SiDOR
Statistical Inference, Decision
& Operations Research Group
Universidade Vigo

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Cofinanciado por la Unión Europea
AGENCIA ESTATAL DE INVESTIGACIÓN

Workshop on Statistics

May 28 and 29, 2026

Facultade de Ciencias Económicas e Empresariais
– Aula-Seminario 8 –
Universidade de Vigo

Thursday, May 28th 2026

10:30–11:00. Conditional C-index for survival data with a cure fraction. Juan Carlos Pardo-Fernández (Universidade de Vigo)

11:00–11:30. Modelando la curva ROC con variables dependientes del tiempo. Arís Fanjul-Hevia (Universidade de Oviedo)

11:30–12:00. Coffee break

12:00–12:30. A nonparametric test for the significance of a quantitative covariate in conditional transition probabilities. Nora M. Villanueva (Universidade de Vigo)

12:30–13:00. Métodos de diagnóstico en modelos de fragilidad basados en residuos normalizados. María del Carmen Iglesias Pérez (Universidade de Vigo)

13:00–13:30. Estimación non paramétrica da moda dunha variable espacial. Tomás R. Cotos Yáñez (Universidade de Vigo)

13:30–15:30. Lunch break

15:30–16:00. Goodness-of-fit tests under dependent censoring. Adrián Lago (Universidade de Vigo)

16:00–16:30. Goodness-of-fit tests for copula selection under dependent double truncation with interval sampling. Carla Moreira (Universidade do Porto)

Friday, May 29th 2026

10:30–11:00 Estimación M-convexa eficiente en modelos lineales con respuesta censurada. Paula Soto Rodríguez (Universidade de Vigo)

11:00–11:30. A computationally efficient alternative for clustering survival curves. Marta Sestelo (Universidade de Vigo)

11:30–12:00. Coffee break

12:00–12:30. Multivariate regression models proposal of new estimation and variable selection. Sara Rodríguez Pastoriza (Universidad intercontinental de la Empresa)

12:30–13:00. Proposal of a general framework to categorize continuous predictor variables. Javier Roca Pardiñas (Universidade de Vigo)

Abstracts

Conditional C-index for survival data with a cure fraction

Juan Carlos Pardo-Fernández (Universidade de Vigo)

Abstract: In survival analysis, cure models are employed to study situations where some individuals never experience the event of interest. Such individuals are called cured. In this talk we will consider a mixture cure model, which combines the probability of being uncured (also called incidence) and the survival function of the uncured patients (also called latency). For more information about these models, see for example the review in Amico and Van Keilegom (2018).

In practice, risk scoring systems of latency and incidence are crucial elements for identifying relevant biomarkers and treatment strategies. Concordance measures that discriminate higher-risk subjects from lower-risk subjects are valuable tools to evaluate the overall performance of risk scoring systems. In contrast to conventional concordance measures, conditional concordance measures are proposed in this talk to provide comprehensive assessment of fitted cure models for particular values of a set of covariates. Specifically, we will consider the conditional version of the concordance index or C-index to evaluate the discrimination capacity of risk factors for both the latency and the incidence. Non- and semi-parametric modelling strategies are proposed to estimate and perform inferences about the conditional C-index. An application to real data is presented for illustrating the methodology.

This is joint work with Bo Han and Ingrid Van Keilegom.

References:

Amico, M. and Van Keilegom, I. (2018). *Cure models in survival analysis. Annual Review of Statistics and Its Application*, 5, 311–342.

Modelando la curva ROC con variables dependientes del tiempo

Arís Fanjul-Hevia (Universidade de Oviedo)

Abstract: En problemas de clasificación con dos poblaciones (como puede ser un método diagnóstico) el uso de la curva ROC está bastante extendido para medir su capacidad de discriminación. El tipo de variables que están involucradas en estos trabajos son las que se usan como marcadores (para hacer la clasificación), las que indican la población a la que se pertenece y otras covariables que pueden influir en su comportamiento.

Cuando alguna de estas variables (o todas) depende del tiempo (ya sea en formato de datos longitudinales, en datos funcionales o considerando el tiempo como una covariable más) surge la noción de curva ROC tiempo dependiente. En este trabajo se revisan las diversas alternativas que existen en la literatura para modelar y estimar este tipo de curvas.

A nonparametric test for the significance of a quantitative covariate in conditional transition probabilities

Nora M. Villanueva (Universidade de Vigo)

Abstract: In many clinical studies, it is often of interest to estimate transition probabilities conditional on a covariate (e.g., age, biomarkers), making it natural to investigate how these covariates affect the transitions. This work proposes a fully nonparametric testing procedure to assess the statistical significance of a quantitative covariate in the estimation of conditional transition probabilities for multi-state models. Specifically, the method evaluates whether the probability of moving from one state to another between two time points depends on the value of a quantitative covariate. The validity and behaviour of the proposed method was evaluated through simulation studies.

This is joint work with Marta Sestelo and Luís Meira-Machado.

Métodos de diagnóstico en modelos de fragilidad basados en residuos normalizados

María del Carmen Iglesias Pérez (Universidade de Vigo)

Abstract: Este trabajo aborda el diagnóstico del modelo para modelos de supervivencia con un término de fragilidad mediante el uso de residuos normalizados. Primeramente, proponemos utilizar probabilidades de supervivencia aleatorias normalizadas (residuos NRSP introducidos por Li, Wu y Feng en 2021) en regresión con y sin fragilidad y censura para estudiar la presencia de fragilidad, es decir, evaluar la homogeneidad del modelo. A continuación, estudiamos la bondad de ajuste de la distribución de riesgo y el efecto lineal de las covariables en un modelo de regresión con fragilidad. Para ello, generalizamos los residuos NRSP a modelos de fragilidad definiendo residuos condicionales (Z-residuos de Wu, Li y Feng, 2025) y marginales, y proponemos pruebas estadísticas y gráficos de bondad de ajuste basados en dichos residuos. Los métodos propuestos se evalúan mediante simulaciones, analizando los dos tipos de residuos por tamaño de muestra y de grupo. Finalmente, se aporta un ejemplo ilustrativo.

Este es un trabajo conjunto con Tomé Rodríguez Rodríguez.

Estimación non paramétrica da moda dunha variable espacial

Tomás R. Cotos Yáñez (Universidade de Vigo)

Abstract: Aínda que a literatura recente propón diversos procedementos non paramétricos indirectos para estimar a moda en datos xeoestatísticos –como a maximización da función de

densidade, como por exemplo, mediante o método do núcleo–, existen alternativas directas. Este traballo céntrase nun enfoque baseado na relación da moda cos cuantís da distribución. A partir desta idea, proponse un método non paramétrico directo para a estimación modal, demostrando a súa consistencia baixo certas condicións de regularidade. O estudo complétase cunha extensa análise de simulación, que ilustra o comportamento do procedemento proposto.

Este é un traballo conxunto con Pilar García Soidán.

Goodness-of-fit tests under dependent censoring

Adrián Lago (Universidade de Vigo)

Abstract: In survival analysis, the variable of interest is frequently only partially observed due to censoring, which biases the information provided by the sample. This has yielded extensive methodological developments for inference under these conditions. A common feature of many statistical methods is the assumption that the censoring mechanism is independent of the target variable. However, this assumption can be unrealistic in some practical applications. For example, in medical studies assessing new treatments, patients in worse health may be more likely to discontinue participation or experience events related to their condition, inducing a dependence between survival and censoring times. When this occurs, classical techniques designed for right-censored data may no longer be valid and may produce misleading results.

In this talk, we first introduce the issues caused by dependent censoring when estimating the distribution function, both parametrically and nonparametrically. Following estimation methodology proposed recently in the literature, we propose Kolmogorov–Smirnov and Cramér–von Mises tests for model assessment when data are subject to dependent censoring and the dependency structure is modelled by an Archimedean copula. The asymptotic behaviour of the new tests is addressed theoretically. Following the difficult application of the results derived, we propose a bootstrap resampling plan to approximate the null distribution of the proposed tests. Their finite-sample performance under the null and alternative hypotheses is investigated in a simulation study. A real data example is discussed as well.

This is joint work with Ingrid Van Keilegom and Juan Carlos Pardo-Fernández.

Goodness-of-fit tests for copula selection under dependent double truncation with interval sampling

Carla Moreira (Universidade do Porto)

Abstract: Dependence modelling under double truncation with interval sampling is a challenging statistical problem, since the truncation mechanism induces selection on the observable sample and complicates inference on the dependence structure. Recently, Moreira et al. (2021)

proposed a nonparametric framework for estimating marginal distributions and copula dependence structures under dependent double truncation with interval sampling. Building on this methodology, we propose goodness-of-fit procedures for copula selection in this setting. The proposed approach compares the empirical observable joint distribution with its fitted model-based counterpart through Kolmogorov?Smirnov and Cramér?von Mises type statistics, using bootstrap samples generated directly from the fitted observable truncated distribution. A simulation study is conducted to investigate the finite-sample performance of the proposed tests under different dependence structures and truncation levels, followed by an application to real doubly truncated data. The methodology is motivated by copula- based goodness-of-fit ideas introduced by Emura and Pan (2020) for dependently truncated data.

This is joint work with Jacobo de Uña-Álvarez.

References:

Moreira C, de Uña-Álvarez J, and Braekers R (2021). Nonparametric estimation of a distribution function from doubly truncated data under dependence. *Computational Statistics*.

Emura T, Pan C-H (2020). Parametric likelihood inference and goodness-of-fit for de- pendently left-truncated data, a copula-based approach. *Statistical Papers*, 61, 479?501.

Estimación M-convexa eficiente en modelos lineales con respuesta censurada

Paula Soto Rodríguez (Universidade de Vigo)

Abstract: En modelos lineales, el procedimiento clásico y predominante de estimación es el de mínimos cuadrados ordinarios. Sin embargo, este método puede perder eficiencia y robustez cuando la distribución de los errores se aleja de la normalidad. Una alternativa atractiva la constituyen los M-estimadores definidos mediante funciones de pérdida convexas, que permiten simplificar el análisis teórico y la implementación del estimador. Desarrollos recientes (véase Feng et al., 2026) construyen de forma adaptativa una pérdida convexa "óptima", en el sentido de minimizar la varianza asintótica dentro de la clase de M-estimadores convexos. Siguiendo esta línea, en este trabajo abordamos el caso en que la variable respuesta está censurada, escenario habitual en Análisis de Supervivencia. Proponemos una adaptación del esquema de estimación de Feng et al. (2026) a este marco y consideramos varios estimadores destinados a mitigar el sesgo inducido por la censura. Además, estudiamos sus propiedades asintóticas, estableciendo condiciones para la consistencia y la normalidad, y evaluamos su comportamiento mediante simulaciones en distintos escenarios de censura.

Este es un trabajo conjunto con Jacobo de Uña-Álvarez y Juan Carlos Pardo-Fernández.

Referencias:

Feng, O. Y., Kao, Y.-C., Xu, M. and Samworth, R. J. (2026). Optimal convex M-estimation via score matching. *Annals of Statistics*, to appear.

A computationally efficient alternative for clustering survival curves

Marta Sestelo (Universidade de Vigo)

Abstract: Survival analysis provides essential tools for studying time-to-event data, with the comparison of survival curves across groups being one of the main objectives. Traditional clustering approaches often rely on bootstrap-based procedures to approximate the null hypothesis distribution. While effective, they impose heavy computational demands and limit scalability in large datasets. The aim is to present a novel method that leverages k-means and the log-rank test to efficiently identify and cluster survival curves. By eliminating the need for intensive resampling, the approach substantially reduces computation time while preserving statistical validity. Through simulation studies, the proposed method is demonstrated to achieve performance comparable to bootstrap-based clustering techniques, while offering a significant gain in efficiency. These findings highlight that the proposed method offers a practical and scalable alternative for the analysis of multiple survival curves.

This is joint work with Nora M. Villanueva and Luis Machado.

Multivariate regression models proposal of new estimation and variable selection

Sara Rodríguez Pastoriza (Universidad intercontinental de la Empresa)

Abstract: The clinical interpretation of diagnostic tests often relies on two or more correlated biomarkers whose joint distribution may vary with demographic covariates such as age or sex. Multivariate reference regions, which characterize where 95% of healthy individuals' results are expected to lie, constitute a fundamental tool for personalized diagnostics. However, their conditional estimation remains methodologically challenging, particularly when covariate effects extend beyond a simple shift in location to a more complex reconfiguration of the joint distribution's geometry and dependence structure.

This thesis addresses these challenges through two complementary contributions. First, we propose a novel nonparametric bivariate regression model based on flexible additive quantile regression with penalized spline functions, which estimates a covariate-adjusted bivariate median and constructs directional quantiles to achieve isotropic coverage. Key strengths include robustness to outliers, flexibility in modeling nonlinear covariate effects, and applicability to both Gaussian and non-Gaussian data. Second, we address whether a given covariate justifies dynamically adjusting the reference region at each quantile level, proposing a dual testing strategy that combines a Wild Bootstrap procedure with a quantile-specific Bayesian Information Criterion. Applied to the joint distribution of fasting plasma glucose and glycated hemoglobin in the AEGIS cohort, both procedures confirm that age significantly reconfigures the bivariate dependence structure in a quantile-specific manner, enabling the identification of metabolic discordance even when individual univariate values appear within normal ranges.

Together, these two works provide a rigorous and practical framework for the estimation and formal validation of personalized bivariate reference regions, with direct implications for the clinical interpretation of correlated diagnostic markers across population subgroups.

This is joint work with Óscar Lado-Baleato, Javier Roca-Pardiñas and Francisco Gude.

Proposal of a general framework to categorize continuous predictor variables

Javier Roca Pardiñas (Universidade de Vigo)

Abstract: The use of discretized variables in the development of prediction models is a common practice, in part because the decision-making process is more natural when it is based on rules created from segmented models. Although this practice is perhaps more common in medicine, it is extensible to any area of knowledge where a predictive model helps in decision-making. Therefore, providing researchers with a useful and valid categorization method could be a relevant issue when developing prediction models.

In this talk, we propose a new general methodology that can be applied to categorize a predictor variable in any regression model where the response variable belongs to the exponential family distribution. Furthermore, it can be applied in any multivariate context, allowing to categorize more than one continuous covariate simultaneously. In addition, a computationally very efficient method is proposed to obtain the optimal number of categories, based on a pseudo-BIC proposal. Several simulation studies have been conducted in which the efficiency of the method with respect to both the location and the number of estimated cut-off points is shown.

Finally, the categorization proposal has been illustrated in a real data set of 543 patients with chronic obstructive pulmonary disease (COPD) from Galdakao Hospital's five outpatient respiratory clinics, who were followed up for 10 years. Exercise capacity is known to be an important predictor of adverse events in patients with COPD. In this paper, we applied the proposed methodology to jointly categorize the continuous variables six-minute walking test and forced expiratory volume in one second in a multiple Poisson generalized additive model for the response variable rate of the number of hospital admissions by years of follow-up. The location and number of cut-off points obtained were clinically validated as being in line with the categorizations used in the literature.

This is joint work with Irantzu Barrio, Cristóbal Esteban and María Durbán.