Rect@ Vol 10 Diciembre 2009. Pp 223 - 252

Recibido

10/10/2009

Revisado

2/12/2009

Aceptado

18/12/2009

Especificación de modelos econométricos utilizando minería de datos

Fernández-Rodríguez, Fernando

Acosta-González, Eduardo

Andrada-Félix, Julián

Departamento de Métodos Cuantitativos en Economía y Gestión

Universidad de Las Palmas de Gran Canaria

RESUMEN

En este trabajo presentamos un nuevo procedimiento para seleccionar modelos econométricos. Está basado en un enfoque heurístico, empleando algoritmos genéticos, que permite explorar el universo de modelos disponibles a partir de un modelo general sin restricciones. Este proceso de búsqueda del modelo óptimo está guiado únicamente por el criterio de información de Schwarz, que actúa como la función pérdida de un algoritmo genético empleado para seleccionar el modelo óptimo. Este procedimiento muestra buen comportamiento en relación a otras metodologías alternativas. A modo de ejemplos de su utilidad se muestran tres problemas donde el algoritmo ha sido empleado con éxito: la selección de variables que explican el crecimiento económico, la predicción del fracaso empresarial y la formación de una cartera de pocos activos que siga el comportamiento de un índice bursátil como el IBEX35 Español.

***Palabras claves*:** selección de modelos; algoritmos genéticos; crecimiento económico; fracaso empresarial; seguimiento de un índice;

ABSTRACT

. In this paper we present a new procedure for selecting econometric models. It is based on a heuristic approach, called genetic algorithms, which permits us to explore the universe of available models starting from a general model without restrictions. This search process for the optimal model is guided only by the Schwarz Information Criterion, which acts as the lost function of the genetic algorithm employed for selecting the optimum. This procedure shows good performance with respect to other methodologies. As examples of its utility three problems where the algorithm was successfully employed are presented: the selection of variables that explain economic growth, the prediction of failure of firms and the construction of a portfolio with few actives that follows the behaviour of a stock exchange index such as the Spanish IBEX35.

Keywords: model selection; genetic algorithms; economic growth; failure of firms; index tracking.

1. INTRODUCCIÓN: EL PROBLEMA DE LA SELECCIÓN DE REGRESORES

Un aspecto crucial en la construcción de un modelo de regresión múltiple es el de la selección de los regresores, o variables explicativas, que deben ser incluidos. Se trata de un antiguo e importante problema en estadística, y especialmente en econometría. Es bien conocido que la sobreparametrización de los modelos estadísticos genera el problema de una buena predicción intra-muestral, pero una mala predicción extra-muestral. Así, un modelo con muchos parámetros "memoriza" la muestra pero tiene una mala capacidad de generalización.

En la modelización econométrica, en muchas ocasiones la teoría económica puede servir de ayuda a la hora de tomar este tipo de decisiones, pero no siempre es así. Los modelos económicos suelen ser menos precisos que los econométricos, de esta manera se corre el riesgo de especificar modelos con variables explicativas irrelevantes, o por el contrario con la omisión de variables explicativas relevantes. Estas circunstancias tendrán determinadas repercusiones en el modelo.

Dada una variable dependiente Y y un conjunto de posibles de regresores potenciales $X_1,....,X_K$, el problema que se plantea es encontrar el mejor modelo de la forma:

$$Y = \beta_0 + \beta_1 X_{i_1} + ... + \beta_K X_{i_K} + \varepsilon, \text{ donde } \{i_1, i_2, ..., i_K\} \subseteq \{1, 2, ...K\} \quad (1)$$

Como hay 2^K posibles modelos, desde el punto de vista computacional se trata de un problema que presenta intratabilidad, debido al carácter exponencial del número de alternativas[1].

Desde el punto de vista de la práctica econométrica el problema ofrece serias dificultades y, como veremos, se ha propuesto una amplia variedad de procedimientos de selección de las posibles modelos que incluyen el criterio de información de Akaike Akaike (1973), el criterio de Mallows (1973), el criterio de información de Schwarz (1978) o el criterio de información de Hannan-Quinn (1979).

[1] La teoría de la complejidad computacional califica este problema como NP-Hard, es decir, al menos tan difícil como un problema NP (non deterministic polynomial time) que se pueda resolver en tiempo polinómico respecto al tamaño del problema, aunque puede resultar todavía más difícil.

En cualquier caso, la selección de regresores esta sujeta a errores y la utilización de criterios para llevar a cabo este propósito no nos inmuniza ante ello. Por esta razón, es importante tener presente que ante la duda "estadística" de la inclusión de un determinado regresor, se ha de ser consciente de las repercusiones de su omisión frente a su inclusión.

1.1.- CONSECUENCIAS DE LA INCLUSIÓN DE VARIABLES IRRELEVANTES EN EL MODELO

La inclusión de variables irrelevantes en un modelo significa que estamos introduciendo variables cuyos coeficientes-parámetros son cero en el modelo verdadero. Si el verdadero modelo queda representado por la siguiente relación

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{2}$$

y sin embargo se especifica el modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{3}$$

necesariamente $\beta_3 = 0$, con lo que la variable X_{3i} debería ser excluida.

Es bien conocido que a efectos de insesgadez, la introducción de variables irrelevantes no tiene ninguna repercusión, con lo que los estimadores MCO de los coeficientes, como el de la varianza de las perturbaciones, serán insesgados y consistentes [ver Johnston, 1984]. Sin embargo, el modelo perdería grados de libertad innecesariamente, cuestión a tener en cuenta sobre todo si se trabaja con muestras pequeñas. Igualmente existe un coste de eficiencia pues la existencia de correlación entre las variables explicativas (multicolinealidad) supone un incremento en la varianza de los estimadores, salvo en el caso especial en el que la correlación entre las variables relevantes y las irrelevantes fuera cero, la inclusión de éstas últimas supondrá la obtención de estimadores menos eficientes en comparación con el modelo verdadero. De esta manera se incrementa la probabilidad de rechazar la hipótesis nula de igualdad a cero de los coeficientes de las verdaderas variables.

1.2.- CONSECUENCIAS DE LA OMISIÓN DE VARIABLES RELEVANTES EN EL MODELO

A diferencia de lo que ocurre con la inclusión de variables irrelevantes, la omisión de variables relevantes en el modelo implica estimaciones sesgadas e inconsistentes de los coeficientes y de la varianza de las perturbaciones. El sesgo de los coeficientes depende de dos aspectos: (1) de la importancia que las variables explicativas relevantes omitidas tengan en la explicación de la variable endógena, y (2) de la relación entre las variables incluidas en el modelo y las relevantes excluidas del mismo. Cuanto mayor sea esta última relación mayor será el sesgo, en el caso especial de que no exista relación alguna el sesgo desaparece.

En síntesis, si se selecciona una menor cantidad de regresores en el modelo, la estimación de los parámetros no será consistente y si se selecciona una cantidad elevada, su varianza se incrementará. Por tanto, en caso de duda, hay que tener presente que la pérdida de potencia que supone la inclusión de variables irrelevantes en el modelo es menos perjudicial que el sesgo que produce la exclusión de variables relevantes. Así, por ejemplo, si para el caso de que un determinado coeficiente no se rechaza la hipótesis nula de igualdad a cero para un nivel del 5% pero sí para un nivel del 10%, y se tiene la duda de su inclusión debido a que su relación de causalidad con la endógena parece evidente o lógica, siempre será mejor incluirla, a pesar de la pérdida de potencia que ello suponga, que correr el riesgo de no incluirla y sesgar las estimaciones.
El resto de este trabajo ha sido organizado como sigue. En la sección 2 se presenta el estado del arte de la selección regresores y los procedimientos vigentes. En la sección 3 se describen los problemas econométricos suscitados por la minería de datos. En la sección 4 se presenta una breve reseña de los algoritmos genéticos, destacando sus aplicaciones para el problema de la selección de regresores por medio del algoritmo GASIC. La sección 5 muestra que el algoritmo GASIC puede mejorar los anteriores procedimientos para seleccionar modelos. La sección 6 ofrece ejemplos específicos del uso del algoritmo GASIC en los cuales ha sido empleado con éxito. La sección 7 presenta las conclusiones relevantes extraídas.

2.- CRITERIOS DE SELECCIÓN DE REGRESORES

Por lo general, la selección de regresores se obtiene a partir de probar múltiples alternativas y contrastarlas. Si el tema del trabajo elegido tiene una teoría económica precisa, ésta nos servirá de ayuda para determinar la especificación del modelo. En caso contrario, tal especificación queda muy abierta, sobre todo si el número de regresores candidatos a formar parte del modelo es elevado. El problema crucial con que nos encontramos en este caso es que, dependiendo de la "senda" que se siga a la hora de probar los diferentes regresores, se tendrán diferentes modelos finales. Por esta razón en la literatura se han desarrollado numerosos criterios de selección de modelos. Cada uno de dicho criterios debería aplicarse a cada uno de los posibles modelos, hecho que es sólo viable, en la práctica, cuando el número total de posibles candidatos a regresores es pequeño. Los criterios más utilizados son los siguientes:

Criterio del coeficiente de determinación corregido ($\bar{R}^2$)

Este criterio consistiría en maximizar el coeficiente de determinación corregido $\bar{R}^2$ propuesto por Theil (1971),

$$\bar{R}^2 = 1 - \frac{e'e/N-k}{Y'Y - N\bar{Y}^2/N-1} \tag{4}$$

Theil propone este coeficiente frente al coeficiente de determinación R^2 que no tiene en cuenta los grados de libertad de la estimación. Sin embargo, este criterio no garantiza que la especificación con mayor $\bar{R}^2$ sea el modelo "verdadero" en tanto en cuanto es posible que el modelo seleccinado contenga variables irrelevantes. Para más detalles sobre este aspecto ver Ebbeler (1975). Una forma de proceder en este caso podría consistir en eliminar secuencialmente (de menor a mayor significatividad) las variables no significativas del modelo que maximiza $\bar{R}^2$. Esta última forma de proceder se asemejaría al criterio de selección *stepwise* que veremos en detalle más adelante.

Criterio de Mallows (1973)

Consiste en obtener un modelo que minimice la estimación del error cuadrático medio de predicción. Concretamente este criterio queda definido por la siguiente expresión:

$$C_p = \frac{2k}{N}\sigma^2_{NR} + \frac{e'e}{N} \tag{5}$$

donde

$$\sigma_{NR}^2 = \frac{e'e_{NR}}{N - k_{NR}} \tag{6}$$

siendo $e'e_{NR}$ y k_{NR} la suma del cuadrado de los errores y el número de parámetros del modelo no restringido respectivamente. El modelo no restringido hace referencia a aquel donde se incluyen todas las variables explicativas candidatas a formar parte del mismo, también denominado "especificación general". Por otro lado, el modelo restringido hace referencia a aquel donde sólo se incluye el subconjunto de variables explicativas seleccionadas.

Criterio de Amemiya (1980)

Al igual que el criterio de Mallows, consiste en minimizar la estimación del error cuadrático de predicción. Concretamente en este caso

$$PC = \frac{2k}{N}\sigma_R^2 + \frac{e'e}{N} \tag{7}$$

La diferencia con el criterio de Mallows está en σ_R^2, que en este caso se obtiene a partir del modelo restringido.

$$\sigma_R^2 = \frac{e'e}{N - k} \tag{8}$$

Criterio de información de Schwarz (1978)

Este criterio, que denominaremos en lo sucesivo SIC, tomando sus iniciales de la lengua inglesa (Schwarz Information Criterion), consiste en buscar aquel modelo que minimiza el estadístico

$$SIC = \log\left(\frac{e'e}{N}\right) + \frac{k}{N}\log(N) \tag{9}$$

Es bien conocido que SIC es asintóticamente consistente como criterio de selección. Ello significa que dada una familia de modelos, dentro de la cual está incluido el modelo verdadero, la probabilidad de que SIC seleccione el modelo verdadero tiende a uno cuando el tamaño muestral tiende hacia infinito. De esta forma SIC evita la sobreparametrización y resuelve el problema de trade-off entre ajuste de los datos intramuestrales y la capacidad de generalización extramuestral.

Criterio de información de Akaike (1973)

Se utiliza para modelos que se pueden estimar por máxima verosimilitud. Su expresión es la siguiente:

$$AIC = \frac{-2l}{N} + \frac{2k}{N} \tag{10}$$

donde l es el logaritmo de la función de verosimilitud. Para el caso de modelos de regresión esta expresión se convierte en

$$AIC = \log\left(\frac{e'e}{N}\right) + \frac{2k}{N} \tag{11}$$

Regresión *Stepwise*

Los anteriores criterios de selección de regresores resultan poco prácticos cuando el número K de variables explicativas potenciales es elevado, pues existe entonces un número de 2^N posibles especificaciones del modelo[2].

Con el fin de resolver este problema intratable, existen varios procedimientos heurísticos que disminuyen, de alguna manera, las alternativas a probar en cada criterio. Así, será posible dirigir la atención a un pequeño número de regresores potenciales. Tales procedimientos heurísticos, en lugar de explorar entre todos los posibles modelos, buscan un buen camino a través de ellos. Uno de los procedimiento más simples y populares es la regresión *stepwise*, tanto de selección hacia delante como de eliminación hacia atrás de variables explicativas. En este procedimiento se incluyen o excluyen, secuencialmente, variables en el modelo basándose en consideraciones del estadístico *t*. No obstante, este procedimiento para excluir regresores no significativos da lugar a resultados que son altamente dependientes del orden de exclusión de la variables regresoras, sin que exista una orientación de por qué variable comenzar.

De modo más formal, el procedimiento *stepwise* utiliza el estadístico *t* del contraste de significación individual $\left(H_0 : \beta_j = 0\right)$, para determinar la entrada o salida del modelo de una determinada variable explicativa. Por tanto, antes de iniciar el proceso habrá que fijar el nivel de significación que determine la región de aceptación o rechazo de la hipótesis nula. No rechazar la hipótesis nula implicará la eliminación correspondiente de

[2] Se puede pensar que raramente se trabaja con modelos donde el número de variables explicativas sobrepase la decena, sin embargo, si se considera la posibilidad de estar frente a especificaciones funcionales distintas de la lineal en las variables, el número de variables explicativas se puede incrementar considerablemente, al incluir como tales el cuadrado de las mismas, efectos interacción, etc. Además, en el contexto de modelos dinámicos, la presencia de variables explicativas retardadas, o incluso la propia endógena retardada, pueden incrementar considerablemente el número de regresores candidatos iniciales del modelo.

la variable explicativa del modelo, mientras que su rechazo, la continuidad de dicha variable en el modelo. Existen dos variantes del algoritmo:

Regresión *Stepwise forward* : en este caso el procedimiento en ningún caso se replantea que una variable pueda salir del modelo una vez ha entrado a formar parte del mismo.

Regresión *Stepwise backward*: este procedimiento se inicia con la estimación de un modelo donde se incluyen todas las variables explicativas con las que se cuenten. A partir de este modelo, se irán eliminando una a una las variables explicativas no significativas. En primer lugar se elimina la variable explicativa con el ratio *t* más pequeño, en valor absoluto, de entre los que caigan en la región de aceptación. Una vez eliminada esta variable, se vuelve a estimar el modelo y se procede de la misma manera que en la estimación anterior. Este proceso continua hasta que todos los ratios *t* de las variables explicativas que se mantienen en el modelo caigan en la región de rechazo.

Aún cuando el modelo seleccionado mediante los diferentes procedimientos *stepwise* tiene muchas posibilidades de ser el modelo correcto, o al menos un modelo que contenga al verdadero, nunca se tendrá la seguridad total de que así sea. Hay decisiones previas al inicio del proceso, como la determinación de los niveles de confianza del contraste *t* de significación individual, que afectan claramente a la especificación final del modelo. Además, una vez finalizado el proceso, en la mayoría de los casos, el número de modelos chequeados es inferior al número de modelos plausibles. Para reducir el efecto de este inconveniente, Hoover y Perez (1999) proponen que del paso 1 de la regresión *stepwise* salgan más de un modelo candidato, a partir del cual seguir el procedimiento descrito. Cada uno de estos modelos iniciales configuraría lo que estos autores denominan “sendas” que terminarán en propuestas factibles de ser estudiadas con posterioridad mediante contrastes de abarcamiento.

Finalmente, hay que tener en cuenta que la selección de regresores está íntimamente ligada a la forma funcional de la especificación. Ésta cambiará en la medida que se decida introducir como regresores el cuadrado de alguna de las variables explicativas, los efectos de su interacción, o la transformación logarítmica de alguna de ellas, por señalar tan sólo algunos ejemplos.

3. SELECCIÓN DE REGRESORES Y MINERÍA DE DATOS

El hecho de que la selección del modelo final dependa de la "senda" que hagamos con los regresores candidatos cuestiona los procedimientos de minería de datos consistentes en la búsqueda de una especificación consistente del modelo sobre la base de ir introduciendo o eliminando regresores al emplear como criterios la significatividad dada por el estadístico *t* o valores altos del estadísitico R^2. Esta idea también cuestiona la posibilidad de extracción de información predictiva a partir de grandes bases de datos mediante el uso de diversas estrategias de investigación frecuentemente automatizadas como es el caso típico del *stepwise*.

Utilizar como criterios de selección de modelos el estadístico de significación individual *t* o el coeficiente de determinación R^2 puede conducir, fácilmente, a la selección de modelos incorrectos. Un trabajo seminal que puso de manifiesto estas serias limitaciones de los procedimientos de minería de datos se debe a Lovell (1983). En este trabajo se evalúa la capacidad de algunos métodos de selección de modelos. Para llevar a cabo este objetivo se genera un "modelo verdadero" a partir de algunas de las variables de la base de datos[3] que se utiliza. Cada uno de los métodos de selección alternativos se evalúa por su capacidad de identificar el "modelo verdadero". En tal caso, el número posible de modelos candidatos es $2^{40} = 1.099.511.627.776$.

Lovell (1983) muestra los problemas que pueden introducir en la modelización econométrica las técnicas de minería de datos advirtiendo que el proceso de modelización tiene que ir acompañado de un aumento proporcional en nuestro conocimiento de cómo funciona realmente la economía. Sus conclusiones fueron demoledoras. Las aparentemente inocuas actividades de la minería de datos, que realizan muchos investigadores, consistentes en ir introduciendo y sacando variables del modelo con los criterios *t* o R^2, erosionan los niveles de significación en los contrastes de hipótesis cuando el investigador opta por considerar el mejor resultado aisladamente.

Más recientemente, la *London School of Economics* (LSE) presentó una nueva aproximación al problema desarrollando una variedad de metodologías econométricas conocidas como modelización "de lo general a lo específico" que permitían reconsiderar la reconstrucción de modelos, dando lugar a una literatura favorable a la utilización de

[3] Esta base de datos contienen 40 regresores macroeconómicos incluyendo algunos retardos.

la minería de datos. La especificación de modelos en el marco de la LSE consiste en buscar modelos que son restricciones de un modelo general completo que incluye todas las variables candidatas a formar parte del mismo. Un trabajo seminal en este sentido fue el de Hoover y Perez (1999) (HP).

HP desarrollaron un algoritmo mecánico que imita algunos aspectos de los procedimientos de búsqueda utilizados por la LSE, simulando una selección de lo general a lo específico para modelos dinámicos de regresión lineal. La característica esencial del algoritmo de HP y sus derivados es que la estrategia de selección del modelo descansa sobre la elección de una batería de diagnósticos sobre los residuos y contrastes de hipótesis sobre los coeficientes, seleccionando el mejor modelo entre aquellos modelos que no son rechazados por los contrastes. Como ya se ha mencionado, el camino de búsqueda es esencial en el algoritmo para evitar tener que examinar individualmente la totalidad de los 2^K distintos modelos. En el trabajo de HP, las variables de la especificación general son ordenadas de forma ascendente de acuerdo con su estadístico *t*. Para cada replicación, se examinan 10 caminos de búsqueda o sendas. Cada senda comienza con la eliminación de una de las variables en el subconjunto con los 10 estadísticos *t* menos significativos. La primera búsqueda comienza eliminando la variable con el estadístico *t* más bajo en valor absoluto y re-examinando la regresión. Esta re-estimación de la regresión se considera y se convierte, entonces, en la especificación corriente o actual. La búsqueda continúa hasta que alcanza una especificación terminal. Las múltiples sendas que se consideran pueden conducir a múltiples modelos. HP seleccionan entonces el modelo final a partir de la utilización de contrastes de abarcamiento.
Posteriormente, Hendry y Krolzig (1999, 2001, 2003) propusieron mejoras potenciales al procedimiento de búsqueda de HP. Este algoritmo está disponible comercialmente en forma de software informático denominado PcGets.

Una parte esencial de los trabajos citados sobre la selección automática de modelos es la senda empleada para poder encontrar el modelo final dentro del universo de modelos. Cuando el número de regresores potenciales es elevado, dicho universo de modelos se vuelve prohibitivamente grande. En los trabajos previamente citados, el procedimiento de selección de variables se basa principalmente en el estadístico *t* y en el procedimiento

stepwise con eliminación *backward*, de modo que en cada paso se emplean numerosos contrastes de diagnóstico sobre los residuos para verificar los modelos.

Existe otra línea de investigación que no centra la estrategia de selección de modelos en diagnósticos sobre los residuos y contrastes de hipótesis sobre los coeficientes. Este es el caso del trabajo de Hansen (1999) quien sugiere simplificar el algoritmo de HP, proporcionando un camino de búsqueda simple aunque no siempre adecuado. Basado en la evidencia numérica, Hansen sostiene que el simple y elegante criterio del SIC funciona al menos tan bien, si no mejor, que los complicados algoritmos desarrollados sobre la metodología de la LSE.

Aunque el SIC es uno de los criterios de selección más prometedores en la selección de modelos, el talón de Aquiles de la metodología de Hansen está en la reducción del máximo número de regresores a $K = 10$. Tal reducción no tiene justificación ni desde el punto de vista computacional, ni desde el punto de vista metodológico. Aún cuando esta reducción se base en criterios de significatividad de los coeficientes, es arbitraria y no puede justificarse de antemano.

Otro procedimiento alternativo de selección de modelos ha sido propuesto por Pérez-Amaral et al. (2003), quienes usan una medida de *performance* extramuestral para la estrategia de selección. Se trata de una herramienta flexible para la construcción de modelos basada en la *Relevant Transformation of the Inputs Network Approach*, llamada RETINA.

Finalmente, Acosta-González y Fernández-Rodríguez (2007) han proporcionado un nuevo procedimiento de selección de modelos basado en un algoritmo genético (AG) que emplea como función de pérdida el SIC. Dicho algoritmo denominado GASIC (*Genetic Algorithm Schwarz Information Criterion*) muestra una alta capacidad en la selección de modelos cuando se compara con las otras metodologías existentes y previamente citadas.

Otros autores también han propuesto procedimientos heurísticos basados en algoritmos evolutivos para abordar el problema de selección de variables en el análisis discriminante. En este sentido cabría señalar a Pacheco et al. (2007) junto con otros trabajos allí citados.

También existen otros trabajos que proponen el uso de los AG para estimar modelos no lineales de regresión y buscar formas funcionales en Economía. Tales son los trabajos

Schmertmann (1996), Szpiro (1997) y Beenstock y Szpiro (2002). Tales trabajos están basados en una variación de los algoritmos genéticos introducida por Koza (1992) y que se denomina Programación Genética (PG). La PG hace uso de complejos cromosomas que son, a su vez, programas de ordenador que evolucionan según unas reglas preestablecidas. El algoritmo de PG conduce finalmente a encontrar el programa óptimo capaz de resolver un determinado problema propuesto de antemano, susceptible de ser escrito en un lenguaje algorítmico. La PG es un buen procedimiento para estimar e identificar un proceso generador de datos con formas funcionales complicadas. Sin embargo, este procedimiento adolece de graves problemas derivados de la complejidad que se produce en las formas funcionales a las que conduce. El primer problema son los altos tiempos de computación, que frecuentemente obligan al algoritmo a parar sin seguir ningún criterio de convergencia. El segundo se refiere a la pobre capacidad de predicción extramuestral y de generalización a que puede conducir la sobreparametrización y unas complejas formas funcionales que se adaptan a los datos intramuestrales de forma casi perfecta. Finalmente cabría destacar que, aunque la aproximación de la PG es aparentemente más general que el algoritmo GASIC, porque permite encontrar modelos de regresión no lineales, uno de los mayores inconvenientes de la metodología basada en la PG es que no se produce convergencia hacia un único modelo "verdadero", cuando se vuelve a repetir el algoritmo con diferente semilla. Con el fin de evitar esta dependencia del modelo final de la semilla inicial del algoritmo, Beenstock y Szpiro (2002) usan un contraste de abarcamiento para seleccionar un único modelo final. Sin embargo, el procedimiento GASIC es más simple y directo porque está basado en la versión binaria de los AGs, propuesta originalmente por Holland (1975), desarrollada para ser usada en la selección de modelos de regresión.

GASIC tiene las siguientes ventajas:

- Se ha mostrado como una herramienta parsimoniosa y robusta capaz, tanto de obtener predicciones extramuestrales, como de llevar a cabo un simple análisis estructural.
- Además de la transformación inicial de regresores y sus combinaciones, GASIC permite la selección de formas funcionales no lineales. Este hecho se pone de manifiesto cuando GASIC trabaja sobre los datos de Pérez-Amaral et al. (2003),

donde todos los regresores usados son obtenidos como combinación de dos variables.

- Por tanto, GASIC proporciona un procedimiento directo de selección de regresores usando un AG en el contexto de la econometría clásica.

4. LA SELECCIÓN DE REGRESORES CON ALGORITMOS GENETICOS

La metodología GASIC propuesta por Acosta-González y Fernández-Rodríguez (2007) sustentada en el empleo de un algoritmo genético para la selección del modelo óptimo, se basa en transformar cada posible modelo candidato a óptimo en un cromosoma. Aplicando los procedimientos de selección natural y cruzamiento, el algoritmo genético hará que un conjunto inicial de soluciones tomadas al azar vaya evolucionando y mejorando su comportamiento hasta encontrar el óptimo.

Para un problema que parte de un conjunto inicial de K regresores potenciales, un cromosoma es un vector de K variables binarias que toman los valores 0 ó 1, es decir,

$cromosoma = [g_1, g_2, ..., g_K]$, $g_i = 0 \, ó \, 1$

Por ejemplo, para $K = 5$ y un modelo general de X_1, X_2, X_3, X_4, X_5 regresores, el cromosoma (1,0,1,0,1) se corresponde con el modelo cuyos regresores son X_1, X_3, X_5.

El procedimiento GASIC para la selección de un modelo a partir de datos reales puede ser descrito como sigue:

1. Examinar la congruencia del modelo general no restringido formado por la totalidad de regresores potenciales $X_1, X_2, ..., X_K$ usando una batería de contrastes de mala especificación como aconsejan Hendry y Krolzig (2001). El modelo general no restringido podría ser revisado empíricamente si tales contrastes fueran rechazados.
2. Buscar el modelo óptimo conteniendo un subconjunto de variables seleccionadas a partir del modelo general usando el algoritmo GASIC.

Los pasos básicos del algoritmo GASIC son los siguientes:

Paso 1. Población Inicial. Se genera una población inicial de modelos. Por lo tanto, el algoritmo comienza con la selección aleatoria de una población inicial de cromosomas binarios que representan modelos anidados en un modelo general sin restricciones. Estos cromosomas actúan como semillas del proceso. En este sentido, es importante tener en cuenta que, tal como se ha comprobado mediante simulación (Acosta-González y Fernández-Rodríguez, 2007), el algoritmo es robusto ante el cambio de dichas semillas y conduce, en la inmensa mayoría de las ocasiones, al mismo modelo final.

Paso 2. Ranking. Se evalúa la función de pérdida (SIC)

$$SIC(m) = \log \hat{\sigma}^2(m) + c\frac{\log(T)k(m)}{T} \tag{12}$$

en cada uno de los modelos correspondientes a los diferentes cromosomas de la generación corriente. En tal caso *k* minúscula es el número de unos en cada cromosoma *m*, que representa los regresores seleccionados, *T* es el tamaño muestral y $\hat{\sigma}^2(m)$ representa la varianza del residuo del modelo correspondiente al cromosoma *m*. Esto permite realizar un ranking de los cromosomas de la actual generación según el valor que toma para cada uno de ellos el estadístico SIC. El factor de corrección c evita la posibilidad de sobreparametrización del modelo. Cuanto mayor es el valor de *c*, mayor es la penalización que se produce por la inclusión de regresores en el modelo.

Paso 3. Selección Natural. Con el fin de simular el proceso de selección natural darwiniana, se desecha la mitad de los cromosomas de la actual generación que tengan un valor más alto del estadístico SIC. A la mitad de cromosomas que se conserva se le denomina conjunto de emparejamiento.

Paso 4. Emparejamiento. Se seleccionan parejas de cromosomas del conjunto de emparejamiento con el fin de producir una nueva generación de modelos candidatos. La selección de las parejas de cromosomas que van a cruzarse en el siguiente paso puede realizarse de muchas formas. Es frecuente llevar a cabo un proceso aleatorio de emparejamiento que asigna igual probabilidad a cada cromosoma.

Paso 5. Cruzamiento. Este proceso se llama recombinación genética o *crossover* y permite el intercambio de material genético entre dos cromosomas. Se basa en seleccionar, al azar, un determinada posición, o punto de corte, en cada pareja que va a cruzarse. Dicho punto de corte se emplea para separar el vector binario de cada cromosoma en dos subvectores. Los dos subvectores a la derecha del punto de corte son entonces intercambiados entre la pareja de cromosomas, obteniéndose dos nuevos

cromosomas que contienen material genético de ambos progenitores. Por ejemplo, consideremos un par de cromosomas llamados madre y padre:

Madre=(0,1,0|,1,0) , Padre = (1,0,1,|0,1),

donde el punto de corte ha sido seleccionado después de la tercera posición en cada cromosoma. A partir de estos progenitores es posible crear otros dos descendientes de la forma:

Descendiente1=(0,1,0|,0,1) y Descendiente2=(1,0,1,|1,0).

Cada uno hereda parte del material genético de sus padres, lo que significa que si recombinamos los modelos que contienen los subconjuntos de regresores $\{X_2, X_4\}$ y $\{X_1, X_3, X_5\}$, obtenemos como descendencia los modelos $\{X_2, X_5\}$ y $\{X_1, X_3, X_4\}$.

Paso 6. Mutación. La mutación es un proceso que consiste en producir cambios aleatorios del código genético binario de un cromosoma. Para ello se selecciona, al azar, un elemento particular de la estructura binaria de un cromosoma. Si el elemento seleccionado es un “cero” (un “uno”), se cambia por un “uno” (un “cero”). Esto ocurre con muy baja probabilidad con el fin de no alterar las áreas más prometedoras del espacio de búsqueda. Las mutaciones evitan que el AG converja demasiado rápido a un óptimo local.

Paso 7. Convergencia. Se vuelve al paso 2 y se repite consecutivamente este proceso obteniendo sucesivas generaciones de soluciones hasta que sea satisfecho algún criterio de convergencia. Un criterio de convergencia que suele establecerse es que la población converja a una única solución. Otro posible criterio de parada es que el algoritmo alcance un determinado número máximo de generaciones.

5. COMPARACIÓN DE GASIC CON PROCEDIMIENTOS PREVIOS DE SELECCIÓN DE MODELOS

GASIC muestra un buen comportamiento relativo cuando se le compara con otras metodologías en un entorno de simulación. A fin de investigar el rendimiento relativo de GASIC, Acosta-González y Fernández-Rodríguez (2007) lo comparan con otras dos recientes metodologías desarrollas para la selección de modelos. Concretamente con las metodologías de Hoover y Perez (1999) (HP) y de Pérez-Amaral et al. (2003).

HP habían utilizado un procedimiento basado en la búsqueda de modelos yendo de lo general a lo específico, típica de la LSE, en la base de datos propuesta por Lovell

(1983), consistente en diversas variables macroeconómicas que miden la actividad real, los flujos fiscales del gobierno, los agregados monetarios, las rentabilidades fiscales, las condiciones del mercado de trabajo y una tendencia temporal. El algoritmo de selección de HP está formado por siete diagnósticos residuales y contrastes de hipótesis sobre un particular camino de búsqueda basado en el algoritmo *stepwise*, que actúa como un procedimiento de eliminación de variables en el modelo general. Considerando las 11 especificaciones construidas por el algoritmo HP, excepto en 2 de los modelos, GASIC mejora la especificación del modelo verdadero, mejorando dramáticamente el algoritmo *stepwise*, que en la base de datos de Lovell tiende a sobre identificar los modelos, siendo la selección de variables significativas falsas muy alta.

Por otra parte, también es posible comparar GASIC con otra metodología muy flexible para la construcción de modelos llamada RETINA propuesta por Pérez-Amaral et al. (2003). RETINA se basa en el uso de medidas extramuestrales de ejecución para la selección del modelo. Como señalan Acosta-González y Fernández-Rodríguez (2007), GASIC y RETINA son asintóticamente similares. Las simulaciones señalan que ambas metodologías muestran gran capacidad para seleccionar el modelo verdadero en el contexto de variables con alto grado de multicolinealidad. No obstante, GASIC suele mejorar a RETINA para pequeños tamaños muestrales con bajos R^2.

6. UTILIDAD PRÁCTICA DE GASIC

En este apartado se muestran tres ejemplos donde el algoritmo GASIC ha sido empleado con éxito: la selección de variables que explican el crecimiento económico, la predicción del fracaso empresarial y la formación de una cartera, de pocos activos, cuyo objetivo sea tener un rendimiento lo más parecido posibles al índice bursátil español IBEX35.

6.1 SELECCIÓN DE FACTORES QUE EXPLICAN EL CRECIMIENTO ECONÓMICO

El trabajo seminal de Barro (1991) ha dado lugar a una literatura empírica sobre el crecimiento económico tratando de identificar las posibles variables que están correlacionadas parcialmente con las tasas de crecimiento económico. Estos trabajos

intentan identificar los factores que explican las tasas de crecimiento de los diferentes países regresando el crecimiento del PIB observado sobre diversas características que podrían explicar dicho crecimiento. La metodología básica consiste entonces en realizar regresiones de la forma

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon \ ,$$

donde Y es el vector de las tasas de crecimiento económico, y $X_1,...,X_n$ son vectores de variables explicativas que varían según los investigadores que traten el tema. Entre los diferentes trabajos desarrollados en este sentido es común encontrar variables como el nivel inicial de renta, diversas medidas relativas a la calidad del sistema educativo y del capital humano del país, la tasa de inversión, los impuestos, distorsiones del mercado, grado de monopolio, indicadores sobre el sistema político y mantenimiento de los derechos de propiedad, indicadores relativos a aspectos culturales, históricos e ideosincráticos del país (tales como su sistema democrático, la religión, el sistema legal o el haber sido antigua colonia), variables que describen el nivel de la tecnología, acceso a determinadas materias primas como el agua, y muchas otras variables que parecen estar correlacionadas con el crecimiento y que no siempre existe una forma clara para medirlas (tales son, por ejemplo, la eficiencia gubernamental, el nivel de corrupción, la actitud hacia el trabajo, etc.). De todas estas variables puede formarse un conjunto de aproximadamente 60 regresores potenciales que explican el crecimiento. El problema al que se enfrentan los investigadores es que la teoría del crecimiento no especifica claramente qué variables explicativas incluir en el modelo.

En este sentido, existe una amplia controversia a la hora de identificar los factores que explican las diferentes tasas de crecimiento de los diversos países. Así, en ausencia de una teoría económica que guíe dicha búsqueda, han proliferando numerosos trabajos de contenido contradictorio sobre cuáles deberían ser las variables explicativas más relevantes en el crecimiento.

Levine and Renelt (1992) han concluido que son muy pocos los regresores que han pasado el contraste de cotas extremas de Leamer (1983, 1985) para identificar relaciones empíricas “robustas”. Dicho contraste funciona como sigue: supongamos un conjunto *X* de *K* variables explicativas que han sido previamente identificadas como potencialmente explicativas del crecimiento y que estamos interesados en averiguar si una determinada variable *z* es “robusta”. En tal caso estimaremos las posibles regresiones de la forma

$$Y = \alpha_j + \beta_{yj} y + \beta_{zj} z + \beta_{xj} x_j + \varepsilon$$

donde *y* es un vector de cuatro variables fijas que aparecen en todas las regresiones (el nivel inicial de renta, la tasa de inversión, la tasa de matriculación en la escuela secundaria y la tasa de crecimiento de la población). La *z* es la variable de interés y $x_j \in X$ es un conjunto de tres variables tomadas a partir de las restantes *K-4* variables disponibles. Entonces es necesario estimar esta regresión para todas las posibles combinaciones de variables $x_j \in X$.

Para cada modelo j se encuentra entonces una estimación de β_{zj} y su correspondiente desviación standard σ_{zj}. La cota extrema inferior se define como valor más bajo de $\beta_{zj} - 2\sigma_{zj}$, y la cota extrema superior como el valor más alto de $\beta_{zj} + 2\sigma_{zj}$. El contraste de cotas extremas para una variable *z* dice que si la cota extrema inferior es negativa y la cota extrema superior es positiva, entonces la variable *z* no es robusta. Levine y Renelt (1992) concluyen que muy pocas variables están correlacionadas de forma robusta y sistemática con el crecimiento. Por el contrario, Sala-i-Martin (1997) afirma que el contraste es demasiado estricto para que las variables que explican el crecimiento puedan pasarlo. Así, empleando un contraste menos severo para la selección de variables explicativas, identifica un amplio conjunto de variables significativas.

De esta forma, el debate sobre las variables que explican las tasas de crecimiento de los diferentes países ha sido transferido al terreno de la selección de modelos, dando lugar a una competición entre metodologías alternativas: la modelización Bayesiana de Fernández et al. (2001), la aproximación de la *London School of Economics*, representada por Hoover y Perez (2004) y el PcGets de Hendry y Krolzig (2004). GASIC también ha entrado en dicha competición metodológica, presentando una buena ejecución frente a las metodologías alternativas. Así, el trabajo de Acosta-González y Fernández-Rodríguez (2007) ha ilustrado la potencia de GASIC en la selección de variables explicativas del crecimiento. Cuando se emplea el procedimiento estadístico de imputación múltiple de King et al (2001) para rellenar los *missing* de la base de datos de Sala-i-Martin, el algoritmo de Hoover y Perez (2004) selecciona cuatro variables explicativas: “el número de años de economía abierta”, “la inversión en equipo”, “la fracción de Confucionismo”, “las revoluciones y golpes de estado” y “la fracción protestante”. Por el contrario, GASIC selecciona un modelo anidado en el anterior que

contiene tan solo dos variables, “el número de años de economía abierta” y ”la fracción de Confucionismo”.

6.2 SELECCIÓN DE VARIABLES QUE EXPLICAN LA QUIEBRA EMPRESARIAL

Dado un amplio conjunto de ratios financieros, la metodología GASIC permite la construcción de un modelo LOGIT con el objetivo de predecir la quiebra empresarial. Este modelo se construye mediante la selección de un reducido número de dichos ratios financieros.

La predicción de la quiebra es un tema recurrente en la literatura financiera. Se trata de un problema de interés teórico y práctico que ha atraído la atención de los reguladores, profesionales y también de los académicos. Ha emergido así una literatura donde numerosos investigadores han intentado mostrar la utilidad de los modelos predictivos basados en la contabilidad publicada anualmente por las empresas. Por otra parte, como señalan Altman (2001) y Duffie y Singleton (2003), entre otros, las metodologías desarrolladas para la predicción de quiebras también pueden ser empleadas con numerosos propósitos en las finanzas, incluyendo el control de solvencia de las instituciones financieras por parte de los reguladores, la valoración de la concesión de préstamos, la medida del riesgo de una cartera, la clasificación de bonos y la valoración de derivados de crédito y otros activos expuestos al riesgo de crédito. Todo ello ha dado lugar al estudio de lo que se conoce como “modelos de fracaso”.

Uno las técnicas estadísticas más usadas en los modelos de fracaso es el modelo LOGIT, usado por primera vez por Ohlson (1980), quizás sólo superado en la frecuencia de su uso por el análisis discriminante múltiple de Altman (1968). El modelo LOGIT presenta ciertas ventajas que lo hacen superior. Un resumen de sus principales ventajas e inconvenientes de estos modelos puede encontrarse en Balcane y Ooghe (2004).

Igualmente se han propuesto numerosas metodologías alternativas con diferentes suposiciones subyacentes y complejidad computacional. Cabría citar, entre otras: Sistemas Expertos (Messier y Hansen, 1988), Modelos no Paramétricos como los Splines de Regresión Adaptativa Multivariante (MARS) (Friedman, 1991), Redes Neuronales Artificiales Artificial (Tam y Kiang, 1992), Clasificadores Híbridos para

combinar procedimientos (Olmeda y Fernández, 1997) y modelos LOGIT mixtos (Jones y Hensher, 2004). Sin embargo, existen un número de dificultades comunes en la aplicación de estas técnicas todavía no resueltas.

Uno de estos problemas es la inexistencia de una teoría económica sobre la solvencia de las empresas, de modo que no existe consenso en los estudios empíricos de cuáles deben ser los ratios financieros capaces de explicar la quiebra. Por otra parte, la evidencia empírica acumulada sólo es capaz de permitir la formulación de hipótesis sobre los ratios financieros capaces de explicar y predecir la situación financiera futura de una empresa. Por tanto, no existe un modelo restringido, universalmente aceptado, al que el investigador empírico pueda recurrir para la selección de los ratios financieros, ni una metodología estadística universalmente aceptada.

El trabajo de Acosta-González y Fernández-Rodríguez (2009) aborda, entre otras cosas, el problema de la selección de los ratios financieros que explican el fracaso empresarial en el contexto de un modelo LOGIT. Dicha selección se realiza empleando el algoritmo GASIC. Para ello, se comienza seleccionando un amplio conjunto de variables explicativas que formarán el modelo general no restringido. Siguiendo a Beaver (1966) en uno de los trabajos seminales sobre la utilidad de la información contable en la predicción del fracaso empresarial, se consideraron 32 tipos de ratios financieros:

- Ratios empleados frecuentemente en la literatura para medir la solvencia de la empresa.
- Ratios que se han comportado bien en estudios previos.
- Ratios definidos en términos del concepto de “cash flow”.

Igualmente se han añadido a las 32 ratios iniciales sus cuadrados. De esta forma GASIC parte de un conjunto de 64 regresores en el modelo general no restringido. Este tipo de información es muy redundante y la alta correlación entre variables puede producir problemas de multicolinealidad en las estimaciones, dando lugar a elevados errores estándar en los coeficientes estimados, lo que afecta negativamente a la precisión de las estimaciones. No obstante, tal como señala Benishay (1971), se perdería un gran contenido informativo si sólo se incluyesen en el modelo econométrico un reducido conjunto de variables poco dependientes. GASIC da respuesta a este problema en el contexto de la construcción de un modelo LOGIT, posibilitando la identificación automática de un subconjunto de ratios financieros explicativos del fracaso empresarial, conteniendo la información más completa posible, pero intentando evitar su duplicidad.

Esto es así debido principalmente a que el criterio de selección que utiliza GASIC está basado en una medida global del modelo, como es el SIC, y no en criterios individuales de significación como es el caso del estadístico *t*, cuyos resultados están fuertemente condicionados por el alto grado de multicolinealidad que generan los ratios financieros.

Denotemos por X_{ji}, al valor de un ratio financiero j en la empresa i, y consideremos una variable no observable y_i^* que puede ser interpretada como "propensión hacia el fracaso". Como y_i^* es una variable no observable, puede llevarse a cabo el análisis LOGIT definiendo una variable binaria y_i, de la forma:

$$y_i = \begin{cases} 1 \quad si \quad y_i^* = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ji} + u_i > 0 \quad La\ empresa\ \ i\ \ fracasa \\ 0 \quad en\ otro\ caso. \quad La\ empresa\ \ i\ \ no\ \ fracasa \end{cases} \tag{13}$$

En tal caso, la probabilidad de fracaso de la empresa *i* es

$$\begin{aligned} P_i = P(y_i = 1) = P\left(u_i > -(\beta_0 + \sum_{j=1}^{k} \beta_j X_{ji}) \right) = \\ 1 - F\left(-(\beta_0 + \sum_{j=1}^{k} \beta_j X_{ji}) \right) = F\left(\beta_0 + \sum_{j=1}^{k} \beta_j X_{ji} \right) \end{aligned} \tag{14}$$

donde F es la función de distribución acumulada de u_i, que suponemos que se trata de la función logística.

Por tanto, llamando $Z_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ji}$ resulta

$$P_i = F(Z_i) = \frac{\exp(Z_i)}{1 + \exp(Z_i)} = \frac{1}{1 + \exp(-Z_i)} \tag{15}$$

Los parámetros del modelo de regresión logistica se estiman maximizando la verosimilitud de las observaciones contenidas en la muestra, siendo la función de verosimilitud

$$L = \prod_{y_i=1} P_i \prod_{y_i=0} (1 - P_i) \tag{16}$$

Con el fin de obtener la probabilidad de fracaso de una empresa *i* como una función de los ratios financieros $X_{i1}, ..., X_{iK}$, GASIC se enfrenta al problema de encontrar el mejor modelo de la forma:

$$P_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 X_{i j_1} + ... + \beta_K X_{i j_k})} \quad \text{donde} \quad \{ j_1, j_2 ...., j_k \} \subseteq \{1, 2, ...K\} \tag{17}$$

Como se mencionó previamente, K es el número total de ratios financieros de las que se parte inicialmente, existiendo un total de 2^K posibles modelos potenciales que examinar, algo prohibitivo cuando K es grande. En el caso de un modelo LOGIT no hay porque suponer que los errores del modelo se distribuyan como una normal. En tal caso en el paso 2 del algoritmo GASIC, en vez de utilizar la expresión (12), es conveniente emplear como función de pérdida la siguiente formulación de SIC:

$$SIC = -\frac{2l}{N} + \frac{k \log(N)}{N} \tag{18}$$

donde l es el logaritmo de la verosimilitud obtenida a partir del estimador de máxima verosimilitud del modelo LOGIT, N es el tamaño muestral y k es el número de regresores del modelo, siendo $k \leq K$.

Así, la metodología GASIC, desarrollada inicialmente para la selección de modelos con variable endógena continua y estimados por mínimos cuadrados ordinarios, fue adaptada en Acosta-González y Fernández-Rodríguez (2009) a un modelo LOGIT, estimado por máxima verosimilitud, con el fin de estudiar el fracaso empresarial. En dicho trabajo se realiza un ejercicio de predicción del fracaso empresarial, durante el año 2004, de empresas españolas de la construcción, empleando información de uno, dos, tres y cuatro años antes de la quiebra, es decir, empleando información procedente de los años 2000, 2001, 2002 y 2004. Las variables seleccionadas por GASIC para explicar el fracaso empresarial durante 2004 (a partir de las 32 ratios financieros originales sin incluir sus cuadrados) fueron las siguientes:

Con información procedente del año 2000: productividad, gastos financieros, tesorería, capacidad de pago, margen de beneficios e ingresos de explotación sobre fondos propios mas pasivo fijo.

Con información procedente del año 2001: gastos financieros, tesorería, crédito a proveedores y margen de beneficios.

Con información procedente del año 2002: rotación de activos, endeudamiento, liquidez inmediata y rendimiento sobre capital empleado.

Con información procedente del año 2003: rentabilidad económica y fondo de maniobra.

La capacidad predictiva de los modelos seleccionados, tanto intramuestral como extramuestral, es elevada.

Hay que señalar que GASIC es especialmente adecuado para tratar el problema del fracaso empresarial, donde la información de las variables es muy redundante, debido al

buen comportamiento que ha mostrado dicha metodología al seleccionar modelos cuando los regresores muestran una elevada multicolinealidad, tal como se señaló en el apartado anterior.

Finalmente, debe observarse que muchas decisiones empresariales implican una clasificación binaria, existiendo una extensa literatura para estudiar este problema en varios contextos tales como clasificación crediticia, predicción de fracaso, fusión y adquisición y calificación de bonos, entre otros. En todos esos problemas de clasificación se pueden emplear las técnicas desarrolladas en el problema de la predicción de la quiebra. En este sentido, la metodología propuesta puede emplearse en el contexto muy general de realización de decisiones financieras multicriterio.

6.3 SELECCIÓN DE ACTIVOS FINANCIEROS EN UNA CARTERA QUE SIGUE UN ÍNDICE BURSATIL

Los gestores de fondos cuentan, básicamente, con dos tipos de estrategias de inversión, la activa y la pasiva. Las estrategias de inversión activa se basan en la convicción de que la habilidad de un inversor con sus actividades de compra y venta de acciones puede batir los rendimientos del mercado. En las estrategias de inversión pasiva el gestor de fondos tiene muy poca flexibilidad en la toma de sus decisiones inversoras y su actividad ha de limitarse a obtener, aproximadamente, los mismos rendimientos que un determinado índice bursátil que se toma como referencia. Las estrategias de inversión pasiva descansan en la creencia de que los mercados son eficientes y que es, por tanto, imposible mejorar los rendimientos agregados del mercado de forma consistente. Este tipo de estrategias se han hecho muy populares porque su coste es muy inferior al de las estrategias activas. En este sentido, Elton et al. (1996) han demostrado en un exhaustivo análisis empírico que, históricamente, la mayoría de los fondos gestionados de forma activa no mejoran los rendimientos del mercado.

De cara a la formación de una estrategia de inversión pasiva, hay dos posibles formas de replicar el índice bursátil que se tome de referencia. Por un lado, se puede acudir a la replicación plena que consiste en la compra de todos los activos que conforman el índice, en la misma proporción en que figuran en él. Esta replicación plena tiene numerosas desventajas tales como que ciertos activos de la cartera de réplica pueden aparecer en muy pequeñas cantidades, altos costes de transacción y dificultades en la

recomposición de la cartera cuando cambian los pesos en los activos financieros que componen el índice. Por esa razón el seguimiento de un índice bursátil se suele realizar de esta otra forma, eligiendo un pequeño número de títulos del índice empleando algún tipo de criterio.

En el trabajo de Fernández-Rodríguez et al. (2009) se muestra como GASIC puede contribuir a la selección de estos títulos de forma muy eficiente, para el caso del IBEX35. Se trata de resolver un complejo programa de programación no lineal mixta asociada con la selección de tales activos a la hora de replicar el índice.

Específicamente, la técnica que se propone con el propósito de seleccionar las acciones en la cartera de seguimiento del índice descansa en el siguiente modelo de regresión

$$r_{It} = \alpha_0 + \sum_{i=1}^{k} \alpha_i r_{it} + u_i \tag{19}$$

donde r_{It} representa el rendimiento del índice en el momento *t* y r_{it} el rendimiento del activo financiero i en el momento *t*. Todos los rendimientos han sido obtenidos a partir de la diferencia logarítmica de los precios. Con el fin de obtener la cartera de réplica se añaden al modelo las restricciones adicionales:

$$\text{(a) } \alpha_0 = 0 \tag{20}$$

$$\text{(b) } \sum_{i=1}^{k} \alpha_i = 1 \tag{21}$$

La primera restricción expresa que el rendimiento de la cartera está sólo asociado con sus propios títulos. La segunda restricción asegura la completa inversión de todo nuestro presupuesto sin permitir que pidamos dinero prestado para financiar la compra de activos financieros.

Existen dos problemas a la hora de especificar el modelo (19). Por un lado, es necesario determinar el valor de k, es decir, el número de activos financieros en la cartera de réplica. Por otra parte, es necesario establecer que activos, específicamente, constituyen la cartera de réplica. Como estamos intentando replicar un índice con 35 títulos, en número total de posibles carteras de réplica es de $2^{35} = 34.359.738.368$. En el caso de un índice tal como el S&P500, el número de posibles especificaciones del modelo (19) será enorme.

La base de datos empleada en este trabajo contiene los rendimientos diarios del IBEX35, al cierre, desde el 1/9/2005 al 19/6/2008, y de cada uno de los 35 activos que lo componen, lo que representa un total de 712 días de cotización. El objetivo del trabajo consiste en construir una cartera de réplica durante el último año del periodo

muestral, de modo que la cartera de réplica debería ser actualizada trimestralmente con el fin de actualizar la información. Así, con el fin de obtener predicciones extra-muestrales, las estimaciones de la cartera de réplica, cada trimestre, se han realizado usando observaciones del periodo muestral previo. Por tanto, durante el último año, la cartera de réplica se reestructura cada tres meses, repitiendo el procedimiento GASIC para la selección de títulos y pesos. El número de empresas seleccionadas por GASIC durante el último año fue de, 15 el primer cuatrimestre, 16 para el segundo, 16 para el tercero y 14 para el cuarto. La composición de la cartera de réplica es bastante estable con respecto a los títulos que contiene. Así, ha habido 11 compañías que siempre han formado parte de la cartera de réplica, mientras que cuatro han sido seleccionadas para tres de los cuatro cuatrimestres; dos compañías fueron seleccionadas en dos cuatrimestres; y una compañía fue solo seleccionada un cuatrimestre. Además, los pesos que tenía cada título en la cartera de réplica, durante los cuatro cuatrimestres, es muy estable y su variación nunca excede del 2.5% para dos cuatrimestres consecutivos. Tal estabilidad de la cartera de réplica trae consigo la reducción de los costes de transacción a la hora de hacer el seguimiento del índice con estas carteras de réplica, respecto a la realización de una replicación plena basada en la adquisición de todos los activos que forman parte del índice. Finalmente, tal como cabría esperar, los títulos con peso más alto en las carteras de réplica han coincidido con aquellos que tienen un peso más elevado en el IBEX35, es decir, Telefónica, Santander y BBVA.

7. CONCLUSIONES

El algoritmo GASIC proporciona una nueva metodología de la minería de datos que permite la selección de modelos econométricos cuando el número de posibles candidatos a regresores sea elevado.

Dicho algoritmo se basa en un procedimiento heurístico de optimización llamado Algoritmo Genético que se emplea para explorar el universo de los modelos disponibles a partir de un modelo general no restringido. La principal contribución de GASIC es proporcionar un método de selección de regresores que mejora considerablemente el algoritmo *stepwise*, tanto en sus versiones *forward* como *backward*.

Según han mostrado las simulaciones de Monte Carlo, GASIC se caracteriza por una extraordinaria capacidad para encontrar el modelo original o verdadero, incluso en el

caso en que las múltiples variables candidatas presenten una elevada correlación entre sí. Además, pese a tratarse de un procedimiento aleatorio que parte de una población de soluciones obtenidas inicialmente al azar, el algoritmo resulta muy robusto en el sentido en que en casi el 100% de las veces obtiene el mismo modelo final, con independencia de la semilla empleada.

GASIC ha sido empleado con éxito en tres problemas prácticos, completamente diferentes de la literatura económica: la selección de variables que explican el crecimiento económico, la predicción del fracaso empresarial y la formación de una cartera de pocos activos que siga el comportamiento de un índice bursátil como el IBEX35.

Esta herramienta de selección de modelos presenta, por tanto, una amplia capacidad de aplicaciones potenciales en numerosos otros campos de la economía.

REFERENCIAS BIBLIOGRÁFICAS

- Acosta-González, E. y Fernández-Rodríguez, F. (2007). Model selection via genetic algorithms illustrated with cross-country growth data. *Empirical Economics*, 33 313-337.
- Acosta-González, E. y Fernández-Rodríguez, F. (2009). Financial Ratios Selection for Predicting Failure of Firms via Genetic Algorithms. Aceptado en *The Journal of Credit Risk*.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N., Csake, F., editors. *2nd International Symposium on Information Theory*. Budapest, 1973; 267-281.
- Altman, E. (1968). Financial ratios, discriminate analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23, 589-609.
- Altman, E. (2001). *Bankruptcy, Credit Risk and High Yield Junk Bonds*. Blackwell Publishers. New York.
- Amemiya, T. (1980). Selection of Regressors, *Internacional Economic Review*, 21, 331-354.
- Balcane, S. y Ooghe, H. (2004). 35 Years of studies on business failure. *Working Paper* 2004/248. Faculteit Economie en Bedrijfskunde.
- Barro, R. (1991). Economic Growth in a Cross Section of Countries. *Quarterly Journal of Economics*, 106, 2(May), 407-443.

- Beaver, W. H. (1966). Financial ratios as predictor of failure. *Journal of Accounting Research* 4, 71-111.
- Beenstock, M. y Szpiro, G. (2002). Specification search in nonlinear time-series models using the genetic algorithm. *Journal of Economic Dynamics y Control* 26, 811-835.
- Benishay, H. (1971). Econometric information in financial ratio analysis. *Accounting and Business Research.* Spring. 174-179.
- Duffie, D. y Singleton, K. (2003). *Credit Risk: Pricing, Measurement and Management*. Princeton, Princeton University Press.
- Elton, E., Gruber, G. y Blake, C. (1996). Survivorship bias and mutual fund performance. *Review of Financial Studies* 9, 1097-1120.
- Ebbeler, D.H. (1975). On the probability of Correct Model Selection Using the Maximim $\bar{R}^2$ Choice Criterion. *International Economic Review*, junio, 516-520.
- Fernández-Rodríguez, F., Acosta-González, E. y Andrada-Félix, J. (2009). Model selection using Data Mining. In *Data Mining and Management*, Lawrence I. Spendler (ed.). Series: Computer Science, Technology and Applications. Nova Science Publishers, Nueva York.
- Fernández, C., Ley, E. y Steel, M. F. (2001). Model Uncertainty in Cross-Country Growth Regressions. *Journal of Applied Econometrics* 16, 563-576**.**
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* 19, 1-141.
- Hannan, E. J. y Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society B* 41, 190-195.
- Hansen, B. E. (1999). Discussion of "Data mining reconsidered". *Econometrics Journal* 2, 192-201.
- Hendry, D. F. y Krolzig, H. M. (1999). Improving on "Data mining reconsidered" by K.D. Hoover and S.J. Perez. *Econometrics Journal* 2, 202-219.
- Hendry, D. F. y Krolzig, H. M. (2001). *Automatic Econometric Model Selection Using PcGets 1.0*. London: Timberlake Consultants Press.
- Hendry, D. F. y Krolzig, H. M. (2003). The Properties of Automatic Gets Modelling. *Working Paper*, Economics Department, Oxford University.
- Hendry, D. F. y Krolzig H. M. (2004). We Ran One Regression. *Oxford Bulletin of Economics and Statistics* 665, 799-810.

- Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor: The University of Michigan Press.
- Hoover, K. D. y Perez, S. J. (1999). Data mining reconsidered: encompassing and the general- to-specific approach to specification search. *Econometrics Journal* 2, 167-191.
- Hoover, K. D. y Perez S. J. (2004). Truth and robustness in cross-country growth regressions. *Oxford Bulletin of Economics and Statistics* 66 (5), 765-798.
- Johnston, J. (1984). *Econometric Methods*. MacGRAW-HILL, New York.
- Jones, S. y Hensher, D. A. 2004. Predicting firm financial distress: A mixed logit model. *The Accounting Review* 79, 1011-1038.
- King, G., Honaker, J, Joseph, A. y Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review*, 95 49-69.
- Koza, J. (1992). *Genetic Programming*. Cambridge, MA: MIT Press.
- Leamer, E.E. (1983) Let's take the con out of econometrics. *American Economic Review* 73(3), 31-43.
- Leamer, E.E. (1985). Sensitivity analyses would help. *American Economic Review* 75(5), 308-313.
- Levine, R. y Renelt, D. (1992) A sensitivity analysis of cross-country growth regressions. *American Economic Review* 82: 942-963.
- Lovell, M.C. (1983) Data mining. *Review of Economic and Statistics*, 65 1-12.
- Maddala (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press. Cambridge.
- Mallows, C. L. (1973). Some Comments on C_p. *Technometrics*, November, 661-676.
- Messier, W. F. y Hansen, J. V. (1988). Inducting rules for expert system development: an example using default and bankruptcy data. *Management Science* 34, 1403-1415.
- Miller, A. (2002). *Subset Selection in Regression*. Chapman y Hall/CRC.London.
- Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18, 109-131.

- Olmeda, I. y Fernández, E. (1997). Hybrid classifiers for financial multicriteria decision making: The case of bankruptcy prediction. *Computational Economics* 10, 317-335.
- Pacheco-Bonrostro, J., Casado-Yusta, S. y Núñez Letamendía, L. (2007). Algoritmos meméticos para selección de variables en el análisis discriminante. *Estadística Española*, 45(165), 333-347.
- Pérez-Amaral, T, Gallo, G. M., y White, H. (2003). A flexible tool for model building: The relevant transformation of the inputs network approach (RETINA). *Oxford Bulletin of Economics and Statistics* 65, 821-838.
- Sala-i-Martin, X. (1997). I just ran two million regressions. *American Economic Review* 87, 178-183.
- Schmertmann, C. P. (1996). Functional search in economics using genetic programming. *Computational Economics* 9, 275-298.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.
- Szpiro, G. G. (1997). A search for hidden relationship: Data mining with genetic algorithms. *Computational Economics* 10, 267-277.
- Tam, K. Y. y Kiang, M.Y. (1992). Managerial Applications of Neural Networks: The Case of Bank Failure Prediction. *Management Science* 38, 926-947.
- Theil, H. (1971). *Principles of Econometrics*. Ed. Wiley.