Volumen 24 Número 1 febrero 2008

Boletín de la Sociedad de Estadística e Investigación Operativa

Editorial del Secretario de Estado de Universidades

El Rincón del (nuevo) Presidente

A Present Overview on Functional Data Analysis

On Solution Concepts for Multi-Choice Cooperative Games

Some Probability Applications to the Risk Analysis in Insurance Theory

Imputation in the Survey on Living Conditions

Sociedad de Estadística e Investigación Operativa

Boletín de la SEIO
Volumen 24, número 1
FEBRERO 2008

SEIO:
Facultad de CC. Matemáticas, Despacho 502
Universidad Complutense de Madrid
Plaza de Ciencias, 3
28040 Madrid (Ciudad Universitaria)
oficina@seio.es, http://www.seio.es
Tel: (+34) 91 544 91 02

Imprime SEROTEL
Pº de la Castellana, 87.
Dep. Legal: M-13647-1995
ISSN: 1699-8871

Normas para los envíos de colaboraciones:

Los artículos se enviarán por correo electrónico al editor asociado correspondiente o al editor del Boletín. Se escribirán en estilo *article* de LateX. Cada artículo ha de contener el título, el resumen y las palabras clave en inglés sin traducción al castellano. Desde la página Web de la SEIO, www.seio.es, pueden descargarse varios modelos editados con WinEdt y con Scientific WorkPlace, tanto en español como en inglés, que los autores pueden utilizar, si lo desean, como plantillas para la elaboración de sus artículos.

Las cartas al editor se le dirigirán por correo electrónico. El resto de colaboraciones y noticias se dirigirán al corresponsal más cercano o directamente al editor de la parte informativa o al editor del Boletín. Las referencias bibliográficas y de software se acompañarán de los datos necesarios para su localización y una reseña no superior a 120 palabras. Los resúmenes de tesis se limitarán a 200 palabras y contendrán: título, autor, directores, departamento, universidad y la fecha de lectura. Con relación a congresos y cursos bastará una breve reseña semejante a las publicadas en el Boletín. El formato preferible para estas colaboraciones es MS–Word.

ÍNDICE

EDITORIAL

Francisco Marcellán
Secretario General de Política Científica y Tecnológica
Ministerio de Educación y Ciencia

Agradezco la oportunidad que me brinda el Boletín de la Sociedad Española de Estadística e Investigación Operativa para reflexionar e inducir un debate colectivo sobre los retos centrales que debe abordar nuestro sistema de ciencia y tecnología en un futuro inmediato.

Disponemos de una Estrategia Nacional de Ciencia y Tecnología, que en el horizonte de 2015 pretende alcanzar seis grandes objetivos:

1.- Posicionar a nuestro país en la vanguardia del conocimiento.

2.- Incrementar la solidez y competitividad de nuestro tejido empresarial.

3.- Consolidar la articulación de nuestro sistema en base a la imbricación y coordinación de los agentes del mismo, en particular en el marco interministerial y en la relación Administración General del Estado y Comunidades Autónomas.

4.- Potenciar la dimensión internacional de la ciencia y la tecnología españolas.

5.- Disponer de un entorno favorable para la inversión en Investigación, Desarrollo e innovación.

6.- Crear las condiciones adecuadas para la difusión de la Ciencia y la Tecnología entre nuestros ciudadanos.

Sin embargo, debemos abordar importantes cambios estructurales que requieren una respuesta rápida a los retos de un sistema consolidado pero que dispone de instrumentos inadecuados. En primer lugar, la necesaria actualización de la Ley de la Ciencia de 1986, motor en su momento del gran impulso reformista en nuestras estructuras, pero que debe adecuarse a una realidad cambiante y sobre todo proyectada hacia el futuro. Señalaría como elementos centrales:

1.- La definición de una trayectoria investigadora, con etapas delimitadas en el tiempo, con una retribución adecuada a la realidad internacional que facilite un trabajo digno y reconocido socialmente no sólo a los investigadores sino también al personal de apoyo en gestión y a los técnicos.

2.- Nuevas estructuras de asentamiento de los investigadores tanto en el sector público como en el privado, caracterizadas por su agilidad en la gestión, rigor en la evaluación de sus políticas, compromiso con la excelencia y el impacto económico y social de la ciencia y tecnología en el bienestar de los ciudadanos.

3.- Esquemas de financiación que contemplen el trabajo de los investigadores individuales, el de los grupos consolidados y emergentes, el de las instituciones tanto a nivel macro como micro sobre la base de un principio de transparencia en la rendición de cuentas y de la confianza en el desarrollo de su capacidad autónoma para definir sus propias líneas de actuación.

4.- Medir a los investigadores de acuerdo con sus capacidades, su vinculación a las directrices estratégicas de los centros de investigación en los que trabajan sobre la base de excelencia en sus resultados, su capacidad formativa, la de transferencia y la de divulgación a un público usuario intensivo de ciencia y tecnología en su vida cotidiana pero desconocedor de la raíz intelectual de la misma.

5.- Apoyar un sistema educativo que en sus etapas no universitarias debe estimular la pasión por el conocimiento, el descubrimiento y la experimentación. Legitimar y reconocer el valor del profesorado de esos niveles escolares como dinamizador del aprendizaje científico y técnico.

6.- Mejorar sustancialmente la coordinación entre los diferentes gestores de los programas de ciencia y tecnología, favoreciendo una comprensión global del mismo y evitando la compartimentación suicida de la actividad de apoyo a la gestión.

7.- Estabilizar la financiación sobre una base sostenida y sostenible, priorizando acciones de riesgo basadas en la evaluación de oportunidad y la capacidad de los grupos para abordarlas.

8.- Crear un organismo de evaluación, financiación y prospectiva, contemplado en la Ley de Agencias, que integre actuaciones y facilite el trabajo a los usuarios y agentes de nuestro sistema, asentado en principios de profesionalidad, eficacia, eficiencia, transparencia y flexibilidad en la gestión.

9.- Apoyar los canales de divulgación de la Cien-

cia y Tecnología a través de redes de museos pero también a aquellas revistas científicas que contribuyen a consolidar las diversas comunidades temáticas. Su proyección internacional es un elemento clave para visibilizar nuestros científicos y sus resultados.

10.- Favorecer e impulsar la presencia de investigadores españoles en los centros de decisión de las políticas no sólo a nivel europeo sino en ámbitos más extensos. Reconocimiento de esa presencia como estratégica para los intereses de nuestro país.

Todos estos requisitos implican compromisos individuales y colectivos, de instituciones como las universitarias, que deben explicitar de una manera más clara sus objetivos en materia de investigación, de sociedades científicas que deben articular sus respectivas comunidades para estructurarse como auténtica sociedad civil, por investigadores que deben abordar la necesidad de salir de sus laboratorios y despachos para mostrar la necesidad del apoyo social a su actividad, de organizaciones políticas y sociales que deben apoyar el papel de la Ciencia y la Tecnología como identidad colectiva al margen de disputas coyunturales sobre los modelos a seguir y, en último lugar, de una Administración facilitadora y dinamizadora de la capacidad creativa de nuestros científicos y emprendedores.

El gran esfuerzo realizado en los últimos veinte años por situar la Matemática y, en particular, en el campo de la Estadística y la Investigación Operativa, como ciencia de vanguardia de nuestro país, atendiendo no sólo a cifras de productividad cuantitativa sino también desde el punto de vista de impacto científico y social, debe ser reconocido por nuestros ciudadanos y también por los gestores de las políticas de Ciencia y Tecnología como un valor preferencial.

EL RINCÓN DEL PRESIDENTE

Ignacio García Jurado
Departamento de Estadística e Investigación Operativa
Universidad de Santiago de Compostela

Como sabéis, un nuevo equipo se ha incorporado a la dirección de nuestra sociedad en septiembre de 2007. Normalmente, un nuevo equipo trae consigo nuevos proyectos, y éste también es nuestro caso. Nuestros proyectos no pretenden aportar grandes cambios a la SEIO, pero sí consolidar y reforzar aquellas actividades que ésta ya realiza con éxito (las revistas, los congresos, el boletín, la representación de los socios en organismos e instituciones nacionales e internacionales, la difusión de la información a través de la web), y también lanzar nuevas líneas de actuación. En este último sentido pretendemos, por ejemplo, poner en marcha actividades de cooperación internacional, buscar nuevas fuentes de financiación, o llevar a cabo algunas iniciativas de divulgación de la estadística y la investigación operativa. También deseamos implicar más a los socios en la vida de la sociedad, tratando de que ésta sea más activa y pueda producir respuestas rápidas a los desafíos que se nos presenten. Esperamos poder alcanzar algunos de nuestros objetivos. Para ello necesitamos vuestra ayuda, vuestras sugerencias y vuestra crítica constructiva. Nosotros aportaremos nuestro trabajo ilusionado.

1. ARTÍCULOS DE ESTADÍSTICA

A PRESENT OVERVIEW ON FUNCTIONAL DATA ANALYSIS

Manuel Febrero Bande*
Dpto. de Estadística e Investigación Operativa
Universidad de Santiago de Compostela

Abstract

In this work, a general overview on the state of the art in functional data analysis (FDA) is given. After an essentially historical introduction, the basic concepts in FDA are defined in the Section 2. The third section is devoted to some comments on several recent and interesting works for exploratory functional data analysis, regression with functional variables and other techniques.

Keywords: Functional Data Analysis, Functional Linear Regression, Functional Space.

1. Introducción

En los últimos tiempos la computación aplicada a diversas áreas ha provocado un cambio tecnológico muy importante. Este cambio tecnológico viene de incorporar equipos de medición más rápidos y precisos que son capaces de proporcionar información más fiable y más rápidamente. Esta evolución tecnológica cambia o cambiará algunos de los paradigmas en los que se ha basado la estadística clásica, por ejemplo, el paradigma de que en un conjunto de datos siempre el número de datos es mayor que el número de variables. En muchas áreas se ha empezado a trabajar con grandes bases de datos, que cada vez con más frecuencia, corresponden a observaciones de una variable aleatoria tomadas a lo largo de un intervalo continuo (o en discretizaciones cada vez más extensas de este intervalo continuo). Así, por ejemplo, en campos como la espectrometría, el resultado de la medición es una curva que representa a la muestra concreta que al menos se ha evaluado en una centena de puntos. Este tipo de datos, que llamaremos datos funcionales, surgen de manera natural en muchas disciplinas. En Economía podríamos hablar de curvas intra-día de cotizaciones en bolsa, en Ingeniería podríamos hablar de curvas minutales de producción o demanda eléctrica, en Medio Ambiente se dispone de mediciones continuas de redes de vigilancia atmosférica, fluvial o meteorológica y es bien conocido el auge del reconocimiento de imágenes o de la información espacial. Ante estos nuevos retos surge como respuesta la estadística de datos funcionales que originalmente identificaba dato funcional con función en un intervalo continuo.

Básicamente, los problemas a los que se debe enfrentar la estadística con datos funcionales responde a las mismas necesidades que la estadística clásica. Estos se podrían categorizar de la siguiente manera:

- Explorar y describir el conjunto de datos funcionales resaltando sus características más importantes.
- Explicar y modelar la relación entre una variable dependiente y una independiente (modelos de regresión)
- Métodos de Clasificación Supervisada o no Supervisada de un conjunto de datos respecto a alguna característica.
- Contraste, validación y predicción.

Sin duda alguna, el libro que más ha contribuido a popularizar las técnicas estadísticas para datos funcionales es el de Ramsay y Silverman [18] (en adelante RS2002) cuya primera edición de 1997 trata muchos de los problemas básicos de la estadística funcional con un lenguaje muy asequible dirigido tanto a investigadores del área de estadística como de otras áreas. A este libro le sigue otro de carácter muy aplicado [19] donde se explora el uso de las técnicas en conjuntos de datos interesantes. En ambos casos todas las técnicas incluidas están

*Corresponding Author. E-mail: mfebrero@usc.es

restringidas al espacio de funciones L^2 que como veremos más adelante es un espacio con características específicas que lo hacen especialmente tratable. La página web http://www.functionaldata.org/ mantenida por Jim Ramsay puede ser un buen comienzo para familiarizarse con su trabajo. El otro hito bibliográfico relevante es el reciente libro de Ferraty y Vieu [15] (en adelante FV2006) donde se tratan los datos funcionales desde un punto de vista no paramétrico y se establecen marcos teóricos apropiados para su tratamiento. En este caso, el planteamiento es más general, considerando espacios funcionales normados o semi-normados que en algunos casos pueden ser más apropiados para describir la realidad. Estos autores forman parte del grupo francés STAPH que mantienen la página http://www.lsp.ups-tlse.fr/staph/ donde se puede encontrar más información sobre sus actividades y trabajo. Por supuesto, el número de artículos dedicados al tema en los últimos años es muy importante. Una búsqueda en scholar.google.com con la frase exacta "functional data analysis"proporciona más de 1800 resultados y si nos restringimos a entradas posteriores al 2003 aparecen más de 900. El libro de Ramsay y Silverman ha sido citado según esta herramienta más de 800 veces.

2. ¿Qué es un dato funcional?

Seguiremos la definición de FV2006 por su generalidad y facilidad de comprensión.

Definición 2.1. *Una variable aleatoria $\mathcal{X}$ se dice que es una variable funcional si toma valores en un espacio funcional $\mathcal{E}$ (Espacio normado o semi-normado completo)*

Definición 2.2. *Un conjunto de datos funcionales $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ es la observación de n variables funcionales $\mathcal{X}_1, \ldots, \mathcal{X}_n$ identicamente distribuidas.*

Estas definiciones se pueden aplicar a muchos tipos de espacios. En particular, $\mathbb{R}^p$ con las métricas usuales es un espacio funcional y por tanto puede deducirse que toda técnica que se desarrolle para datos funcionales puede ser aplicada con ciertas garantías en el entorno multivariante. El espacio más comunmente usado cuando se habla de datos funcionales es el espacio $L^2[\mathcal{S}]$, esto es, las funciones de cuadrado integrable en el intervalo $\mathcal{S} = [a, b] \subset \mathbb{R}$. Desde un punto de vista más general podemos tener datos funcionales en la familia: $L^p[\mathcal{S}, \mu] = \{f : \mathcal{S} \to \mathbb{R}$ tal que $\int |f(t)|^p d\mu < \infty\}$, donde $(\mathcal{S}, \mu)$ es un espacio de medida y $1 < p < \infty$. Estos espacios son semi-normados salvo el caso $p = 2$ que es el único de esta familia que es un espacio de Hilbert separable. Cuando se desarrolla una nueva técnica para datos funcionales la primera preocupación es siempre determinar en qué espacio funcional vamos a trabajar. Esto determinará decisivamente el conjunto de herramientas que podremos usar. Una preocupación similar la tendremos al aplicar una técnica de datos funcionales a un conjunto de datos. La métrica del espacio funcional que se elija para encuadrar estos datos debe ser coherente con la interpretación física del fenómeno que describan. Por ejemplo, en el conjunto de datos Tecator (http://lib.stat.cmu.edu/datasets/tecator) que se usa extensamente en FV2006 como hilo conductor, las curvas de absorbancia registradas en el intervalo $[850, 1050]$ para el análisis de trozos de carne presentan formas similares aunque se aprecia un cambio de escala entre ellas.

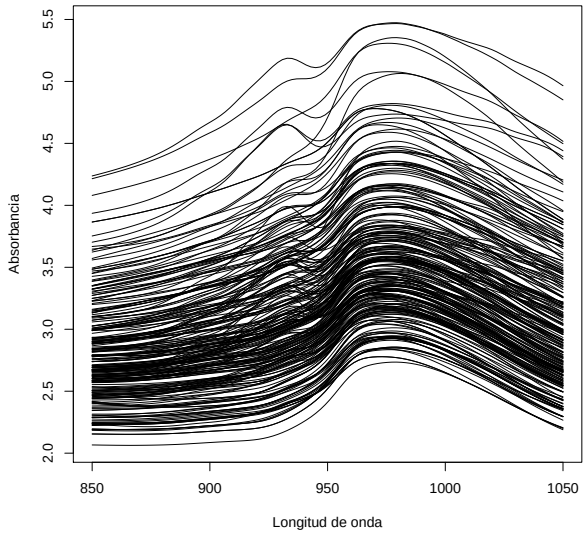


Figura 1: Ejemplo curvas espectrométricas

La primera dificultad que siempre tendremos al analizar datos funcionales, es encontrar una representación adecuada para los datos. Habitualmente, como se refleja también en la Figura 1, la representación de las curvas en el clásico eje X-Y podría esconder las características interesantes. Así, es difícil apreciar el cambio de escala que se mencionaba si no limpiamos un poco la muestra. Si entendemos que este cambio de escala es informativo, la elección L^2 como espacio de referencia sería la más aconsejable. Si por el contrario este cam-

bio de escala es irrelevante y la información está en la forma, una semi-norma como la de las derivadas $\left(d(f,g) = \sqrt{\int_{\mathcal{S}}(f'(t) - g'(t))^2 dt}\right)$ sería una elección más adecuada. Este conjunto de datos presenta como posible variable respuesta el porcentaje de grasa en la muestra. En este caso, atendiendo a esta variable y como se puede ver en la figura siguiente, parece que el cambio de escala es menos importante y el análisis se debe focalizar en el cambio de forma.

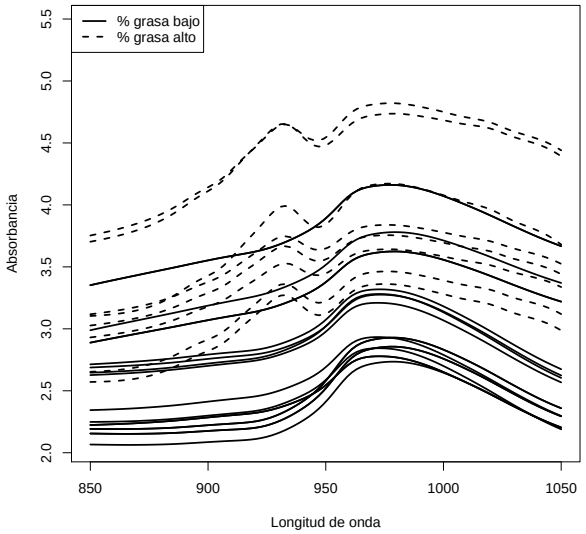


Figura 2: Curvas más extremas por su contenido en grasa

Sin embargo, ambas figuras están diseñadas para apreciar distancias en L^2 entre los datos y no distancias con otro tipo de norma o semi-norma como sería el de la primeras derivadas. Dependiendo del espacio elegido el conjunto de herramientas disponibles cambia notablemente. El caso del espacio más utilizado $L^2[\mathcal{S}]$ es el más favorable. Éste, por ser separable, dispone de bases ortonormales que dan mucho juego a la hora de diseñar procedimientos. En general, la representación de un dato funcional en una base ortonormal proporcionará ventajas tanto desde el punto de vista teórico como práctico sirviendo de puente entre la inevitable discretización del dato funcional y su verdadera forma funcional.

Definición 2.3. *Una base es un conjunto de funciones conocidas e independientes* $\{\phi_k\}_{k\in\mathbb{N}}$ *tales que cualquier función puede ser aproximada, tan bien como se quiera, mediante una combinación lineal de* K *de ellas con* K *suficientemente grande. De esta forma, la observación funcional puede aproximarse como* $\mathcal{X}(t) \approx \sum_{k=1}^{K} c_k \phi_k(t)$.

Básicamente, la idea clave cuando se pueden usar bases ortonormales es representar cada dato funcional en la base usando aquellas coordenadas que son más significativas. Debido a la alta dimensión de los datos funcionales, se elige en general un número K para representar los datos en el subespacio, convirtiendo el problema de dimensión infinita en un problema multidimensional. La elección del parámetro K y de la base más adecuada para los datos observados se antoja crucial y, en principio, no hay ninguna regla que permita hacer una selección óptima de forma universal. El parámetro K es, en cierto modo, un parámetro de suavización de los datos funcionales. Si K es bajo tendremos un modelo muy manejable pero posiblemente habremos perdido información relevante. Si K es alto representaremos muy bien los datos pero el problema de la dimensión cobra importancia. Si atendemos a la elección de la base, para datos periódicos se suele emplear la base de Fourier y para datos no periódicos la base B-spline o la wavelet. Una base muy popular está basada en la expansión de Karhunen-Loève que no es más que la extensión del análisis de componentes principales multivariante a procesos estocásticos y por añadidura a datos funcionales. Calculando a partir del operador momento de segundo orden muestral las correspondientes autofunciones y autovalores es posible construir específicamente una base ortonormal adaptada para cada conjunto de datos. Esta técnica se denomina Componentes Principales Funcionales (FPCA) y ha dado lugar a muchas técnicas interesantes para datos funcionales. Sin embargo, y por incluir alguna sombra, esta técnica puede ser muy sensible a la aparición de datos atípicos y la representación del dato funcional puede no ser relevante para el objetivo del estudio como podría ser la relación con otra variable funcional o no. La decisión sobre qué base elegir debe tomarse en función del objetivo del estudio y los datos y aprovechando las ventajas e inconvenientes que presenta cada tipo de base. Si se trunca cualquiera de estas bases en un número determinado de elementos obtendremos una semimétrica que también podremos usar para manejar los datos funcionales. En este caso, cualquier métrica o semi-métrica en el espacio no es más que una forma de determinar qué elementos del espacio están cercanos y cuáles lejanos.

La estadística con datos funcionales tiene fron-

tera con otros campos relevantes de la estadística como el análisis multivariante, el análisis de datos longitudinales o las series temporales. Como se comentó anteriormente, una técnica de datos funcionales puede aplicarse con ciertas garantías a datos multivariantes. El reverso, en general, no es cierto. Para la mayoría de las técnicas multivariantes que basan mucho de su trabajo en propiedades del álgebra matricial puede ser un problema casi insalvable tratar datos funcionales de alta frecuencia con seguramente, muy fuerte colinealidad. Según aumenta el grado de resolución con el que somos capaces de ver una curva, más difícil resulta para las técnicas multivariantes obtener un resultado convirtiendo el aumento de resolución en una dificultad más que en una oportunidad de obtener mejor información. Algo similar podría decirse del análisis de datos longitudinales. En este campo se obtienen medidas repetidas a lo largo del tiempo para el mismo sujeto, pero en general, éste es un número pequeño y las técnicas multivariantes pueden adaptarse para trabajar con ellas. La principal dificultad para tratar datos longitudinales como datos funcionales suele ser precisamente la baja calidad de representación de las curvas. La relación con el campo de las series temporales es totalmente diferente. Así, ejemplos clásicos de datos funcionales se han construido a base de cortar una serie temporal en ciclos homogéneos. Por ejemplo, en RS2002 se usan los datos de un índice bursátil estadounidense troceados por años (como unidad funcional) para deducir a partir de la forma de cada curva anual la tipología de los distintos años (de expansión, de crisis, ...). Considerados los datos como una serie temporal, el objetivo es predecir alguno de los periodos del próximo año. Como conjunto de datos funcionales, el objetivo es resumir la información y el resultado será siempre un dato funcional, esto es, un ciclo anual completo. Por supuesto, se pueden mezclar ambos mundos para obtener herramientas para series de tiempo funcionales (véase por ejemplo, [14]). Por tanto, la relación entre estos dos campos es peculiar. Muchas veces trabajan sobre la misma información pero desde ópticas completamente diferentes.

3. Estado de la cuestión

3.1. Técnicas exploratorias para Datos Funcionales

Aunque probablemente cualquier estudio estadístico de un conjunto de datos debiera empezar por un análisis descriptivo, sin duda alguna este apartado no ha merecido demasiada atención hasta el momento. En RS2002 en el Capítulo 2 sólo se recogen como herramientas para resumir los datos: la media funcional, la varianza funcional y la función de covarianza. En un capítulo posterior se emplean las componentes principales funcionales como herramientas del análisis descriptivo. Básicamente esto era todo el análisis descriptivo de un conjunto de datos funcionales. Sin embargo, el análisis descriptivo se revela decisivo para el tratamiento de datos funcionales. Como decíamos, un vistazo rápido a la Figura 1 demuestra que el gráfico por defecto de un conjunto de datos funcionales puede ser enormemente no informativo. Esto no ocurre con los típicos gráficos de nube de puntos en $\mathbb{R}^2$ donde una mirada entrenada puede encontrar características relevantes de la población. Para datos funcionales la cuestión se complica si pensamos que nuestros datos pueden estar sujetos a métricas no usuales y por tanto, las representaciones usuales engañarían nuestra mirada. En este campo se echan en falta herramientas descriptivas que en otros ámbitos como el multivariante se han desarrollado expresamente. Esta falta de atención está cambiando en los últimos años donde han aparecido varios trabajos que hacen más hincapié en este apartado. Así en el trabajo de Dabo-Niang et al [10] se definen extensiones de la moda a datos funcionales. Manejando diferentes conceptos sobre profundidad estadística también se han definido extensiones de medidas robustas para datos funcionales como las referidas en los trabajos [17] y [8], incluyendo en este último el bootstrap para datos funcionales como herramienta para analizar la variabilidad de los distintos estimadores. También se han hecho avances en la detección de outliers (véase por ejemplo [12] y [13]).

3.2. Regresión

El apartado de regresión, es el que ha recibido más atención por parte de la comunidad científica. Se pueden establecer distintos modelos bajo las diferentes condiciones que deben cumplir la variable respuesta y las variables regresoras. En sus distintas variantes el problema es inicialmente estudiado en RS2002 al que dedica varios capítulos. En su análisis cobra mucha importancia la necesidad de

penalizar la falta de suavidad de los estimadores y con esta visión revisa varios problemas de regresión sin adentrarse en el marco teórico. El caso más estudiado es posiblemente el de variable respuesta escalar y variable regresora funcional. Bajo diseño aleatorio, los trabajos de Cardot et al ([2],[3]) se han centrado en el modelo lineal funcional, esto es, $y = \alpha_0 + \int_{\mathcal{C}} \alpha(t)\mathcal{X}(t)dt + e(t)$. En este caso, la métrica del espacio funcional siempre es L^2. Una extensión natural de este modelo viene dada por considerar $y = r(\mathcal{X}(t)) + e(t)$ donde r es una función suave general y no está restringida a un operador lineal. Este modelo se conoce como regresión funcional no paramétrica y ha sido extensamente estudiado en FV2006 y trabajos anteriores de los mismos autores (véase por ejemplo [14]). La idea subyacente es usar un estimador similar al de regresión no paramétrica pero adaptado a la norma o semi-norma de los datos funcionales. Para ello se deben definir funciones núcleo apropiadas así como determinar condiciones teóricas del espacio funcional para la convergencia del estimador. Una característica importante de este modelo es que no sufre el desastre de la dimensionalidad al usar las funciones núcleo sobre la métrica (unidimensional).

Para el modelo con respuesta funcional y variables regresoras funcionales no se dispone en general de tantas herramientas fuera de L^2. Este modelo que viene dado por $Y(t) = \alpha(t) + \int_{\mathcal{S}} \mathcal{X}(s)\beta(s,t)ds + e(t)$ está resuelto en RS2002 representando en bases restringidas tanto la respuesta como la variable regresora. En este modelo todavía parece un problema abierto las condiciones teóricas necesarias para la convergencia de los estimadores. En el contexto de diseño fijo algunas condiciones sobre convergencia de los estimadores se puede encontrar en el trabajo de Cuevas et al [6]. Otra referencia reciente de interés es [23].

Las técnicas de análisis de la varianza pueden considerarse como un modelo de regresión con respuesta funcional y variable regresora discreta. Este es el punto de vista que se usa en RS2002 para la obtención de estimadores. Siguiendo la estela marcada por este libro, han aparecido varios trabajos que, básicamente, transforman los datos funcionales en un conjunto de datos multivariante (mediante una base ortonormal truncada) y resuelven un MANOVA. Desde otro punto de vista, el trabajo de Cuevas et al [7] está enfocado al estudio tanto teórico como aplicado de un contraste ANOVA general para datos funcionales con bootstrap. Usando la idea de los modelos de regresión en otros contextos podemos destacar el trabajo de Ferraty y Vieu [14] donde se usa el estimador de la regresión funcional no paramétrico en el contexto de discriminación de curvas o de predicción de series de tiempo. En series de tiempo es destacable el trabajo de Aguilera et al [1] como un primer precedente de aplicación a series de tiempo funcionales. Sería muy largo y prolijo comentar extensiones de los modelos de regresión en datos funcionales siguiendo las ideas del libro de Ramsay y Silverman y sólo a modo de ejemplo citaré aquí el trabajo de Escabias et al [11] como aplicación al campo de la regresión logística.

3.3. Otros métodos

Fundamentalmente, en la literatura abundan trabajos que se dedican de una u otra manera al problema de la clasificación bien supervisada (discriminación) o no supervisada (cluster). Como uno de los trabajos más clásicos en el tema se puede citar el de James y Sugar [16] que está basado en los B-splines y que posteriormente fue seguido por el trabajo de Yao et al [22] aplicando componentes principales funcionales al problema de discriminación. Otro clásico es el trabajo de Tarpey y Kinateder [21]. Más recientemente en FV2006 el tercer capítulo está integramente dedicado a estas cuestiones. En este libro, el problema de discriminación se resuelve, como se ha comentado, como un problema de regresión calculando la esperanza de una variable indicadora. El problema de la clasificación no supervisada es abordado mediante la elaboración de un índice de heterogeneidad que sirve para ir particionando la muestra mediante un método jerárquico. El trabajo de Cuesta-Albertos y Fraiman [4] proporciona una versión robusta del algoritmo de k-medias para datos funcionales mientras que el trabajo de Cuevas et al [9] presenta varios métodos para discriminación basados en conceptos de profundidad. El análisis de datos funcionales está llegando a todos los ámbitos de la estadística y como ejemplo se cita aquí el trabajo de Rossi y Conan-Guez [20] dedicado a redes neuronales con entradas funcionales. En mayor o menor medida los trabajos citados se han dedicado más a la estimación que al contraste. Un trabajo reciente que posiblemente cambiará este planteamiento es el trabajo de Cuesta-Albertos

et al [5] que, en pocas palabras, permite caracterizar poblaciones funcionales mediante el uso de proyecciones aleatorias. Esto abre la puerta a poder utilizar un amplio abanico de herramientas uni o multi-dimensionales para el contraste de características en datos funcionales con técnicas muy simples.

4. Conclusiones

El análisis de datos funcionales es una disciplina emergente en la estadística actual. Las causas de esta eclosión hay que buscarlas en la necesidad creciente de tratar realidades cada vez más complejas que evolucionan rápidamente gracias a las nuevas tecnologías de medición. De este interés creciente da fe el que revistas del máximo nivel hayan dedicado números especiales al tratamiento de datos funcionales como por ejemplo *Statistica Sinica* (Vol. 14, nº3, 2004), *Computational Statistics & Data Analysis* (Vol. 51, 10, 2007) o *Computational Statistics* (Vol. 22, nº 3, 2007). Otra iniciativa que demuestra este interés es la creación de un grupo de trabajo especializado en estadística funcional dentro del grupo de trabajo de Computing & Statistics en el ERCIM (European Research Consortium for Informatics and Mathematics). Puede consultarse la información disponible en el siguiente vínculo http://www.dcs.bbk.ac.uk/ercim/TrackSFD.html. Se ha recorrido mucho camino en poco tiempo fruto de este interés creciente pero todavía queda mucho más por recorrer. A medida que estas herramientas de datos funcionales se vayan popularizando surgirán nuevos tipos de datos funcionales que necesitarán de desarrollos específicos o adaptación de los ya existentes para su correcto tratamiento. Esto augura un inmenso campo de trabajo para los próximos años.

Referencias

[1] Aguilera A.M., Ocaña F.A. y Valderrama M.J. (1999). Forecasting time series by functional PCA. Discussion of several weighted approaches. *Computational Statistics*, **14**, 442-467.

[2] Cardot H., Ferraty F. y Sarda P. (1999). Functional linear model. *Statistics and Probability Letters*, **45**, nº 1, 11-22.

[3] Cardot H., Ferraty F., Mas A. y Sarda P. (2003). Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics*, **30**, nº 1, 241-255.

[4] Cuesta-Albertos J. y Fraiman, R. (2007). Impartial Trimmed k-means for Functional Data. *Computational Statistics and Data Analysis*, **51**, 4864-4877.

[5] Cuesta-Albertos J., Fraiman, R. y Ransford, T. (2007). A sharp form of the Cramer-Wold theorem. *Journal of Theoretical Probability*, **20**, 201-209.

[6] Cuevas A., Febrero M. y Fraiman, R. (2002). Linear functional regression: the case of fixed design and functional response. *The Canadian Journal of Statistics*, **30**,2 285-300.

[7] Cuevas A., Febrero M. y Fraiman, R. (2004). An ANOVA test for functional. *Computational Statistics and Data Analysis*, **47**, 111-222.

[8] Cuevas A., Febrero M. y Fraiman, R. (2007). On the use of bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis*, **51**, 1063-1074.

[9] Cuevas A., Febrero M. y Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, **22**, 1481-496.

[10] Dabo-Niang S., Ferraty F. y Vieu, Ph. (2007). On the using of modal curves for radar waveforms classification. *Computational Statistics and Data Analysis*, **51**, 4957-4968.

[11] Escabias M., Aguilera A.M. y Valderrama, M.J. (2005). Modelling environmental data by functional component logistic regression. *Environmetrics*, **16**, 1, 95-107.

[12] Febrero M., Galeano, P. y González-Manteiga, W. (2007). Outlier detection in functional data by depth measures with application to identify abnormal NOx levels. *Environmetrics*, DOI:10.1002/env.878.

[13] Febrero M., Galeano, P. y González-Manteiga, W. (2007). A functional analysis of NOx levels:location and scale estimation and outlier detection. *Computational Statistics*, **22**, 411-427.

[14] Ferraty F. y Vieu, Ph. (2004). Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. *Nonparametric Statistics*, **16**, 111-125.

[15] Ferraty F. y Vieu Ph. (2006). *Nonparametric Functional Data Analysis*, Springer, New York.

[16] James, G.M. y Sugar, C.A.(2003). Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical Association*, **98**, 397-408.

[17] López-Pintado S. y Romo, J (2007). Depth-based inference for functional data. *Computational statistics and data analysis*, **51**, 4878-4890.

[18] Ramsay J.O. y Silverman, B.W. (2002). *Functional Data Analysis* Second Edition, Springer, New York.

[19] Ramsay J.O. y Silverman, B.W. (2002). *Applied Functional Data Analysis Methods Case and Studies*, Springer, New York.

[20] Rossi, F. y Conan-Guez, B. (2006). Theoretical Properties of Projection Based Multilayer Perceptrons with Functional Inputs. *Neural Processing Letters*, **23**, 55-70.

[21] Tarpey, T. y Kinateder, K. (2003). Clustering Functional Data. *Journal of Classification*, **20**, 93-114.

[22] Yao, F., Müller, H.G. y Wang J.L. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, **100**, nº 470, 577-590.

[23] Zhang J.T. y Chen J. (2007). Statistical inferences for functional data. *Annals of Statistics*, **35**, nº 3, 1052-1079.

Manuel Febrero Bande es profesor del Departamento de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela. Entre sus líneas de investigación se incluye el análisis de series temporales, la estadística espacial, el bootstrap y la estadística con datos funcionales con aplicaciones fundamentalmente en el campo medioambiental, industrial o económico. Es autor/coautor de una treintena de artículos en revistas del JCR y actualmente es co-chair del grupo especializado de Estadística para Datos Funcionales en el grupo de trabajo Computing & Statistics dentro del ERCIM (European Research Consortium for Informatics and Mathematics). Más detalles se pueden encontrar en su página web http://eio.usc.es/pub/febrero/.

2. ARTÍCULOS DE INVESTIGACIÓN OPERATIVA

ON SOLUTION CONCEPTS FOR MULTI-CHOICE COOPERATIVE GAMES

R. Branzei*
Faculty of Computer Science
Alexandru Ioan Cuza University, Iaşi, Romania

Abstract

This paper deals with the model of multi-choice games, a natural extension of the traditional model of cooperative games with transferable utility. Cooperative multi-choice game theory is a booming research topic with many recent developments on which this paper intends to offer a brief overview.

Keywords: Cooperative games, multi-choice games, solution concepts.

1. Solution concepts for arbitrary multi-choice games

Multi-choice cooperative games are introduced in Hsiao and Raghavan [10], [11] to allow cooperating players to be active at more than one level, under the assumption that the same number of participation levels is available for all players. The model of multi-choice games is considered in a more general setting in Nouweland [15] and Nouweland, Potters, Tijs, and Zarzuelo [16], where the number of participation levels for different players may be different. Building blocks for multi-choice games are the so-called multi-choice coalitions which are players' participation profiles available when a maximal participation profile is known. A real-valued characteristic function on the set of multi-choice coalitions quantifies the benefit of cooperation according to any participation profile; it is assumed that overall abstention from cooperation (i.e. cooperation at level 0) generates worth 0. A multi-choice game is a triplet $\langle N, m, v\rangle$ specifying the set $N = \{1, ..., n\}$ of players, their maximal participation profile $m = (m_1, ..., m_n)$ with $m_i \in \mathbb{Z}_+$ for each $i \in N$, and the characteristic function $v : \mathcal{M}^N \to \mathbb{R}$, $v(0) = 0$, where $\mathcal{M}^N$ stands for the set of multi-choice coalitions, that is the set of participation profiles s smaller than or equal to m.

Here is an example of a multi-choice game $\langle N, m, v\rangle$, where N=$\{1, 2\}$, m=$(2, 1)$, $v((0, 0))$=0, $v((1, 0))$=5, $v((2, 0))$=6, $v((0, 1))$=3, $v((1, 1))$=9, $v((2, 1)) = 13$.

Often, a multi-choice game is identified with its characteristic function. Let us denote by $MC^{N,m}$ the set of multi-choice games with a fixed finite set of players N and maximal participation profile m. Examples of application of the multi-choice game model to various situations can be found in Nouweland [15], Calvo and Santos [6], Peters and Zank [17]. Multi-choice cooperative games have been a useful tool for modeling interaction of players in economic and operations research situations in which they may have different options for cooperation, varying from non-cooperation (participation level 0) to a maximal participation level which is greater than or equal to 1. In particular, multi-choice games can be seen as an appropriate analytical tool for modeling cost allocation situations in which commodities are indivisible goods that are only available at certain finite number of levels. Clearly, when all the players can only abstain from cooperation or be active at level 1 we obtain the traditional model of cooperative games. Consequently, solution concepts on $MC^{N,m}$ appear as natural extensions of well-known solution concepts on the set G^N of traditional cooperative games with player set N. A basic notion for defining various solutions for multi-choice games is that of (level) payoff vector. A (level) payoff vector is a function $x : M \to \mathbb{R}$, where $M = \{(i, j) \mid i \in N, j \in M_i^+\}$ with $M_i^+ = \{1, ..., m_i\}$, which specifies for each player $i \in N$ and each of his levels $j \in M_i^+$

*Corresponding Author. E-mail: branzeir@infoiasi.ro

the payoff to player i corresponding to a change of its activity level from $j-1$ to j. By convention, we define $x_{i0}=0$ for all $i\in N$. An example of a (level) payoff vector for the two-person multi-choice game previously presented is $(5,1,7)$, where $x_{11}=5$, $x_{12}=1$, $x_{21}=7$ ($x_{10}=x_{20}=0$). For each $s\in\mathcal{M}^N$, the payoff of s according to x is $X(s)=\sum_{i\in N}\sum_{j=1}^{s_i}x_{ij}$, and the payoff of player i for acting at level s_i is $X_{is_i}=\sum_{j=1}^{s_i}x_{ij}$. Appealing properties for (level) payoff vectors are: *efficiency*, i.e. $X(m)=v(m)$; *level-increase rationality*, i.e. for each $i\in N$ and each $j\in M_i^+$, x_{ij} is at least the increase in payoff that player i can obtain working alone when he changes his activity level from $j-1$ to j; *coalitional stability*, i.e. $X(s)\geq v(s)$ for each $s\in\mathcal{M}^N$. Let $v\in MC^{N,m}$.

The *imputation set* $I(v)$ of v consists of all efficient and level-increase rational (level) payoff vectors, that is

$$I(v)=\{x:M\to\mathbb{R}\mid X(m)=v(m);\\ x_{ij}\geq v(je^i)-v((j-1)e^i),\ i\in N,\ j\in M_i^+\},$$

where e^i is the unitary vector with $e^i_k=0$ for all $k\neq i$ and $e^i_i=1$.

The *core* $C(v)$ of v consists of those imputations which are coalitional stable, that is

$$C(v)=\{x\in I(v)\mid X(s)\geq v(s)\text{ for all }s\in\mathcal{M}^N\}.$$

The *precore* $\mathcal{PC}(v)$ of v consists of all efficient and coalitional stable (level) payoff vectors, that is

$$\mathcal{PC}(v)=\{x:M\to\mathbb{R}\mid X(m)=v(m);\\ X(s)\geq v(s)\text{ for all }s\in\mathcal{M}^N\}.$$

The *minimal core* $C_{\min}(v)$ of v consists of those core elements x for which do not exist other elements $y\in C(v)$ which are weakly smaller than x in the sense that $Y(s)\leq X(s)$ holds for each $s\in\mathcal{M}^N$, that is

$$C_{\min}(v)=\{x\in C(v)\mid \nexists y\in C(v)\text{ s.t. }y\neq x\\ \text{and }y\text{ is weakly smaller than }x\}.$$

By considering a domination relation on $I(v)$ based on players' levels of activity, the notions of *dominance core* and *stable set* are introduced in Nouweland et al. [16] as natural extensions of their traditional counterparts.

Let $s\in\mathcal{M}^N\setminus\{0\}$ and $x,y\in I(v)$. We say that the imputation y dominates the imputation x via coalition s, denoted by $y\,dom_s\,x$, if $Y(s)\leq v(s)$ and $Y_{is_i}>X_{is_i}$ for all $i\in car(s)=\{i\in N\mid s_i>0\}$. Further, we say that the imputation y dominates the imputation x if there exists $s\in\mathcal{M}^N\setminus\{0\}$ such that $y\,dom_s\,x$.

The *dominance core* $DC(v)$ of $v\in MC^{N,m}$ consists of all $x\in I(v)$ for which there exists no $y\in I(v)$ such that y dominates x, that is

$$DC(v)=\{x\in I(v)\mid \nexists y\in I(v)\text{ s.t. }y\,dom\,x\}.$$

A set $A\subset I(v)$ is a *stable set* if it is internally stable, that is $A\cap D(A)\neq\emptyset$, and it is externally stable, that is $I(v)\setminus A\subset D(A)$. Here, $D(A)=\{x\in I(v)\mid\exists a\in A\text{ s.t. }a\,dom\,x\}$.

Relations among the core, the dominance core and stable sets in the traditional cooperative game model still hold in the multi-choice model. In particular, the core of v is a subset of the dominance core of v; every stable set contains the dominance core as a subset; if the dominance core of v is a stable set, then there are no other stable sets. For additional results on cores and stable sets for multi-choice games the reader is referred to Part III in Branzei, Dimitrov and Tijs [1].

The *equal division core* $EDC(v)$ of $v\in MC^{N,m}$ is introduced in Branzei, Llorca, Sánchez-Soriano and Tijs [3] based on the (per-unit level) average worth, $\alpha(s,v)=v(s)/\sum_{i\in N}s_i$, of a multi-choice coalition $s\in\mathcal{M}^N\setminus\{0\}$ for the game v, that is

$$EDC(v)=\{x:M\to\mathbb{R}\mid X(m)=v(m);\\ \nexists s\in\mathcal{M}^N\setminus\{0\}\text{ s.t. }\alpha(s,v)>x_{ij}\\ \text{for all }i\in car(s),\ j\in M_i^+\}.$$

It holds $C(v)\subset\mathcal{PC}(v)\subset EDC(v)$ for each $v\in MC^{N,m}$.

Another set-valued solution concept on $MC^{N,m}$ is the *equal split-off set* introduced in Branzei, Dimitrov and Tijs [2] as a straightforward generalization of the equal split-off set on G^N.

The multi-choice version of the *Weber set* $W(v)$ of $v\in G^N$ is defined as the convex hull of the (level) marginal vectors $w^{\sigma,v}$ corresponding to admissible orderings σ of players in $v\in MC^{N,m}$, i.e. orderings which take into account the fact that each player can reach a higher level of participation only via one-unit level increases starting from level 0. Thus, each admissible ordering σ of players gene-

rates a path from the participation profile $(0, ..., 0)$ to $(m_1, ..., m_n)$, along which the differences in worth for each one-unit level increase become $w_{ij}^{\sigma,v}$ for $i \in N$ and $j \in M_i^+$. We notice that marginal vectors $w^{\sigma,v}$ are not necessarily imputations. Given $v \in MC^{N,m}$ and $x \in C(v)$ it is proved by Nouweland et al. [16] that there is a $y \in W(v)$ that is weakly smaller than x. Thus, the relation between the core and the Weber set in the multi-choice game theory is different from that existing in the classical cooperative game theory, where the Weber set is a core catcher.

On the class of multi-choice games several solution concepts, which we call here *Shapley-like values*, are inspired basically by the *Shapley value* (cf. Shapley [19]). We briefly present here the most important ones.

The *Shapley value* Φ, a natural extension of the Shapley value on G^N, is introduced by Nouweland et al. [16] as the average of (level) marginal vectors, and axiomatically characterized by additivity, the carrier property and the hierarchical strength property. This value is further studied in Calvo and Santos [6] where the focus is on players' total payoffs instead of (level) payoff vectors. It is shown that this value corresponds to the discrete Aumann-Shapley method proposed in Moulin [14].

In Hsiao and Raghavan [11] the *Shapley value* Ψ^w is introduced, where w is a weight vector corresponding to players' levels under the assumptions of equal number of levels for all players and increasing ordering of weights with respect to levels. The Shapley values Ψ^w extend ideas of weighted Shapley values (cf. Kalai and Samet [12]). An axiomatic characterization of Ψ^w is provided using additivity, the carrier property, the minimal effort property and the weight property.

The *Shapley value* Θ is introduced by Derks and Peters [7]. In Klijn, Slikker and Zarzuelo [13] it is proved that Θ can be seen as the (level) payoff vector of average marginal contributions of the elements in $\mathcal{M}^N \setminus \{0\}$. The Shapley value Θ is axiomatically characterized in Nouweland [15] in the spirit of Young [22]; other axiomatic characterizations of it can be found in Klijn, Slikker and Zarzuelo [13].

The *Shapley value* ε, called the *egalitarian multi-choice solution*, is introduced by Peters and Zank [17] and axiomatically characterized by the properties of efficiency, zero-contribution, additivity and level-symmetry.

We also mention here the *multi-choice Shapley value* introduced by Grabisch and Lange [8].

More about the foregoing solution concepts is known on special classes of multi-choice games.

2. Solution concepts for convex multi-choice games

Convex multi-choice games are introduced in Nouweland et al. [16] as games whose characteristic function is supermodular. Formally, a game $v \in MC^{N,m}$ is convex if $v(s \wedge t)+v(s \vee t) \geq v(s)+v(t)$ for all $s, t \in \mathcal{M}^N$, where $(s \wedge t)_i = \min\{s_i, t_i\}$ and $(s \vee t)_i = \max\{s_i, t_i\}$ for all $i \in N$. It is shown that the core of a convex multi-choice game is the unique stable set of the game, and that a multi-choice game v is convex if and only if its Weber set equals the convex hull of the minimal core of the game, i.e. $W(v) = co(C_{\min}(v))$ holds. Consequently, the Shapley value $\Phi(v)$ of v belongs to the core $C(v)$ of v, in case v is convex. In Grabisch and Xie [9] notions related to the core and the Weber set for multi-choice games are defined in such a way that the equality between the core of a convex multi-choice game and the Weber set of that game still holds true.

Convexity of a multi-choice game proved to be a sufficient condition for the existence of monotonic allocation schemes (cf. Sprumont [20]) in a multi-choice setting. Such schemes, called (level-increase) monotonic allocation schemes (limas), are introduced and studied in Branzei, Tijs and Zarzuelo [5]. Let $v \in MC^{N,m}$ be a convex game.

A scheme $a = [a_{ij}^t]_{i \in N,\ j \in \{1,...,t_i\}}^{t \in \mathcal{M}^N \setminus \{0\}}$ is called a (*level-increase*) *monotonic allocation scheme* (*limas*) for v if it satisfies a stability condition, i.e. $a^t \in C(v_t)$ for each subgame v_t of v with $t \in \mathcal{M}^N \setminus \{0\}$, and a (level) monotonicity condition, i.e. $a_{ij}^s \leq a_{ij}^t$ for all $s, t \in \mathcal{M}^N \setminus \{0\}$ with $s \leq t$, each $i \in car(s)$, and each $j \in \{1, ..., s_i\}$. The subgame of $v \in MC^{N,m}$ with respect to $t \in \mathcal{M}^N \setminus \{0\}$ is defined by $v_t(s) := v(s)$ for each $s \in \mathcal{M}^N \setminus \{0\}$ such that $s \leq t$. We denote by $\mathcal{M}_t^N$ the subset of $\mathcal{M}^N \setminus \{0\}$ consisting of multi-choice coalitions $s \leq t$ and by M_i^t the set $\{1, ..., t_i\}$.

In particular, the total Shapley value (cf. Nouweland et al. [16]) of a convex multi-choice game, which is the scheme $[\Phi_{ij}(v_t)]_{i \in N,\ j \in \{1,...,t_i\}}^{t \in \mathcal{M}^N \setminus \{0\}}$ with the Shapley value of the multi-choice subgame

t in each row t, is a (level-increase) monotonic allocation scheme for v. It turns out that each element of the Weber set of a convex multi-choice game is extendable to a limas, that is there exists a limas $[a_{ij}^t]_{i\in N,\ j\in\{1,...,t_i\}}^{t\in\mathcal{M}^N\setminus\{0\}}$ such that $a_{ij}^m = x_{ij}$ for each $i \in N$ and $j \in M_i^+$.

The *constrained egalitarian solution* is introduced on the class of convex multi-choice games in Branzei, Llorca, Sánchez-Soriano and Tijs [3] by using an adjusted version of the Dutta-Ray algorithm for traditional convex games based on the (per one-unit level-increase) average worth of a multi-choice coalition $s \in \mathcal{M}^N \setminus \{0\}$. Here, a key role is played by a proposition showing that there exists a unique multi-choice coalition with the largest aggregate number of levels of players among all coalitions with the highest (per one-unit level-increase) average worth. Then, a sequence of marginal games, each of which is a convex multi-choice game, is considered, that corresponds to the unique sequence of multi-choice coalitions in line with the above mentioned proposition. The marginal game of $v \in MC^{N,m}$ based on $u \in \mathcal{M}^N\setminus\{0\}$ is defined by $v^{-u}(s) := v(s+u)-v(u)$ for each $s \in \mathcal{M}^N \setminus \{0\}$ such that $s \leq m-u$, that is for each $s \in \mathcal{M}_{m-u}^N$. Now, we formulate the Dutta-Ray algorithm for convex multi-choice games.

Step 1: Consider $m^1 := m$, $v_1 := v$. Select the unique element in $\arg\max_{s\in\mathcal{M}_{m^1}^N\setminus\{0\}}\alpha(s,v_1)$ with the maximal aggregate number of levels, say s^1. Define $d_{ij} := \alpha(s^1, v_1)$ for each $i \in car(s^1)$ and $j \in M_i^{s^1}$. If $s^1 = m$, then stop; otherwise, go on.

Step p: Suppose that $s^1, s^2, ..., s^{p-1}$ have been defined recursively and $s^1+s^2+...+s^{p-1} \neq m$. Define a new multi-choice game with player set N and maximal participation profile $m^p := m - \sum_{i=1}^{p-1} m^i$. For each multi-choice coalition $s \in \mathcal{M}_{m^p}^N$, define $v_p(s) := v_{p-1}(s + s^{p-1}) - v_{p-1}(s^{p-1})$. The game $v_p \in MC^{N,m^p}$ is convex. Denote by s^p the (unique) largest element in $\arg\max_{s\in\mathcal{M}_{m^p}^N\setminus\{0\}}\alpha(s,v_p)$ and define $d_{ij} := \alpha(s^p, v_p)$ for all $i \in car(s^p)$ and $j \in \left\{\sum_{k=1}^{p-1} s_i^k + 1, ..., \sum_{k=1}^{p} s_i^k\right\}$.

In a finite number of steps, say P, where $P \leq |M|$, $M = \{(i,j) \mid i \in N,\ j \in M_i\}$, and $|M|$ is the cardinality of the set M, the algorithm will end, and the constructed (level) payoff vector $(d_{ij})_{(i,j)\in M^+}$ is called the (*Dutta-Ray*) *constrained egalitarian solution* $d(v)$ of the convex multi-choice game v. It is proved that the constrained egalitarian solution for convex multi-choice games has similar properties as the constrained egalitarian solution for traditional convex games. Specifically, the constrained egalitarian allocation is a Lorenz undominated element of the precore, and also belongs to the equal division core of the game. We notice that the role of the core for a convex game in G^N is played now by the precore of a convex multi-choice game. However, it is still an open question whether the constrained egalitarian solution of a convex multi-choice game possesses a population monotonicity property regarding players' levels of participation. It turns out that for each convex multi-choice game v the equal split-off set $ESOS(v)$ consists of a unique equal split-off allocation which equals the constrained egalitarian solution of that game, i.e. $ESOS(v) = \{d(v)\}$ for each convex game $v \in MC^{N,m}$.

3. Solution concepts for multi-choice total clan games

Multi-choice clan games are introduced in Branzei, Llorca, Sánchez-Soriano and Tijs [4] to extend the model of traditional clan games (cf. Potters, Poos, Tijs and Muto [18]). In a multi-choice clan game the set N of players consists of two disjoint groups: a fixed (powerful) clan C with 'yes-or-no' choices, and a group of (nonpowerful) non-clan members having more possibilities for being active. Multi-choice clan games are defined using the veto power of clan members, the monotonicity property of the characteristic function, and a (level) union property regarding non-clan members' participation in multi-choice coalitions containing at least all clan members at participation level 1. We denote by $\mathcal{M}^{N,C}$ the set of multi-choice coalitions with player set N and fixed clan C, and by $\mathcal{M}^{N,1_C}$ the set of all multi-choice coalitions containing at least all clan members at participation level 1. For each $s \in \mathcal{M}^{N,C}$ we denote its restrictions to $N \setminus C$ and C, by $s_{N\setminus C}$ and s_C, respectively. Clearly, the maximal participation profile of players in a multi-choice clan game with fixed player set N and fixed clan C is of the form $m = (m_{N\setminus C}, 1_C)$. Formally, a game $< N, (m_{N\setminus C}, 1_C), v >$ is a multi-choice clan game if v satisfies:

(i) Clan property: $v(s) = 0$ if $s_C \neq 1_C$;

(ii) Monotonicity property: $v(s) \leq v(t)$ for all $s, t \in \mathcal{M}^{N,C}$ with $s \leq t$;

(iii) (Level) Union property: For each $s \in \mathcal{M}^{N,1_C}$, $v(m) - v(s) \geq \sum_{i \in N \setminus C} (v(m) - v(m_{-i}, s_i))$, where (m_{-i}, s_i) is the multi-choice coalition where all players $j \in N \setminus C$, $j \neq i$, participate at their maximal level m_j, whereas non-clan member i participates at his level s_i in s.

We denote the set of multi-choice clan games with player set N, fixed clan C and maximal participation profile $m = (m_{N\setminus C}, 1_C)$ by $MC_C^{N,m}$.

The core of a multi-choice game $v \in MC_C^{N,m}$ is explicitly described as

$$C(v) = \{x : M \to \mathbb{R}_+ \mid X(m) = v(m); \sum_{k=j}^{m_i} x_{ik} \leq v(m) - v(m_{-i}, j-1), \text{ for all } i \in N \setminus C,\ j \in M_i^+\}.$$

A multi-choice total clan game is a clan game whose all subgames are also clan games. The subgame of $v \in MC_C^{N,m}$ with respect to $t \in \mathcal{M}^{N,1_C}$ is defined by $v_t(s) := v(s)$ for each $s \in \mathcal{M}_t^{N,1_C}$, where $\mathcal{M}_t^{N,1_C}$ stands for the subset of $\mathcal{M}^{N,1_C}$ with $s_{N\setminus C} \leq t_{N\setminus C}$. The structure of the core of a multi-choice total clan game and that of the core of its subgames play an important role for the existence of bi-monotonic allocation schemes for such games. A (level) total concavity property of multi-choice total clan games also plays a role for the existence of bi-monotonic allocation schemes for such games:

For $s, t \in \mathcal{M}^{N,1_C}$ with $s \leq t$ and for each $i \in car(s_{N\setminus C})$ such that $s_i = t_i$ it holds

$$v(t) - v(t - e^i) \leq v(s) - v(s - e^i).$$

This property reflects the fact that the same one-unit level decrease of a non-clan member in coalitions containing at least all clan members at participation level 1 and where that non-clan member has the same participation level, could be more beneficial in smaller such coalitions than in larger ones. It turns out that, for multi-choice games possessing both the clan property and the monotonicity property, the total concavity property is equivalent with the total (level) union property:

For all $s, t \in \mathcal{M}^{N,1_C}$ with $s \leq t$ it holds:

$$v(t) - v(s) \geq \sum_{i \in car(t_{N\setminus C})} (v(t) - v(t_{-i}, s_i)).$$

A scheme $b = [b_{ij}^t]_{i \in N,\ j \in \{1,\ldots,t_i\}}^{t \in \mathcal{M}^{N,1_C}}$ is called a *bi-(level-increase) monotonic allocation scheme* (*bi-limas*) if it satisfies a stability condition, i.e. $b^t \in C(v_t)$ for each subgame v_t of v with $t \in \mathcal{M}^{N,1_C}$, and a (level) bi-monotonicity property regarding the two types of players, i.e. for all $s, t \in \mathcal{M}^{N,1_C}$ with $s \leq t$ it holds: (i)$b_{i1}^s \leq b_{i1}^t$ for each $i \in C$, and (ii) $b_{ij}^s \geq b_{ij}^t$ for each $i \in car(s_{N\setminus C})$ and each $j \in \{1, ..., s_i\}$.

This kind of bi-monotonic allocation schemes are introduced in Branzei, Llorca, Sánchez-Soriano and Tijs [4] and studied by means of suitably defined compensation-sharing rules $\psi^{\alpha,\beta} : MC_C^{N,m} \to \mathbb{R}^{|M|}$, where $\alpha \in [0,1]^{N\setminus C}$ and $\beta \in \Delta(C)$, with $\Delta(C)$ being the unit simplex whose coordinates correspond to clan members.

The i-th coordinate α_i of the compensation vector α indicates the share, to be given to level 1 of non-clan member i, of i's contribution to the grand coalition m, whereas the i-th coordinate β_i of the sharing vector β determines the share of the remainder for the clan given to clan member i. It turns out that for a subclass of multi-choice total clan games compensation-sharing rules $\psi^{\alpha,\beta}$ with $\alpha \in [0,1]^{N\setminus C}$ and $\beta \in N(C)$ generate bi-(level-increase) monotonic allocation schemes. Furthermore, some elements x in the core of each multi-choice game in that subclass of total clan games are extendable to a bi-limas, that is there exists a bi-limas $[b_{ij}^t]_{i \in N,\ j \in \{1,\ldots,t_i\}}^{t \in \mathcal{M}^{N,1_C}}$ such that $b_{ij}^m = x_{ij}$ for each $i \in N$, $j \in M_i^+$. Clearly, when $s_{N\setminus C} = 1_{N\setminus C}$ a bi-limas coincides with a bi-mas (cf. Voorneveld, Tijs and Grahn [21]), and we obtain as a particular case that each core element of a total clan game in G^N is extendable to a bi-mas.

References

[1] Branzei R., Dimitrov D., and Tijs S. (2005). *Models in Cooperative Game Theory: Crisp, Fuzzy, and Multi-Choice Games*, Lecture Notes in Economics and Mathematical Systems, Springer.

[2] Branzei, R., Dimitrov D., and Tijs S. (2008): *Models in Cooperative Game Theory*, Springer (forthcoming).

[3] Branzei R., Llorca N., Sánchez-Soriano J., and Tijs S. (2007a). Egalitarianism in multi-choice games, *CentER DP 2007-55*, Tilburg University, The Netherlands.

[4] Branzei R., Llorca N., Sánchez-Soriano J., and Tijs S. (2007b). Multi-choice total clan games: characterizations and solution concepts, *CentER DP 2007-77*, Tilburg University, The Netherlands.

[5] Branzei R., Tijs S., and Zarzuelo J. (2007). Convex multi-choice cooperative games and their monotonic allocation schemes, *CentER DP 2007-54*, Tilburg University, The Netherlands.

[6] Calvo E., and Santos J.C. (2000): A value for multi-choice games, *Mathematical Social Sciences*, **40**, 341-354.

[7] Derks, J. and Peters H. (1993): A Shapley value for games with restricted coalitions, *International Journal of Game Theory*, **21**, 351-360.

[8] Grabisch M., and Lange F. (2007). Games on lattices, multichoice games and the Shapley value: A new approach, *Mathematical Methods of Operations Research*, **65**, 153-167.

[9] Grabisch M., and Xie L. (2007). A new investigation about the core and the Weber set of multi-choice games, *Mathematics of Operations Research*, forthcoming.

[10] Hsiao C.-R., and Raghavan TES (1992). Monotonicity and dummy free property for multi-choice cooperative games, *International Journal of Game Theory*, **21**, 301-312.

[11] Hsiao C.-R., and Raghavan TES (1993): Shapley value for multi-choice cooperative games (I), *Games and Economic Behavior* **5**, 240-256.

[12] Kalai E., and Samet D. (1987). On weighted *Shapley values*, *International Journal of Game Theory*, **16**, 205-222.

[13] Klijn F., Slikker M., and Zarzuelo J. (1999). Characterizations of a multi-choice value, *International Journal of Game Theory*, **28**, 521-532.

[14] Moulin H. (1995). On additive methods to share joint costs, *Japanese Economic Review*, **46**, 303-332.

[15] Nouweland van den, A. (1993). *Games and Graphs in Economic Situations*, PhD Thesis, Tilburg University, The Netherlands.

[16] Nouweland van den, A., Potters J., Tijs S., and Zarzuelo J. (1995). Cores and related solution concepts for multi-choice games, *Mathematical Methods of Operations Research*, **41**, 289-311.

[17] Peters H., and Zank H. (2005). The egalitarian solution for multichoice games, *Annals of Operations Research*, **137**, 399-409.

[18] Potters J., Poos R., Tijs S., and Muto S. (1989). Clan games, *Games and Economic Behavior*, **1**, 275-293.

[19] Shapley L.S. (1953). A value for n-person games, *Annals of Mathematics Studies*, **28**, 307-317.

[20] Sprumont Y. (1990). Population monotonic allocation schemes for cooperative games with transferable utility, *Games and Economic Behavior*, **2**, 378-394.

[21] Voorneveld M., Tijs S., and Grahn S. (2002). Monotonic allocation schemes in clan games, *Mathematical Methods of Operations Research*, **56**, 439-449.

[22] Young H.P. (1985). Monotonic solutions of cooperative games, *International Journal of Game Theory*, **14**, 65-72.

Rodica Branzei is Associate Professor at Faculty of Computer Science, Alexandru Ioan Cuza University, Iaşi, Romania (PhD University of Bucharest). Her research interests include game theory, operations research and project management. She has published papers in International Journal of Game Theory, International Game Theory Review, European Journal of Operational Research, Mathematical Methods of Operations Research, Annals of Operations Research, Fuzzy Sets and Systems, Mathematical Social Sciences, TOP, Journal of Operations Research Society of Japan, Fuzzy Economic Review, Journal of Public Economic Theory, Theory and Decision, International

Journal of Uncertainty, Fuzziness and Knowledge-based Systems, Journal of Mathematical Analysis and Applications, Economics Bulletin, Libertas Mathematica, Journal of Mathematical Economics, Balkan Journal of Geometry and its Applications, Italian Journal of Pure and Applied Mathematics, Scientific Annals of the Alexandru Ioan Cuza University (Computer Science Section), Revue Roumaine de Mathématiques Pures et Appliquées, Mathematical Reports, Bulletin Mathématique de la Société de Sciences Mathématiques de Roumanie, Annals of University of Bucharest, Nihonkai Mathematical Journal, Banach Center Publications.

R. Branzei has also published textbooks for her students, books and chapters in Proceedings of conferences. Since 2005, R. Branzei has taught her students the Game Theory elective course and the Banking Mathematics elective course at the Faculty of Computer Science.

3. ARTÍCULOS DE APLICACIÓN

SOME PROBABILITY APPLICATIONS TO THE RISK ANALYSIS IN INSURANCE THEORY

Jinzhi Li
College of Science
Central University for Nationalities, China

Shixia Ma*
School of Sciences
Hebei University of Technology, China

Abstract

This work concerns with some probability applications in insurance theory. The problem to determine the ruin probability of an insurance company is considered. We show that by using some stochastic models and considering several probabilistic procedures such a probability can be approached. As illustration, an application to car insurance is provided.

Keywords: Risk analysis, Stochastic modelling, Insurance theory applications.

1. Introduction

Risk analysis in insurance theory is an important area for probability and statistical applications. In particular, the probabilistic modelling of the surplus evolution in an insurance company has received some attention in the specialized literature. In fact, denoting by $U(t)$ the surplus of an insurance company at time t, the classical model establishes that:

$$U(t) = u + ct - S(t),\ t \geq 0, \tag{1.1}$$

where $u > 0$ is the initial surplus, c is the constant rate at which the premiums are received per unit time and $S(t)$ is the aggregate amount concerning the claims reported in the time interval $[0, t]$. It is assumed that $\{S(t), t \geq 0\}$ is a compound Poisson process,

$$S(t) = \sum_{i=1}^{N(t)} X_i,\ t \geq 0$$

$\{N(t), t \geq 0\}$ being a Poisson process and $\{X_n, n = 1, 2, \ldots\}$ a sequence of independent and identically distributed random variables, both assumed to be independent. The variable $N(t)$ represents the number of reported claims in $[0, t]$ and X_i is the amount corresponding to the ith claim.

In general, from model (2.1), some results in risk theory have been derived but such a model it is not flexible enough in order to describe the probabilistic evolution of $U(t)$ in more complex real situations. In an attempt to contribute some solution to this problem, several classes of stochastic models have been introduced and some theory and applications about them developed. We will quote, for example:

(a) Stochastic models considering claim arrivals governed by non-Poissonian processes. See e.g. [4] where it is assumed that the number of claims is described through a Cox process.

(b) Stochastic models allowing a rate $c(\cdot)$ which change through a Markov process. See e.g. [1] or [6] where it is considered that the rate changes modulated by an underlying irreducible Markov chain or a premium rate in a Markovian environment, respectively.

(c) Stochastic models including a diffusion component which represents uncertainties in both the premium income and the costs. Firstly studied by Gerber, such models have been applied in several risk problems, see e.g. [2], [3] or [8].

In this work, we will focus our interest in a class of risk models which allows premium rate and

*Corresponding Author. E-mail: mashixia1@163.com

diffusion component depending on an underlying continuous-time Markov chain. Moreover, the occurrence of claims is assumed to be well-described by a Cox process. It is usually referred as *the class of risk models with Markov modulated speed*. We will center our attention in the determination of the ruin probability for an insurance company. This problem is an important objective in many research developed in actuarial risk theory. We will show by using some classical probabilistic techniques that it is possible to obtain the ruin probability.

The paper is organized as follow: In Section 2, we provide the probabilistic descriptions of such a class of risk models. In Section 3, we develop a probabilistic procedure to determine the ruin probability. Finally, Section 4 is devoted to considering an application in car insurance.

2. Stochastic modelling

Let us consider the following stochastic modelling for $U(t)$, $t \geq 0$:

$$U(t) = u + \int_0^t c(I(s))ds - \sum_{i=1}^{N(t)} X_i + \int_0^t \sigma(I(s))dW_s \tag{2.1}$$

where:

(a) $\{N(t), t \geq 0\}$, $N(0) = 0$, is a Cox point process. $N(t)$ represents the number of claims received for the insurance company during the interval $[0, t]$.

(b) $\{X_n, n = 0, 1, \ldots\}$ is a sequence of independent and identically distributed random variables which represents the amount corresponding to the successive claims received by the company in $[0, t]$. Let us write $F(x) = P(X_1 \leq x)$, $x > 0$, $F(0) = 0$, and $\mu = E[X_1]$.

(c) $\{W(t), t \geq 0\}$ is a standard Wiener process with diffusion coefficient $\sigma(\cdot)$. This process represents the disturbances originated from several tiny stochastic factors.

(d) $\{I(t), t \geq 0\}$ is an underlying continuous-time homogeneous Markov chain with state space $S = \{1, \ldots, n\}$ assumed to be irreducible.

Notice that in (2.1) the premium rate $c(\cdot)$ and the diffusion coefficient $\sigma(\cdot)$ depend on the current state of the Markov chain $\{I(t), t \geq 0\}$. In fact, if at time t it is verified that $I(t) = i$ then, by simplicity, $c(I(t))$ and $\sigma(I(t))$ will be denoted, respectively, as c_i and σ_i, assumed to be positive. On the other hand, we are considering that the number of reported claims is governed by a Cox process, hence if $I(s) = i$, $s \in [0, t]$ then the number of claims received in $[0, t]$ has a Poisson distribution with mean $\lambda_i > 0$.

According to [5], we shall denote by q_i the rate at which the Markov chain $\{I(t), t \geq 0\}$ leaves the state i, and by q_{ij} and p_{ij}, respectively, the transition intensity and the transition probability that it leaves the state i for the first time and enter into the state j immediately. Assuming that $p_{ii} = 0$, $i \in S$, one deduces that $q_{ij} = q_i p_{ij}$ for $i \neq j$ and $q_{ii} = -q_i$. Since all the states communicate,

$$\pi_i q_i = \sum_{j=1}^{n} \pi_j q_j p_{ji}. \tag{2.2}$$

where $\pi_1, \pi_2, ..., \pi_n$ denotes a stationary distribution corresponding to $\{I(t), t \geq 0\}$.

Also, we will assume that the named safety loading is positive, namely $c - \lambda\mu > 0$ where:

$$c = \sum_{i=1}^{n} \pi_i c_i \quad \text{and} \quad \lambda = \sum_{i=1}^{n} \pi_i \lambda_i.$$

3. Ruin probability

In this section we are interested in the determination of the ruin probability defined by:

$$\psi(u) = \sum_{i=1}^{n} \pi_i \psi_i(u) \tag{3.1}$$

where $\psi_i(u) = P(U(t) \leq 0 \mid U(0) = u, I(0) = i)$ for some $t \geq 0$.

First, by considering the evolution of $U(t)$ in a short interval $[0, h)$, $h > 0$, we shall determine a system of equations for $R_i(u) = 1 - \psi_i(u)$, $i \in S$. In fact, assuming that $I(0) = i$, we can consider the following possibilities during the interval $[0, h)$:

(a) There is not reported claims and $I(s) = i$ for $s \in [0, h)$.

(b) One claim is produced but the amount to be paid for such a claim does not cause ruin and $I(s) = i$ for $s \in [0, h)$.

(c) There is not reported claims and, from the state i a change to other state it is produced.

(d) At least one claim is reported in and at least one change of state, from i, is produced.

Consequently, see for more details [2], on has for $t \in [0,h)$,

$$\begin{aligned} R_i(u) &= (1-\lambda_i h - q_i h + o(h))E[R_i(\varphi_i(u,h)] \\ &+(\lambda_i h + o(h))(1-q_i h + o(h)) \\ &E[\int_0^{\varphi_i(u,h)} R_i(\varphi_i(u,h)-x)dF(x)] \\ &+(1-\lambda_i h + o(h))(q_i h + o(h)) \\ &\sum_{j=1}^{n} p_{ij} E[R_j(\varphi_i(u,h))] + o(h). \end{aligned} \tag{3.2}$$

where $\varphi_i(u,h) = u + c_i h + \sigma_i W_h$.

By considering the Taylor expression in u of $E[R_i(\varphi_i(u,h))]$, dividing by h, and taking limit as $h \downarrow 0$, it is matter of some straightforward calculation to deduce,

$$\begin{aligned} \frac{\sigma_i^2}{2} R_i^{''}(u) + c_i R_i^{'}(u) &= (\lambda_i + q_i) R_i(u) \\ &- \lambda_i \int_0^u R_i(u-x)dF(x) \\ &- q_i \sum_{j=1}^{n} p_{ij} R_j(u). \end{aligned} \tag{3.3}$$

By integration of (3.3) on $[0,t]$ and using the fact that $R_i(0) = 0$,

$$\begin{aligned} \frac{\sigma_i^2}{2} R_i^{'}(t) + c_i R_i(t) &= \frac{\sigma_i^2}{2} R_i^{'}(0) \\ &+ (\lambda_i + q_i) \int_0^t R_i(u)du \\ &- \lambda_i \int_0^t \int_0^u R_i(u-x)dF(x)du \\ &- q_i \sum_{j=1}^{n} p_{ij} \int_0^t R_j(u)du. \end{aligned}$$

and, taking into account that

$$\int_0^t \int_0^u R_i(u-x)dF(x)du = \int_0^t R_i(u)du + \int_0^t R_i(t-x)F^*(x)dx.$$

where $F^*(x) = 1 - F(x)$, considering the fact that $\psi_i(t) = 1 - R_i(t)$, $i = 1, \ldots, n$,

$$\begin{aligned} &\frac{\sigma_i^2}{2}\psi_i^{'}(t) + c_i\psi_i(t) = \\ &c_i + \frac{\sigma_i^2}{2}\psi_i^{'}(0) + \lambda_i \int_0^t F^*(x)dx \\ &-\lambda_i \int_0^t \psi_i(t-x)F^*(x)dx \\ &+q_i \int_0^t \psi_i(u)du - q_i \sum_{j=1}^{n} p_{ij} \int_0^t \psi_j(u)du. \end{aligned} \tag{3.4}$$

Finally, taking limit as $t \uparrow \infty$, we deduce the following system of equations for $\psi_i(u)$, $i = 1, \ldots, n$:

$$\begin{aligned} \psi_i^{'}(0) &= \frac{2}{\sigma_i^2} \left(-c_i - \lambda_i \mu - q_i \int_0^{\infty} \psi_i(u)du \right. \\ &\left. +q_i \sum_{j=1}^{n} p_{ij} \int_0^{\infty} \psi_j(u)du \right). \end{aligned} \tag{3.5}$$

Using the numerical solutions of (3.5), from (3.1) we may determine the corresponding ruin probability.

Note that when $c_i = c$ and $\sigma_i = \sigma$, $i = 1, \ldots, n$, by (3.4) and (3.5), taking into account (2.2) we deduce:

$$\frac{\sigma^2}{2}\psi^{'}(t) + c\psi(t) = c + \frac{\sigma^2}{2}\psi^{'}(0) + \lambda \int_0^t F^*(x)dx$$

$$-\sum_{i=1}^{n} \pi_i \lambda_i \int_0^t \psi_i(t-x)F^*(x)dx,$$

and

$$\psi^{'}(0) = \frac{-2}{\sigma^2}(c + \lambda\mu)$$

respectively.

4. Application to car insurance

It is well-known the influence that certain factors, for example the inclemency of the weather, the conditions of the roads, and so on, have in the occurrence of traffic accidents. Next we shall consider an application of the model (2.1) in car insurance.

In a first approximation, we will assume that

the underlying Markov chain $\{I(t), t \geq 0\}$ has a two-states space, namely $S = \{1, 2\}$, where:

- The state 1 represents the risk under normal conditions.
- The state 2 represents the risk under bad conditions (for e.g. slippery roads, foggy days or high traffic volume.

We refer the reader to [5] and [7] for more details. Also, we will consider that the variable X_i has exponential distribution with mean μ and that $p_{12} = p_{21} = 1$, $p_{11} = p_{22} = 0$. Then,

$$q_{11} = -q_1, \quad q_{22} = -q_2, \quad q_{12} = q_1, \quad q_{21} = q_2$$

and

$$\pi_1 = q_2(q_1 + q_2)^{-1}, \quad \pi_2 = q_1(q_1 + q_2)^{-1}.$$

Let us denote by

$$\phi_i(s) = \int_0^\infty e^{-st}\psi_i(t)dt$$

and

$$\phi^*(s) = \int_0^\infty e^{-st}F^*(t)dt.$$

Taking into account that

$$\int_0^\infty e^{-st}\psi_i^{'}(t)dt = s\phi_i(s) - 1,$$

and

$$\int_0^\infty e^{-st}\int_0^t \psi_i(u)dudt = \frac{1}{s}\phi_i(s).$$

by using the Laplace transformation in (3.4), ones deduces, for $i = 1, 2$,

$$\left(c_i + \frac{\sigma_i^2}{2}s - \frac{q_i}{s} + \lambda_i\phi^*(s)\right)\phi_i(s) + \frac{q_i}{s}\sum_{j=1}^{n} p_{ij}\phi_j(s) =$$

$$\frac{\sigma_i^2}{2} + \frac{1}{s}(c_i + \frac{\sigma_i^2}{2}\psi_i^{'}(0)) + \frac{\lambda_i}{s}\phi^*(s)$$

hence,

$$\left(c_1 + \frac{\sigma_1^2}{2}s - \frac{q_1}{s} + \frac{\lambda_1\mu}{s\mu + 1}\right)\phi_1(s) + \frac{q_1}{s}\phi_2(s) =$$

$$\frac{\sigma_1^2}{2} + \frac{1}{s}\left(c_1 + \frac{\sigma_1^2}{2}\psi_1^{'}(0)\right) + \frac{\lambda_1\mu}{s(s\mu + 1)}$$

and

$$\left(c_2 + \frac{\sigma_2^2}{2}s - \frac{q_2}{s} + \frac{\lambda_2\mu}{s\mu + 1}\right)\phi_2(s) + \frac{q_2}{s}\phi_1(s) =$$

$$\frac{\sigma_2^2}{2} + \frac{1}{s}\left(c_2 + \frac{\sigma_2^2}{2}\psi_2^{'}(0)\right) + \frac{\lambda_2\mu}{s(s\mu + 1)}.$$

Conclusion:
The surplus evolution corresponding to an insurance company could be suitably described in terms of the general model given in (2.1). From a practical point of view, by solving the system given in (3.5) and taking into account expression (3.1), it is possible to determine the ruin probability. This parameter plays a crucial role in research about risk analysis in insurance theory.

Acknowledgements: The authors would like to thank Professor Molina from the Department of Mathematics at Extremadura University for his constructive suggestions which have improved this paper. This research has been supported by the Funded Projects of the Education Department of Hunnan Province, grant number 07C745.

References

[1] Asmussen, S. and Kella, O. (1996). Rate modulation in dams and ruin problems, *Journal of Applied Probability* **33**, 523-535.

[2] Dufresne, F. and Gerber, H. U. (1991). Risk theory for the compound poisson process that is perturbed by diffusion. *Insurance: Mathematics and Economics.* **10**, 51-59.

[3] Gerber, H. U. and Landry, B. (1998). On the discounted penalty at ruin a jump-diffusion and the perpetual put option. *Insurance: Mathematics and Economics.* **22**, 263-276.

[4] Gerber, H. U. and Shiu, E. S. W. (1998). On the time value of ruin. *North American Actuarial Journal.* **2**(1), 48-78.

[5] Grandell, J. (1991). Aspects of Risk Theory. Springer, Berlin.

[6] Jasiulewicz, H. (2001). Probability of ruin with variable premium rate in a Markovian environ-

ment. *Insurance: Mathematics and Economics.* **29**, 291-296.

[7] Reinhard, J.M. (1984). On a class of semi-Markov risk models obtained as classical risk models in Markovian environment. *ASTIN Bulletin XIV*, 23-43.

[8] Wang, G. (2001). A decomposition of the ruin probability for the risk process perturbed by diffusion. *Insurance: Mathematics and Economics.* **28**, 49-59.

4. ESTADÍSTICA OFICIAL

IMPUTATION IN THE SURVEY ON LIVING CONDITIONS

José María Méndez Martín*
Instituto Nacional de Estadística

Abstract

The Spanish "European Statistics on Income and Living Conditions" (EU-SILC) is one of the statistical operations that has been harmonised to EU standards. In this household survey there are two kinds of non-response: unit non-response (one or several household or individual questionnaires are missing) and item non-response (no questionaire is missing but some variables are). The main target variable is the total household income, which is defined as the aggregate of different income components. Components with missing values are imputed when they cannot be estimated with the help of other variables or other information in the questionnaire of the current or previous surveys. The procedure applied to the data preserves the variability of the variables and the correlations between them. The statistical software used for imputation is the IVEware. The IVEware implements a multivariate model involving a multiple regression sequence where imputation is carried out variable by variable generating draws from the predictive distribution specified by the regression model. An iterative imputation scheme is used, updating previous imputed values in order to better preserve the correlation among variables.

Keywords: EU-SILC, imputation, item non-response, IVEware.

1. Introducción

La Encuesta de Condiciones de Vida (ECV) es una encuesta anual del tipo panel rotatorio que recoge variables relacionadas con los ingresos y las condiciones de vida de los hogares. Se inició en 2004 y es una operación estadística armonizada de ámbito europeo. En INE (2007) se detalla la metodología de dicha encuesta.

Cada hogar que participa en la encuesta ha de cumplimentar un cuestionario de hogar y un cuestionario individual para cada persona adulta. Las áreas que abarca el cuestionario de hogar son: componentes de ingresos que son propios de la unidad hogar (ayudas sociales a la vivienda o al hogar, rentas de la propiedad, etc.), vivienda, equipamiento del hogar, etc. El cuestionario de personas adultas recoge información sobre componentes de ingresos que son propios de la unidad persona (salarios, renta de autónomos, prestaciones sociales, etc.), situación en la actividad, trabajo actual, estudios que realiza, máximo nivel de estudios alcanzado, salud, etc. En consonancia con los dos tipos de cuestionarios también hay dos factores de elevación, uno de hogares y otro de adultos, empleándose en las estimaciones uno u otro dependiendo de la unidad considerada.

En la ECV se pueden presentar tres tipos de falta de respuesta: total, individual y parcial. Cada tipo de falta de respuesta se trata de forma diferente. La falta de respuesta total consiste en que un hogar no colabora y, por tanto, no se recoge ningún cuestionario. El tratamiento de este tipo de falta de respuesta se realiza recalculando los factores de elevación de los hogares que sí han respondido. La falta de respuesta individual consiste en que en un hogar colaborador no se obtienen algunos de los cuestionarios de los adultos del hogar. La falta de respuesta individual se corrige ajustando los factores de elevación de los adultos de los que sí se obtiene cuestionario. Sin embargo este tipo de falta de respuesta también afecta a las variables del hogar que se obtienen por agregación de variables de adultos (por ejemplo los ingresos del trabajo por cuenta ajena del hogar, que es la suma de los ingresos por cuenta ajena de los miembros del hogar). Por ello, se calcula, para cada hogar, un factor multiplicador que intenta corregir la renta que falta de los cuestionarios individuales no cumplimentados. Fi-

*Corresponding Author. E-mail: jmmendez@ine.es

nalmente, la falta de respuesta parcial consiste en que no falta ningún cuestionario, pero algunas variables no están debidamente cumplimentadas. Para tratar este tipo de falta de respuesta se aplica el método de regresión secuencial multivariante, utilizando el software IVE, desarrollado por el *Institute for Social Research* de la Universidad de Michigan.

A partir de la información recogida en la encuesta se obtienen unos ficheros de trabajo que reorganizan la información, haciéndola más manejable para el investigador. En estos ficheros se trata la falta de respuesta en sus distintas modalidades. En particular, en el caso de la falta de respuesta parcial de las variables relacionadas con los ingresos, se lleva a cabo una imputación.

2. Imputación de los ingresos en la ECV

Los ingresos totales del hogar se calculan a partir de sus componentes. No sería adecuado dar por perdidos todos los ingresos cuando falta sólo algún componente. Por tanto resulta esencial realizar las imputaciones de componentes de los ingresos en aquellos casos donde razonablemente se puede llevar a cabo.

Un Reglamento de la Comisión Europea sobre aspectos de trabajo de campo y procedimientos de imputación da algunas recomendaciones. En concreto especifica que "el procedimiento aplicado a los datos debería preservar la variabilidad de las variables y la correlación entre ellas. Los métodos que incluyan un «componente de error» en los valores imputados serán preferibles a los que imputen simplemente un valor determinado. Los métodos que tengan en cuenta la estructura de las correlaciones (u otras características de la distribución conjunta de las variables) serán preferibles al enfoque marginal o univariante.". Como se verá más adelante estos principios están contemplados en el método de imputación de la ECV, que sigue un modelo similar al utilizado por Eurostat en el Panel de Hogares de la Unión Europea (EUROSTAT 2001).

La imputación en la ECV se lleva a cabo después de la depuración de los datos, que ha de ser realizada minuciosamente, tanto a nivel de microdatos como en los resultados agregados. La depuración se inicia en la entrevista personal durante la fase de recogida de la información, ya que la aplicación informática tiene incorporados una serie de controles que detectan posibles errores e inconsistencias. En los Servicios Centrales se aplican unos controles exhaustivos desarrollados por la unidad promotora del INE y también unos programas de chequeo de Eurostat. La depuración manual se realiza mediante una aplicación que visualiza el contenido de los datos recogidos en los cuestionarios de los diferentes años.

La depuración permite corregir una parte de la falta de respuesta parcial. También se detectan y eliminan los "outliers" antes de realizar las imputaciones. Con este fin se fijan unos límites mediante observación de la distribución de los importes extremos declarados y se ponen como valores perdidos los que no estén dentro de estos límites.

A partir del segundo año de producción de la ECV, las imputaciones de los ingresos se pueden realizar utilizando los datos disponibles del año anterior. Aprovechando la componente longitudinal de la encuesta, cuando un importe falta en el año t y no falta en el año t-1, se imputa el importe de t multiplicando por un valor el importe de t-1.

Cuando no es posible obtener el valor del ingreso en la depuración o con datos disponibles del año anterior, se realiza la imputación a partir de la información disponible. En algunos casos se dispone del tramo en el que está situado el importe que falta. En este caso se imputará el importe con la restricción del intervalo proporcionado. Cuando no se dispone ni siquiera del tramo entonces se imputa el importe con una restricción construida a partir de los percentiles 10 y 90 de la distribución de los respondientes. La imputación en esta fase se realiza aplicando el método de imputación de regresión secuencial multivariante, imputando solamente las variables de ingresos.

3. Método de imputación de regresión secuencial multivariante

En la imputación se utiliza una técnica regresión multivariante basada en unos modelos que implementa el software IVE. Es un procedimiento de imputación general multivariante que puede tratar datos con una estructura compleja y que permite añadir residuos aleatorios. La descripción completa del método se puede encontrar en Raghunathan, Lepkowski, Van Hoewyk y Solenberger (2001) y en las referencias que contiene esta publicación. Es posible descargar el software de imputación desde la

página web "www.isr.umich.edu/src/smp/ive".

El procedimiento se basa en crear imputaciones por medio de una secuencia de regresiones. Se pretende recoger la correlación de todas las variables, tanto las completas como de las que tienen valores perdidos. El programa permite distintos tipos de regresiones (lineal, logística, logística generalizada y de Poisson). Sin embargo, en el caso de la ECV solamente se utiliza regresión lineal para imputar los ingresos, previa aplicación a éstos de una transformación logarítmica. Las variables explicativas pueden ser discretas, continuas o binarias.

En el modelo de regresión la aleatoriedad en la estimación se introduce por dos vías: por una parte se considera el término correspondiente al residuo aleatorio y por otra se incorpora una perturbación en los coeficientes de regresión estimados. La distribución que se obtiene con este enfoque se puede consultar en Gelman, Carlin, Stern y Rubin (1995).

Con este software es posible considerar el intervalo en el que está el valor imputado, es decir, se puede imponer un valor mínimo y máximo al valor imputado para cada registro. Este aspecto es importante en la imputación de ingresos ya que en el cuestionario, para muchos componentes de la renta, se solicita en primer lugar el importe exacto y, cuando éste se desconoce, se solicita el intervalo.

El procedimiento general de imputación sigue la siguiente estrategia. Supongamos que X es la matriz de datos construida con todas las variables completas (es decir, que no tienen ningún valor perdido). X se compone de variables explicativas como sexo, edad, región, nivel de estudios y otras que pueden ser continuas (ya transformadas si es necesario), binarias o categóricas.

Por otra parte sean Y1 , Y2 , Y3 , Y4 , ... , Yk las variables que tienen valores perdidos. Suponemos que las variables Y están ordenadas de menor a mayor falta de respuesta. En total se tienen las variables:

X1 , X2 , X3 , ... , Y1 , Y2 , Y3 , Y4 , ... , Yk

En la iteración inicial se imputa según las siguientes distribuciones condicionales:

[Y1 /X]
[Y2 /X,Y1]
...
[Yk /X,Y1 ... Yk-1]

donde [A /B] denota la distribución de A condicionada por B.

Por tanto, en esta iteración inicial se empieza haciendo la regresión de la variable con menos falta de respuesta Y1 sobre las variables explicativas X. Una vez obtenida una "predicción" de Y1 se incorpora esta variable a la matriz X de las variables completas y se obtiene la matriz [X,Y1]. A continuación se repite el proceso con la siguiente variable (Y2) tomando como variables explicativas [X,Y1] Se repite el proceso con Y3 , Y4 , ... , Yk hasta que todas las variables han sido imputadas.

Una vez que se ha realizado esta iteración de regresiones se tiene una primera imputación de todos los valores perdidos. Esta imputación mantiene la estructura multivariante de las variables imputadas con X y con algunas de las Y. Así en la imputación de Ym se tienen en cuenta las X y las variables Y1 ... Ym-1. Sin embargo las variables Ym+1 ... Yk no se han tenido en cuenta en la imputación de Ym.

En las iteraciones siguientes, utilizando la estrategia de esta iteración inicial, lo que se hace es repetir esta iteración pero incluyendo como variables explicativas todas las variables, ya que ahora no hay valores perdidos en ninguna de ellas.

Iteración 2:
[Y1 /X,Y2 ... Yk]
[Y2 /X,Y1,Y3,Y4 ... Yk]
...
[Yk /X,Y1 ... Yk-1]
...
...
Iteración "n":
[Y1 /X,Y2 ... Yk]
[Y2 /X,Y1,Y3,Y4 ... Yk]
...
[Yk /X,Y1 ... Yk-1]

En cada iteración se actualizan las imputaciones hechas en la iteración anterior. Así se obtienen actualizaciones que van recogiendo de una manera más completa la estructura de correlaciones del conjunto de las variables. Este proceso se detiene cuando se alcanza el número de iteraciones especificado por el usuario.

Referencias

[1] EUROSTAT (2001). Imputation of Income in the ECHP. Doc. Pan 164/2001.

[2] Gelman, Carlin, Stern y Rubin (1995). *Bayesian Data Analysis*, Chapman and Hall, London.

[3] INE (2007). Encuesta de Condiciones de Vida. Metodología.

[4] Raghunathan, Lepkowski, Van Hoewyk y Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models, Survey Methodology. *Statistics Canada*, **27(1)**.

Corresponsales

Luz Braña Rey
luzmari@ine.es
Instituto Nacional de Estadística

Luis Felipe Rivera Galicia
luisf.rivera@uah.es
Universidad de Alcalá de Henares

Fernando Reche Lorite
freche@ual.es
Universidad de Almería

Ana Justel
ana.justel@uam.es
Universidad Autónoma de Madrid

Jordi Ocaña
jocana@ub.edu
Universitat de Barcelona

Luis Antonio Sarabia Peinador
lsarabia@ubu.es
Universidad de Burgos

Gabriel Ruiz Garzón
gabriel.ruiz@uca.es
Universidad de Cádiz

Araceli Tuero
tueroma@unican.es
Universidad de Cantabria

Isabel Molina Peralta
imolina@est-econ.uc3m.es
Universidad Carlos III de Madrid

Licesio Rodríguez Aragón
L.RodriguezAragon@uclm.es
Universidad de Castilla-La Mancha

Susana Muñoz López
smunoz@estad.ucm.es
Universidad Complutense de Madrid

José María Caridad y Ocerín
ccjm@uco.es
Universidad de Córdoba

José Antonio Vilar Fernández
eijoseba@udc.es
Universidade da Coruña

Miguel González Velasco
mvelasco@unex.es
Universidad de Extremadura

Vera Pawlowsky-Glahn
vera.pawlowsky@ima.udg.es
Universitat de Girona

Rocío Raya Miranda
rraya@ugr.es
Universidad de Granada

Beatriz Hernández Jiménez
beatriz.hernandez@dmat.uhu.es
Universidad de Huelva

Emilio Lozano Aguilera
elozano@ujaen.es
Universidad de Jaén

David Alcaide López de Pablo
dalcaide@ull.es
Universidad de la Laguna

María Eva Vallejo Pascual
eva.vallejo@unileon.es
Universidad de León

Carles Capdevila Marques
ccm@matematica.udl.es
Universitat de Lleida

Carmen Morcillo Aixelá
aixela@uma.es
Universidad de Málaga

Marc Almiñana Alemany
marc@umh.es
Universidad Miguel Hernandez

José Fernández Hernández
josefdez@um.es
Universidad de Murcia

Susana Montes Rodríguez
montes@uniovi.es
Universidad de Oviedo

Dolores Romero Morales
Dolores.Romero-Morales@sbs.ox.ac.uk
University of Oxford

Maria del Pilar Moreno Navarro
mpmornav@upo.es
Universidad Pablo de Olavide

Pilar Muñoz
pilar.munyoz@upc.edu
Universitat Politécnica de Catalunya

Javier Alcaraz Soria
jalcaraz@eio.upv.es
Universitat Politécnica de Valencia

Ana Fernández Militino
militino@unavarra.es
Universidad Pública de Navarra

Antonio Alonso Ayuso
antonio.alonso@urjc.es
Universidad Rey Juan Carlos

Juan Carlos Fillat Ballesteros
juan-carlos.fillat@dmc.unirioja.es
Universidad de la Rioja

María José Lombardía Cortiña
mjoselc@usc.es
Universidade de Santiago de Compostela

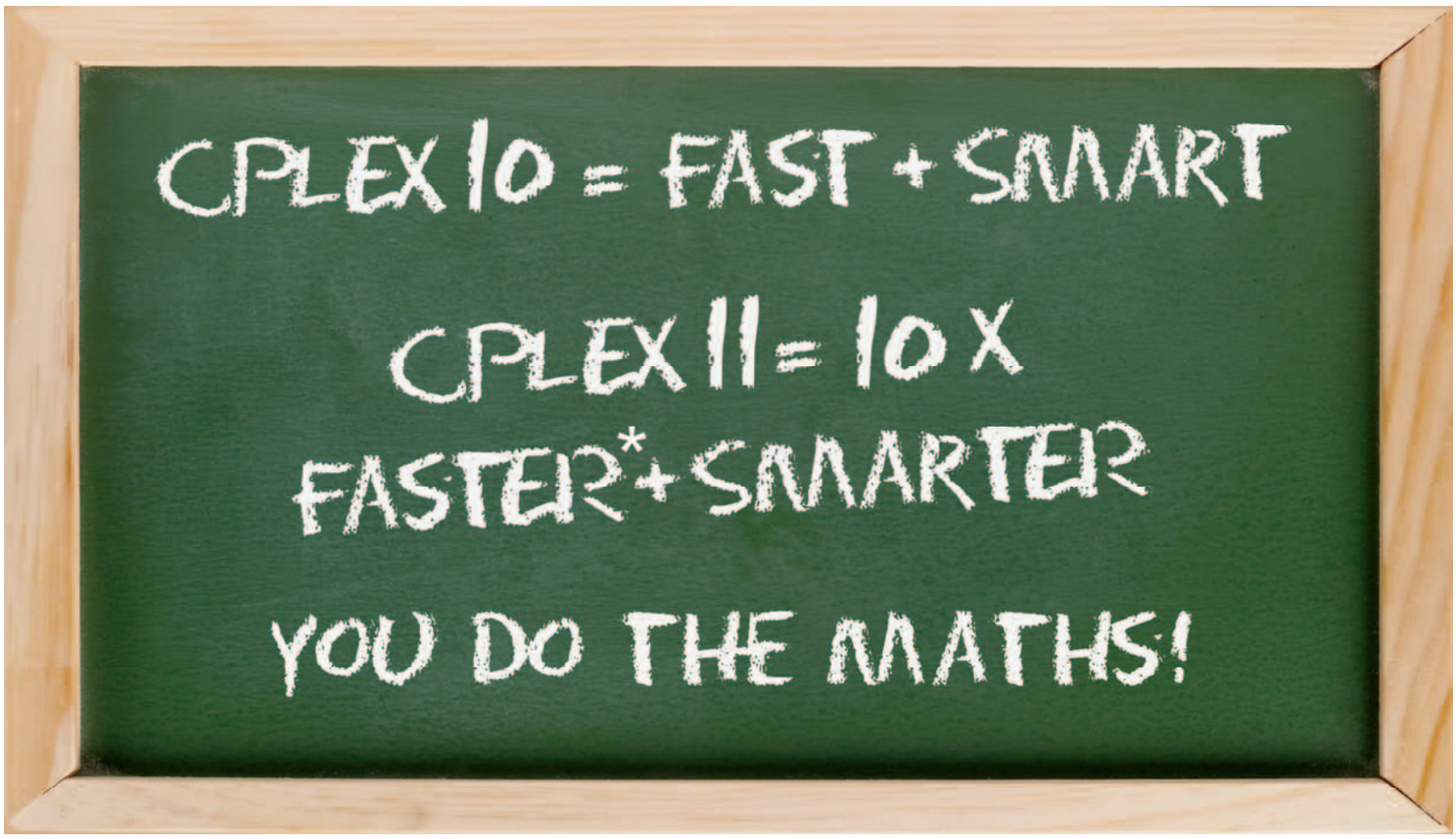
Antonio Beato Moreno
beato@us.es
Universidad de Sevilla

José Manuel Belenguer
jose.belenguer@uv.es
Universitat de Valencia

María Cruz Valsero Blanco
mcruz@eio.uva.es
Universidad de Valladolid

Leticia Lorenzo Picado
leticiap@uvigo.es
Universidade de Vigo

Fernando Plo
fplo@unizar.es
Universidad de Zaragoza



ILOG CPLEX 11

Performance like nothing before

Breakthrough MIP Performance:

*Take advantage of improved time to optimality, which is achieved on average 10 times faster on problems taking more than five minutes with CPLEX 10.

Enhanced Parallel MIP:

Leverage your multi-core machine and the new deterministic parallel MIP mode to get repeatable invariant solution paths.

Multiple MIP Solutions:

Generate and store multiple solutions to a MIP model, allowing you to consider subjective preferences on solutions.

Performance Tuning:

Get better performance from CPLEX and improve the performance of your optimization applications with the new performance tuning utility.

Learn more at: http://cplex.ilog.com

ILOG | CPLEX 11