VI Congreso de Jóvenes Investigadores de la RSME - 6CJI 2023

León, 6 al 10 de febrero de 2023

Sesión Especial en

TEORÍA DE LA PROBABILIDAD

VI Congreso de Jóvenes Investigadores de la RSME - 6CJI 2023

León, 6 al 10 de febrero de 2023

Programa del 6CJI

	Lunes	Martes	Miércoles	Jueves	Viernes
9:00 - 10:00	**Bienvenida** Salón de Actos	**Plenaria** Salón de actos	**Plenaria** Salón de actos	**Plenaria** Salón de actos	**Plenaria** Salón de actos
10:00 - 11:00	**Plenaria** Salón de actos	**S. Paralelas** M1	**S. Paralelas** X1	**S. Paralelas** J1	**S. Paralelas** V1
11:00 - 12:00	**Café** Hall de la Escuela de Ingenierías				
12:00 - 13:00	**Charla** Salón de Actos	**S. Paralelas** M2	**S. Paralelas** X2	**S. Paralelas** J2	**S. Paralelas** V2
13:00 - 14:00	**S. Paralelas** L1				**Clausura** Salón de Actos
14:00 - 15:30	**Comida** Cafetería Dani y José				
15:30 - 16:30	**Junta RSME** Salón de actos	**Plenaria** Salón de actos	**Visita Guiada**	**Plenaria** Salón de actos	
16:30 - 17:30	**Mesa Redonda** Salón de actos	**Mesa Redonda** Salón de actos		**Café** Hall	
17:30 - 18:30	**Café** Hall	**Café** Hall		**S. Paralelas** J3	
18:30 - 20:00	**S. Paralelas** L2	**S. Paralelas** M3		**Charla** Salón de actos	

VI Congreso de Jóvenes Investigadores de la RSME - 6CJI 2023

León, 6 al 10 de febrero de 2023

Programa de la Sesión Especial en Probabilidad

Martes 7 de febrero		
09:00 a 10:00	Conferencia pleanaria	
10:00 a 11:00	**S.Paralelas M1: Procesos Estocásticos**	
	10:00 a 10:30	Trabajo: Procesos de ramificación defectuosos en ambiente variable Autores: *C.Minuesa* y G.Kersting
	10:30 a 11:00	Trabajo: Un paseo aleatorio entre modelos epidemiológicos Autora: *G.Binotto*
11:00 a 11:30	Pausa Café	
11:30 a 13:30	**S.Paralelas M2: Probabilidad y Aplicaciones**	
	11:30 a 12:00	Trabajo: On approximate validation of models: a Kolmogorov−Smirnov based approach Autores: E.del Barrio, *H.Inouzhe* y C.Matrán
	12:00 a 12:30	Trabajo: Coeficientes de asimetría estadística asintóticamente normales Autores: M.Iturrate-Bobes y *R.Pérez-Fernández*
	12:30 a 13:00	Trabajo: Making data fair through optimal transport Autores: *P.Gordaliza*, E.del Barrio, F.Gamboa, H.Inouzhe y J-M.Loubes
	13:00 a 13:30	Trabajo: Un análisis de robustez Bayesiana multivariante con enfoque en principios de primas Autores: F.Ruggeri, *M.Sánchez-Sánchez* y A.Suárez-Llorens
13:30 a 15:30	Comida	
15:30 a 16:30	Conferencia Plenaria	
16:30 a 17:45	Mesa Redonda	
17:45 a 18:15	Pausa Café	
18:15 a 19:45	**S.Paralelas M3: Probabilidades Imprecisas**	
	18:15 a 18:45	Trabajo: Conjuntos credales representables mediante intervalos de probabilidad alcanzables y funciones de creencia Autor: *S.Moral-García*
	18:45 a 19:15	Trabajo: Determinando si un conjunto aleatorio es conexo Autor: *J.J.Salamanca*
	19:15 a 19:45	Trabajo: EM for Approximating Unidentifiable Counterfactual Queries Autores: *R.Cabañas*, M.Zaffalon y A.Antonucci

VI Congreso de Jóvenes Investigadores de la RSME - 6CJI 2023

León, 6 al 10 de febrero de 2023

Ponentes e instrucciones

- La sesión especial en Teoría de la Probabilidad tendrá lugar el martes 7 de febrero.
- La sesión estará dividida en tres partes, dos en horario de mañana y una en horario de tarde.
- La duración de las presentaciones será de 30 minutos, divididos en 25 minutos para la presentación y 5 minutos de discusión, aproximadamente.
- Se recuerda que el 15 de diciembre termina el plazo para la inscripción temprana con cuota reducida, y que miembros de la RSME tienen un descuento especial.

Ponente	**Procedencia**
Carmen Minuesa	Universidad de Extremadura
Giulia Binotto	Universitat Autònoma de Barcelona
Hristo Inouzhe	Basque Center for Applied Mathematics
Raúl Pérez-Fernández	Univeresidad de Oviedo
Paula Gordaliza	Basque Center for Applied Mathematics
Marta Sánchez-Sánchez	Universidad de Granada
Serafín Moral-García	Universidad de Granada
Juan Jeasús Salamanca	Univeresidad de Oviedo
Rafael Cabañas	Universidad de Almería

VI Congreso de Jóvenes Investigadores
Real Sociedad Matemática Española
León, Febrero de 2023

Procesos de ramificación *defectuosos* en ambiente variable

C. Minuesa,* G. Kersting**

En este trabajo nos centramos en la modelización de poblaciones hasta la aparición por primera vez de una característica mostrada por los individuos, como puede ser una mutación o enfermedad, supuesto que la población comienza con individuos que no la presentan. En concreto, consideramos el caso en el que en cada generación todos los individuos se reproducen de manera independiente al resto y de acuerdo a la misma distribución de probabilidad, conocida como ley de reproducción, la cual puede variar a lo largo de las generaciones.

Esta situación puede describirse a través de procesos estocásticos conocidos como *procesos de ramificación* defectuosos *en ambiente variable* (PRDAV, o *defective Galton-Watson processes in a varying environment* en inglés) introducidos en [1]. Un PRDAV es una cadena de Markov que queda determinada por las leyes de reproducción en cada generación o equivalentemente, a través de $v = \{f_1, f_2, \ldots\}$, donde f_n denota la función generatriz de probabilidad de la ley de reproducción, posiblemente impropia, en la generación n-ésima. De esta forma, $1 - f_n(1)$, puede interpretarse como la probabilidad de que en la generación n, un individuo envíe el proceso a un estado absorbente Δ en la generación $n+1$, donde el proceso permanece para siempre. Como consecuencia, su espacio de estados es $\mathbb{N}_\Delta = \mathbb{N}_0 \cup \{\Delta\}$, donde dos de estos estados son absorbentes: 0 and Δ. En este trabajo estudiamos el comportamiento asintótico de un PRDAV. Se presentarán los dos resultados centrales relativos al comportamiento límite de un PRDAV: la convergencia casi segura del proceso a una variable aleatoria con valores en $\mathbb{N}_\Delta \cup \{\infty\}$ y dos caracterizaciones de la dualidad extinción-absorción en Δ. Estos extienden los resultados ya presentados en [2] para el caso en que la ley de reproducción es constante a lo largo de las generaciones.

Agradecimientos Esta línea de investigación forma parte del proyecto PID2019-108211GB-I00, concedido por la Agencia Estatal de Investigación MCIN/AEI/10.13039/501100011033/.

Este tema de investigación fue iniciado cuando Carmen Minuesa visitó el Instituto de Matemáticas, de la Universidad Goethe Fráncfort del Meno, y está agradecida por la hospitalidad y colaboración

Referencias

[1] G. Kersting, C. Minuesa. *Defective Galton-Watson processes in a varying environment*. Bernoulli. **28(2)** (2022) 1408-1431.

[2] S. Sagitov, C. Minuesa. *Defective Galton-Watson processes*. Stochastic Models. **22** (2017) 451–472.

*Departamento de Matemáticas, Universidad de Extremadura, Avda. de Elvas, s/n, 06006, Badajoz, España. Email: cminuesaa@unex.es
**Institute of Mathematics, Goethe University Frankfurt, Robert-Mayer-Str. 6-10 60325, Frankfurt, Alemania. Email: kersting@math.uni-frankfurt.de

VI Congreso de Jóvenes Investigadores
Real Sociedad Matemática Española
León, Febrero de 2023

Un paseo aleatorio entre modelos epidemiológicos

G. Binotto*

Los modelos matemáticos aplicados al campo de la biología y, en particular, de la epidemiología tienen actualmente un gran interés a causa del brote generado por el COVID-19. No obstante, su estudio para describir la propagación de una enfermedad tiene una larga historia. Se remonta a la investigación sobre la inoculación de la viruela llevada a cabo por Bernoulli en el siglo XVIII [1]. Desde entonces se han considerado tanto modelos deterministas como estocásticos y se han tenido en cuenta muchos factores: agentes infecciosos, modo de transmisión, periodos de latencia, inmunidad temporal o parcial, periodos de cuarentena, etc.
El objetivo de esta ponencia es presentar algunos de los modelos epidemiológicos más importantes que se han presentado al largo de la historia, hasta llegar a un modelo estocástico más reciente, propuesto por Tuckwell y Williams en 2007 [2]. Este hace parte de los modelos de tipo SIR, que dividen la población en tres diferentes clases: Susceptibles, Infectados y Recuperados (*Removed*, en inglés), respectivamente. Es un modelo a tiempo discreto en el cual la población total permanece constante y los individuos se encuentran con un número aleatorio de otros individuos en cada paso de tiempo.
Para terminar, se verán extensiones de este modelo en los cuales se añaden nuevos factores y se tiene en cuenta la dependencia del tiempo de algunos parámetros. Con la ayuda de algunas simulaciones, se mostrará como la evolución de una enfermedad se ve afectada por esta dependencia temporal. Estos últimos resultados forman parte de un trabajo realizado con M. Besalú [3].

Referencias

[1] D. Bernoulli. *Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'Inoculation pour la prévenir*. Histoire de l'Académie royale des sciences, Paris (1760) 1-45.

[2] H.C. Tuckwell, R.J. Williams *Some properties of a simple stochastic epidemic model of SIR type*. Math. Biosci. **208** (2007) 76-97.

[3] M. Besalú, G. Binotto *Time-dependent non-homogeneous stochastic epidemic model of SIR type*. `https://arxiv.org/search/math?searchtype=author&query=Binotto%2C+G`

*Department de Matemàtiques, Universitat Autònoma de Barcelona. Email: gbinotto@mat.uab.cat

VI Congreso de Jóvenes Investigadores

Real Sociedad Matemática Española

León, Febrero de 2023

On approximate validation of models: a Kolmogorov–Smirnov-based approach

E. del Barrio [*], H. Inouzhe[†], C. Matrán[**] [‡]

Classical tests of fit typically reject a model for large enough real data samples. In contrast, often in statistical practice, a model offers a good description of the data even though it is not the 'true' random generator. We consider a more flexible approach based on α-contamination neighbourhoods which were introduced by Huber in [1].

An α-contamination neighbourhood of a probability distribution P_0 is the set of probability distributions

$$\mathcal{V}_\alpha(P_0) = \{(1-\alpha)P_0 + \alpha Q : Q \in \mathcal{P}\},$$

where $\mathcal{P}$ is the set of all probability distributions in the space (throughout this work probabilities on the real line $\mathbb{R}$). Contamination neighbourhoods are nicely related to trimmings by (see [2])

$$P \in \mathcal{V}_\alpha(P_0) \iff P_0 \in R_\alpha(P),$$

where $R_\alpha(P)$ denotes the set of α-trimmings of the probability distribution P,

$$R_\alpha(P) = \left\{Q \in \mathcal{P} : Q \ll P,\ \tfrac{dQ}{dP} \leq \tfrac{1}{1-\alpha}\ P\text{-a.s.}\right\}.$$

Using trimming methods and the well-known Kolmogorov metric between two cumulative distribution functions F and G

$$d_K(F,G) = \sup_{x\in\mathbb{R}} |F(x) - G(x)|,$$

we introduce a functional statistic measuring departures from a contaminated model. We show how the plug-in estimator $d_K(F_0, R_\alpha(F_n))$ allows testing of fit for the (slightly) contaminated model (F_0) vs sensible deviations from it, i.e.,

$$H_{0,n} : d_K(F_0, R_\alpha(F)) = 0 \quad \text{vs.} \quad H_{1,n} : d_K(F_0, R_\alpha(F)) > \rho_n,$$

for $\rho_n = \rho(n) > 0$, with uniformly exponentially small type I and type II error probabilities.

We also address the asymptotic behaviour of the estimator showing that, under suitable regularity conditions,

$$\sqrt{n}\,(d_K(F_0, R_\alpha(F_n)) - d_K(F_0, R_\alpha(F)))$$

asymptotically behaves as the supremum of a Gaussian process.

As an application, we explore methods of comparison between descriptive models based on the paradigm of model falseness. We also include some connections of our approach with the false discovery rate setting, showing competitive behaviour when estimating the contamination level, and being applicable in a wider framework. The interested reader can find further details in [3] and [4].

References

[1] P. J. Huber. *Robust estimation of a location parameter*. The Annals of Mathematical Statistics, **35**, (1964) 73-101.

[2] P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. *Uniqueness and approximate computation of optimal incomplete transportation plans*. Ann. Inst. H. Poincaré Probab. Statist, **47**, (2011) 358-375.

[3] E. del Barrio, H. Inouzhe, and C. Matran. *On approximate validation of models: a Kolmogorov–Smirnov-based approach*. TEST, **29**(4), (2020) 938-965.

[4] E. del Barrio, H. Inouzhe, and C. Matran. *Box-Constrained Monotone Approximations to Lipschitz Regularizations, with Applications to Robust Testing*. Journal of Optimization Theory and Applications, **187**(1), (2020) 65-87.

[*]Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Valladolid Email: eustasio.delbarrio@uva.es
[†]Basque Center for Applied Mathematics (BCAM), Bilbao. Email: hinouzhe@bcamath.org
[‡]Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Valladolid. Email: carlos.matran@uva.es

VI Congreso de Jóvenes Investigadores
Real Sociedad Matemática Española
León, Febrero de 2023

Coeficientes de asimetría estadística asintóticamente normales

Marina Iturrate-Bobes* Raúl Pérez-Fernández*

La *simetría* es una característica de un objeto que permanece invariante tras ser expuesto a una transformación geométrica. La presencia de esta característica ha sido estudiada desde tiempos inmemoriales en contextos a priori tan distintos como pueden ser el arte, la naturaleza y las matemáticas. En el campo de la probabilidad y la estadística, la simetría de una variable aleatoria X implica que la distribución de $X - x_0$ y de $x_0 - X$ es la misma para un cierto $x_0 \in \mathbb{R}$. Un ejemplo común en el que esta propiedad resulta ser de gran utilidad es en el uso de tablas estadísticas de distribuciones simétricas, como es por ejemplo el caso de la distribución normal, donde podemos calcular fácilmente probabilidades acumuladas en la cola derecha a partir de las probabilidades acumuladas en la cola izquierda.

Los primeros esfuerzos en medir la asimetría (i.e., la falta de simetría) de una variable aleatoria se suelen atribuir a Pearson en el siglo XIX, donde se presentan los primeros ejemplos de lo que hoy en día son conocidos como coeficientes de asimetría. Sin embargo, la formalización axiomática del concepto de coeficiente de asimetría es fruto de numerosos estudios posteriores donde destacan los trabajos de van Zwet [5] y Oja [4]. De manera formal, se entiende como coeficiente de asimetría a una función $\gamma : \mathcal{F} \to \mathbb{R}$ que asigna a cada variable aleatoria X un grado de asimetría cumpliendo, al menos, las tres propiedades siguientes: (i) $\gamma(X) = 0$ siempre que X sea una variable aleatoria simétrica; (ii) $\gamma(a\,X + b) = \gamma(X)$, para todo $a, b \in \mathbb{R}$ con $a > 0$; y (iii) $\gamma(-X) = -\gamma(X)$. Otros axiomas adicionales han sido explorados por diversos autores, sin embargo no suelen englobar a todas las funciones utilizadas en la literatura con el fin de medir el grado de asimetría de una variable.

Una gran cantidad de coeficientes de asimetría han sido estudiados en la literatura, desde ejemplos más clásicos como los coeficientes de asimetría de Fisher, Pearson o Bowley, hasta ejemplos más recientes como es el caso de la medcouple [1]. Todos estos coeficientes de asimetría tienen asociada una versión muestral que, bajo condiciones más o menos restrictivas, tiene una distribución asintótica normal, centrada en el valor poblacional. Esta normalidad asintótica, unida al axioma (ii) de la definición de coeficiente de asimetría, nos permite utilizar un coeficiente de asimetría para definir un test de bondad de ajuste a una familia de localización-escala (véase [1]). Desafortunadamente, puesto que la varianza asintótica depende de la distribución subyacente, definir un test de simetría no es tarea sencilla. Diferentes autores han tratado de evitar este problema considerando una distribución de referencia o simplemente estimando la varianza asintótica (véase [2]). En esta última dirección, en el presente trabajo se persigue el uso del bootstrap [3, 6] para estimar dicha varianza asintótica en combinación con diferentes coeficientes de asimetría. Este proceso resulta en la definición de diversos tests de simetría, la mayoría de ellos no estudiados en la literatura con anterioridad, para los que se realiza un estudio comparativo de la potencia en distintas distribuciones (tanto simétricas como asimétricas).

Referencias

[1] G. Brys, M. Hubert, A. Struyf. *A robust measure of skewness*. Journal of Computational and Graphical Statistics **13** (2004) 996–1017.

[2] P. Cabilio, J. Masaro. *A simple test of symmetry about an unknown median*. The Canadian Journal of Statistics **24** (1996) 349–361.

[3] V. Lyubchich, X. Wang, A. Heyes, Y. R. Gel. *A distribution-free m-out-of-n bootstrap approach to testing symmetry about an unknown median*. Computational Statistics & Data Analysis **104** (2016) 1–9.

[4] H. Oja. *On location, scale, skewness and kurtosis of univariate distributions*. Scandinavian Journal of Statistics **8** (1981) 154–168.

[5] W. R. van Zwet. *Convex Transformations of Random Variables*. Mathematisch Centrum, Amsterdam, 1964.

[6] T. Zheng, J. L. Gastwirth. *On bootstrap tests of symmetry about an unknown median*. Journal of Data Science **3** (2010) 413–427.

*Departamento de Estadística e Investigación Operativa y Didáctica de la Matemática, Universidad de Oviedo, Asturias. Email: perezfernandez@uniovi.es

VI Congreso de Jóvenes Investigadores
Real Sociedad Matemática Española
León, Febrero de 2023

Making data fair through optimal transport

A. Paula Gordaliza*† B. Eustasio del Barrio† C. Fabrice Gamboa‡ D. Hristo Inouzhe*
E. Jean-Michel Loubes‡

The aim of this talk is two-fold. The first part is devoted to the asymptotic theory of the empirical transportation cost where optimal transportation methods are studied for statistical inference purposes. We provide a Central Limit Theorem for the Monge-Kantorovich distance between two empirical distributions with different sizes n and m, $\mathcal{W}_p(P_n, Q_m),\ p \geq 1$, for observations on $\mathbb{R}$. In the case $p > 1$ our assumptions are sharp in terms of moments and smoothness. We prove results dealing with the choice of centering constants. We provide a consistent estimate of the asymptotic variance which enables to build two sample tests and confidence intervals to certify the similarity between two distributions. Additionally, a moderate deviation principle for the empirical transportation cost is obtained in general dimension.

In the second part, we present an application of this theoretical results to the recent fair learning problem. Algorithmic fairness is one of the main concerns of today's scientific society due to the generalization of predictive algorithms in all aspects of human life. We motivate the fairness problem by presenting some comprehensive results from the study of the *demographic parity* (DP) criterion through the analysis of the *disparate impact* index on the real and well-known *Adult Income* dataset. Importantly, we show that trying to make fair machine learning models may be a particularly challenging task, especially when the training observations contain bias. Then a review of mathematical models for fair learning is given in a general setting, with some novel contributions in the analysis of the price for fairness in regression and classification. Moreover, we recast the links between fairness and predictability in terms of probability metrics. After this review part, we present our contributions to fair learning using optimal transport.

The goal is to check if there is group bias in the response variable Y with respect to a sensitive information S present in the data. However, not all individuals in S are comparable, and some differences in the target Y may arise from genuine differences in the data. We study two different kind of methods: repairing and trimming. In the first case, we analyze data repair methods based on mapping conditional distributions to the Wasserstein barycenter that ensure total fairness in terms of DP. Additionally, we propose a *random repair* which yields a trade-off between minimal information loss and fairness. Finally, in the classification setting, we build a tool for global fairness assessment of a dataset based on the two-sample similarity tests mentioned above. In contrast, the second approach aims at making data fair though optimal trimmed matching. Precisely, we propose to eliminate such cases by trimming an α proportion of the input data as a pre-processing step to any further learning mechanism in order to obtain the two closest possible marginal distributions (w.r.t. S). On this population that is 'similar enough' we can check for discrimination, in the sense of DP. We solve a trimmed matching problem subject to fairness constraints that is a linear program that can be addressed with well-known techniques. We present some successful results of application to synthetic and real data.

Agradecimientos This research was funded by the Basque Government through the BERC 2022-2025 program and Elkartek project 3KIA (KK-2020/00049), and by the Spanish Ministry of Science, Innovation, and Universities (BCAM Severo Ochoa accreditation SEV-2017-0718).

*BCAM, Alameda de Mazarredo, 14, 48009, Bilbao Email: pgordaliza@bcamath.org, hinouzhe@bcamath.org
†Instituto de Matemáticas de la Universidad de Valladolid, Campus Miguel Delibes, 47011, Valladolid. Email: eustasio.delbarrio@uva.es
‡Institut de Mathématiques de Toulouse, 118 Rte de Narbonne, 31400 Toulouse, Francia. Email: fabrice.gamboa@math.univ-toulouse.fr , loubes@math.univ-toulouse.fr

VI Congreso de Jóvenes Investigadores
Real Sociedad Matemática Española
León, Febrero de 2023

Un análisis de robustez Bayesiana multivariante con enfoque en principios de primas

F. Ruggeri [*], M. Sánchez-Sánchez[**] A. Suárez-Llorens[***]

En esta comunicación utilizaremos una nueva metodología que nos permitirá estudiar la sensibilidad Bayesiana incorporando incertidumbre a través de las clases de distribuciones a priori que encontramos en [1], donde usan densidades ponderadas para inducir incertidumbre sobre la información a priori. En un enfoque actuarial y debido a la particular forma de dichas clases, podremos obtener cotas superiores e inferiores tanto para las primas colectivas o primas priori como para las primas Bayesianas o primas a posteriori. Esto será posible gracias a las propiedades de preservación del orden de las distribuciones a priori y posteriori a través de ordenaciones estocásticas multivariantes que tienen algunos principios de primas. Adicionalmente, veremos como esta metodología tiene aplicaciones en los sistemas de tarificación Bonus-Malus. Para finalizar, podremos ver la aplicación de estos resultados en un conjunto de datos actuariales.

Referencias

[1] F. Ruggeri, M. Sánchez-Sánchez, M.A. Sordo, A. Suárez-Llorens. *On a New Class of Multivariate Prior Distributions: Theory and Application in Reliability*. Bayesian Analysis. **16 (1)** (2021) DOI: 10.1214/19-BA1191.

[*]IMATI(Istituto di Matematica Applicata e Tecnologie Informatiche "Enrico Magenes") – CNR(Consiglio Nazionale delle Ricerche), Milano, Italy. Email: fabrizio@mi.imati.cnr.it
[**]Universidad de Granada, Dpto. Estadística e Investigación Operativa. Email: martasanchez@ugr.es
[***]Universidad de Cádiz, Dpto. Estadística e Investigación Operativa. Email: alfonso.suarez@uca.es

VI Congreso de Jóvenes Investigadores
Real Sociedad Matemática Española
León, Febrero de 2023

Conjuntos credales representables mediante intervalos de probabilidad alcanzables y funciones de creencia

A. Serafín Moral García*

La teoría clásica de la Probabilidad (TP) es la forma estándar de representar la información probabilística sobre un conjunto finito de alternativas proporcionada por un experto o conjunto de datos. Sin embargo, en muchos casos, una única distribución de probabilidad no es adecuada porque no hay información suficiente debido a errores o incompletitud en los datos.

Por este motivo, se han desarrollado en la literatura teorías basadas en *probabilidades imprecisas*. Un resumen de ellas puede verse en [1]. Todas ellas generalizan la TP. Hay teorías de probabilidades imprecisas que son más generales que otras, y también hay pares de teorías de probabilidades imprecisas tales que ninguna de ellas generaliza la otra. Dado que estas teorías tienen propiedades matemáticas específicas, algunas de ellas son más adecuadas que otras en situaciones concretas.

Por un lado, la *Teoría de la Evidencia* (TE) [2, 3] se ha usado frecuentemente para tratar información basada en incertidumbre en aplicaciones prácticas. En la TE, el conocimiento probabilístico sobre el conjunto de alternativas viene determinado por una función de creencia.

Por otro lado, los *intervalos de probabilidad alcanzables* [4], que proporcionan una cota de probabilidad inferior y superior sobre la probabilidad de cada alternativa, son fáciles de entender y manejar y tienen gran poder expresivo. Se han usado con éxito en aplicaciones prácticas como clasificación.

Como se demuestra en [5], la TE y los intervalos de probabilidad alcanzables son paralelos en cuanto a generalidades, ya que un conjunto de intervalos de probabilidad alcanzables no siempre se puede representar mediante funciones de creencia ni una función de creencia tiene asociada siempre un conjunto de intervalos de probabilidad alcanzables.

En este trabajo, estudiamos la relación entre la TE y los intervalos de probabilidad alcanzables. Para esto, proporcionamos un conjunto de condiciones necesarias y suficientes bajo las cuales un conjunto de intervalos de probabilidad alcanzables es representable mediante una función de creencia, así como una condición necesaria y suficiente para que una función de creencia tenga asociado un conjunto de intervalos de probabilidad alcanzables.

Agradecimientos Este trabajo ha sido financiado por los fondos UGR-FEDER con el proyecto A-TIC-344-UGR20, y por la JUnta "Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidadesçon el proyecto Project P20_00159.

Referencias

[1] Peter Walley (1991). *Statistical reasoning with imprecise probabilities*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

[2] A. P. Dempster. *Upper and Lower Probabilities Induced by a Multivalued Mapping*. The Annals of Mathematical Statistics **38.2** (1967) 325-339.

[3] Glenn Shafer (1976). *A mathematical theory of evidence*. Princeton University Press

[4] Luis M. De Campos, Juan F. Huete, and Serafín Moral. *Probability intervals: a tool for uncertain reasoning*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **02.02** (1994) 167-196.

[5] Joaquín Abellán. *Uncertainty measures on probability intervals from the imprecise Dirichlet model*. International Journal of General Systems, **35.5** (2006) 509-528.

*Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, C/ Periodista Daniel Saucedo Aranda, Granada Email: seramoral@decsai.ugr.es

VI Congreso de Jóvenes Investigadores
Real Sociedad Matemática Española
León, Febrero de 2023

Determinando si un conjunto aleatorio es conexo

J.J. Salamanca*

Los conjuntos aleatorios son modelos probabilísticos que han aparecido en diversos contextos (véase, por ejemplo, [3] para profundizar sobre el papel de estos modelos en diversas situaciones estadísticas).

Fíjese un espacio topológico $\mathcal{M}$ con ciertas propiedades topológicas. Un *conjunto aleatorio* (*cerrado*) $\widehat{X}$ sobre $\mathcal{M}$ es una aplicación desde un espacio de probabilidad completo hasta la clase de conjuntos cerrados de $\mathcal{M}$ y que ha de satisfacer cierta hipótesis de medibilidad, [2]. La definición de conjunto aleatorio replica la idea de elemento aleatorio como la de variable aleatoria o vector aleatorio, estando aquí el espacio muestral representado por la clase de cerrados de $\mathcal{M}$. El axioma de medibilidad es el necesario para que estén bien definidas las probabilidades de que $\widehat{X}$ interseque a cualquier conjunto compacto K de $\mathcal{M}$, [2]. Matemáticamente, $P(\widehat{X} \cap K \neq \emptyset)$ siempre estará bien definido para todo compacto K. De este modo podemos definir la *capacidad* $T_{\widehat{X}}$ de $\widehat{X}$ como la aplicación que asocia a cada compacto K de $\mathcal{M}$ la probabilidad anterior. El famoso teorema de Choquet-Kendall-Matheron establece que todo conjunto aleatorio queda caracterizado por su capacidad. Es decir, que la capacidad de un conjunto aleatorio es a un conjunto aleatorio lo que la función de distribución es a una variable aleatoria. Equivalentemente, que toda la información probabilística de un conjunto aleatorio la preserva su capacidad. En consecuencia, en teoría se ha de caracterizar toda propiedad topológica de un conjunto aleatorio a través de la capacidad de dicho conjunto aleatorio. En particular, si es conexo.

Nuestro objetivo es determinar si un conjunto aleatorio dado es conexo atendiendo únicamente a su capacidad. Observar que la capacidad es una aplicación sobredimensionada, en el sentido de que su espacio inicial lo conforma la clase de compactos de $\mathcal{M}$. Esto supone un escollo grave en principio, pues idealmente deseamos poder resolver el problema planteado sobre conexión de una manera algorítmica y en tiempo finito.

El resultado principal que expondremos caracteriza la conexión de un conjunto aleatorio mediante su capacidad. De hecho, se demuestra que, bajo ciertas hipótesis razonables, un conjunto aleatorio es conexo si y sólo si su capacidad satisface cierta ecuación lineal cuyos términos son las capacidades evaluadas en cierto número finito de conjuntos, [4, 5].

La presentación pretenderá exponer algunas ideas sobre la demostración de los resultados matemáticos principales.

Agradecimientos Proyecto PGC2018-098623-B-I00.

Referencias

[1] A. Hatcher. *Algebraic Topology*, Cambridge University Press, 2002.

[2] Ilya S. Molchanov. *Theory of Random Sets*, Springer, 2005.

[3] H. T. Nguyen. *An Introduction to Random Sets*. Chapman and Hall/CRC, 2006.

[4] J. J. Salamanca. *On the connectedness of random sets of R*. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **29** (2021) 57–64.

[5] J. J. Salamanca, J. Herrera, R. M. Rubio. *On the connectedness of a random closed set of a Euclidean space*. Fuzzy Sets and Systems **443** (2022) 127–136.

*Departamento de Estadística, Escuela Politécnica de Ingeniería, Universidad de Oviedo, 33071 Gijón, España. Email: salamancajuan@uiovi.es

VI Congreso de Jóvenes Investigadores
Real Sociedad Matemática Española
León, Febrero de 2023

EM for Approximating Unidentifiable Counterfactual Queries

Rafael Cabañas [*] Marco Zaffalon [†] Alessandro Antonucci [†]

Causality is an emerging direction for data science with applications in diverse domains which allows to reason about hypothetical scenarios. Structural causal models (SCMs)[1] are a class of probabilistic graphical model for causality made of endogenous and exogenous variables. Frequently, the exogenous probabilities are not available (due to the lack of data for these variables). Thus, many causal and counterfactual queries are not identifiable. For solving this, a previous work [2] proposes an exact mapping between SCMs and credal networks by specifying a set of linear constraints for the exogenous probabilities. However this method is not always applicable. Here we summarize a recent work [3] proposing an alternative procedure for specifying the exogenous credal sets based on the *expectation-maximization (EM)* algorithm.

In a SCM, the causal relations are expressed in the form of a directed acyclic graph (DAG) in which the exogenous ones are the root nodes. Figure 1 shows an example of a SCM with two endogenous variables X and Y. The relation between the exogenous and endogenous variables is given by the structural equations (i.e., deterministic functions). We assume that data from the endogenous variables is available and hence the empirical distributions $\tilde{P}(Y|X)$ and $\tilde{P}(X)$ can be estimated.

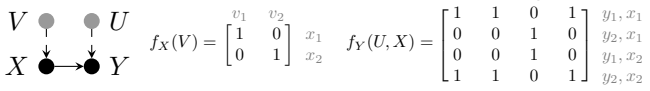

Figure 1: Example of a simple SCM.

The exact transformation of a SCM requires to specify a set of linear constraints for each exogenous variable. In the previous example, the credal set $K(V)$ is defined by the constraints given in Figure 2 (left). Let's assume that in the limit of infinite data we have that $\tilde{P}(y_1|X) = [0.4, 0.6]$, then the feasible solution is the line depicted in Figure 2 (right). When dealing with finite sampled data we might obtain $\tilde{P}(y_1|X) = [0.4, 0.58]$, for which the constraints cannot be solved. In this case, we can proceed as follows: $P(U)$ is randomly initialized and EM is executed . If the process is repeated multiple times, $K(V)$ is defined by all the values for $P(V)$ obtained at convergence. The ten red points in Figure 2 (right) are the result of applying EM with the incompatible data. Once the specification for the credal set is available, any causal query can be calculated if the proper graphical transformations are done. In case of the EM-based solution, the causal query is run independently for each precise model and then combined. In the example, the bounds for the counterfactual query $P(Y_{x_1} = y_1|x_1)$ are $[0.41, 0.59]$ while with the exact approach these are $[0.4, 0.6]$.

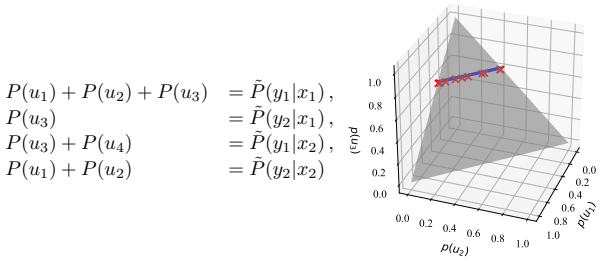

Figure 2: Specification of the credal set $K(V)$.

References

[1] J. Pearl. *Causality*. Cambridge university press, 2009.

[2] M. Zaffalon, A. Antonucci, and R. Cabañas. Structural causal models are (solvable by) credal networks. In *PGM*, 2020.

[3] M. Zaffalon, A. Antonucci, and R. Cabañas. Causal expectation-maximisation. *WHY*, 2021.

[*]Department of Mathematics, University of Almería, Ctra. Sacramento, s/n, 04120 La Cañada, Almería (Spain) Email: rcabanas@ual.es
[†]IDSIA, Via la Santa 1, 6962 Lugano (Switzerland) Emails: marco.zaffalon@idsia.ch, antonucci@idsia.ch