Repensando la estadística en un mundo digitalizado: la utilidad de los modelos estadísticos multivariantes de variables latentes en la industria y la salud 4.0

Alberto Ferrer
Box Medal Award 2025
Grupo de Ingeniería Estadística Multivariante,
Departamento de Estadística Aplicada, Investigación Operativa y Calidad,
Universitat Politècnica de València (ESPAÑA)

Kenko Imalytics S.L. (Co-founder and Chief Scientific Officer)
Kensight Solutions S.L. (Co-founder and Scientific Advisor)

Palabras clave: Digitalization; Data Science; Latent variables models; Principal Component Analysis; Partial Least Squares Regression; Machine Learning

La era del 4.0 representa la Cuarta Revolución Industrial, una nueva era de innovación marcada por la integración de tecnologías digitales como la Inteligencia Artificial (IA), el Internet de las cosas (IoT) y el análisis de grandes bases de datos (Big Data), para crear sistemas interconectados, automatizados y centrados en datos, que pueden mejorar la eficiencia, personalización y toma de decisiones en fábricas (Industria) y servicios sanitarios (Salud).
En esta nueva era de la digitalización, sin embargo, existe la creencia predominante de que, debido a la enorme cantidad y velocidad con la que se generan los datos, estas tecnologías digitales emergentes pueden resolver los problemas importantes de la sociedad, industria, o de la sanidad, únicamente mediante el análisis de datos empíricos, sin necesidad de modelos científicos, teoría, experiencia o conocimiento del sector. La idea es que la causalidad ya no importa, solo la simple correlación basta. Algunos incluso llegan a afirmar que la Estadística ha muerto, debido al surgimiento de la Ciencia de Datos. Consideran la Estadística obsoleta y a los estadísticos como profesionales excesivamente especializados en técnicas que ya no son útiles, y preocupados por cuestiones que parecen irrelevantes ante la complejidad de los desafíos del siglo XXI.
En este minicurso, a través de un análisis DAFO (Debilidades-Amenazas-Fortalezas y Oportunidades) de la Estadística en el mundo digitalizado, trataré de desvelar los porqués de estas creencias sobre la obsolescencia de la Estadística, y compartir mi perspectiva sobre los factores clave que podrían ayudar a la Estadística a recuperar su reconocimiento social y su papel fundamental como componente clave para abordar con éxito muchos de los problemas que enfrenta nuestra sociedad.
Muchas las ideas planteadas provienen de las enseñanzas del profesor George EP Box, eminente estadístico inglés fallecido hace una década, y del profesor John MacGregor, destacado ingeniero químico y estadístico canadiense, discípulo del profesor Box, quienes han sido unas de las personas más influyentes en mi carrera profesional.
Destacaré el potencial de los Modelos Estadísticos Multivariantes de Variables Latentes, y en particular del Análisis de Componentes Principales (PCA) y de la Regresión en Mínimos Cuadrados Parciales (PLS), para abordar algunos de los principales retos de la industria y la salud en entornos altamente digitalizados (Industria y Salud 4.0). Entre ellos, destacaré el desarrollo de biomarcadores de imagen para el diagnóstico precoz del cáncer utilizando PCA no ortogonal, el diseño de esquemas de Control Estadístico Multivariante de Procesos (MSPC) y el análisis multivariante de imágenes basados en PCA, así como la optimización de procesos, la definición del espacio de diseño de materias primas y el desarrollo de índices de capacidad multivariante basado en variables latentes utilizando datos históricos mediante PLS. Estos enfoques aprovechan la capacidad del PLS para modelar la causalidad en el espacio latente, incluso a partir de datos históricos, típicos en entornos altamente digitalizados. También

plantearé ventajas/inconvenientes de estas técnicas con otras procedentes del *Machine Learning* como *Random F*orests, Redes Neuronales, etc, y con las técnicas de Regresión Lineal Múltiple y Regresión Logística.
Todos los temas abordados en el minicurso se ilustrarán a través de ejemplos reales de la industria (química, farmacéutica y de alimentación) y de la salud (diagnóstico de cáncer por imagen) en los que hemos estado involucrados en los últimos años, fruto de nuestra colaboración en proyectos de investigación y en labores de consultoría con varias empresas y hospitales.

Referencias:
Borràs-Ferrís J, Palací-López D, Duchesne C, Ferrer A. Defining multivariate raw material specifications in industry 4.0. *Chemom Intel Lab Syst*. 2022;225, 104563.

Borràs-Ferrís J, Duchesne C, Ferrer A. Defining multivariate raw material specifications via SMB-PLS. *Chemom Intel Lab Syst.* 2023;240, 104912.

Borràs-Ferrís J, Duchesne C, Ferrer A. A latent space-based multivariate capability index: A new paradigm for raw material supplier selection in industry 4.0. *Chemom Intel Lab Syst*. 2025;258, 105339.

Box GEP. Science and Statistics. *J. Am. Stat. Assoc.* 1976;71(356):791-799.

Box GEP, Hunter JS, Hunter WG. *Statistics for Experimenters: Design, Discovery, and Innovation*. 2nd edn. John Wiley & Sons, Inc. 2005.

Ferrer, A. (2007). Multivariate statistical process control based on principal component analysis (MSPC-PCA): Some reflections and a case study in an autobody assembly process. *Qual Eng.* 2007*;*19:311–325.

Ferrer A. Latent structures-based multivariate statistical process control: a paradigm shift. *Qual Eng.* 2014;26(1):72-91.

Ferrer A. Discussion of "A review of data science in business and industry and a future view" by Grazia Vicario and Shirley Coleman. *Appl Stochastic Models Bus Ind*. 2020;36:23–29.

Ferrer A. Multivariate six sigma: A key improvement strategy in industry 4.0. *Qual Eng.* 2021*;*33:758–763.

Ferrer A, Borràs-Ferrís J, García-Carrión S. *Data Analytics Strategies to Exploit Historical Databases for Process Optimization and Innovation in Digitalized Industry 4.0.* In: di Bella, E., Gioia, V., Lagazio, C., Zaccarin, S. (eds) Statistics for Innovation I. SIS 2025. Italian Statistical Society Series on Advances in Statistics. Springer, Cham. 2025:33-37.

Jaeckle CM, MacGregor JF. Industrial applications of product design through the inversion of latent variable models. *Chemom Intel Lab Syst.* 2000;50:199-210.

MacGregor, JF. Empirical models for analyzing "big" data - what's the difference. In: Spring AIChE Conf., Orlando, Florida, USA 2018.

Tomba E, Barolo M, García-Muñoz S. General framework for latent variable model inversion for the design and manufacturing of new products*. Ind Eng Chem Res.* 2012;51:12886-1290*0.*