las intervenciones médicas⋆
27	predem	Grado de premeditación⋆
28	est_civ	Estado civil
29	hijos	Número de hijos
30	hijos1	Tiene o no hijos
31	tabaco	Consumo de tabaco
32	niv_edu	Nivel educativo
33	prof	Profesión
34	sit_lab	Situación laboral
35	age1a	Edad durante el primer intento
36	vio	Intento violento de suicidio

Nº var.	Etiqueta	Descripción de la variable
37	dd_mood	Trastornos depresivos
38	depre_bip	Trastorno bipolar o depresivo
39	dd_anxiety	Ansiedad
40	dd_ocd	Trastorno obsesivo-compulsivo
41	dd_al_drug	Consumo de alcohol o drogas
42	dd_psychosis	Trastorno psicótico
43	dd_eating	Trastorno alimentario
44	IS	Número de intentos de suicidio

Cuadro 3.1: Listado de variables del problema psiquiátrico. †Factores de riesgo de Weisman y Worden. ‡Factores de rescate de Weisman y Worden. ⋆Escala Beck.

La escala de intencionalidad suicida de Beck (Beck et al., 1974) es un test de 15 entradas puntuables, que provee un indicador de la gravedad del intento de suicidio (variables 3 a 12). La escala de riesgo-rescate de Weissman y Worden (Weissman y Worden, 1974) es un cuestionario de 10 preguntas acerca de la letalidad y seriedad del intento (variables 13 a 27); intenta medir el grado de riesgo que asume el paciente y la probabilidad de ser rescatado durante el IS.

3.2. Metodología causal de los intentos de suicidio

Se ha enfrentado el problema de búsqueda causal de manera progresiva; en primer lugar se han buscado las variables más significativas, aquellas con más posibilidades de formar parte del "Markov Blanket" (MB), para después obtener un árbol causal con los métodos de inferencia causal. El crecimiento super-exponencial del número de grafos acíclicos con el número de variables, como se detalla en la Sección 1.3.1, hace inviable la obtención del árbol causal completo.

3.2.1. Preprocesado

Algunas de las variables sólo pueden tener valor en ciertos pacientes, como por ejemplo en suicidas y mujeres. Únicamente se han cogido para cada prueba las variables presentes en todos los pacientes.

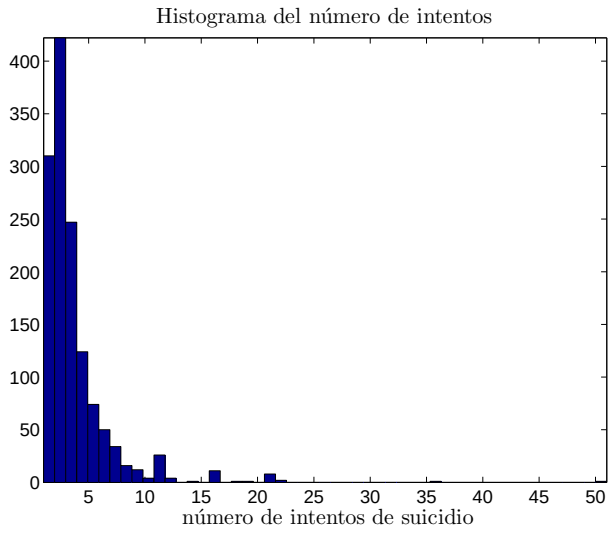


Figura 3.1: Histograma del número de intentos de suicidio.

La base de datos presentaba huecos en los registros, es decir, variables en las que no se ha registrado su valor. Se ha empleado un doble criterio para rellenar estos valores perdidos. En primer lugar, se ha purgado de la base de datos aquellos pacientes con un porcentaje de valores perdidos superior al 10 %.

Por otro lado, cuando se han empleado métodos causales, se ha incluido un nuevo estado más. En los métodos que requieren conocer un valor concreto, se ha realizado una imputación sustituyendo el valor perdido por la moda de la variable.

3.2.2. Selección de variables más relevantes

Se han ensayado varios métodos para seleccionar las variables más relevantes. Fisher, Kolmogorov-Smirnov y Búsqueda hacia delante ("Forward Selection").

El discriminante de Fisher es un mecanismo de puntuación de las variables basado en un criterio de separación lineal. Para cada variable, se calcula la siguiente medida:

$$\phi_i = \frac{|\mu_{i+} - \mu_{i-}|}{\sigma_{i+}^2 + \sigma_{i-}^2} \tag{3.1}$$

donde μ_{i+} es la media y σ_{i+}^2 la varianza de las muestras de la variable i que pertenecen la clase $y_i = 1$; análogamente, μ_{i-} y σ_{i-}^2 son media y varianza para

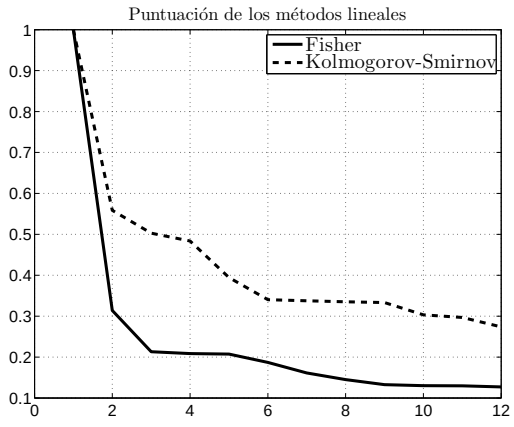
la clase $y_i = 0$. El discriminante de Fisher puntúa más alto aquellas variables cuyas medias ponderadas se encuentran más alejadas para cada clase.

El criterio de Kolmogorov-Smirnov ordena las variables en función del supremo de la distancia entre las funciones de distribución empíricas:

$$\kappa_i = \sqrt{N} \sup \left(\hat{F}(X \leq x_i) - \hat{F}(X \leq x_i | y_i = 1) \right) \tag{3.2}$$

Estos métodos valoran cada variable de manera independiente del resto. Esto puede ocasionar que aparezcan en los primeros lugares variables relevantes, pero que sean muy colineales. Para evitar este inconveniente, se ha empleando una búsqueda hacia delante con máquinas de vectores soporte.

El listado de variables, así como la puntuación normalizada de cada método, se muestra en el Cuadro 3.2 para las 12 primeras variables. En la gráfica se comprueba que a partir de la cuarta variable, todas tienen una puntuación similar. Ambos métodos coinciden en un alto porcentaje en las variables más importantes.



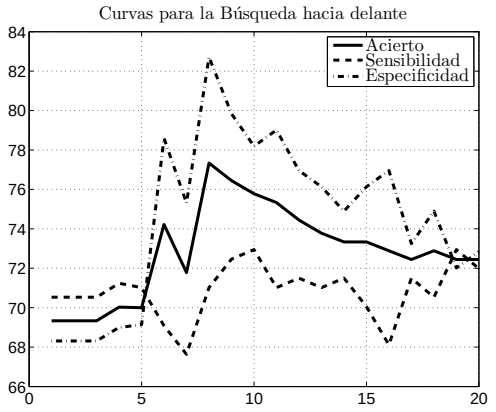
Fisher	K-S
dd_anxiety	dd_anxiety
ri_conci	ri_conci
ri_tto	ri_tto
age1a	niv_edu
depre_bip	dd_al_drug
dd_al_drug	actit_mu
inter_me	comunic
tabaco	prof
actit_mu	inter_me
comunic	tabaco
ri_rever	ri_rever
niv_edu	ri_lesio
ri_lesio	age1a

Cuadro 3.2: Selección de variables de los métodos lineales. Listado de las mejores variables, y puntuación normalizada.

La búsqueda hacia delante y el cálculo del "Markov Blanket" (MB) se muestran en el Cuadro 3.3. El método de búsqueda incremental es un algoritmo iterativo que elige en cada paso la variable que más incrementa el promedio entre sensibilidad y especificidad. El clasificador empleado ha sido una máquina de vectores soporte, con el esquema de entrenamiento habitual por validación cruzada. Las variables que sean relevantes, pero muy correlacionadas con las ya elegidas, no se tendrán en cuenta. Este es el motivo

por el que la búsqueda incremental escoge variables distintas de los otros métodos. En la gráfica del Cuadro 3.3 también se comprueba que ofrece las mejores probabilidades de acierto; alcanza un acierto promedio del 77 % con ocho variables.

El MB de la variable de IS múltiples engloba las variables que independizan IS del resto de información de la base de datos (Ver Sección 1.4.2). El método causal selecciona directamente el número de variables óptimo, sin necesidad de establecer un umbral como ocurre en el resto de métodos. Junto con el listado de variables del MB, se muestra la probabilidad de acierto de una máquina entrenada con las variables de manera incremental. Puesto que el MB no impone ningún orden a sus componentes, la ordenación se ha realizado según las prestaciones del clasificador al que da lugar.



BhD	MB	Ac
dd_anxiety	dd_anxiety	67.6
his_fam_s	est_civ	70.0
age	age1a	70.2
tabaco	dd_al_drug	70.7
ri_conci	predem	72.2
age1a	re_acces	72.4
prof	niv_edu	72.4
letal	ri_tto	72.9
depre_bip	expect	74.9
dd_psychosis	age	75.8
re_retra		
re_ubica		
re_acces		
inter_me		
predem		

Cuadro 3.3: Variables más relevantes de IS múltiples. Método de Búsqueda hacia delante (BhD) y "Markov Blanket" (MB). En la tabla del MB se indican también las probabilidades de acierto (Ac) en porcentaje. El MB ha sido obtenido mediante MMHC.

3.2.3. Generación de la red bayesiana causal

Una vez obtenido el "Markov Blanket" (MB) de la variable de IS múltiples, se ha generado el árbol causal. Sin embargo, las variables *expect*, *re_acces*, *est_civ* y *predem* se han eliminado, pues o bien estaban fuera del árbol causal o bien eran descendientes en segundo grado (efectos de efectos de IS). Estas

variables han sido incluidas en el MB como falsas alarmas por el método MMHC.

Con tal fin, se han empleado dos algoritmos de inferencia causal: el MMHC (Tsamardinos et al., 2006) y el ccKnn (de-Prado-Cumplido y Artés-Rodríguez, 2008). Se discute en primer lugar la optimización del MMHC, que ha sido el método empleado en (López-Castromán et al., 2011) para inferir el grafo causal.

Para optimizar el MMHC se ha variado el parámetro del peso Dirichlet. El modelo elegido es el que ofrece una mayor verosimilitud. La curva se muestra en la Figura 3.2.

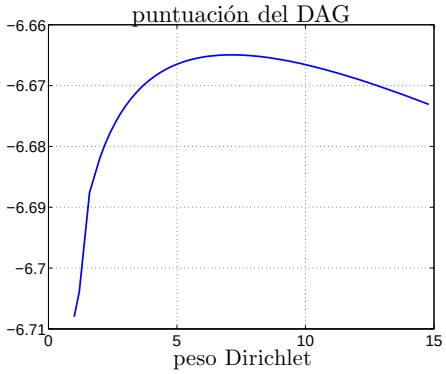


Figura 3.2: Puntuación de los grafos en función del peso Dirichlet del BDEu.

El árbol causal completo para la variable de IS repetidos se puede ver en la Figura 3.3. Las relaciones obtenidas resultan bastante razonables en general. El enlace entre las variables *edad*→*nivel educativo* puede ser el que menos sentido causal aporte; es comprensible que personas muy jóvenes o muy mayores no tengan acceso a una titulación superior; más que una relación causa / efecto, parece que entre ambas variables hay un alto nivel de correlación. Semejante argumento se puede emplear con la relación entre *edad*→*edad del* 1^{er} *intento*.

El resto de relaciones, por el contrario, tienen una interpretación causal muy interesante. Los padres de la variable IS múltiples son la edad, la edad durante el primer intento y los trastornos de ansiedad. Las variables de edad se han cuantificado en tres segmentos: menores de 35 años, entre 35 y 64 y, finalmente, mayores de 65 años. Los descendientes de IS son el consumo de drogas o alcohol y la variable de tratamiento requerido, uno de los factores de riesgo en el test de Weissman y Worden, que tiene por valor esta terna: primeros auxilios o emergencias, ingreso hospitalario y Unidad de Cuidados Intensivos (UCI). El nivel educativo es un ascendiente de los trastornos de

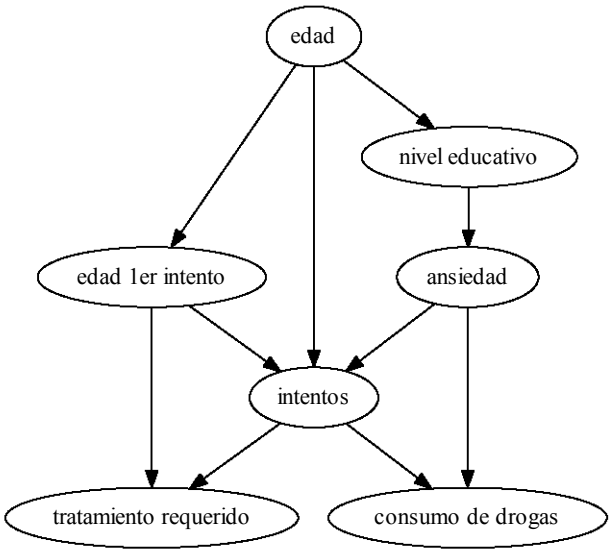


Figura 3.3: Modelo gráfico causal de las variables relacionadas con intentos de suicidio repetidos, calculado con MMHC.

ansiedad, y está dividido entre analfabetos, personas con el graduado escolar, con formación profesional (FP), con bachillerato y con estudios universitarios.

Las relaciones entre estas variables y la de IS múltiples han aparecido en diversos artículos de la literatura médica. Antes de explicar con más detalle el modelo causal, se procede a calcular las probabilidades condicionales de los nodos del grafo.

Modelo causal con ccKnn

Se ha probado a emplear el método de inferencia causal *ccKnn* desarrollado en la Sección 2.1.1. Es preciso notar que la primera fase del algoritmo, la búsqueda del esqueleto del grafo, es la misma que en el MMHC. Sin embargo, este último algoritmo orienta los enlaces maximizando la verosimilitud de los datos.

El método *ccKnn*, por el contrario, busca estructuras en "V" y orienta los enlaces para no crear nuevas colisiones ni ciclos en el grafo. El árbol causal de la Figura 3.3 es denso, por lo que hay numerosos grupos de tres nodos completamente conectados. Esto fuerza a que el número posible de colisiones sea bajo; en concreto, sólo un 60 % de los posibles caminos de tres nodos son colisiones potenciales. Más aún, en el modelo causal generado por MMHC sólo se identifican dos colisiones, a saber, las formadas por *edad* 1^{er} *intento* $\rightarrow$ *intentos* $\leftarrow$ *ansiedad* y *edad* $\rightarrow$ *intentos* $\leftarrow$ *ansiedad*.

En este contexto con pocas colisiones es difícil para el *ccKnn* orientar los enlaces. El resultado es el mostrado en la Figura 3.4. El algoritmo discrepa con el MMHC en una colisión, al identificar como tal las variables *edad* → *intentos* ← *tratamiento requerido*, si bien ha coincidido en las dos estructuras en "V" presentes en el grafo de la Figura 3.3.

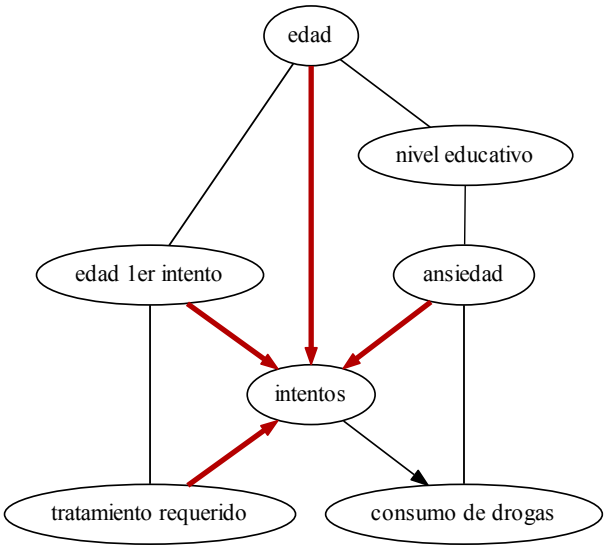


Figura 3.4: Modelo causal con ccKnn. En rojo, y trazo de mayor grosor, se indica los enlaces identificados mediante estructuras en "V". El enlace *intentos*→*consumo* se orienta para generar nuevas colisiones. No es posible determinar la orientación del resto de enlaces del esqueleto.

Para ahondar más en qué esta ocurriendo, en el Cuadro 3.4 se muestran todos los caminos de tres nodos, junto al valor del test de Wilcoxon sobre el umbral. Hay más caminos que superan el umbral del test, pero debido al proceso de post-tratamiento, el algoritmo los elimina de la lista de colisiones válidas (ver Sección 2.1.1). Por ejemplo, el camino formado por *nivel educativo* → *edad* ← *intentos*, que supera ampliamente el umbral, ha sido excluido como colisión. Esto se debe a que los caminos *edad* → *intentos* ← *ansiedad* y *edad* → *intentos* ← *tratamiento requerido* orientan doblemente el enlace *edad* → *intentos*. Un razonamiento similar motiva la exclusión del camino *nivel educativo* → *ansiedad* ← *intentos*. Los caminos que forman un grupo de tres variables completamente conectados tampoco son aceptados. El procedimiento de purga de orientaciones contradictorias elimina la estructura en "V" completa, no sólo el enlace conflictivo.

La fase de completitud del grafo mediante heurísticos es capaz de orientar un enlace más, el de *intentos*→*consumo* (ver paso 3 del Algoritmo 1 en

Caminos de tres nodos			Test %
edad→	nivel educativo	←ansiedad	0
edad→	edad 1er intento	←intentos	0
edad→	edad 1er intento	←tto requerido	0
edad→	intentos	←edad 1er intento	0
edad→	*intentos*	*←ansiedad*	**109**
edad→	intentos	←tto requerido	**368**
edad→	intentos	←consumo de drogas	89
nivel educativo→	ansiedad	←intentos	110
nivel educativo→	ansiedad	←consumo de drogas	21
nivel educativo→	edad	←edad 1er intento	0
nivel educativo→	edad	←intentos	437
edad 1er intento→	*intentos*	*←ansiedad*	**176**
edad 1er intento→	intentos	←tto requerido	0
edad 1er intento→	intentos	←consumo de drogas	0
edad 1er intento→	tto requerido	←intentos	0
edad 1er intento→	edad	←intentos	0
ansiedad→	intentos	←tto requerido	3
ansiedad→	intentos	←consumo de drogas	0
ansiedad→	consumo de drogas	←intentos	0
intentos→	edad 1er intento	←tto requerido	0
intentos→	ansiedad	←consumo de drogas	0
tto requerido→	intentos	←consumo de drogas	0

Cuadro 3.4: Caminos de tres nodos y valores del test de Wilcoxon asociados. Las colisiones según el modelo MMHC se indican en cursiva. Los valores del test de Wilcoxon (en % sobre el umbral) correspondientes a las estructuras en "V" aceptadas se resaltan en negrita.

Sección 1.6.1). En la Figura 3.4 se muestra en trazo más grueso y color rojo los enlaces obtenidos directamente por la detección de colisiones. La otra conexión se ha orientado siguiendo la primera regla heurística del algoritmo PC (ver Sección 1.6.1): se orienta $X_j - X_k$ como $X_j \to X_k$ en los casos en los que exista $X_i \to X_j$ y no sean adjuntos X_i y X_k.

La probabilidad de acierto en orientación de enlaces es del 80 %. Sin embargo, a diferencia de los errores en clasificación, una equivocación en el sentido de una conexión supone un cambio muy importante en el grafo. Una medida deseable es una indicación de la confianza que se tiene sobre el sentido de los enlaces. Con el método *ccKnn* se puede emplear un criterio derivado de la batería de clasificadores. Así, aquellos enlaces que más veces fueran orientados en un mismo sentido tendrían más confianza que los demás. En el

ejemplo, los enlaces *edad→intentos* y *ansiedad→intentos* aparecen dos veces cada uno en sendos clasificadores k-vecinos-más-próximos. Estos resultados parecen apoyar los resultados clínicos, pues esas relaciones se han presentado en la literatura psiquiátrica. Varias de las relaciones que el algoritmo *ccKnn* no ha orientado, como *edad-edad 1er intento* o *edad-nivel educativo*, son también las más discutibles desde un punto de vista lógico. Por supuesto, es necesario validar esta hipótesis con nuevas simulaciones y pruebas específicas. Los experimentos con bases de datos sintéticas, presentados en la Sección 2.1.2, presentan en cualquier caso resultados prometedores.

Modelo causal para intentos de suicidio múltiples. Tras el análisis realizado, el modelo causal para IS múltiples, atendiendo sólo a las causas (edad, edad en el primer intento, nivel educativo y ansiedad), se ha empleado para generar una máquina de vectores soporte. El modelo ofrece unas probabilidades de generalización de 69.3 % de sensibilidad y 70.1 % de especificidad.

En la Figura 3.5 se muestra la curva ROC[3] del modelo. El área bajo la curva (AUC[4]) es del 71.7 %, que indica el rendimiento general del clasificador. También puede interpretarse este valor como la probabilidad de que una muestra de la clase positiva tenga una salida en el clasificador mayor que la clase negativa. La curva ROC se ha obtenido variando el parámetro de sesgo en el funcional de la máquina de vectores soporte (ver Apartado 1.4.1).

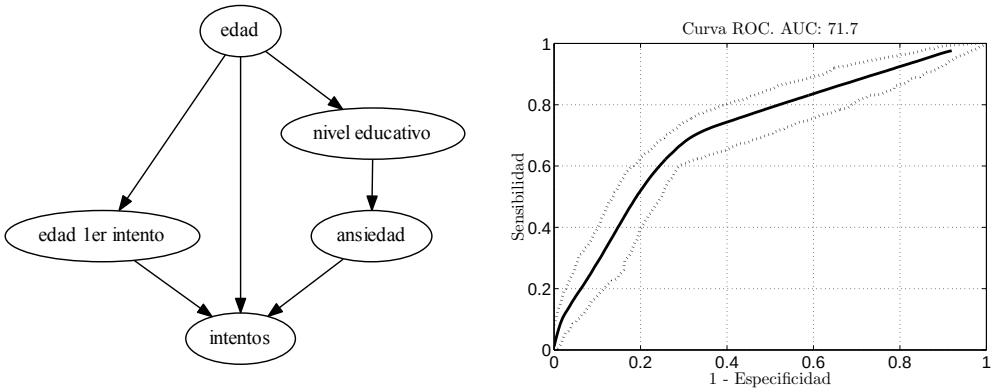


Figura 3.5: Curva ROC del modelo de predicción de intentos de suicidio múltiples. Las líneas punteadas indican el intervalo de confianza al 5 % (calculado mediante remuestreo "bootstrap").

[3] "Receiver Operational Curve".
[4] "Area Under the Curve".

3.2.4. Cálculo de riesgos y probabilidades condicionales

Para profundizar en la interpretación de los resultados, se han calculado las probabilidades condicionales de las variables presentes en el modelo de la Figura 3.3. Además de las probabilidades condicionales, se han calculado los "Odd Ratio" (OR) (Morris y Gardner, 1998). En la comunidad médica esta medida es ampliamente utilizada. Expresa el cociente entre la probabilidad de que ocurra un evento en un grupo frente a la probabilidad de que ocurra en otro. Generalmente estos grupos son dicotómicos. En el problema que se está estudiando, la variable de intentos de suicidio es IS = 1 si el número de intentos de suicidio es mayor de 2, e IS = 0 si los intentos son #IS ≤ 2.

La tabla de probabilidades para la variable de trastorno de ansiedad es la recogida en el Cuadro 3.5.

	Ansiedad = 1	Ansiedad = 0
#IS > 2	p_1	$1 - p_1$
#IS ≤ 2	p_0	$1 - p_0$

Cuadro 3.5: Tabla de probabilidades para calcular el OR de la variable ansiedad.

El riesgo OR se define como:

$$\begin{aligned} \text{OR} &= \frac{P(A=1|\text{IS}=1)/P(A=0|\text{IS}=1)}{P(A=1|\text{IS}=0)/P(A=0|\text{IS}=0)} = \\ &= \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{p_1(1-p_0)}{p_0(1-p_1)} \end{aligned} \tag{3.3}$$

donde $A = 1$ si el paciente sufre ansiedad y $A = 0$ en caso contrario. Los OR están relacionados con otra medida ampliamente usada en psiquiatría, el riesgo relativo, y son asintóticamente idénticos para el caso de probabilidades pequeñas. El riesgo relativo se define como $RR = \frac{p_1}{p_0}$, y cuantifica el cociente de probabilidad de que se dé ansiedad en los pacientes del grupo donde es cierta la hipótesis frente al grupo de pacientes con baja tasa de IS. Se puede comprobar en (3.3) que si $p_1 \ll 1$ y $p_0 \ll 1$, el riesgo relativo converge en el OR.

Se han calculado las tablas de probabilidad para las variables del árbol causal del modelo 3.3, y a partir de ellas los OR con sus respectivos intervalos de confianza (IC) mediante remuestreo "bootstrap". El número de repeticiones para el cálculo de los IC se ha fijado en $B = 1000$, siguiendo las recomendaciones de (Andrews y Buchinsky, 2000).

Los valores más representativos de los riesgos se muestran en el Cuadro 3.6. Aquellos OR cuyo intervalo de confianza circunscriben la unidad se han marcado como no significativos con un asterisco.

Variable Ppal.	Etiqueta	OR	IC 95 %	Contraste
Edad	1 (<35)	1	-	IS frecuentes
	2 (35-64)	1.59	[1.28,1.99]	
	3 ($\geq$ 65)	0.52	[0.33,0.81]	
Edad	1 (<35)	1	-	IS frecuentes
(sin trast. ans.)	2 (35-64)	2.34	[1.66,3.31]	
	3 ($\geq$ 65)	0.37*	[0.12,1.07]	
Edad	1 (<35)	1	-	IS frecuentes
(con trast. ans.)	2 (35-64)	1.25*	[0.90,1.72]	
	3 ($\geq$ 65)	1.32*	[0.48,3.60]	
Edad en el	1 (<35)	1	-	IS frecuentes
1^{er} intento	2 (35-64)	0.67	[0.54,0.84]	
	3 ($\geq$ 65)	0.10	[0.06,0.17]	
Educación	1 (Iletrado)	1	-	Trast. ansiedad
	2 (Graduado)	0.67*	[0.23,1.95]	
	3 (FP)	0.80*	[0.26,2.44]	
	4 (Bachiller)	1.89*	[0.64,5.51]	
	5 (Universidad)	1.72*	[0.60,4.92]	
Trast. ansiedad	0 (No)	1	-	IS frecuentes
	1 (Sí)	4.06	[3.19,5.16]	
Trast. ansiedad	0 (No)	1	-	IS frecuentes
(edad <35)	1 (Sí)	5.46	[3.83,7.79]	
Trast. ansiedad	0 (No)	1	-	IS frecuentes
(edad 35-64)	1 (Sí)	2.91	[2.13,3.98]	
Trast. ansiedad	0 (No)	1	-	IS frecuentes
(edad $\geq$ 65)	1 (Sí)	19.51	[4.70,80.75]	
IS frecuentes	0 (1-2 IS)	1	-	Uso de drogas
	1 (>2 IS)	1.96	[1.50,2.56]	
IS frecuentes	0 (1-2 IS)	1	-	Uso de drogas
(sin ansiedad)	1 (>2 IS)	3.21	[2.19,4.69]	
IS frecuentes	0 (1-2 IS)	1	-	Uso de drogas
(con ansiedad)	1 (>2 IS)	1.10*	[0.77,1.58]	
IS frecuentes	0 (1-2 IS)	1	-	Tratamiento†
	1 (>2 IS)	1.46	[1.08,1.96]	
IS frecuentes	0 (1-2 IS)	1	-	Tratamiento‡
	1 (>2 IS)	2.25	[1.66,3.05]	

Variable Ppal.	Etiqueta	OR	IC 95 %	Contraste
IS frecuentes	0 (1-2 IS)	1	-	Tratamiento$^{\bullet}$
	1 (>2 IS)	3.28	[2.31,4.67]	

Cuadro 3.6: Riesgo (OR) asociado a las variables más significativas del modelo causal para la predicción de intentos de suicidio. *Valores no significativos. †Primeros auxilios y/o emergencias vs. Ingreso hospitalario. ‡UCI vs. Primeros auxilios. $^{\bullet}$UCI vs. Ingreso hospitalario.

3.3. Interpretación de los resultados y conclusiones

Con el modelo causal obtenido y las probabilidades condicionales y los riesgos del Cuadro 3.6, ya es posible realizar un estudio en profundidad de los resultados.

Atendiendo a la edad de los pacientes, y comparando con los más jóvenes, el riesgo de intentos de suicidio (IS) frecuentes en la población de edad intermedia aumenta (OR = 1.6), si bien disminuye en la tercera edad (OR = 0.5). Respecto a la edad en el primer intento, el riesgo de repetición del IS decrece de manera proporcional: OR = 0.7 en edad intermedia y OR = 0.1 en tercera edad.

El trastorno de ansiedad es un factor que incrementa notablemente el riesgo de realizar IS (OR = 4.1). En estos casos, y estratificando por edades, el riesgo es muy notable en la tercera edad (OR = 19.5), seguido en gravedad por los pacientes jóvenes (OR = 5.5). En la edad intermedia el riesgo es de OR = 2.9. Entre aquellos pacientes sin trastorno de ansiedad, el riesgo es mayor en los casos de edad intermedia (OR = 2.34, IC = [1.66,3.31]).

También se compara el riesgo de ser ingresado en la UCI frente a tener una atención de primeros auxilios o emergencias y frente a sufrir una baja hospitalaria, para los pacientes con IS múltiples. En ambos casos, el riesgo de recaer en la UCI es mayor, con OR de 2.25 y 3.3, respectivamente. Sin embargo, los reincidentes múltiples ofrecen más riesgo (OR = 1.46) de ser atendidos de primeros auxilios que de tener una baja hospitalaria.

En cuanto al incremento de riesgo por consumo de alcohol u otras drogas, y para aquellos pacientes sin trastorno de ansiedad, el OR asciende a OR = 3.21.

Tras el estudio realizado, es posible extraer algunas conclusiones. La comprensión y predicción de los intentos de suicidio repetidos puede mejorar atendiendo a las variables de edad, edad en el primer intento y trastorno de

ansiedad. Las consecuencias de un alto número de IS se reflejan en el consumo de drogas y alcohol, así como en la severidad de los daños autoinfringidos, como indica la variable de tratamiento requerido.

La edad del paciente condiciona de manera significativa la repetición de los IS. Aquellos con edades mayores de 35 años no suelen recaer con frecuencia; esto se confirma de manera especial en el grupo de mayores de 64 años. Parece razonable, puesto que los IS son más frecuentes entre los jóvenes, además de que un IS a edades avanzadas tiene más probabilidades de acabar con la vida del paciente.

El trastorno de ansiedad incrementa de manera significativa el riesgo de múltiples IS, especialmente en la tercera edad y la población joven. Una hipótesis a verificar es que la edad module la asociación entre IS y ansiedad, independientemente de otros trastornos mentales. Una carencia del modelo es que no incluye variables relacionadas con los trastornos de personalidad (trastorno bipolar o depresión). En el modelo de búsqueda hacia delante sí aparecen entre las más significativas las variables *depre-bip* y *dd_psychosis*, que indican, respectivamente, si el paciente presenta depresión o trastorno bipolar y alguna psicosis. Es posible que una base de datos más completa permitiera realizar un modelado causal que incluyera más variables. Por otro lado, los trastornos bipolares están asociados en la literatura médica con intentos de mayor gravedad, que provocaría intentos consumados en mayor proporción que intentos repetidos.

La relación entre consumo de drogas o alcohol con un número alto de IS también ha sido reflejada en numerosos estudios. El consumo puede deberse tanto a una escapatoria para aliviar el problema mental, como a una medida contra los estados de ansiedad. Finalmente, la gravedad del IS, medido como atención en UCI frente a atención en emergencias o ingreso hospitalario, aumenta cuando el número de intentos es mayor. Estas conclusiones son coherentes con los resultados de otros estudios psiquiátricos.

Los datos de este estudio pueden ser empleados para detectar pacientes con riesgo de repetir intentos de suicidio o conductas autolesivas.

Conclusiones. Mediante métodos de aprendizaje máquina y técnicas de inferencia causal se han seleccionado las variables relevantes y se ha generado un modelo que explica la interacción entre las variables más importantes de la base de datos disponible. De esta manera ha sido posible construir un árbol causal, jerarquizando los factores relacionados con la repetición de intentos de suicidio. El árbol causal facilita la interpretabilidad de los resultados, más allá de los valores de probabilidad de error. Empleando únicamente las causas de la variable de intentos de suicidio se ha obtenido una sensibilidad,

especificidad y AUC de aproximadamente el 70 %.

Los modelos causales obtenidos mediante los algoritmos MMHC y ccKnn son coincidentes en las relaciones más importantes según el punto de vista clínico, lo cual apoya la verosimilitud de que las relaciones halladas son correctas. Sin embargo, sería deseable disponer de índices de confianza sobre las relaciones causales. En la Sección 3.2.3 se ha esbozado un esquema de cómo obtenerlos a partir de la batería de clasificadores de los métodos ccKnn o ccMSVM.

La topología del grafo es un factor importante a la hora de realizar inferencia causal, en especial en los métodos ccKnn y ccMSVM, que realizan búsqueda de estructuras en "V". En el problema de los IS repetidos, el esqueleto del modelo causal es muy denso, por lo que se dan pocas colisiones. La ausencia de información en algunas variables de la base de datos también ha podido dificultar esta tarea.

La aplicación de los métodos de inferencia causal a nuevos problemas ayudará a contrastar la validez de los algoritmos presentados.

Capítulo 4

Determinación de focos en fibrilaciones auriculares

Las técnicas estadísticas y de tratamiento de señal clásicas presentan deficiencias a la hora de determinar el origen y el sentido de propagación del flujo de información. Cuando la información está contenida en series de variación temporal, es posible emplear el concepto de causalidad de Granger para generar un modelo del mecanismo que subyace al sistema. En este capítulo se procede a aplicar las técnicas de inferencia causal para series temporales en un problema real, en el ámbito de la cardiología.

Se desea estudiar la propagación de señales durante episodios de fibrilación auricular. A partir de electrogramas intracardiacos, se ha desarrollado un procedimiento para obtener mapas causales dinámicos de la propagación de señales por la superficie del corazón.

El tratamiento de las fibrilaciones auriculares mediante ablación del miocardio es habitual tras ensayar otros tratamientos previos. Una hipótesis de trabajo, con bastante aceptación, consiste en eliminar los puntos o áreas de la aurícula con mayor frecuencia de estimulación. Se asume que las zonas con *frecuencia dominante* son las causantes de la fibrilación. Sin embargo, no existe evidencia de que la relación entre frecuencia dominante y foco de la fibrilación auricular sea correcta. Un modelo basado en causalidad tipo Granger ofrece una nueva perspectiva del problema. En este capítulo se presenta un trabajo en común con personal médico del servicio de electrofisiología cardíaca del Hospital General Universitario Gregorio Marañón de Madrid (HGGM). Parte de este capítulo ha sido publicado en el congreso (de-Prado-Cumplido y Artés-Rodríguez, 2010).

En primer lugar se presentará el problema clínico: el análisis de señales de fibrilación auricular obtenidas con un catéter intracardiaco. A continuación se expone un método de análisis de estas señales, que realiza una búsqueda

de las frecuencias dominantes presentes en las formas de onda. Se muestran las limitaciones de este método, que carece de fundamentación teórica para identificar causas y efectos, y se propone el uso de métodos de inferencia causal. Se emplean dos métodos de inferencia causal, uno basado en la pendiente de la fase del espectro ("Phase Slope Index") y el otro, el cMultiSVARMA, desarrollado en la Sección 2.2. Para presentar las interacciones de manera dinámica, se detalla el desarrollo de una herramienta de visualización. Finalmente, se exponen la interpretación de los resultados y las conclusiones del capítulo.

4.1. Hipótesis de generación de fibrilaciones

La fibrilación auricular (FA) es la arritmia cardíaca más frecuente, que si bien no es de gravedad extrema, reduce la calidad de vida, puede ocasionar otras patologías más graves e incrementa el riesgo de mortalidad. Los mecanismos que generan la FA no se comprenden en su totalidad actualmente.

Se barajan principalmente dos hipótesis acerca de la generación y mantenimiento de estas fibrilaciones: puede ser o bien debido a varios frentes de onda que se desplazan aleatoriamente por la aurícula, o bien, a rotores estables que generan una jerarquía de frecuencias en la superficie auricular (Mansour et al., 2001; Sanders et al., 2005; Ng et al., 2006). Entre otras medidas terapéuticas, la ablación mediante catéter mejoraría si se comprendiera en profundidad el mecanismo de activación y sostenimiento de estas arritmias. Hay evidencias de que pacientes con FA paroxística[1], tratados con ablación, sufren menos episodios de fibrilación un año después de la operación que los pacientes no tratados (Pappone y Santinelli, 2009). Es posible estudiar las FA mediante señales registradas intracardialmente.

Sin embargo, la mayoría de técnicas estadísticas no permiten identificar la dirección en la que fluye la información entre diferentes series temporales, tan sólo la relación entre ellas. Técnicas como el espectro cruzado es un ejemplo de esta situación.

Se van a emplear técnicas de inferencia causal para estudiar las relaciones entre señales de fibrilación auricular. Se pretende obtener una herramienta para visualizar y cuantificar la dirección de propagación de las ondas en la superficie de la aurícula durante episodios de fibrilación. Se estudiará la estabilidad y reproducibilidad de la propagación eléctrica.

Particularizando en el problema médico de este trabajo, verificar la existencia y localizar los focos de generación de FA puede mejorar el éxito te-

[1]Las FA se dividen en paroxísticas (terminan por sí mismas), persistente (tratable con cardioversión) y permanente (no tratable con cardioversión ni medicamentos).

rapéutico de la ablación auricular. Recientemente se ha investigado el uso del concepto de frecuencia dominante para estudiar las FA (Fischer et al., 2007), así como otras técnicas de seguimiento espectral (Sandberg et al., 2008).

4.1.1. Señales de fibrilación auricular

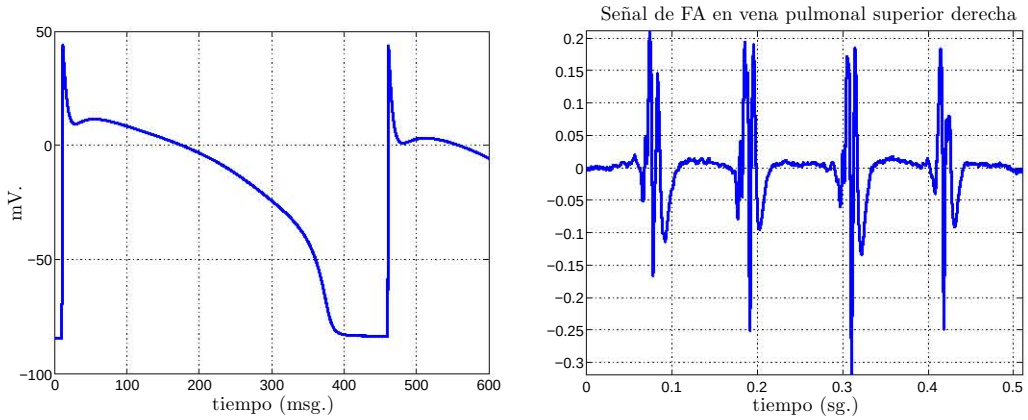
Resulta de interés conocer la fisiología de las células cardíacas (Jalife et al., 1999; Sörnmo y Laguna, 2005), que a la postre son las que generan las señales eléctricas del corazón que se pueden sensar. Las células poseen una pared poco conductiva, pero que es permeable, bajo determinadas condiciones, a ciertas sustancias. Estos canales de las membranas celulares son los que gobiernan las diferencias de potencial entre el interior y el exterior. Si bien las neuronas, las células cardíacas y las del resto de músculos funcionan de una manera similar, se diferencian, entre otros factores, en las sustancias y tipos de canales. En el corazón, los principales canales son los que permiten o niegan la difusión de iones de sodio (Na^+), potasio (K^+) y cloro (Cl^-).

Cuando la célula se encuentra en reposo, se establece un equilibrio electroquímico entre las fuerzas de difusión, que mueven iones a través de los canales, y las fuerzas eléctricas originadas por las diferentes cargas de los iones. Por convenio, se toma como referencia la tensión en el exterior de la célula. Habitualmente el potencial transmembrana en reposo está acotado entre -60 y -100 mV.

Si la célula recibe una corriente que la estimula, cambia la permeabilidad de la membrana y, por tanto, su potencial. Este es el conocido como *potencial de acción*. La propiedad de *excitabilidad* de estas células es la que las diferencia del resto de células del ser humano. La característica decisiva de la corriente de estimulación no es su forma de onda, sino que supere un cierto umbral, por debajo del cual no hay respuesta celular. Por tanto, el potencial de acción es una respuesta no lineal a los impulsos de corriente.

Una simulación del potencial de acción se puede ver en la Figura 4.1(a). Se pueden observar dos fases, a saber, un periodo de despolarización rápido, en el que la diferencia de tensión tiende a 0 V., seguido de una repolarización más lenta, que lleva a la célula a su potencial de reposo al final del mismo. Los canales responsables de esta variación son principalmente los de sodio y potasio.

Una vez que la célula ha producido un potencial de acción, se establece un periodo refractario, durante el que la célula no puede activarse de nuevo. Esta refractariedad determina la frecuencia de propagación máxima. En (Nattel et al., 2008), se muestra cómo las FA remodelan las células auriculares, convirtiendo las paroxísticas, menos graves, en persistentes o permanentes. La remodelación induce periodos refractarios más cortos, lo cual a su vez hace



(a) Potencial de acción. (b) Señal de fibrilación auricular.

Figura 4.1: (a)Ejemplo de potencial de acción simulado según modelo de (Luo y Rudy, 1994). (b)Señal de fibrilación capturada en la pared auricular.

más probable la aparición de fibrilaciones.

El potencial de acción generado en una célula se propaga hacia las colindantes, dando lugar a frentes de onda que atraviesan el miocardio. Una terapia habitual para controlar las FA consiste en ablacionar zonas de la aurícula, para así evitar la propagación no deseada de frentes de onda.

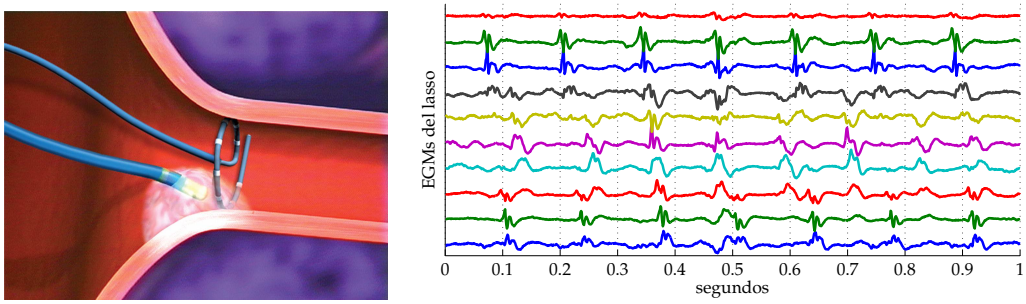
La actividad conjunta de las células del miocardio se puede registrar a nivel local, mediante catéteres intracardiacos, o a nivel global, mediante ECG no invasivos. En la Figura 4.1(b) se representa una señal de fibrilación real, que es la resultante de la suma de los potenciales de acción de muchas células. Existen modelos de generación de potenciales de acción (Rush y Larsen, 1978; Victorri, 1985; Luo y Rudy, 1994; Kléber y Rudy, 2004), que son muy útiles a la hora de estudiar las arritmias y proponer nuevas técnicas de tratamiento.

4.2. Registro de electrogramas con polígrafo

Para realizar este estudio se dispone de señales de electrogramas (EGM) obtenidos durante estudios médicos, en venas pulmonares, techo y orejuela de la aurícula izquierda. Las medidas se han registrado con catéteres circulares[2] marca Lasso™ (Biosense Webster, Inc.) y Optima™ (St. Jude Medical, Inc.), de 10 sensores con registro simultáneo. La disposición de los sensores se engloba en una circunferencia de diámetro variable entre 15 y 25 mm. En la Figura 4.2 se muestra una recreación virtual del catéter circular, situado en

[2]No confundir con el método de optimización LASSO discutido en la Sección 2.2.

una vena pulmonar de la aurícula.



(a) Catéter circular en vena pulmonar, junto a la sonda de ablación.

(b) Señales del Lasso™ en fibrilación auricular.

Figura 4.2: Catéter circular Lasso™ multipolar situado en una vena pulmonar, junto al catéter de ablación. Imagen extraída de (Ghaye et al., 2003). A la derecha se muestra un ejemplo de electrogramas de fibrilación sensados con el catéter circular.

Estas señales $x_i(t),\ i = 1, \ldots, 10$, están muestreadas a 977 Hz. Los pacientes han dado su consentimiento al estudio y éste ha sido aprobado por el comité de ética del Hospital Gregorio Marañón.

4.3. Análisis de FA mediante frecuencias dominantes

En la última década se ha investigado en una técnica conocida en el mundo médico como de *frecuencias dominantes*. Se asume la hipótesis de que las zonas de la pared cardíaca con una mayor velocidad en la generación de impulsos eléctricos son el foco de la arritmia; la búsqueda de las zonas con frecuencias dominantes y su ablación se ha mostrado como un método eficaz para el mantenimiento del ritmo sinusal en pacientes con FA (Ng et al., 2006; Atienza et al., 2009). Con el objeto de identificar estas áreas se ha empleado un tratamiento de señal básico (Fischer et al., 2007), que se recoge en los bloques presentados en la Figura 4.3.

Estos bloques de tratamiento permiten identificar la frecuencia fundamental de la FA, obviando la forma de onda de cada estimulación.

Para un primer análisis, se modela la fibrilación auricular como un tren de ondas como la de la Figura 4.4, equiespaciadas cada $T = 130$ msg.

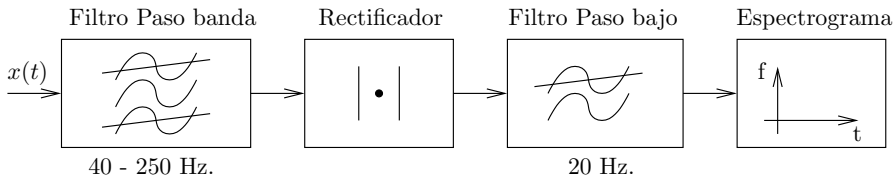
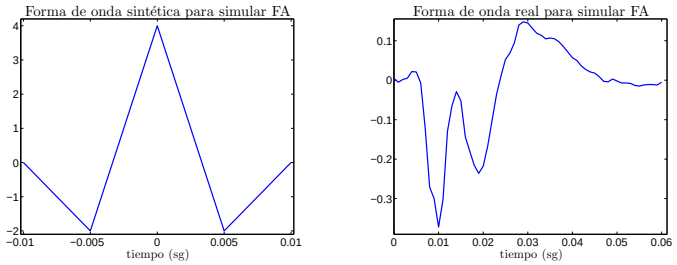


Figura 4.3: Bloques del tratamiento de señal empleado para detectar las frecuencias dominantes.



(a) Forma de onda bipolar, $b(t)$, para simular una FA.

(b) Forma de onda real.

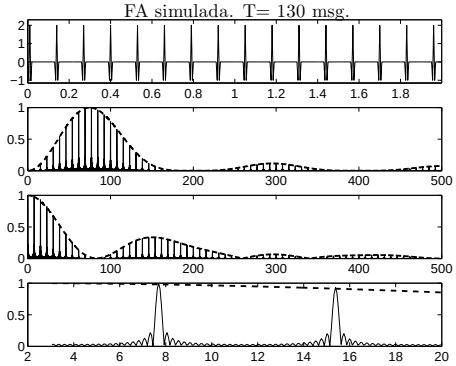
Figura 4.4: Patrón de la señal de fibrilación auricular.

La señal se modela por tanto como N réplicas desplazadas del patrón básico: $B(t) = \sum_{n=1}^{N} b(t - nT)$. A continuación se realiza el procesado explicado anteriormente. Si se calcula el espectro del tren de ondas $B(t)$ se obtiene el producto de la envolvente del espectro de $b(t)$ por un tren de deltas equiespaciadas cada $1/T$ Hz. La siguiente ecuación desarrolla esta idea:

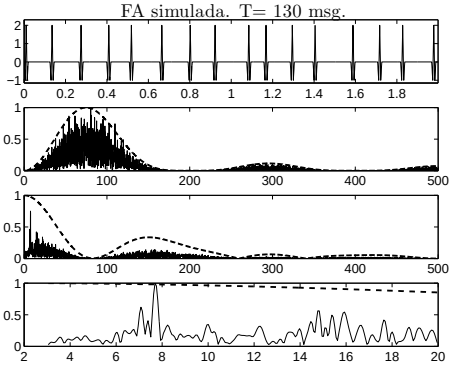
$$\mathcal{F}\{B(t)\} = \sum_{n=1}^{N} \mathcal{F}\{b(t - nT)\} = \sum_{n=1}^{N} \mathcal{F}\{b(t)\} e^{iwnT} = \\ = \mathcal{F}\{b(t)\} \sum_{n=1}^{N} e^{iwnT} \tag{4.1}$$

Como puede verse en la Figura 4.5(a), la transformación de Fourier de la onda, $\mathcal{F}\{b(t)\}$, es la envolvente del espectro, dibujada en rojo.

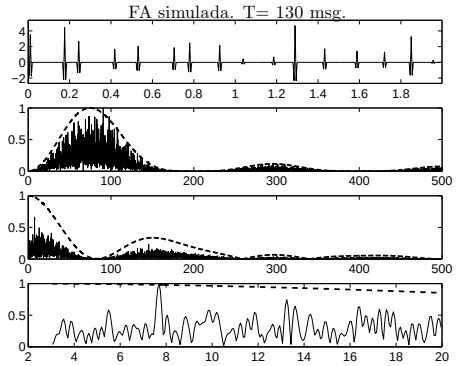
La secuencia de picos equiespaciados del espectro están relacionados con la inversa del ciclo de la fibrilación. Se puede comprobar que los picos están a $1/T$ Hz unos de otros. La transformación no lineal (rectificación o valor absoluto) hace que el espectro se desplace a frecuencias bajas, dónde se recupera finalmente la frecuencia dominante. Esta frecuencia dominante coincide con el inverso del ciclo $1/T = 7.7$ Hz.



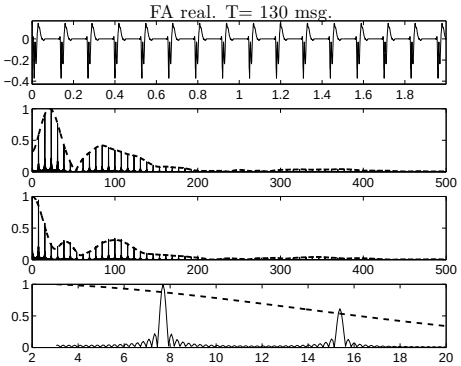
(a) $B_1(t)$, ciclo constante.



(b) $B_2(t)$, variabilidad en el ciclo.



(c) $B_3(t)$, variabilidad en ciclo y amplitud.



(d) $B_4(t)$, modelo ideal con forma de onda real.

Figura 4.5: Para cada secuencia $B_i(t)$ se representa, de arriba a abajo, i/ Señal de fibrilación (en sg.), ii/ Espectrograma, iii/ Espectrograma de la señal rectificada y iv/ Zoom de las frecuencias bajas del espectrograma previo (en Hz).

A este modelo ideal se le van a incluir dos modificaciones para aproximar mejor el funcionamiento real: añadir variabilidad en los instantes de aparición de las ondas y la aleatorización de las amplitudes de las mismas.

En lugar de introducir una onda cada T msg., se hará cada $T + r$, donde r es una variable aleatoria gausiana de $\sigma = 10$. Los resultados se pueden ver en la Figura 4.5(b) para la secuencia $B_2(t)$. La envolvente (curva roja) se mantiene invariante, pero los picos subtendidos bajo ella pierden su estructura en gran medida. Sin embargo, en el espectrograma de la Figura 4.5(b), puede comprobarse que es posible recuperar la frecuencia dominante, que se

aproxima razonablemente a la media de los intervalos entre latidos.

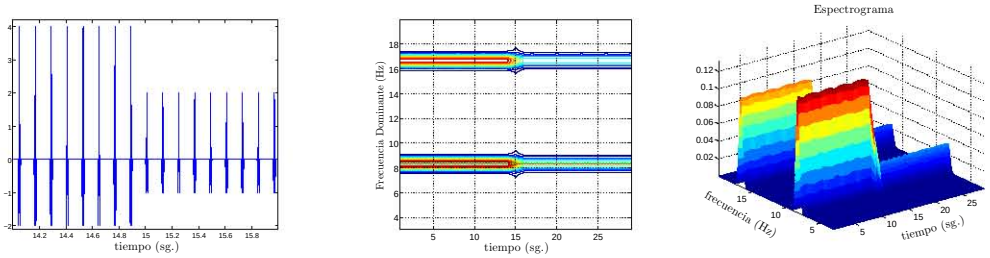
Se modifica el modelo añadiendo ruido gausiano a las amplitudes de las ondas, generando $B_3(t)$. Esta perturbación sobre el modelo ideal no parece afectar significativamente, como se comprueba en la Figura 4.5(c). Finalmente, se repite el modelo ideal, pero sustituyendo la forma de onda por otra extraída de una fibrilación real, que puede verse en la Figura 4.4(b). El espectro de esta fibrilación se muestra en la Figura 4.5(d).

Espectrograma de señales simuladas. Con el fin de comprender mejor cómo se reflejan en el espectrograma los diversos ritmos y formas de señal, se han simulado diferentes escenarios. En la Figura 4.6(a-d) se muestran diversas secuencias simuladas, junto al resultado de aplicar el tratamiento de la Figura 4.3.

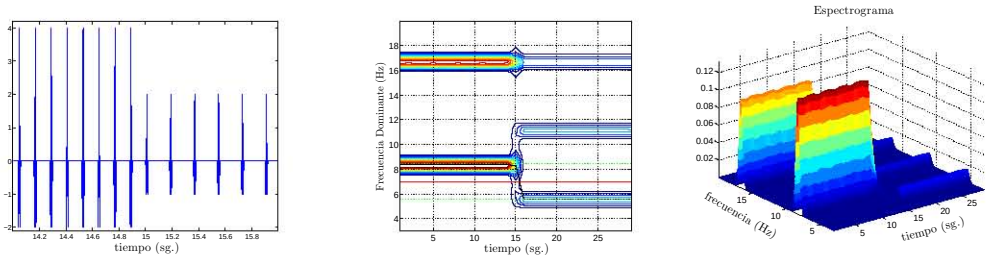
La primera de ellas es una señal de periodo $T = 120$ msg. (frecuencia $f = \frac{1}{T} = 8.3$ Hz.), formada por dos segmentos, el primero de amplitud doble. Como se ve en las gráficas del espectrograma (Figura 4.6), la amplitud no afecta a la ubicación de la frecuencia principal de la señal. En la segunda secuencia, a partir del segundo 15, se ha cambiado el periodo de la señal de $T_1 = 120$ msg. a $T_2 = 180$ msg., por lo que la frecuencia en la segunda parte de la señal es de 5.5 Hz. Se pueden observar los armónicos en las frecuencias de 11 y 16.5 Hz. La curva de color negro marca la frecuencia dominante, que pasa de ser 8 Hz. en la primera mitad de la señal a 5.5 Hz. en la segunda mitad, como era de esperar.

La tercera señal consiste en la superposición de dos: una de amplitud unitaria y periodo $T_1 = 180$ msg. (5.5 Hz.) y la otra de amplitud mitad y periodo $T = 120$ msg. (8.3 Hz.). El algoritmo de búsqueda de frecuencias dominantes estimaría, acertadamente, que la señal de 5.5 Hz. es la principal componente de la señal cardíaca. Finalmente, la cuarta simulación presenta durante la primera mitad de la señal la misma combinación que en el caso anterior, mientras que la segunda mitad es la adición de dos componentes de igual ciclo ($T = 120$ msg.) pero distinta amplitud, de 1 y 0.5. La frecuencia dominante, como es de esperar, se sitúa en 5.5 y 8.3 Hz. en la primera y segunda mitad de la señal, respectivamente. La curva negra marca este recorrido.

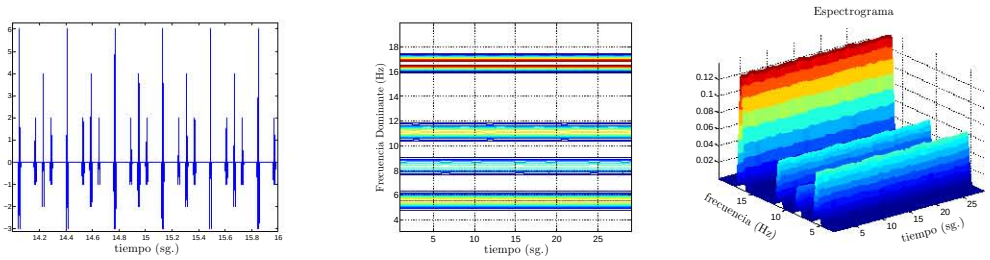
Espectrograma de señales reales. Se ha aplicado esta metodología a la base de datos de fibrilaciones del Hospital Gregorio Marañón. Sin embargo los resultados no han sido satisfactorios, pues las señales son demasiado complejas para atacar el problema con esta herramienta. La Figura 4.7 recoge un ejemplo.



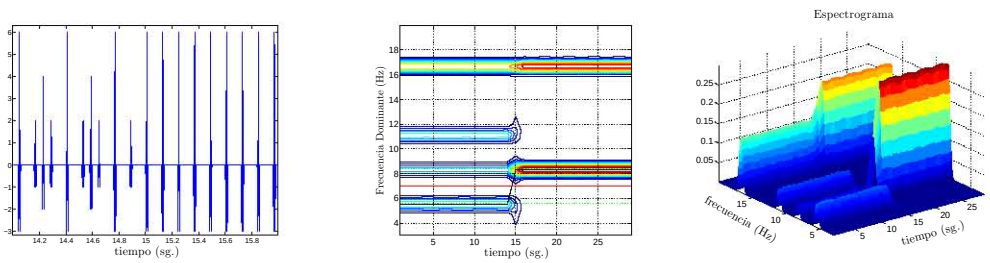
(a) Prueba 1.



(b) Prueba 2.



(c) Prueba 3.



(d) Prueba 4.

Figura 4.6: Diferentes modelos de señal de FA y espectrograma asociado.

La fila superior de la figura muestra la frecuencia dominante en cada ventana del espectrograma, así como su valor medio. Las distintas columnas representan el espectrograma de 4 de los 10 sensores del Lasso™, situado en la vena pulmonar superior derecha. Las gráficas abarcan un segmento de FA

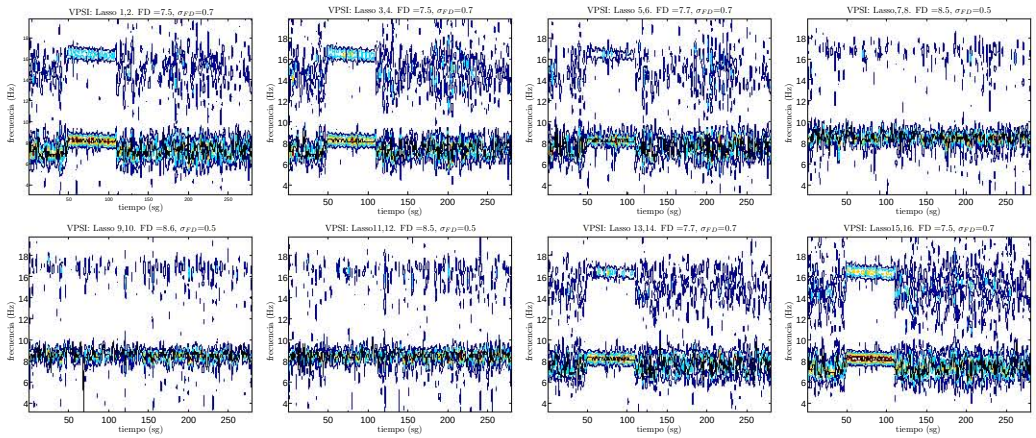


Figura 4.7: Espectrogramas de las señales reales de fibrilación auricular. La curva negra es la Frecuencia Dominante para cada instante. Puede verse un ritmo alto sostenido entre los segundos 50 y 100.

de 100 sg.

La simplicidad del método hace que el estudio de frecuencias dominantes no sea válido para identificar focos de actividad en la aurícula. Si bien puede ser una hipótesis válida, su cálculo no permite analizar si dicha hipótesis es correcta. Además de esto, la determinación de la frecuencia dominante, debido a la complejidad de las señales, es un método ruidoso y de poca resolución.

Una inferencia causal en las señales EGM proveería a los resultados de una fundamentación teórica para establecer un mapa de propagación causal.

No se pretende crear un método de representación tridimensional como los sistemas CARTO™ (Biosense Webster, Inc.) que se emplean en las intervenciones cardíacas, sino un mapa de medidas causales, del que poder inferir el mecanismo que gobierna el flujo de frentes de onda en la aurícula.

En las siguientes secciones se propone usar métodos de inferencia causal para series temporales. Ello permitirá representar un mapa causal del que será posible extrapolar la dirección de propagación de los frentes de onda.

4.4. Análisis y representación de relaciones causales

Además de aplicar el método cMultiSVARMA para obtener las medidas causales, se va a emplear el PSI. Uno de los métodos de búsqueda de causalidad entre señales es el “Phase Slope Index” (Nolte et al., 2008). Este

método emplea no sólo la máxima frecuencia de las señales, sino toda la fase del espectro cruzado. Ha sido empleado anteriormente con éxito en señales de fMRI y electroencefalogramas.

El algoritmo PSI está basado en la estima de la pendiente promedio de la fase de la coherencia cruzada de las series temporales. De manera intuitiva, esta medida dará un valor alto si una de las series precede a la otra; si las señales son independientes, la parte imaginaria será nula. La coherencia entre dos señales $x_i(t)$ y $x_j(t)$ se define como el espectro cruzado normalizado:

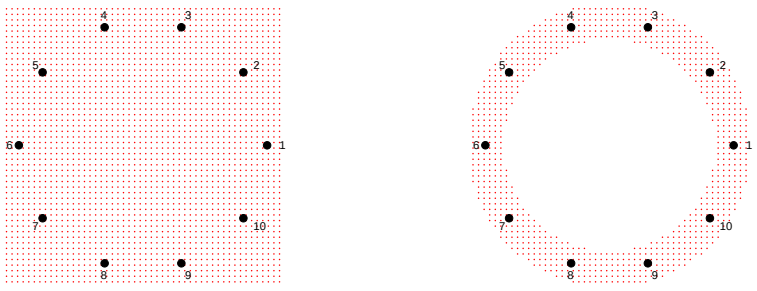
$$C_{ij}(f) = \frac{(S_{ij}(f))}{\sqrt{(S_{ii}(f)S_{jj}(f))}} \tag{4.2}$$

donde $S_{ij}(f)$ es el espectro cruzado entre dichas señales.

Finalmente, la influencia causal entre dos señales, $\hat{\Psi}_{ij}$, viene cuantificada por la parte imaginaria de esta correlación: $\hat{\Psi}_{ij} = \Im(\sum_f C^*_{ij}(f)C_{ij}(f+\delta f))$, donde δf es la resolución frecuencial del algoritmo.

4.4.1. Interpolación de relaciones causales y representación

Para la visualización de las relaciones causales, se ha generado un modelo circular, en el que los sensores del catéter circular se muestran como los puntos negros de la Figura 4.8. El mapa causal se mide en la rejilla formada por los puntos de menor tamaño. En la literatura se han ensayado otras herramientas de representación a partir de registros de electrogramas (Weber et al., 2010; Richter et al., 2011), pero no se basan en medidas causales.



(a) Mapa global. (b) Mapa local.

Figura 4.8: Rejilla de posicionamiento espacial de los sensores del catéter.

Las etapas del procedimiento se resumen en el Algoritmo 7 y se desarrollan en las siguientes secciones.

Algoritmo 7 Generación de secuencias de mapas causales.

1: Calcular el índice causal entre pares de sensores, para las ventanas temporales $c_{i\to j}(t)$.
2: Interpolar $c_{i\to j}(t)$ en los puntos de la rejilla: $\vec{IC}_g(\vec{y};t)$ y $\vec{IC}_l(\vec{y};t)$ para, respectivamente, mapa global y local.
3: Generar los fotogramas. El módulo del mapa causal $|\vec{IC}_x(\vec{y};t)|$ se codifica en color, y la dirección en cada punto mediante un conjunto de vectores.
4: Finalmente, secuenciar los fotogramas para generar un vídeo.

Medidas de causalidad

A partir de las señales grabadas con el Lasso™ (ver Figura 4.2), se obtiene la matriz de influencias causales, para ventanas de señal de 250 msg. y 50 % de solapamiento: $c_{i\to j}(t)$, con $i,j = 1,\ldots,10$ y siendo t el índice temporal de las ventanas. Las medidas causales entre pares de sensores se cuantificarán de dos formas, mediante el índice PSI, $\Psi_{ij}(t)$ y mediante la salida del cMultiSVARMA, $M_{ij}(t)$.

Generación de mapas

El catéter suele situarse en la confluencia de las venas con la pared auricular. Para representar de manera gráfica las relaciones causales se han generado dos tipos de representaciones o gráficas:

1. Mapa Global. A pesar de que el espacio interior del Lasso™ no es una superficie real al sensar las venas, para una mejor representación y visualización de la propagación de las ondas se ha realizado la representación en la zona interior del catéter circular. En los casos en los que el catéter se sitúa en paredes u orejuela esta representación tiene pleno sentido.

2. Mapa Local. Sólo las medidas de causalidad entre los 3 sensores más cercanos del Lasso™ son tenidas en cuenta para la representación. La topografía de las rejillas de cálculo se muestra en la Figura 4.8.

En ambos casos, para interpolar el efecto de las relaciones causales sobre un punto concreto de la rejilla, $\vec{y}$, se deben calcular dos distancias (ver Figura 4.9) entre el segmento $\overline{\vec{x}_i, x_j}$ e $\vec{y}$:

- d_o: distancia al origen del segmento. El punto origen lo indica el signo de $c_{i\to j}(t)$.

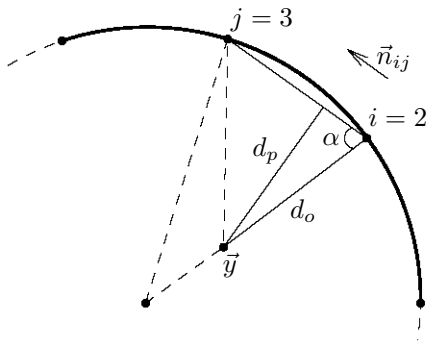


Figura 4.9: Distancias entre los puntos de la rejilla y los segmentos entre pares de señales, para interpolar la relación causal general.

- d_p: distancia perpendicular al segmento.

Las ecuaciones para determinar estas distancias y vectores incluyen el cálculo del vector unitario n_{ij} de dirección entre el foco $\vec{x}_i$ y el destino $\vec{x}_j$:

$$\vec{n}_{ij} = \frac{\vec{x}_j - \vec{x}_i}{\|\vec{x}_j - \vec{x}_i\|} \tag{4.3}$$

$$\alpha_{ij} = \angle(\vec{y} - \vec{x}_i) - \angle(\vec{x}_j - \vec{x}_i) \tag{4.4}$$

$$d_{p,ij}(\vec{y}) = \|\vec{x}_i - \vec{y}\| \sin(\alpha_{ij}) \tag{4.5}$$

$$d_{o,ij}(\vec{y}) = \|\vec{y} - \vec{x}_i\| \tag{4.6}$$

En el caso del mapa global, la interpolación se realiza mediante la siguiente ecuación:

$$\vec{IC}_g(\vec{y};t) = \sum_{i=1}^{10} \sum_{j=i+1}^{10} e^{-1.2 d_{o,ij}(\vec{y})} e^{-1.6 d_{p,ij}(\vec{y})} c_{i\to j}(t) \vec{n}_{ij} \tag{4.7}$$

$$\vec{IC}_l(\vec{y};t) = \sum_{i=1}^{2} \sum_{j=i+1}^{3} e^{-1.6 d_{o,k(i)k(j)}(\vec{y})} c_{k(i)\to k(j)}(t) \vec{n}_{k(i)k(j)} \tag{4.8}$$

donde $k(i)$ es la lista ordenada de sensores más cercanos a $\vec{y}$. De esta forma se promedian las influencias causales de las tres señales del Lasso™ más cercanas al punto de medida. Los factores de suavizado de las exponenciales se han determinado heurísticamente.

Un ejemplo de las contribuciones de cada par de señales y de la resultante se muestra en la Figura 4.10.

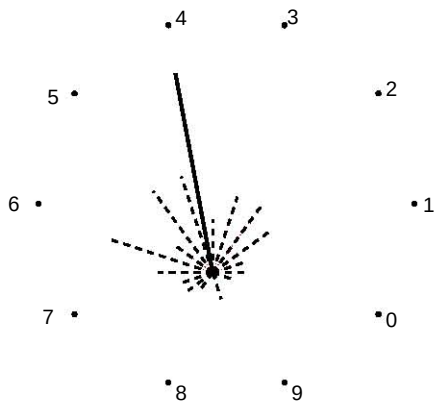


Figura 4.10: Contribuciones parciales y resultante de las influencias causales.

Finalmente, este procedimiento se repite para cada ventana temporal y se construye una secuencia visual. La intensidad de la relación causal (módulo de $\vec{IC}_g(\vec{y};t)$) se representará como un código de color y la dirección mediante flechas superpuestas.

Ángulo promedio causal

Con el objeto de poder sintetizar la información generada en los vídeos o secuencias visuales, se ha cuantificado el ángulo causal promedio (con respecto al eje de abscisas) como:

$$\beta_x(t) = \langle \angle \left(\vec{IC}_x(\vec{y};t) \right) \rangle \tag{4.9}$$

tanto para los mapas globales como para los mapas con relaciones locales.

Esta medida cuantifica en promedio, y asumiendo la existencia de un frente de onda, en qué sentido se propagaría. En la Figura 4.13 se representa un ejemplo de $\beta(t)$ para una señal sinusal.

Asumiendo que el comportamiento de la propagación de las ondas es estable, el ángulo $\beta_x(t)$ debe ser cicloestacionario. En la figura se comprueba como, de manera síncrona con el ritmo sinusal, se produce una variación periódica del ángulo. Según se propaga el frente de onda, se comprueba cómo la dirección promedio comienza en 180°, se desplaza hasta los $-90°$ y finaliza en aproximadamente 0°.

Esta medida se puede emplear para identificar etapas estables en la estimulación auricular.

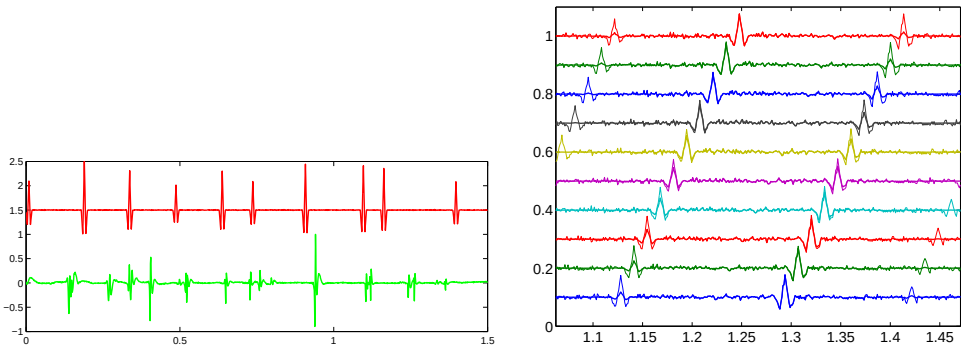
4.5. Resultados con señales sintéticas

Con el fin de validar la metodología empleada, se ha generado una señal sintética que simula una fibrilación auricular. La ecuación que la genera es la siguiente:

$$s_0(t) = (A + \alpha_1 \rho_n) \sum_{n=1}^{N} p\left(t - nT(1 + \alpha_2 \tau_n)\right) + e(t) \tag{4.10}$$

donde $p(t)$ es el patrón base, y $e(t)$, ρ_n y τ_n son procesos estocásticos gaussianos normales para simular, respectivamente, ruido de media, variaciones en amplitud y variaciones temporales. El resto de señales son copias retardadas: $s_i(t) = s_0(t - iT_0)$, $i = 1, \ldots, 10$, donde $T_0 = 10/T$ para que haya continuidad entre $x_{10}(t)$ y $x_1(t)$.

La Figura 4.11 muestra la señal sintética generada mediante el procedimiento descrito y la compara con uno de los registros sensado en el catéter circular. En la Figura 4.11(b) se recoge una ventana de procesado temporal, con las señales del Lasso™. Se ha aplicado un enventanado tipo "hamming" para evitar artefactos no deseados en el espectro. En trazo fino se dibujan las señales sin enventanar, y en trazo grueso la salida del filtro "hamming".

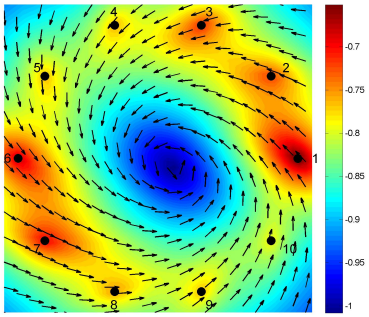


(a) Señal de FA y simulada. (b) Señales del Lasso™.

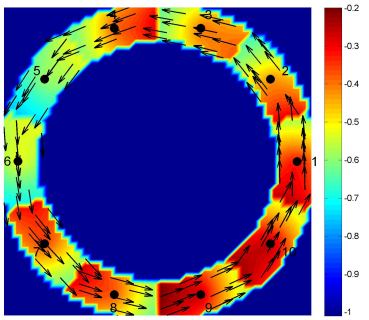
Figura 4.11: Señal sintética y señal real de fibrilación auricular.

Si se aplica el método de generación de mapa causal descrito en el Algoritmo 7 (pág. 136) a estas señales simuladas, se obtiene una secuencia de fotogramas, una de las cuales (valor en un instante del mapa de causalidad) puede verse en la Figura 4.12(a-d), para las configuraciones de mapa global y local y con los métodos PSI y cMultiSVARMA. Como era de esperar, se puede comprobar que existe una relación circular entre las 10 señales del

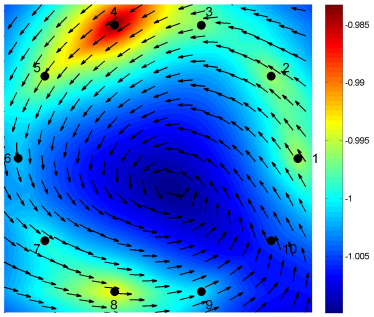
Lasso™. El fotograma se corresponde con el tratamiento de las señales de la Figura 4.11(b).



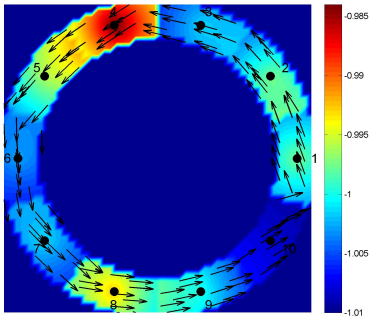
(a) Mapa causal global (PSI).



(b) Mapa causal local (PSI).



(c) Mapa causal global (cMultiSVARMA).



(d) Mapa causal local (cMultiSVARMA).

Figura 4.12: Mapa causal de las señales de FA sintéticas.

Si se representa el esquema local, el resultado es el que puede verse en la Figura 4.12(b,d). Si bien las intensidades del mapa causal, codificado en la escala de color, varían entre los métodos PSI y cMultiSVARMA, la orientación de las relaciones es similar.

4.6. Resultados con señales reales

Esta metodología se ha aplicado en señales auriculares en ritmo sinusal y en fibrilación. Para la primera de ellas, en la Figura 4.13, se muestran los principales parámetros de la relación causal: ángulo e intensidad máxima en función del tiempo. La curva roja indica el módulo de la influencia causal.

La curva negra es $\beta_g(t)$, el ángulo promedio (en grados) de las relaciones causales con respecto al eje de abscisas en cada fotograma, calculado en base al mapa global.

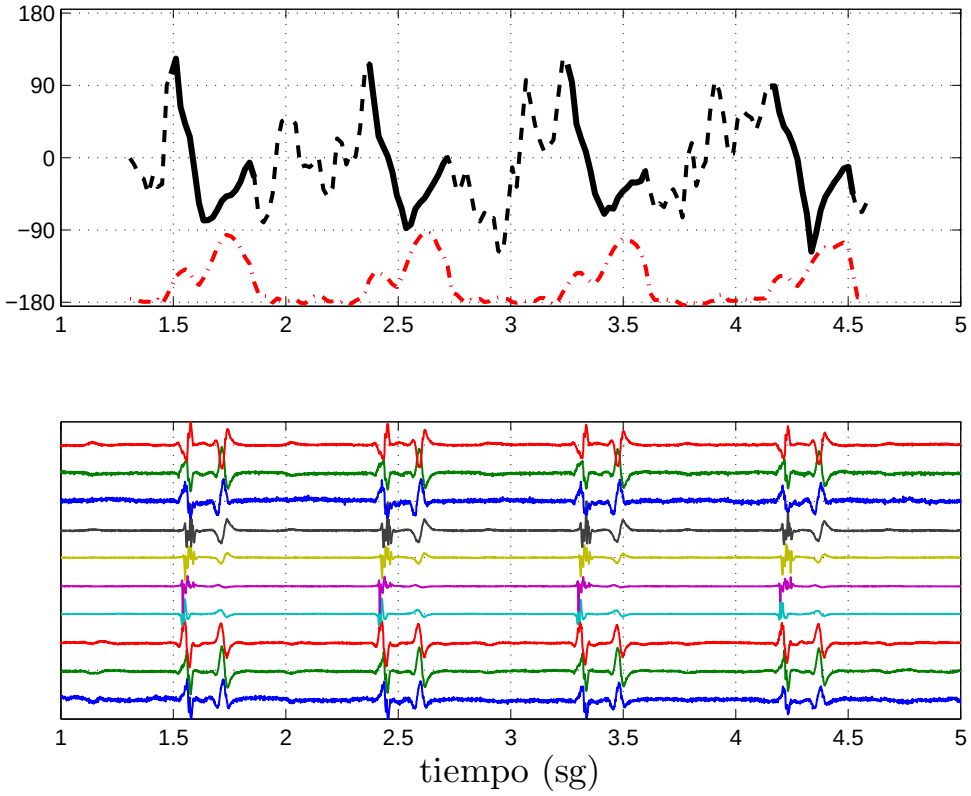


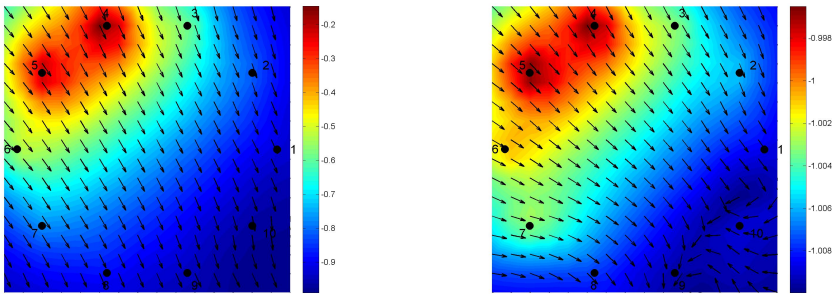
Figura 4.13: Arriba, ángulo promedio (en grados) de la dirección causal $\beta_g(t)$ en negro. Se han resaltado las zonas donde la influencia causal (curva roja punteada) es significativa. Abajo, señales del Lasso™ en la misma escala temporal.

Puede verse que de manera repetida y coincidente con el ciclo sinusal, la variación en el ángulo promedio causal mantiene el mismo patrón. En el tiempo transcurrido entre máximos de las señales auriculares la intensidad causal es irrelevante, por lo que la función del ángulo $\beta_g(t)$ en esos intervalos no aporta ninguna información de interés.

Si se deseara estudiar únicamente los momentos de activación auricular se podría emplear la intensidad en las relaciones causales para identificarlos y filtrarlos, de forma que se mostraría únicamente información relevante.

La Figura 4.14 muestra el mapa de relaciones causales en el momento de máxima influencia causal, correspondiente al segundo 2.6. Como se aprecia

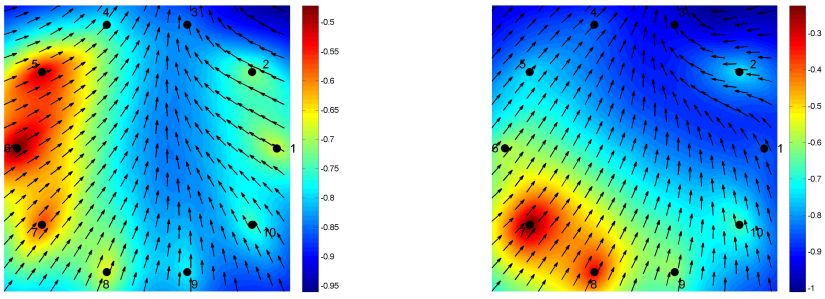
en la Figura 4.13, tiene lugar aproximadamente cuando la dirección promedio es de $-85°$. Tanto los mapas generados con PSI como con cMultiSVARMA ofrecen una información parecida.



(a) RS auricular (PSI). (b) RS auricular (cMultiSVARMA).

Figura 4.14: Mapas causales para EGM de ritmo sinusal en la aurícula.

También se ha representado un fotograma del mapa causal de una fibrilación auricular, recogido en la Figura 4.15. Se pueden apreciar dos frentes de onda en los sensores 1 y 6, que producen una dirección de propagación promedio hacia la posición de los sensores 3 y 4, en la parte superior.



(a) FA (PSI). (b) FA (cMultiSVARMA).

Figura 4.15: Mapas causales de una fibrilación auricular. La propagación parece dirigirse hacia los sensores 3 y 4, en la parte superior de las imágenes.

4.7. Interpretación de los resultados y conclusiones

Aplicando técnicas de causalidad para series temporales es posible determinar la dirección del flujo de información entre dichas señales. Se han utilizado estas técnicas en el estudio de fibrilaciones auriculares y se ha elaborado una representación gráfica de las relaciones causales. A diferencia de otros métodos empleados para el estudio de las FA, como por ejemplo las frecuencias dominantes, los métodos de inferencia causal se asientan en una base teórica que permite modelar el mecanismo de propagación de las señales. Se han empleado dos métodos de inferencia causal, uno basado en medidas espectrales, como es el PSI, y el otro un método causal tipo Granger, desarrollado en la Sección 2.2, el cMultiSVARMA.

Los ejemplos sintéticos y con ritmo sinusal muestran que el mapa causal obtenido es coherente con el sentido de propagación de las señales. En el primer caso, la señal sintética se desplaza de sensor en sensor de una manera circular. Las secuencias de mapas causales corroboran ese hecho. En el caso del estudio auricular en ritmo sinusal, el ángulo promedio causal ofrece un comportamiento cicloestacionario, síncrono con la propagación de las ondas auriculares.

La aplicación de los mapas causales al estudio de fibrilaciones auriculares se presenta prometedora. Personal médico del Hospital Gregorio Marañón, de donde fueron recogidos los registros, ha evaluado esta herramienta confirmando su potencial uso en el análisis de FA y como apoyo en las prácticas de ablación auricular. En la medida en la que se disponga de más información sobre los sensores, como por ejemplo su disposición concreta en la aurícula, así como más medidas simultáneas en diferentes puntos del miocardio, se podrán abordar nuevos trabajos de investigación.

Capítulo 5

Conclusiones y líneas futuras

En esta tesis se ha estudiado el establecimiento de relaciones causales, a partir de asociaciones estadísticas y métodos de aprendizaje máquina. Se han desarrollado nuevos algoritmos para realizar inferencia causal a partir de muestras, tanto para datos discretos como continuos. Finalmente, se han aplicado estos desarrollos algorítmicos a la resolución de problemas relevantes en el campo de la psiquiatría y de la cardiología.

5.1. Conclusiones

La representación de independencias estadísticas en forma de grafo permite establecer relaciones causales entre variables o señales, y orientar los enlaces según el mecanismo que gobierna el flujo de la información. Por consiguiente, un modelo causal aporta más información que las relaciones estadísticas, así como una comprensión más en profundidad del problema.

Durante la última década, trabajos como los de (Pearl, 2009; Spirtes et al., 2000) han dado una sólida fundamentación teórica al nuevo paradigma causal.

Un área que actualmente tiene mucha actividad investigadora es el desarrollo de algoritmia que realice inferencia causal a partir de conjuntos de muestras o de señales. Los métodos difieren, ya sean datos discretos o señales continuas.

En esta tesis, se han realizado las siguientes aportaciones en el desarrollo de nuevos algoritmos de búsqueda causal:

- Para datos discretos, se ha presentado un método novedoso, que emplea una batería de clasificadores para inferir la relaciones causales. Se han empleado dos tipos de clasificadores, el k vecinos más próximos, que ha sido publicado en (de-Prado-Cumplido y Artés-Rodríguez,

2008) (ccKnn), y máquinas de vectores soporte multiclase (ccMSVM). Los resultados confirman que es competitivo sustituir los test de independencia estadística habituales por múltiples clasificadores.

- Con el fin de agilizar el algoritmo de inferencia causal con clasificadores SVM, se ha empleado la salida probabilística de la SVM, además de emplear permutaciones aleatorias para reducir el número de máquinas a entrenar. Se ha comprobado cómo el núcleo de la SVM debe estar adaptado al tipo de datos que se desea analizar.

- Para series temporales, es habitual emplear el concepto de causalidad de Granger. Los métodos de modelado autorregresivo son más eficaces si son dispersos en el espacio de coeficientes. Se han probado varios algoritmos de modelado disperso. También se ha aplicado el método SVARMA (Rojo-Álvarez et al., 2004) para realizar inferencia causal.

- Se ha comprobado cómo la función de coste del algoritmo de inferencia causal cSVARMA es más robusta frente a ruido de tipo impulsivo.

- El cSVARMA realiza el modelado de manera independiente en cada dimensión. Se ha modificado el método para realizar un modelado multivariable, en la línea de (Sánchez-Férnandez et al., 2004).

- Se ha implementado este modelado multivariable basado en máquinas de vectores soporte mediante programación cónica de segundo orden. El algoritmo converge a la solución óptima haciendo un consumo de recursos razonable.

- El método resultante, el cMultiSVARMA se muestra de los mejores en los experimentos realizados. Su función de coste hace que sea más robusto frente a muestras fuera de rango o ruido impulsivo, pero también frente a ruido de baja intensidad.

Los algoritmos desarrollados se han aplicado a dos problemas clínicos. En el Capítulo 3 se detalla el trabajo realizado junto con un equipo médico del Hospital Jiménez Díaz de Madrid, para la búsqueda de los factores relevantes a la hora de predecir la frecuencia de comisión de intentos de suicidio.

- El problema de psiquiatría, publicado en (López-Castromán et al., 2011) (con métodos de inferencia causal clásicos), ha consistido en la generación de un árbol causal para la variable de intentos de suicidio repetido. Se ha contado con una base de datos sociológicos y clínicos del Hospital Universitario Ramón y Cajal, en Madrid, y el Hospital Universitario Lapeyronie, en Montpellier, Francia.

- Se han seleccionado las variables relevantes del estudio mediante técnicas de aprendizaje máquina y de búsqueda causal del "Markov Blanket".

- El método ccKnn ha validado las principales relaciones obtenidas por el método causal clásico y no ha orientado las relaciones más dudosas.

Se ha estudiado en el Capítulo 4 el problema de la determinación de focos de activación durante episodios de fibrilación en la aurícula. Se han analizado las señales de fibrilación con técnicas de inferencia causal, y se ha diseñado un método de representación visual de las mismas.

- Se ha recogido una base de datos de registros de electrogramas, en pacientes con fibrilaciones auriculares, en el servicio de electrofisiología cardiaca del Hospital Universitario Gregorio Marañón de Madrid.

- Se ha estudiado el uso y limitaciones de las actuales herramientas de discriminación de focos de arritmias, basados en el cálculo de frecuencias dominantes.

- Se han aplicado las herramientas de búsqueda causal para series temporales a señales de fibrilación auricular, como una mejor manera de identificar los mecanismos de generación de las arritmias.

- Se ha diseñado una herramienta de representación visual de las relaciones causales (de-Prado-Cumplido y Artés-Rodríguez, 2010). Esta herramienta se puede emplear para verificar la existencia de focos causantes de las fibrilaciones auriculares.

Se ha comprobado cómo en ambas aplicaciones de bioingeniería, los métodos causales constituyen una herramienta fácil de interpretar y discutir por parte del personal médico.

5.2. Líneas futuras de trabajo

El desarrollo de los algoritmos de inferencia causal, y su aplicación a los problemas psiquiátrico y cardiológico, ha abierto nuevas vías de trabajo.

- Se ha comprobado que un equipo multidisciplinar no se compone de un grupo de expertos, cada uno en su área. Es preciso la hibridación de los profesionales implicados en el trabajo. Estudios de grado como Ingeniería Biomédica es un ejemplo de esto. Por ello, es interesante la

implementación de un modelo de generación de taquiarritmias, sobre los que probar los algoritmos, empleando para ello modelos de conducción celular como los descritos en (Kléber y Rudy, 2004).

- Se ha comprobado lo importante que resulta el tipo de clasificador en función de las características de los datos. Para el algoritmo ccKnn, se puede variar el espacio de cálculo de las distancias, definiendo métricas optimizadas a la tipología de los datos. Esto aliviaría la carga computacional del método y se adaptaría a las características de las muestras. En lugar de calcular las distancias euclídeas, habría que optimizar la matriz A en distancia$(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 - \mathbf{x}_2)^T A(\mathbf{x}_1 - \mathbf{x}_2)$. Un intento de mejorar el k vecinos más próximos en este sentido se ha publicado en (Weinberger y Saul, 2009).

- En la línea del punto anterior, y para el modelo de clasificadores causales con SVMs, el ccMSVM, no sería costoso probar nuevas funciones núcleo, por ejemplo tipo "string" (Lodhi et al., 2002). Este ccMSVM con núcleo "string" sería muy interesante para estudios en documentos de texto y en genética. Por ejemplo, las variables obtenidas en (Baca-García et al., 2010) se podrían tratar de jerarquizar de manera causal.

- Puesto que un error orientando un enlace puede variar drásticamente la interpretación del árbol causal, es interesante obtener medidas de confianza sobre los sentidos de las conexiones. Para los algoritmos de datos discretos basados en clasificadores, ccKnn y ccMSVM, sería interesante estudiar si el número de veces que se ha dirigido un enlace en un mismo sentido puede ser empleado como medida de confianza.

- La función de coste del método multivariable cMultiSVARMA incluye una zona de insensibilidad y un tramo lineal. Sin embargo carece de la zona cuadrática intermedia. Sería de interés ampliar el método para que presentara una función de coste robusto de Huber.

- La medida de causalidad tipo Granger se lleva a cabo a partir de las varianzas de los residuos de los modelados ARMA. Adicionalmente, parece razonable pensar que se mejorarían los resultados si se empleara información de los coeficientes. Cuando no hay ruido, métodos como el "Group LASSO", que fuerzan la dispersidad por grupos, anulan los coeficientes correspondientes a la relación no causal. Sin embargo, el ruido rompe la dispersidad de los coeficientes. Los métodos autorregresivos basados en máquinas de vectores soporte, c(Multi)SVARMA,

mantienen más la baja densidad de los coeficientes en problemas ruidosos. Se podrían combinar medidas de los residuos con medidas de dispersidad para afinar los resultados.

- Continuando esta línea, se puede desarrollar una versión espectral de los c(Multi)SVARMA y aprovechar información de la fase para tener en cuenta la información de retardos entre señales.

- Un camino no explorado es emplear, en lugar del criterio de Granger, un esquema basado en el método PC (Spirtes et al., 2000), para series temporales.

- Finalmente, los mapas causales de fibrilación auricular, que han sido evaluados positivamente para identificar vías de propagación auricular, se pueden mejorar desarrollando medidas de cicloestacionariedad de la señal de ángulo promedio causal. De esta forma se harían más fácilmente interpretables los resultados.

Apéndice A

Optimización cónica del "Group LASSO"

La solución del "Group LASSO" (Haufe et al., 2008) para buscar relaciones causales proviene del funcional:

$$\hat{A}^{Glasso} = \arg\min_{A} ||Y - XA||^2 \tag{A.1}$$
$$\text{restringido a } ||A_{11}(p), \ldots, A_{dd}(P)|| + \sum_{i \neq j} || \left(A_{ij}(1), \ldots, A_{ij}(P)\right) || \leq k$$

Para poder emplear esta herramienta, es preciso ajustar el funcional a minimizar de manera que las restricciones sean de tipo cónico o cónico rotado. Este tipo de optimización también es usado en el método cMultiSVARMA. Las restricciones cónicas son:

$$\mathcal{C}_t \equiv x_1 \geq \sqrt{\sum x_j^2} \tag{A.2}$$
$$\text{Rot}\mathcal{C}_t \equiv 2x_1 x_2 \geq \sum x_j^2 \tag{A.3}$$

Por lo tanto se procede a expresar el problema (A.1) de la siguiente manera:

$$Y = X^S A^S \Rightarrow \left(\begin{array}{c} x_1(P+1) \\ x_1(P+2) \\ \vdots \\ x_1(T) \\ \hline x_2(P+1) \\ x_2(P+2) \\ \vdots \\ x_2(T) \end{array}\right) = \left(\begin{array}{cc} X & \mathbf{0} \\ \hline \mathbf{0} & X \end{array}\right) \left(\begin{array}{c} A_{11}^1 \\ A_{21}^1 \\ A_{11}^2 \\ A_{21}^2 \\ \vdots \\ A_{11}^P \\ A_{21}^P \\ \hline A_{12}^1 \\ A_{22}^1 \\ \vdots \\ A_{12}^P \\ A_{22}^P \end{array}\right) \tag{A.4}$$

donde la matriz X consiste en las porciones de señal retardadas:

$$X = \begin{pmatrix} x_1(P) & x_1(P-1) & \cdots & x_1(1) & x_2(P) & x_2(P-1) & \cdots & x_2(1) \\ x_1(P+1) & x_1(P) & \cdots & x_1(2) & x_2(P+1) & x_2(P) & \cdots & x_2(2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1(T-1) & x_1(T-2) & \cdots & x_1(T-P) & x_2(T) & x_2(T-1) & \cdots & x_2(T-P) \end{pmatrix} \tag{A.5}$$

Resulta beneficioso agrupar los coeficientes de la matriz A^S en tres vectores, que se renombrarán como $\mathbf{q}$, $\mathbf{r}$ y $\mathbf{s}$.

$$\begin{array}{lll} \mathbf{q} = (A_{11}^1, A_{22}^1, \ldots, A_{11}^P, A_{22}^P). & \text{Diagonal:} & x_i(t) \to x_i(t), i = 1, 2 \\ \mathbf{r} = (A_{12}^1, A_{12}^2, \ldots, A_{12}^P). & \text{Relación cruzada:} & x_2(t) \to x_1(t) \\ \mathbf{s} = (A_{21}^1, A_{21}^2, \ldots, A_{21}^P). & \text{Relación cruzada:} & x_1(t) \to x_2(t) \end{array} \tag{A.6}$$

Con este cambio de variables, y añadiendo las variables mudas y_i, el funcional a minimizar se convierte en:

$$\begin{aligned} & \min_{A^S} \|Y^S - X^S A^S\|^2 \\ \text{restringido a } & \|\mathbf{q}\| \leq y_1 \\ & \|\mathbf{r}\| \leq y_2 \\ & \|\mathbf{s}\| \leq y_3 \\ & y_1 + y_2 + y_3 \leq k \end{aligned} \tag{A.7}$$

El siguiente paso consiste en linealizar la parte cuadrática del funcional. Para ello se desarrolla la norma cuadrática de (A.7) y se añaden unas nuevas variables mudas t, v y w:

$$\|Y^S - X^S A^S\|^2 = \underbrace{\|X^S A^S\|^2}_{\text{linealizable con programación cónica}} + \underbrace{(Y^S)^T Y^S}_{\text{constante}} - \underbrace{2((X^S)^T Y^S)^T A^S}_{\text{lineal en } A^S} \tag{A.8}$$

Para la transformación de la parte cuadrática remanente, se emplea el siguiente cambio de variables:

$$\begin{aligned} \min_{A^S} \quad & \|X^S A^S\|^2 && \text{(A.9)} \\ & \Rightarrow \\ \min_{A^S} \quad & \|\vec{t}\|^2 && \text{(A.10)} \\ \text{restringido a} \quad & X^S A^S - \vec{t} = \mathbf{0} \\ & \Rightarrow \\ \min_{A^S} \quad & 2v && \text{(A.11)} \\ \text{restringido a} \quad & X^S A^S - \vec{t} = \mathbf{0} \\ & w = 1 \\ & v \geq 0 \\ & \|\vec{t}\|^2 \leq 2vw \end{aligned}$$

Finalmente, introduciendo las Ecuaciones A.7-A.11 en la Ecuación (A.1), el problema es resoluble mediante programación cónica de segundo orden (SOCP). El funcional resultante es el mostrado en la siguiente ecuación:

$$\begin{aligned} \min_{A^S} \quad & ((X^S)^T Y^S) A^S + 2v + c_{fix} && \text{(A.12)} \\ \text{r.a.} \quad & \|\mathbf{q}\| \leq y_1 \\ & \|\mathbf{r}\| \leq y_2 \\ & \|\mathbf{s}\| \leq y_3 \\ & y_1 + y_2 + y_3 \leq k \\ & X^S A^S - \vec{t} = \mathbf{0} \\ & w = 1 \\ & v \geq 0 \\ & \|\vec{t}\|^2 \leq 2vw \end{aligned}$$

donde se han agrupado los términos constantes en $c_{fix} = (Y^S)^T Y^S$.

Existen varios paquetes de software disponibles para resolver estos problemas de optimización, como por ejemplo (Sturm, 1998; Mosek y ApS, 2010).

Apéndice B

Variables del problema psiquiátrico

El siguiente Cuadro B.1 recoge el listado completa de las variables de la base de datos de los hospitales de Madrid y Montpellier.

#	Etiqueta	#	Etiqueta	#	Etiqueta
1	codemontp	2	codegeneve	3	id
4	idem	5	fecha	6	source01
7	valoraciones	8	o_val	9	tipo
10	int_historia	*11*	*his_fam_s*	*12*	*age*
13	sex	*14*	*ri_agent*	*15*	*ri_conci*
16	*ri_lesio*	*17*	*ri_rever*	*18*	*ri_tto*
19	*re_ubica*	*20*	*re_perso*	*21*	*re_descu*
22	*re_acces*	*23*	*re_retra*	*24*	*aislamie*
25	*tiempo*	*26*	*precauci*	*27*	*ayuda*
28	*acto_fin*	*29*	*prepa_in*	*30*	*nota*
31	*comunic*	*32*	*propo_ii*	*33*	*expect*
34	*letal*	*35*	*seriedad*	*36*	*actit_mu*
37	*inter_me*	*38*	*predem*	39	weis_ri
40	weis_re	41	rrrs_ratio	42	sis_plan
43	sis_exp	44	sis_tot	45	reac_iii
46	vis_muer	47	num_int	48	alco_int
49	drug_int	50	fiabl_iv	51	confusio
52	*des*	53	des_emon	54	oposic
55	bloq	*56*	*est_civ*	*57*	*hijos*
58	*hijos1*	*59*	*tabaco*	60	fecha_int
61	sui_mens	62	ab41	63	ab43
64	ab50	65	mens_men_other	*66*	*niv_edu*

#	Etiqueta	#	Etiqueta	#	Etiqueta
67	*prof*	*68*	*sit_lab*	*69*	*IS*
70	*age1a*	*71*	*vio*	*72*	*dd_mood*
73	*depre_bip*	*74*	*dd_anxiety*	*75*	*dd_ocd*
76	*dd_al_drug*	*77*	*dd_psychosis*	*78*	*dd_eating*
79	dd_som	80	dd_adjust	81	mal_1_max1
82	mal_2_max1	83	mal_4_max1	84	bis1
85	bis2	86	bis3	87	bis4
88	bis5	89	bis6	90	bis7
91	bis8	92	bis9	93	bis10
94	bis11	95	bis12	96	bis13
97	bis14	98	bis15	99	bis16
100	bis17	101	bis18	102	bis19
103	bis20	104	bis21	105	bis22
106	bis23	107	bis24	108	bis25
109	bis26	110	bis27	111	bis28
112	bis29	113	bis30	114	bis31
115	bis32	116	bis33	117	bis34
118	@5httpr_recod	119	htt_sl	120	@_httpa_promoteur
121	@_httpb_promoteur	122	@_httpb_mutation	123	@_httpa_mutation
124	intr?_2_recod	125	i_2_12	126	ht_t_i2
127	gabra3_recod	128	gabra3_c	129	gabra3_a
130	gabra3_d	131	gabra_11	132	gabra_12
133	gabra_13	134	@5ht2a_1438_recod	135	@5ht2a102_recod
136	@5ht2a1354_recod	137	d3dr_recod	138	drd2_recod
139	ssat1_recod	140	maoa_recod		

Cuadro B.1: Listado completo de variables de la base de datos. En cursiva, las variables empleadas en el estudio.

Las variables se agrupan de la siguiente manera:

- Las 9 primeras variables son de identificación. La variable 9 indica si son españoles, franceses o del grupo de control.
- De la 11 a la 13, así como de la 56 a la 66 son de tipo socio-económico.
- De la 14 a la 38 son preguntas de escala de los test de Beck y de Weissman y Worden. Se combinan para generar de la 39 a la 55.
- Las variables 70 a 80 son diversos trastornos psiquiátricos.

- De la 84 a la 117 recogen preguntas de cuestionario previo.
- De las variables 118 a la 140 son de tipo genético.

Apéndice C

Secuencias de mapas causales

La secuencia de imágenes de la Figura C.1 es una selección de fotogramas de un mapa causal durante un episodio de fibrilación auricular. Las relaciones causales se han calculado con el método cMultiSVARMA.

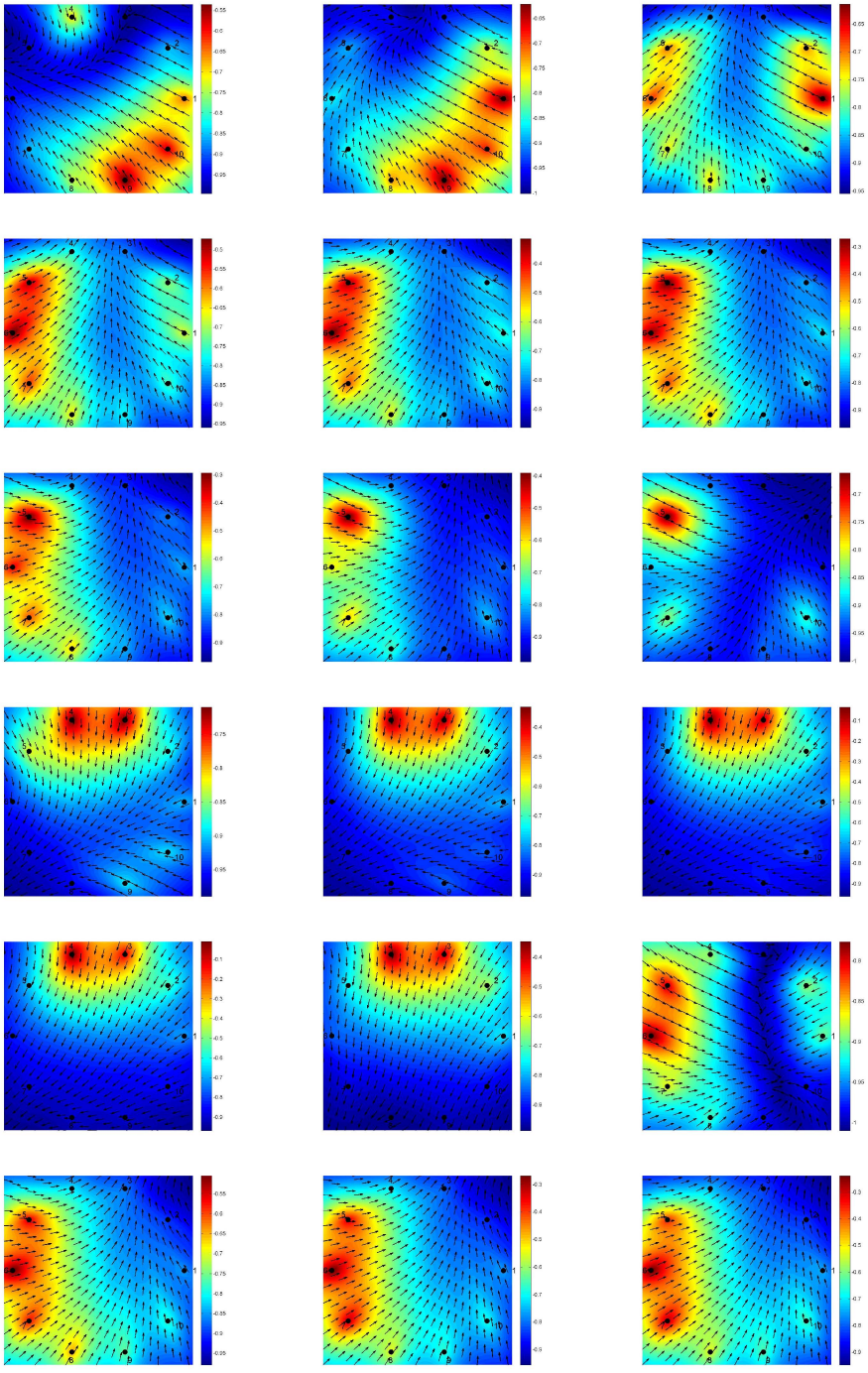


Figura C.1: Secuencia de mapas causales para una fibrilación auricular.

Bibliografía

Constantin F. Aliferis, Ioannis Tsamardinos, y Alexander Statnikov. HITON: A novel Markov Blanket algorithm for optimal variable selection. *Annual AMIA Symposium proceedings*, páginas 21–25, 2003. ISSN 1559-4076.

Etienne Aliot, Rémi Nitzsché, y Alain Ripart. Arrhythmia detection by dual-chamber implantable cardioverter defibrillators. A review of current algorithms. *Europace*, 6(4):273–286, 2004.

Erin L. Allwein, Robert E. Schapire, y Yoram Singer. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal Machine Learning Research*, 1:113–141, 2001. ISSN 1532-4435.

Nicola Ancona, Daniele Marinazzo, y Sebastiano Stramaglia. Radial basis function approach to nonlinear Granger causality of time series. *Physical Review E*, 70(056221), 2004.

Donald W. K. Andrews y Moshe Buchinsky. A three-step method for choosing the number of bootstrap repetitions. *Econometrica*, 68(1):23–51, 2000. ISSN 00129682.

Leonardo Angelini, Daniele Marinazzo, Mario Pellicoro, y Sebastiano Stramaglia. Causality and communities in neural networks. En *15th European Symposium On Artificial Neural Networks Advances in Computational Intelligence and Learning (ESANN)*, páginas 459–464. 2007.

Ángel Arenal-Maíz, Mercedes Ortiz-Patón, Rafael Peinado, Jose L. Merino, Aurelio Quesada, Felipe Atienza, Arcadio García-Alberola, Jose Ormaetxe, Eduardo Castellanos, Juan C. Rodriguez, Nicasio Pérez, Javier García, Luis Boluda, Mario de Prado-Cumplido, y Antonio Artés-Rodríguez. Differentiation of ventricular and supraventricular tachycardias based on the analysis of the first postpacing interval after sequential anti-tachycardia pacing in implantable cardioverter-defibrillator patients. *Heart Rhythm*, 4(3):1547–5271, 2007.

Felipe Atienza, Jesús Almendral, José Jalife, Sharon Zlochiver, Robert Ploutz-Snyder, Esteban G. Torrecilla, Ángel Arenal-Maíz, Jérôme Kalifa, Francisco Fernández-Avilés, y Omer Berenfeld. Real-time dominant frequency mapping and ablation of dominant frequency sites in atrial fibrillation with left-to-right frequency gradients predicts long-term maintenance of sinus rhythm. *Heart Rhythm*, 6(1):33–40, 2009. ISSN 1547-5271.

Enrique Baca-García, María de las Mercedes Pérez-Rodríguez, Ignacio Basurte, Jerónimo Saiz, José Miguel Leiva-Murillo, Mario de-Prado-Cumplido, Ricardo Santiago-Mozos, Antonio Artés-Rodríguez, y José de León. Using data mining to explore complex clinical decisions: A study of hospitalization after a suicide attempt. *Journal of Clinical Psychiatry*, 67:1124–1132, 2006.

Enrique Baca-García, María de las Mercedes Pérez-Rodríguez, Dolores Saiz-González, Ignacio Basurte-Villamor, Jerónimo Saiz-Ruiz, José Miguel Leiva-Murillo, Mario de-Prado-Cumplido, Ricardo Santiago-Mozos, Antonio Artés-Rodríguez, y José de León. Variables associated with familial suicide attempts in a sample of suicide attempters. *Progress Neuropsychopharmacology & Biological Psychiatry*, 31(6):1312–1316, 2007a.

Enrique Baca-García, Concepción Vaquero-Lorenzo, María de las Mercedes Pérez-Rodríguez, Mónica Gratacós, Mónica Bayés, Ricardo Santiago-Mozos, José Miguel Leiva-Murillo, Mario de Prado-Cumplido, Antonio Artés-Rodríguez, Antonio Ceverino, Carmen Díaz-Sastre, Pablo Fernandez-Navarro, Javier Costas, José Fernandez-Piqueras, Montserrat Diaz-Hernandez, José de Leon, Enrique Baca-Baldomero, Jerónimo Saiz-Ruiz, J. John Mann, Ramin V. Parsey, Angel Carracedo, Xavier Estivill, y María A. Oquendo. Nucleotide variation in central nervous system genes among male suicide attempters. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 153(1):208–213, 2010. ISSN 1552-485X.

Enrique Baca-García, María de las Mercedes Pérez-Rodríguez, Ignacio Basurte-Villamor, Jorge López-Castromán, A. L. Fernández del Moral, M. A. Jiménez-Arriero, J. L. González de Rivera, Jerónimo Saiz-Ruiz, José Miguel Leiva-Murillo, Mario de Prado-Cumplido, Ricardo Santiago-Mozos, Antonio Artés-Rodríguez, María A. Oquendo, y José de Leon. Diagnostic stability and evolution of bipolar disorder in clinical practice: A prospective cohort study. *Acta Psychiatrica Scandinavica*, 115:473–480, 2007b.

Aaron T. Beck, Harvey L.P. Resnik, y Dan J. Lettieri. *The Prediction of Suicide*. Charles Press Publishers, 1974. ISBN 978-0913486139.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2 edición, 2007. ISBN 978-0387310732.

Stephen Boyd y Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 978-0521833783.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. ISSN 0885-6125.

Christopher J. C. Burges. A tutorial on Support Vector Machines for pattern recognition. *Knowledge Discovery and Data Mining*, 2(2):121–167, 1998.

Chih-Chung Chang y Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011.

Tianjiao Chu, David Danks, y Clark Glymour. Data driven methods for nonlinear Granger causality: Climate teleconnection mechanisms. Informe Técnico 116, Carnegie Mellon University, Dpt. of Philosophy, 2005.

William J. Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, Inc., 3 edición, 1999. ISBN 978-0521833783.

Mario de-Prado-Cumplido, Ángel Arenal-Maíz, Mercedes Ortiz-Patón, y Antonio Artés-Rodríguez. SVM classification of sparse set of 1:1 ventricular and supraventricular tachycardia. En *The First International Workshop on Biosignal Processing and Classification (BPC, ICINCO)*. Institute for Systems and Technologies of Information, Control and Communication, Barcelona, 2005.

Mario de-Prado-Cumplido y Antonio Artés-Rodríguez. Búsqueda de relaciones causales en señales de fibrilación auricular. En *28 Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB)*. 2010.

Mario de-Prado-Cumplido y Antonio Artés-Rodríguez. Discovery of causation direction by machine learning techniques. En *Neural Information Processing Systems (NIPS) Workshop on Causality*. 2008.

Thomas G. Dietterich y Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995. ISSN 1076-9757.

Dale Dubin. *Rapid Interpretation of EKG's*. Cover Publishing, 6 edición, 2000. ISBN 978-0912912066.

Bradley Efron y Robert J. Tibshirani. *An introduction to the Bootstrap*. Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, 1993. ISBN 978-0412042317.

Gerald Fischer, Markus Stühlinger, Claudia N. Nowak, Leonhard Wieser, Bernhard Tilg, y Florian Hintringer. On computing dominant frequency from bipolar intracardiac electrograms. *IEEE Transactions on Biomedical Engineering*, 54(1):165–169, 2007. ISSN 0018-9294.

Richard R. Fletcher, Kelly Dobson, Matthew S. Goodwin, Hoda Eydgahi, Oliver Wilder-Smith, David Fernholz, Yuta Kuboyama, Elliott B. Hedman, Ming-Zher Poh, y Rosalind W. Picard. iCalm: Wearable sensor and network architecture for wirelessly communicating and logging autonomic activity. *Information Technology in Biomedicine, IEEE Transactions on*, 14(2):215–223, 2010. ISSN 1089-7771.

Jerome H. Friedman, Trevor Hastie, y Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Benoit Ghaye, David Szapiro, Jean-Nicolas Dacher, Luz-Maria Rodriguez, Carl Timmermans, David Devillers, y Robert F. Dondelinger. Percutaneous ablation for atrial fibrillation: The role of crosssectional imaging. *RadioGraphics*, 23:19–33, 2003.

Neil J. Gordon, David J. Salmond, y Adrian F.M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F, Radar and Signal Processing*, 140(2):107–113, 1993.

Clive Granger. Investigating causal relations by econometric models and cross spectral methods. *Econometrica; Journal of the Econometric Society*, 37(3):424–438, 1969.

Arthur Gretton, Olivier Bousquet, Alexander Smola, y Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. Informe técnico, Max Planck Institute for Biological Cybernetics, 2005.

Isabelle Guyon, Constantine Aliferis, y André Elisseeff. Causal feature selection. Informe técnico, Clopinet, California, 2007.

Stefan Haufe, Klaus-Robert Müller, Guido Nolte, y Nicole Krämer. Sparse causal discovery in multivariate time series. En Isabelle Guyon, Dominik Janzing, y Bernhard Schölkopf, editores, *JMLR Workshop and Conference Proceedings*, tomo 1, páginas 1–16. 2008.

David Heckerman, Dan Geiger, y David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

Ralf Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, Cambridge, MA., 2002. ISBN 0-262-08306-X.

José Miguel Hernández-Lobato, Pablo Morales-Mombiela, y Alberto Suárez. Gaussianity measures for detecting the direction of causal time series. En Toby Walsh, editor, *22th International Joint Conference on Artificial Intelligence (IJCAI)*, páginas 1318–1323. 2011. ISBN 978-1-57735-516-8.

Aaron B. Hesselson. *Simplified Interpretation of ICD Electrograms*. Wiley-Blackwell, 2004. ISBN 978-1405127318.

Katerina Hlavackova-Schindler y Pablo F. Verdes. Computational intelligence approaches to causality detection. En *ESANN 2007*, páginas 433–444. 2007.

Patrik O. Hoyer, Dominik Janzing, Joris Mooij, Jonas Peters, y Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. En *22th Annual Conference on Neural Information Processing Systems (NIPS'08)*, páginas 689–696. 2009.

Patrik O. Hoyer, Shohei Shimizu, y Antti Kerminen. Estimation of linear, non-Gaussian causal models in the presence of confounding latent variables. En *Third European Workshop on Probabilistic Graphical Models (PGM)*, páginas 155–162. Prague, Czech Republic, 2006.

Sanqing Hu, Guojun Dai, Gregory A. Worrell, Qionghai Dai, y Hualou Liang. Causality analysis of neural connectivity: Critical examination of existing methods and advances of new methods. *IEEE Transactions on Neural Networks*, 22(6):829–844, 2011.

José Jalife, Mario Delmar, Jorge Davidenko, y Justus Anumonwo. *Basic Cardiac Electrophysiology for the Clinician*. Futura/Wiley-Blackwell, 1999. ISBN 978-0879934170.

Stiliyan N. Kalitzin, Jaime Parra, Demetrios N. Velis, y F.H. Lopes da Silva. Quantification of unidirectional nonlinear associations beetween multidimensional signals. *IEEE Transactions on Biomedical Engineering*, 54(3):454–461, 2007.

André G. Kléber y Yoram Rudy. Basic mechanisms of cardiac impulse propagation and associated arrhythmias. *Physiological Reviews*, 84(2):431–488, 2004.

Ron Kohavi y George John. *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Springer, 1998. ISBN 978-0792381969.

José Miguel Leiva-Murillo. *Extracción de Características mediante Técnicas Basadas en Teoría de la Información*. Tesis Doctoral, Universidad Carlos III de Madrid, 2007. Tutor: Antonio Artés Rodríguez.

Wei Liao, Daniele Marinazzo, Zhengyong Pan, Qiyong Gong, y Huafu Chen. Kernel Granger causality mapping effective connectivity on fMRI data. *IEEE Transactions on Medical Imaging*, 28(11):1825–1835, 2009. ISSN 0278-0062.

Hsuan-Tien Lin, Chih-Jen Lin, y Ruby Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68:267–276, 2007. ISSN 0885-6125.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, y Chris Watkins. Text classification using string kernels. *Machine Learning Research*, 2:419–444, 2002.

Jorge López-Castromán, María de las Mercedes Pérez-Rodríguez, Isabelle Jaussent, Ana Lucía Alegría, Antonio Artés-Rodríguez, Peter Freed, Sébastien Guillaume, Fabrice Jollant, José Miguel Leiva-Murillo, Alain Malafosse, María A. Oquendo, Mario de-Prado-Cumplido, Jerónimo Saiz-Ruiz, Enrique Baca-García, Philippe Courtet, y European Research Consortium for Suicide (EURECA). Distinguishing the relevant features of frequent suicide attempters. *Journal of Psychiatric Research*, 45(5):619–625, 2011.

Ching-Hsing Luo y Yoram Rudy. A dynamic model of the cardiac ventricular action potential. *Circulation Research*, 74:1097–1113, 1994.

Zhi-Quan Luo y Wei Yu. An introduction to convex optimization for communications and signal processing. *IEEE Journal on Selected Areas in Communications*, 24(8):1426–1438, 2006.

Moussa Mansour, Ravi Mandapati, Omer Berenfeld, Jay Chen, Faramarz H. Samie, y José Jalife. Left-to-right gradient of atrial frequencies during acute atrial fibrillation in the isolated sheep heart. *Circulation*, 103(21):2631–2636, 2001.

Daniele Marinazzo, Wei Liao, Huafu Chen, y Sebastiano Stramaglia. Nonlinear connectivity by Granger causality. *Neuroimage*, 58(2):330–338, 2011.

Daniele Marinazzo, Mario Pellicoro, y Sebastiano Stramaglia. Nonlinear parametric model for Granger causality of time series. *Physical Review E*, 73(066216), 2006.

Andrew Moore y Weng keen Wong. Optimal reinsertion: A new search operator for accelerated and more accurate Bayesian network structure learning. En *Proceedings of the 20th International Conference on Machine Learning (ICML)*, páginas 552–559. AAAI Press, 2003.

Julie A. Morris y Martin J. Gardner. Statistics in medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British Medical Journal*, 296:1313–1316, 1998.

Mosek y ApS. The MOSEK optimization tools version 6.0. 2010. Disponible en la URL: http://www.mosek.com (consultado en 2010-11-02).

Nitai D. Mukhopadhyay y Snigdhansu Chatterjee. Causality and pathway search in microarray time series experiment. *Bioinformatics*, 23(4):442–449, 2007.

Yuval Nardi y Alessandro Rinaldo. Autoregressive process modeling via the LASSO procedure. *Journal of Multivariate Analysis*, 102(3):528–549, 2011.

Stanley Nattel, Brett Burstein, y Dobromir Dobrev. Atrial remodeling and atrial fibrillation. Mechanisms and implications. *Circulation: Arrhythmia and Electrophysiology*, 1:62–73, 2008.

Jason Ng, Alan H. Kadish, y Jeffrey J. Goldberger. Effect of electrogram characteristics on the relationship of dominant frequency to atrial activation rate in atrial fibrillation. *Heart Rhythm*, 3:1295–1305, 2006.

Guido Nolte, Andreas Ziehe, Nicole Krämer, Florin Popescu, y Klaus-Robert Müller. Comparison of Granger causality and phase slope index. En Isabelle Guyon, Dominik Janzing, y Bernhard Schölkopf, editores, *Causality: Objectives and Assessment*, tomo 6 de *JMLR Workshop and Conference Proceedings*, páginas 267–276. 2010.

Guido Nolte, Andreas Ziehe, Vadim V. Nikulin, Alois Schlögl, Nicole Krämer, Tom Brismar, y Klaus-Robert Müller. Robustly estimating the flow direction of information in complex physical systems. *Physical Review Letters*, 100(23):234101, 2008.

María A. Oquendo, Enrique Baca-García, Antonio Artés-Rodríguez, Fernando Pérez-Cruz, H.C. Galfalvy, H. Blasco-Fontecilla, D. Madigan, y N. Duan. Hypothesis generation in the 21st century, 2012. Aceptado para su publicación en Molecular Psychiatry.

World Health Organization. Global health risks. Informe técnico, Organización Mundial de la Salud, 2009.

David Owens, Judith Horrocks, y Allan House. Fatal and non-fatal repetition of self-harm: Systematic review. *The British Journal of Psychiatry*, 181(3):193–199, 2002.

Carlo Pappone y Vincenzo Santinelli. Ablación de la fibrilación auricular: ¿dónde estamos? *Revista Española de Cardiología*, 62(10):1087–1091, 2009.

Shyamal Patel, Konrad Lorincz, Richard Hughes, Nancy Huggins, John Growdon, David Standaert, Metin Akay, Jennifer Dy, Matt Welsh, y Paolo Bonato. Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 13(6):864–873, 2009. ISSN 1089-7771.

Judea Pearl. *Causality. Models, Reasoning and Inference*. Cambridge University Press, 2000.

Judea Pearl. *Causality. Models, Reasoning and Inference*. Cambridge University Press, 2 edición, 2009.

Fernando Pérez-Cruz, Gustavo Camps-Valls, Emilio Soria-Olivas, Juan José Pérez-Ruixo, Aníbal R. Figueiras-Vidal, y Antonio Artés-Rodríguez. Multi-dimensional function approximation and regression estimation. En *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, páginas 757–762. Springer-Verlag, London, UK, 2002. ISBN 3-540-44074-7.

Jonas Peters, Dominik Janzing, Arthur Gretton, y Bernhard Schölkopf. Detecting the direction of causal time series. En *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, páginas 801–808. ACM, New York, NY, USA, 2009. ISBN 978-1-60558-516-1.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. En *Advances in Large Margin Classifiers*, páginas 61–74. MIT Press, 1999.

Ulrike Richter, Luca Faes, Alessandro Cristoforetti, Michela Masè, Flavia Ravelli, Martin Stridh, y Leif Sörnmo. A novel approach to propagation pattern analysis in intracardiac atrial fibrillation signals. *Annals of Biomedical Engineering*, 39:310–323, 2011. ISSN 0090-6964.

Robert W. Robinson. Counting unlabeled acyclic digraphs. En Charles Little, editor, *Combinatorial Mathematics V*, tomo 622 de *Lecture Notes in Mathematics*, páginas 28–43. Springer Berlin / Heidelberg, 1977. ISBN 978-3-540-08524-9.

José-Luis Rojo-Álvarez, Manel Martínez-Ramón, Mario de-Prado-Cumplido, Antonio Artés-Rodríguez, y Aníbal Figueiras-Vidal. Support vector method for robust ARMA system identification. *IEEE Transactions on Signal Processing*, 52(1):155–164, 2004. ISSN 1053-587X.

Michael G. Rosenblum, Laura Cimponeriu, Anastasios Bezerianos, Andreas Patzak, y Ralf Mrowka. Identification of coupling direction: Application to cardiorespiratory interaction. *Physical Review E*, 65, 2002.

Volker Roth y Bernd Fischer. The group-LASSO for generalized linear models: Uniqueness of solutions and efficient algorithms. En *ICML '08: Proceedings of the 25th international conference on Machine learning*, páginas 848–855. ACM, New York, USA, 2008. ISBN 978-1-60558-205-4.

Stanley Rush y Hugh Larsen. A practical algorithm for solving dynamic membrane equations. *IEEE Transactions on Biomedical Engineering*, 25:389–392, 1978.

Matilde Sánchez-Férnandez, Mario de-Prado-Cumplido, Jerónimo Arenas-García, y Fernando Pérez-Cruz. SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems. *IEEE Transactions on Signal Processing*, 52(8):2298–2307, 2004.

Frida Sandberg, Martin Stridh, y Leif Sörnmo. Frequency tracking of atrial fibrillation using hidden Markov models. *IEEE Transactions on Biomedical Engineering*, 55(2):502–511, 2008. ISSN 0018-9294.

Prashanthan Sanders, Omer Berenfeld, Mélèze Hocini, Pierre Jaïs, Ravi Vaidyanathan, Li-Fern Hsu, Stéphane Garrigue, Yoshihide Takahashi, Martin Rotter, Fréderic Sacher, Christophe Scavée, Robert Ploutz-Snyder,

José Jalife, y Michel Haïssaguerre. Spectral analysis identifies sites of high-frequency activity maintaining atrial fibrillation in humans. *Circulation*, 112:789–797, 2005. ISSN 0009-7322.

Louis L. Scharf. *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*. Addison-Wesley, 1991. ISBN 978-0201190389.

Björn Schelter, Matthias Winterhalder, y Jens Timmer, editores. *Handbook of Time Series Analysis*. Wiley-VCH, 2006.

Alois Schlögl. The time series analysis toolbox for Octave and MATLAB version 2.43. 2005. Disponible en la URL: http://biosig.sourceforge.net.

Alois Schlögl. A comparison of multivariate autoregressive estimators. *Signal Processing*, 86(9):2426–2429, 2006.

Bernhard Schölkopf y Alexander J. Smola. *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. The MIT Press, 2001. ISBN 978-0262194754.

Jun Shao y Dongsheng Tu. *The Jackknife and Bootstrap*. Springer Series in Statistics. Springer, 1995.

Kai-Quan Shen, Chong-Jin Ong, Xiao-Ping Li, y Einar P. Wilder-Smith. Feature selection via sensitivity analysis of SVM probabilistic outputs. *Machine Learning*, 70:1–20, 2008. ISSN 0885-6125.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, y Antii Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

Leif Sörnmo y Pablo Laguna. *Bioelectrical Signal Processing in Cardiac and Neurological Applications*. Elsevier Academic Press, 2005. ISBN 978-0124375529.

Peter Spirtes, Clark Glymour, y Richard Scheines. *Causation, Prediction and Search*. The MIT Press, 2000.

Jos F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. 1998. Disponible en la URL: http://sedumi.ie.lehigh.edu/.

Xiaohai Sun y Dominik Janzing. Learning causality by identifying common effects with kernel-based dependence measures. En *Proceedings of the 15th European Symposium on Artificial Neural Networks (ESANN)*, páginas 453–458. 2007.

Xiaohai Sun, Dominik Janzing, y Bernhard Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, páginas 1–11, 2005.

Xiaohai Sun, Dominik Janzing, y Bernhard Schölkopf. Inferring causal directions by evaluating the complexity of conditional distributions. En *Neural Information Processing Systems (NIPS) Workshop on Causality and Feature Selection*. 2006.

Xiaohai Sun, Dominik Janzing, y Bernhard Schölkopf. Distinguishing between cause and effect via kernel-based complexity measures for conditional distributions. En *Proceedings of the 15th European Symposium on Artificial Neural Networks (ESANN)*, páginas 441–446. 2007a.

Xiaohai Sun, Dominik Janzing, Bernhard Schölkopf, y Kenji Fukumizu. A kernel-based causal learning algorithm. En *Proceedings of the 24th International Conference on Machine Learning (ICML)*, páginas 855–862. 2007b.

Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1994.

Michael E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal Machine Learning Research*, 1:211–244, 2001. ISSN 1532-4435.

Michael E. Tipping. An efficient MATLAB implementation of the sparse Bayesian modelling algorithm. Versión 2. 2009. Disponible en la URL: http://www.vectoranomaly.com.

Ioannis Tsamardinos, Laura E. Brown, y Constantin F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 2006.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer–Verlag, New York, 1995.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998. ISBN 978-0471030034.

Bernard Victorri. Numerical integration in the reconstruction of cardiac action potentials using Hodgkin-Huxley-type models. *Computers and Biomedical Research*, 18:10–23, 1985.

Frank M. Weber, Christopher Schilling, Gunnar Seemann, Armin Luik, Claus Schmitt, Cristian Lorenz, y Olaf Dössel. Wave-direction and conduction-velocity analysis from intracardiac electrograms; a single-shot technique. *IEEE Transactions on Biomedical Engineering*, 57(10):2394–2401, 2010. ISSN 0018-9294.

Kilian Q. Weinberger y Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009. ISSN 1532-4435.

Avery D. Weissman y J. William Worden. *The Prediction of Suicide*, capítulo Risk-Rescue Rating in Suicide Assessment. Charles Press, Philadelphia, 1974.

Changwon Yoo y Gregory Cooper. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. *Artificial Intelligence in Medicine*, 31(2):169–182, 2004.

Ming Yuan y Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

Michael Zibulevsky y Michael Elad. L1-L2 optimization in signal and image processing. *IEEE Signal Processing Magazine*, 27(3):76–88, 2010.

Abdelhak M. Zoubir y D. Robert Iskander. *Bootstrap Techniques for Signal Processing*. Cambridge University Press, 2004. ISBN 978-0521831277.