Proyecto Fin de Master en Ingeniería Informática para la Industria.
Master en Investigación en Informática.
Facultad de Informática.
Universidad Complutense de Madrid.

ANALISIS PREDICTIVO DE DATOS MEDIANTE TECNICAS DE REGRESION ESTADISTICA

Autor: Augusto Pereira González
Director: Matilde Santos Peñas
Colaborador externo de dirección: Jesús A. Vega Sánchez
Curso académico: 2009-2010

PREDICTIVE DATA ANALYSIS BY MEANS OF STATISTICAL REGRESSION TECHNIQUES

Abstract:

Statistical regression is one of the most widely used technique to find a variable that is function of one or more explanatory variables; however, usually it's used the 'Ordinary Least Square' technique (OLS), but it faces problems when the variables have multicollinearity (linear relation between them). This work describes the troubles of collinearity, the effects on the models achieved and discusses the main diagnostic techniques to solving them and preventing them. 'Ridge Regression' and 'Kernel Ridge Regression' are the most commonly used procedures to mitigate its effects. These can be implemented through different modes of computation, allowing us to quantify and to adjust the results in predictions from the initial conditions of the input data (number of observations and number of dimensions of the variables to be treated).
Finally, experimental results are provided by applying the previous techniques and by comparing the accurate on the predictions for different data sets.

Keywords: Predictive Data Analysis, Statistical Regression, Ridge Regression.

ANALISIS PREDICTIVO DE DATOS MEDIANTE TECNICAS DE REGRESION ESTADISTICA

Resumen:

La regresión estadística es una de las técnicas mas empleadas cuando se busca determinar una variable respuesta en función de una o más variables explicativas; sin embargo, tradicionalmente se emplea la técnica de mínimos cuadrados ordinarios (MCO), la cual enfrenta problemas cuando las variables explicativas presentan multicolinealidad (relación lineal entre ellas). En este trabajo se describe el problema de la colinealidad, sus efectos en los modelos generados y se discuten las principales técnicas de diagnóstico y prevención. Las variantes de regresión sesgada ('*Ridge Regression*' y '*Kernel Ridge Regression*') son los procedimientos más empleados para mitigar dicho efecto. Éstas pueden ser aplicadas mediante diferentes modalidades de cómputo, permitiéndonos cuantificar y ajustar los resultados en las predicciones a partir de las condiciones iniciales de los datos de entrada (número de observaciones y número de dimensiones de las variables a tratar).
Finalmente se muestran y aportan resultados experimentales mediante la aplicación de las técnicas analizadas, comparando las precisiones en las predicciones para diferentes conjuntos de datos.

Palabras clave: Análisis predictivo, regresión estadística, regresión sesgada.

Índice de contenido

Índice de ilustraciones

”If your experiment needs statistics [i.e., inference],
you ought to have done a better experiment.”

(Ernest Rutherford)
Nobel Prize in Chemistry in 1908

With high dimensionality,
complex regularities,
weak prior knowledge and large data sets,
… Can one always do a better experiment?

(Bernhard Schölkopf)
Empirical Inference Department
Max Planck Institute for Biological Cybernetics
Tübingen, Germany

“The brain is nothing but a statistical decision organ”

(Horace B. Barlow)
Australia Prize in Sensory perception theme in 1993

1. INTRODUCCION

El análisis de regresión es una técnica estadística para estudiar la relación entre variables. El término regresión fue introducido por Francis Galton **[Galton, 1886]** . Su trabajo se centró en la descripción de los rasgos físicos de los descendientes (variable A) a partir de los de sus padres (variable B). Estudiando la altura de padres e hijos a partir de más de mil registros de grupos familiares, se llegó a la conclusión de que los padres muy altos tenían una tendencia a tener hijos que heredaban parte de esta altura, pero que revelaban también una tendencia a regresar a la media. Galton generalizó esta tendencia bajo la "ley de la regresión universal": «Cada peculiaridad en un hombre es compartida por sus descendientes, pero en media, en un grado menor.».
Tanto en el caso de dos variables (regresión simple) como en el caso de más de dos variables (regresión múltiple), el análisis puede utilizarse para explorar y cuantificar la relación entre una variable llamada dependiente o criterio (Y) y una o más variables llamadas independientes, predictoras o regresoras (X_1, X_2, …, X_n), así como para desarrollar una ecuación lineal con fines predictivos. En problemas de regresión se dispone de una serie de datos de entrenamiento que representan las entradas y las correspondientes salidas de un sistema lineal o no lineal. El objetivo de la regresión es descubrir la relación funcional entre la entrada y la salida de este sistema, para poder así predecir la salida del sistema cuando se le presenta un dato de entrada nuevo. Tradicionalmente se emplea la técnica de mínimos cuadrados ordinarios (MCO) como método básico de regresión, la cual encuentra problemas cuando las variables independientes presentan multicolinealidad (cuando una variable independiente puede ser explicada como una combinación lineal o correlación de una u otras variables independientes). Este efecto provoca frecuentemente elevados errores puntuales en las predicciones, lo que conduce a generar modelos predictivos con muy poco poder explicativo y de difícil interpretación en las salidas correspondientes a entradas similares que deberían también predecir salidas similares. El procedimiento de eliminar variables correlacionadas del análisis puede ser aceptado por reduccionista y como un modo de simplificar el modelo generado (computacionalmente más eficiente); sin embargo este medio reduce la carga de datos de entrada inicial al sistema y esto lo puede convertir en una técnica que genere un modelo con menor poder predictivo (reduciéndose la tasa de acierto global en las salidas a predecir). Para resolver el problema anterior se propuso la metodología denominada '*Ridge Regression*' (RR) o regresión sesgada. Este método consiste en agregar un parámetro sesgado a los estimadores de mínimos cuadrados ordinarios con la finalidad de reducir el error estándar de éstos que se comete a la hora de predecir el valor de la variable dependiente. Pero esta no es la única ventaja que ofrece este procedimiento; RR nos ofrece dos modalidades de cómputo diferentes (solución primal y dual) que podemos utilizar dependiendo de si la dimensión del espacio de características (el número de variables independientes utilizadas) es menor o mayor que el número total de ejemplos de entrenamiento que se quieren aproximar, consiguiendo así un gasto computacional mas razonable y menos costoso que el obtenido por el método tradicional de regresión utilizando MCO. Pero esto no es todo, la versión dual del procedimiento RR permite realizar regresión no lineal mediante la construcción de una función de regresión lineal en un espacio de características de más alta dimensión (comúnmente conocidas como funciones kernel); dichas funciones permiten obtener resultados sorprendentes en problemas no lineales utilizando solamente operaciones algebraicas sencillas. A esta variante regularizada de la regresión utilizando funciones kernel se le denomina

'*Kernel Ridge Regression*' (KRR) y es computacionalmente muy efectiva incluso cuando el número de dimensiones del sistema de entrada es muy elevado.
En este trabajo se quiere analizar la regresión y sobre todo sus variantes RR y KRR como métodos de aproximación en el ámbito del procesado de señales y la posibilidad de implementarla como funciones kernel para ser capaz de resolver así problemas no lineales de manera eficiente y rápida, independientemente de la dimensionalidad tanto del número de características a utilizar como del número de ejemplos de entrenamiento a tratar.
La primera parte de la memoria consiste en un estudio de la literatura sobre la RR y su implementación en algoritmos mediante métodos kernel, la segunda parte se enfoca más en las aplicaciones de estas técnicas al procesado de señales y a la precisión en la predicción de series temporales no lineales.

2. ANALISIS DE REGRESION

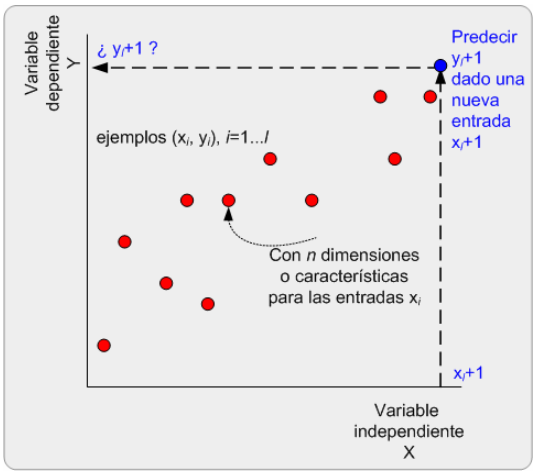

Fig. 1. Variable Y en función de X.

En un análisis de regresión simple existe una variable respuesta o dependiente (y) y una variable explicativa o independiente (x). El propósito es obtener una función sencilla de la variable explicativa, que sea capaz de describir lo más ajustadamente posible la variación de la variable dependiente. La variable explicativa puede estar formada por un vector de una sola característica o puede ser un conjunto de n características, atributos o dimensiones (regresión múltiple). La regresión se utiliza para predecir una medida basándonos en el conocimiento de otra y la intención final es que dado un vector de entrada x_{l+1} se persigue predecir un valor de salida y_{l+1} a partir de una función generada mediante la supervisión previamente observada de un conjunto de entrenamiento inicial de ejemplos (x_i, y_i), i=1…l (Fig. 1) **[NIST, 2003]**.

2.1 Regresión lineal

Como los valores observados de la variable dependiente difieren generalmente de los que predice la función, ésta posee un error. La función más eficaz es aquella que describe la variable dependiente con el menor error posible o, dicho en otras palabras, con la menor diferencia entre los valores observados y predichos. La diferencia entre los valores observados y predichos (el error de la función) se denomina variación

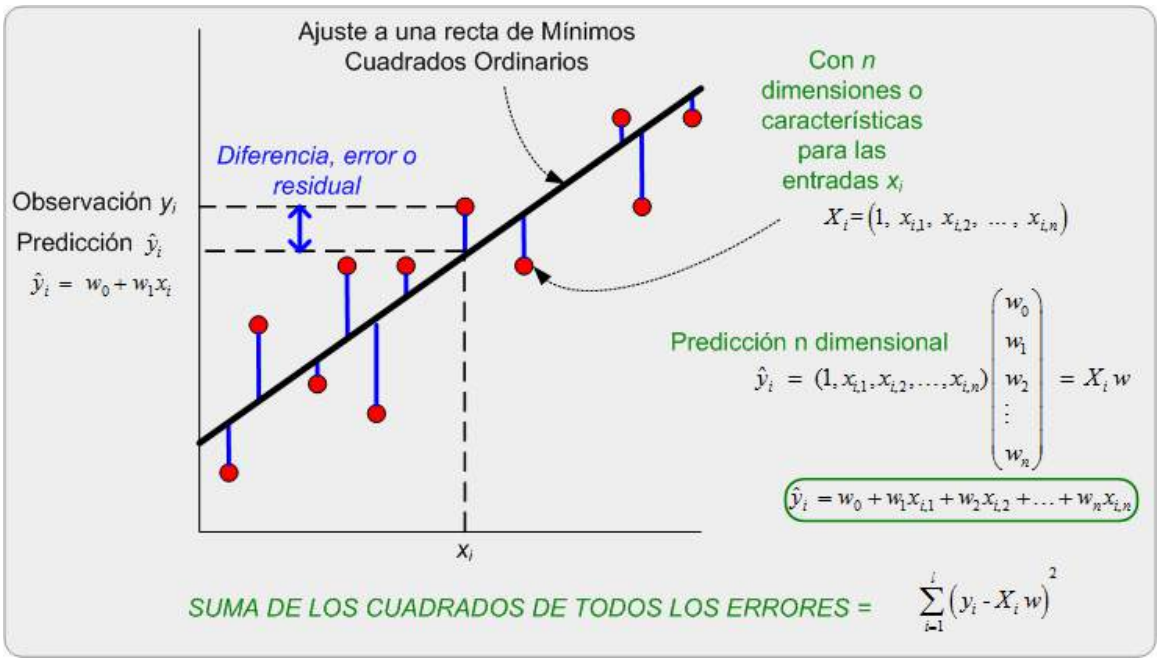

Fig. 2. Ajuste por mínimos cuadrados.

residual o residuos. Para estimar los parámetros de la función se utiliza el ajuste por mínimos cuadrados (Fig. 2) **[NIST, 2003]**. Es decir, se trata de encontrar la función en la cual la suma de los cuadrados de las diferencias entre los valores observados y esperados sea menor. Sin embargo, con este tipo de estrategia es necesario que los residuos o errores estén distribuidos normalmente y que varíen de modo similar a lo largo de todo el rango de valores de la variable dependiente. Estas suposiciones pueden comprobarse examinando la distribución de los residuos y su relación con la variable dependiente.
Cuando la variable dependiente es cuantitativa y la relación entre ambas variables sigue una línea recta, la función es del tipo $\hat{y}_i = w_0 + w_1 x_i$, en donde w_o es el intercepto o valor del punto de corte de la línea de regresión con el eje de la variable dependiente y w_1 es la pendiente o coeficiente de regresión. Pero en el supuesto de que tengamos *n* dimensiones y por tanto un caso de regresión múltiple la función de predicción será la siguiente:

$$\hat{y}_i = X_i\, w = w_o + w_1 x_{i,1} + w_2 x_{i,2} + \ldots + w_n x_{i,n}$$

2.1.1 Regresión lineal múltiple en notación matricial

Encontrar la función en la cual la suma de los cuadrados de las diferencias entre los valores observados y esperados sea menor corresponde a encontrar los coeficientes de regresión *w* para los cuales la función por la cual determinamos dicho error sea un error mínimo, o dicho de otra forma, corresponde a diferenciar la ecuación,

$$E(w) = \sum_i (y_i - X_i w)^2 \qquad (1.1)$$

Dados l ejemplos de entrada (x_i, y_i) para $i = 1 \ldots l$,
donde $X_i = (f_1(x_i)\ f_2(x_i) \cdots\ f_d(x_i))$ con d funciones definidas,

$$\frac{\partial E}{\partial w} = 0 \quad \Rightarrow \quad \sum_i \frac{\partial}{\partial w}(y_i - X_i w)^2 = 0 \quad \Rightarrow \quad \sum_i 2 X_i^T (y_i - X_i w) = 0$$

$$\Rightarrow \quad \left(\sum_i X_i^T X_i \right) w = \sum_i X_i^T y_i$$

Dejando las ecuaciones y los sistemas de ecuaciones lineales e introduciendo una notación plenamente matricial **[Thibaux, 2008]**, podemos continuar la expresión de la siguiente forma:

$$w = \left(X^T X \right)^{-1} X^T y$$

y observamos que la matriz de coeficientes de regresión *w* es función lineal de la matriz de datos observados *y*, asumiendo que $(X^T X)$ tiene inversa para todo,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_l \end{bmatrix}, \quad X = \begin{bmatrix} x_{10} & x_{11} & \dots & x_{1n} \\ x_{20} & x_{21} & \dots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{l0} & x_{l1} & \dots & x_{ln} \end{bmatrix}, \quad w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$$

2.1.2 Calidad del ajuste y su medición

Después de haber ajustado un modelo es importante contar con ciertos valores que nos ofrezcan información de cómo de importante es dicho ajuste con respecto a los datos. Como veremos más adelante, al analizar la correlación existente entre las variables independientes, existen muchos términos cuantitativos que nos dan información muy valiosa respecto a dicha medición. No obstante, una vez obtenidos los coeficientes de MCO, **[Chatterjee, 2006]** sugiere el cálculo de las siguientes cantidades:

$$\text{SST} = \sum (y_i - \bar{y})^2$$
$$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2$$
$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$

Donde SST (*Sum Squared Total*) es el sumatorio de los cuadrados de las diferencias de la variable respuesta *Y* respecto de su media. SSR (*Sum Squared Regression*) representa la suma de los cuadrados de las diferencias de la variable predictiva $\hat{Y}$ respecto a la media de la variable observada *Y*, finalmente SSE (*Sum Squared Errors*) es el sumatorio de los cuadrados de los residuales (los errores observados entre las variables Y e $\hat{Y}$). Una relación fundamental entre estas variables es la siguiente:

$$\text{SST} = \text{SSR} + \text{SSE}$$

Tomando valores ficticios para $y, \hat{y}$ e $\bar{y}$, en la Fig. 3 se representan e ilustran gráficamente las relaciones existentes entre ellas.

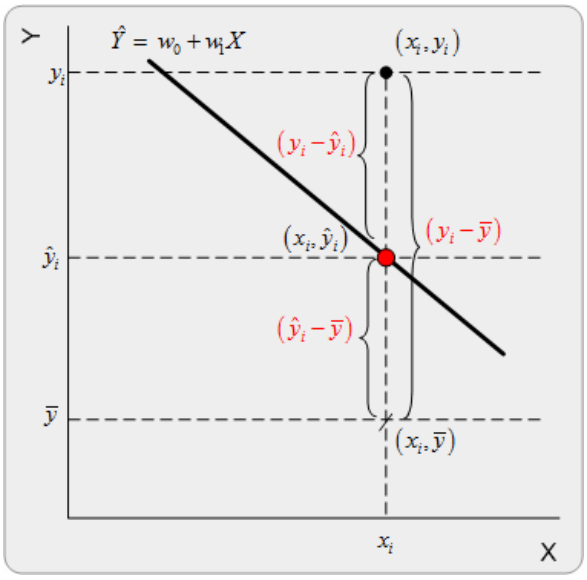

Fig. 3. Ilustración gráfica de la medición del ajuste.

Una vez introducidas las variables que hacen referencia a la suma de cuadrados, es necesario continuar con las variables que utilizan la media cuadrática, habitualmente utilizadas por el análisis de la varianza (ANOVA) en regresión múltiple. Está técnica estudia la igualdad de las medias para diferentes muestras poblacionales bajo la hipótesis de que éstas deben coincidir y por tanto el análisis de varianza sirve para comparar si los valores de un conjunto de datos numéricos son significativamente distintos a los valores de otro o más conjuntos de datos. No obstante la utilidad importante en un análisis de regresión respecto al análisis ANOVA son las variables medias cuadráticas que se utilizan frecuentemente como medida de comparación de los errores que se producen en los ajustes de regresión.
En la siguiente ilustración se puede observar la tabla resultante de un análisis ANOVA y sus equivalencias entre variables:

Fuente	**Suma de cuadrados**	**Media cuadrática**	**Cociente F**
Regresión	SSR	MSR = SSR / *n*	F = MSR / MSE
Residuales	SSE	MSE = SSE / *l*	

Fig. 4. Análisis de la Varianza (ANOVA).

Dónde MSE (Mean Square Error) es la media del cuadrado debido al error de los residuales y MSR (Mean Square Regression) es la media del cuadrado debido a la

regresión. El factor F es el cociente entre MSR y MSE y es la prueba de significación final en un análisis ANOVA.
MSE representa la medición de comparación más común utilizada en los ajustes de regresión y es la que normalmente utilizaremos en los cálculos siguientes a realizar.

2.2 Regresión no lineal

Si la relación no es lineal, pueden transformarse los valores de una o ambas variables para intentar linealizarla. Si no es posible convertir la relación en lineal, puede comprobarse el grado de ajuste de una función polinomial más compleja. La función polinomial más sencilla es la cuadrática $y = w_0 + w_1 x_1 + w_2 x_2^2$ que describe una parábola, pero puede usarse una función cúbica u otra de un orden aun mayor (orden *k*) capaz de conseguir un ajuste casi perfecto a los datos.

$$\hat{y}_i = X_i\, w = w_o + w_1 x_i + w_2 x_i^2 + \ldots + w_k x_i^k$$
$$\text{para } X_i = \left(1,\ x_i,\ x_i^2, \ldots,\ x_i^k\right)$$

Las fronteras de decisión no lineales permiten representar conceptos más complejos al ajustarse más a los datos, no obstante este sobreajuste implica también inconvenientes, instancias de entrenamiento ruidosas (outliers) son también sobreajustadas, desplazando estas fronteras hacia esas instancias equivocadas y ocasionando así confundir al sistema de predicción a la hora de predecir nuevas entradas **[Zhang, 2009]**. Este sobreajuste (overfitting) es un problema muy común y produce un modelo que no es capaz de generalizar. Normalmente, fronteras de decisión muy complejas producen sobreajuste, no funcionando adecuadamente con nuevas instancias.
La regresión lineal suele conseguir fronteras de decisión más correctas y menos artificiales que la regresión no lineal. A pesar de producir mayores errores con los ejemplos de entrenamiento, tiene mayor capacidad de generalización y se comporta mejor ante nuevos ejemplos a predecir.

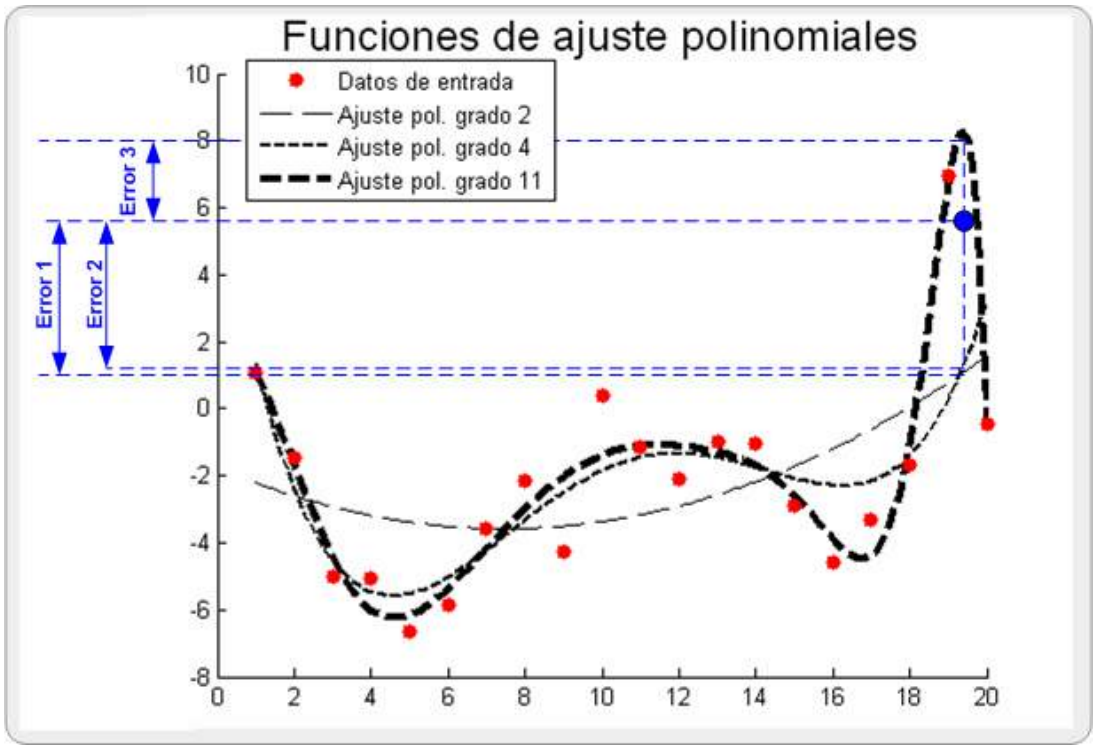

Fig. 5. Funciones de ajuste polinomiales y sobreajuste.

En la (Fig. 5) podemos observar el efecto de una regresión no lineal sobreajustada. Con un polinomio de grado 11 conseguimos ajustar muy bien los datos de entrenamiento. No obstante la presencia de un outlier en los datos de entrenamiento originó el desplazamiento de la frontera de decisión hacia dicho punto, aproximando una nueva instancia de prueba (el punto azul) como menor error en un polinomio de grado 11 cuando el ajuste con los polinomios de grado 2 y de grado 4 le asignan un error mucho mayor que cualquiera de los datos de entrenamiento. Cuanto mayor es el grado del polinomio mejor ajustaremos nuestros datos de entrenamiento, pero tenemos que estar plenamente seguros que nuestros datos de partida no tienen errores, cosa que prácticamente es difícil de encontrar en la práctica cuando se trabajan con datos reales suministrados por los sistemas de adquisición que conllevan errores implícitos no solo en sus sistemas de medida sino en las interferencias externas a las que están expuestos.

2.3 Colinealidad entre variables independientes

Una de las principales premisas a tener en cuenta en el modelado de regresión es que las variables independientes no posean ningún tipo de dependencia lineal entre ellas. Cuando una variable independiente posee alta correlación con otra ú otras ó puede ser explicada como una combinación lineal de alguna de ellas, se dice que el conjunto de datos presenta el fenómeno denominado multicolinealidad **[García, 2006]**.
Según **[Akdeniz, 2001]**, cuando se emplean los mínimos cuadrados ordinarios en la estimación de los parámetros de regresión y existe el problema de multicolinealidad en las variables independientes, se pueden observar problemas de inestabilidad de los mismos, signos incorrectos en los parámetros y frecuentemente elevados errores estándar, lo que conduce a generar modelos con muy poco poder explicativo o de difícil interpretación. Éste fenómeno debe ser investigado antes de generar un modelo de regresión, ya que puede generar errores en los pronósticos y dificultar la interpretación de la importancia de cada una de las variables independientes en el modelo.

2.3.1 Principales técnicas de detección

Las principales técnicas para poder detectar estas colinealidades son las siguientes:

2.3.1.1 Diagramas de dispersión

Si se representa cada par de variables independientes $(x_{i,1},x_{i,2})$ … $(x_{i,1},x_{i,n})$, en unos ejes cartesianos diferentes para cada par, obtendremos tantos diagramas de dispersión o nube de puntos como n-1 características o variables independientes existan para una única variable $x_{i,1}$. Con esta representación conseguiremos visualizar la variable independiente $x_{i,1}$ con respecto a todas las demás variables independientes. De esa forma, podremos obtener una primera idea acerca de la forma estructural que toma esta variable $x_{i,1}$ respecto a las demás y si realmente existe una relación morfológica entre ellas. Esa nube de puntos entre variables la podemos clasificar bien sea como una dependencia funcional perfecta o una dependencia estocástica con un cierto grado de dependencia **[GEA, 2006]**.

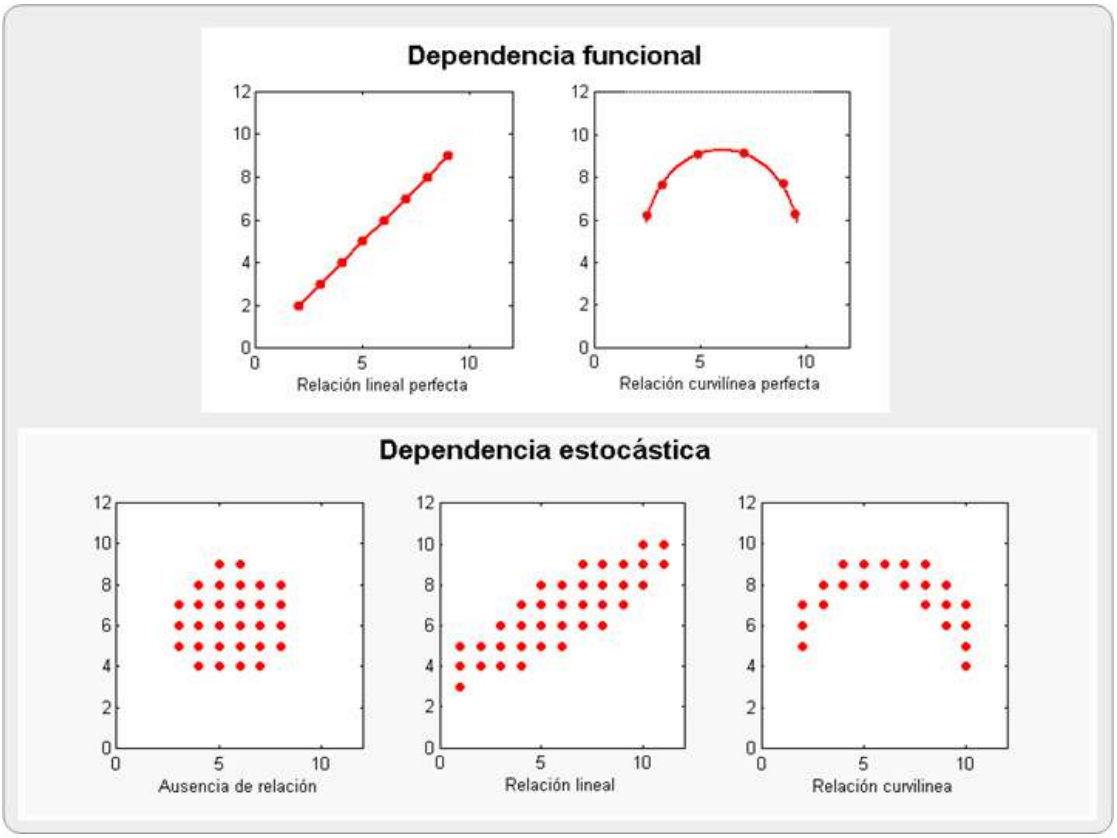

Fig. 6. Diagramas de dispersión.

En el primer caso (Fig. 6) tenemos una dependencia funcional completa y la relación atiende matemáticamente a una expresión del tipo $x_{i,1} = f(x_{i,2})$ sin ningún margen de error y morfológicamente el ajuste puede ser perfectamente lineal o perfectamente no lineal (curvilíneo o polinomial). Esto provoca que la Matriz $X^T X$ tenga determinante 0, y sea singular (no invertible) y en consecuencia no podríamos obtener el estimador MCO. Sin embargo, lo que suele ocurrir casi siempre, es que no se consigue un ajuste tan sumamente perfecto y entonces hablamos de dependencia estocástica entre variables con un determinado grado de relación (la no correlación de dos variables es un proceso idílico, que sólo se podría encontrar en condiciones de laboratorio), las relaciones y las dependencias entre las variables suelen ser menos rigurosas y aunque las tendencias estructurales también suelen ser lineales o no lineales siempre suele existir un error implícito en el ajuste para cada valor que toman las variables tratadas con respecto a su valor real.

Aunque la colinealidad existente entre dos variables independientes no sea exactamente perfecta pero sí casi perfecta, provoca que su determinante sea casi singular y su inversa sea casi infinito, o por lo menos un valor muy elevado que origine que los coeficientes MCO resultantes sean también muy elevados. En esta situación surgen problemas de precisión en la estimación de los coeficientes, ya que los algoritmos de inversión de matrices pierden precisión al tener que dividir por un número muy pequeño, siendo además inestables.

$$\left|X^T X\right| \simeq 0 \Rightarrow \frac{1}{\left|X^T X\right|} \approx \infty$$

Si se trata de buscar alguna relación entre variables independientes es preferible que exista una total ausencia de relación entre ellas o por lo menos una relación no muy alta, cada una de las variables debe aportar por sí misma poder explicativo hacia la variable dependiente y no tener que ser función de ninguna de las variables independientes.

2.3.1.2 Método del factor de inflación de la varianza

Según **[Wang, 1994]**, la principal consecuencia de las altas colinealidades entre las variables independientes es la siguiente.
En un modelo de dos variables, el error estándar de los coeficientes estimados es muy grande; esto es debido a que al coeficiente de variación tiene un factor de la forma $1/(1-r^2)$ denominado FIV (factor de inflación de la varianza), donde r es el coeficiente de correlación de Pearson $r = S_{xy}/\sigma_x\sigma_y$ (un índice que mide la relación lineal entre dos variables aleatorias cuantitativas y su valor está en el intervalo [-1,1]), S_{xy} es la

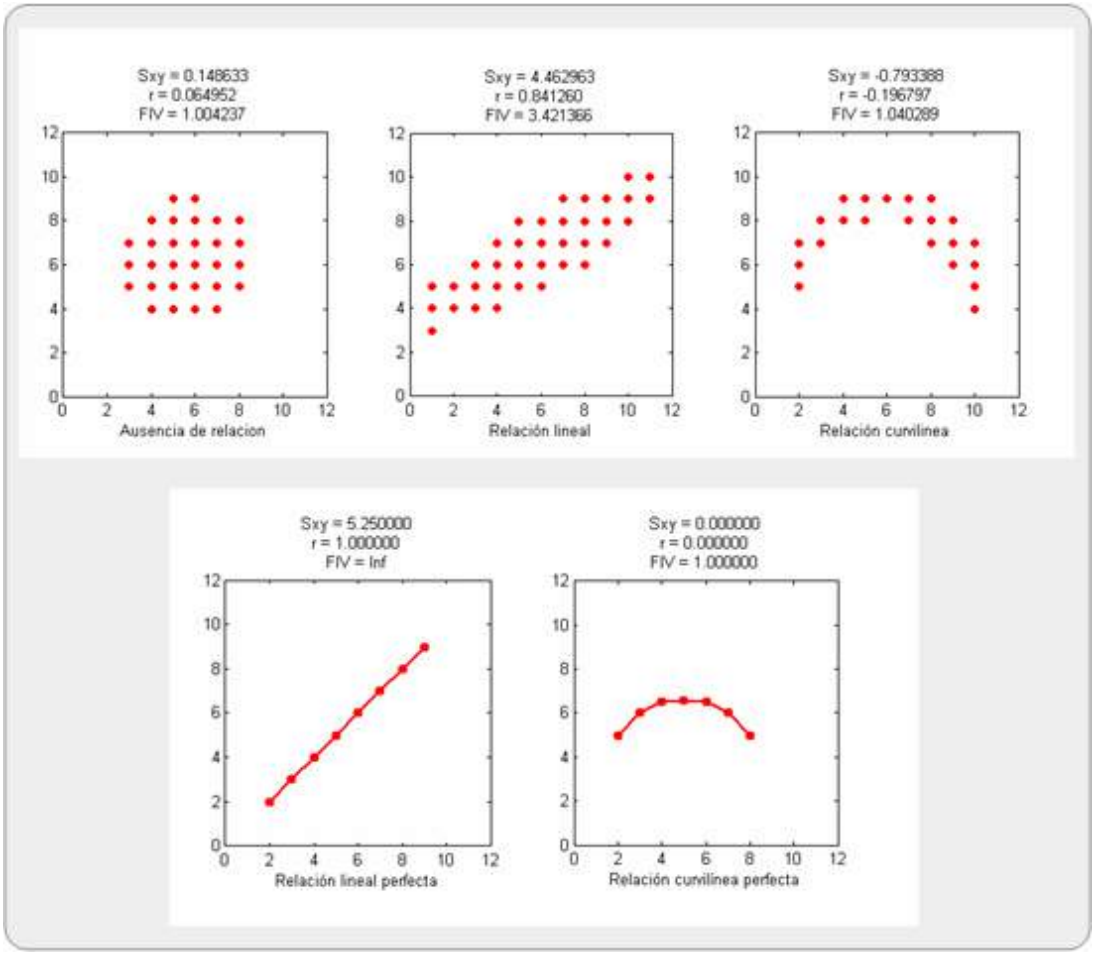

Fig. 7. Factor de inflación de la varianza.

covarianza de las dos variables, y σ_x, σ_y las desviaciones típicas de las distribuciones marginales. El signo de esta covarianza nos determinará el tipo de pendiente de la relación lineal (pendiente positiva o negativa).
A diferencia de la covarianza,

$$S_{xy} = \frac{\sum_{i=1}^{l} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

cuyo término expresa la variación o dispersión conjunta de dos variables *x* e *y* que tienen la misma escala de medida, la correlación de Pearson tiene aún más valor añadido debido a que es independiente de la escala de medida de las variables. La medida más comúnmente utilizada para medir el ajuste de la recta de regresión es este coeficiente de correlación (también se le conoce como medida de bondad de ajuste). Cuando r=0 no existe colinealidad, las variables independientes son ortogonales y su FIV es igual a 1 (pero esto no necesariamente implica que las variables sean independientes, pueden existir todavía relaciones no lineales entre las dos variables). A medida que el valor de *r* se incrementa en valor absoluto, es decir, existe una correlación negativa o positiva entre las variables, el FIV también se incrementa, ya que el denominador tiende a cero a medida que *r* tiende a uno (correlación perfecta). Algunos autores recomiendan que los FIV sean menores a 10, de lo contrario se concluye que existe multicolinealidad. En la Fig. 7 podemos observar como el FIV es igual a la unidad cuando no existe ninguna relación ó cuando la relación existente es no lineal (curvilinea).

2.3.1.3 Matriz de correlaciones

Una forma muy práctica de determinar el grado de colinealidad es la construcción de una matriz de correlación. Las variables se colocan en filas y en columnas y sus intercepciones deben presentar el coeficiente de regresión lineal de Pearson. Inicialmente se puede construir también una matriz de correlación con la covarianza en sus intercepciones, no obstante como ya se comentó anteriormente no suele ser de gran utilidad cuando las variables tienen diferente escala de medida. Asimismo, es de gran utilidad la construcción de una tercera matriz con los diagramas de dispersión de los datos para comprobar visualmente la lejanía o cercanía de dichos datos sobre la tendencia lineal que llegaran a mostrar (Fig. 8). **[Mason, 1991]** recomienda que sea eliminada una de las variables que tenga un coeficiente de correlación mayor a 0.8 con otras.

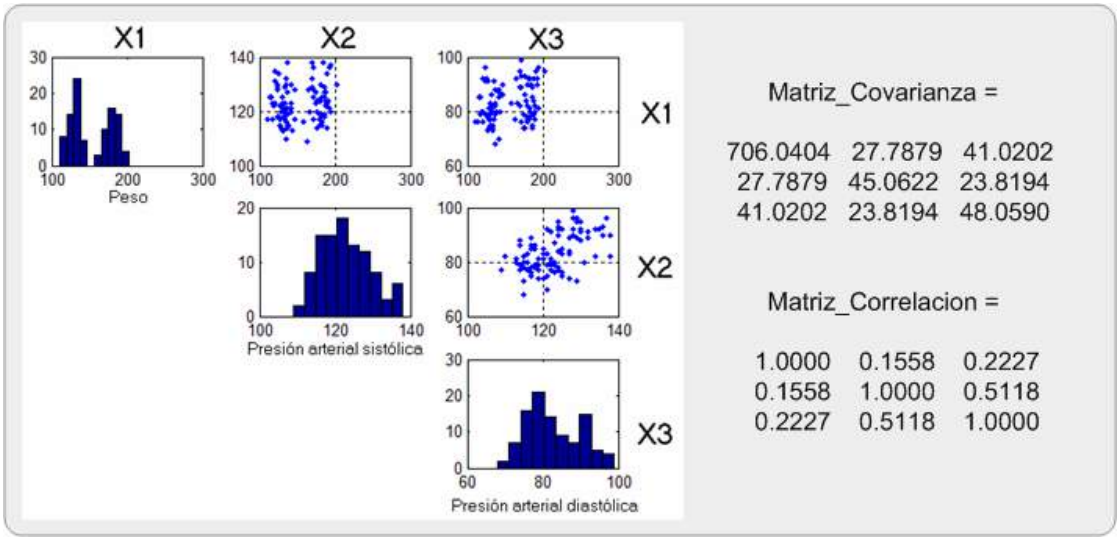

Fig. 8. Matriz de correlación.

2.3.1.4 Análisis del autosistema

También conocido como Análisis de Componentes Principales (ACP). Es una técnica proveniente del análisis exploratorio de datos cuyo objetivo es la síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante una tabla de datos con muchas variables (Fig. 9), el objetivo será reducirlas a un menor número de variables transformadas perdiendo la menor cantidad de información posible. Esta aproximación se basa en el hecho de que cualquier conjunto de *n* variables (X_1, ..., X_n) pueden ser transformadas a un conjunto de *n* variables ortogonales (y por tanto independientes entre sí, sin ninguna relación). Las nuevas variables ortogonales son conocidas como componentes principales (C_1, ..., C_n). Cada variable C_j es una combinación lineal de las variables $\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n$ (las variables originales normalizadas) de la forma:

$$\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & & \vdots \\ X_{l1} & X_{l2} & \dots & X_{ln} \end{bmatrix} \rightarrow \begin{bmatrix} C_{11} & C_{12} & \dots & C_{1n} \\ C_{21} & C_{22} & \dots & C_{2n} \\ \vdots & \vdots & & \vdots \\ C_{l1} & C_{l2} & \dots & C_{ln} \end{bmatrix}$$

100% de la información 80% 16% 0.02%

Fig. 9. Transformación de las variables originales en componentes.

$$C_j = v_{1j}\tilde{X}_1 + v_{2j}\tilde{X}_2 + \dots + v_{nj}\tilde{X}_n, \quad j = 1, \dots, n$$

Estos nuevos componentes principales o factores son calculados como una combinación lineal de las variables originales normalizadas, y además serán linealmente independientes entre sí. Técnicamente, el ACP busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados y construye una transformación lineal que escoge un nuevo sistema de coordenadas para el conjunto original de datos en el cual la varianza de mayor tamaño del conjunto de

datos es capturada en el primer eje (llamado el Primer Componente Principal), la segunda varianza más grande es el segundo eje, y así sucesivamente. La elección de los factores se realiza de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente. Del total de factores se elegirán aquéllos que recojan el porcentaje de variabilidad que se considere suficiente. A éstos se les denominará componentes principales.
La matriz de correlación de los componentes principales resultantes es de la forma:

$$\begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \dots & C_n \end{array} \\ \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} \end{array}$$

Los elementos que no están en la diagonal son ceros debido a que los componentes principales son ortogonales. Los elementos que están en la diagonal se conocen con el sobrenombre de *eigenvalues* o autovalores, de tal forma que cada autovalor λ_j es la varianza de cada variable ortogonal C_j, y cumple la propiedad $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$, debido a que el primer componente principal tiene la varianza más grande y el último componente principal la varianza más pequeña. Los coeficientes involucrados en la creación de cada C_j son conocidos como *eigenvectors* o autovectores y están asociados con el j-ésimo autovalor λ_j.

Para construir esta transformación lineal debe construirse primero la matriz de coeficientes de correlación. Debido a la simetría de esta matriz existe una base completa de vectores propios de la misma. La transformación que lleva de las antiguas coordenadas a las coordenadas de la nueva base es precisamente la transformación lineal necesaria para reducir la dimensionalidad de datos. Además las coordenadas en la nueva base dan la composición en factores subyacentes de los datos iniciales. Una de las ventajas del ACP para reducir la dimensionalidad de un grupo de datos, es que retiene aquellas características del conjunto de datos que contribuyen más a su varianza.

La función de Matlab *pcacov* nos devuelve el ACP a partir de la matriz de correlación (para datos normalizados) o a partir de la matriz de covarianza para datos no escalados.
Si aplicamos esta función dándole como entrada la matriz de correlación generada en el apartado anterior obtenemos los resultados mostrados en la Fig.10.

```
>> [eigen_Vectors,eigen_Values,porcentaje_varianza_total] = pcacov(Matriz_Correlacion)

eigen_Vectors =

   -0.3938    0.9132   -0.1051
   -0.6383   -0.3540   -0.6836
   -0.6614   -0.2021    0.7223


eigen_Values =

    1.6265
    0.8903
    0.4832


porcentaje_varianza_total =

   54.2173
   29.6776
   16.1051
```

Fig. 10. ACP a partir de la Matriz de correlación.

Los componentes principales correspondientes a los datos originales X_1, X_2, X_3 obtenidos a partir de la matriz de correlación anterior son:

$$C_1 = -0.3938\tilde{X}_1 - 0.6383\tilde{X}_2 - 0.6614\tilde{X}_3$$
$$C_2 = 0.9132\tilde{X}_1 - 0.3540\tilde{X}_2 - 0.2021\tilde{X}_3$$
$$C_1 = -0.1051\tilde{X}_1 - 0.6836\tilde{X}_2 + 0.7223\tilde{X}_3$$
$$54.22\% \qquad 29.68\% \qquad 16.11\%$$

Observamos también que la variable $\tilde{X}_1$ es la que más contribuye a la varianza total con un 54.22% de ella y por tanto es el componente principal del nuevo conjunto de variables, seguida de la variables $\tilde{X}_2$ con un 29.68%.

La matriz de correlación correspondiente a estas nuevas variables es:

$$\begin{pmatrix} 1.6265 & 0 & 0 \\ 0 & 0.8903 & 0 \\ 0 & 0 & 0.4832 \end{pmatrix}$$

Según **[Chatterjee, 2006]** , si alguno de los λ, son exactamente igual a cero existe una relación perfectamente lineal entre las variables originales y por tanto es un caso extremo de colinealidad. Si uno de los autovalores es mucho más pequeño que los demás (y cercano a cero), la colinealidad también se hace presente pero en menor grado. En la matriz de correlación de los componentes principales podemos observar como el menor valor de λ no está muy cerca de cero pero si es mucho menor que los otros dos, sobre todo del mayor autovalor, lo que indica algo de colinealidad existe entre las variables X_2 y X_3.

Si quisiéramos obtener los componentes principales a partir de la matriz de datos original sin tener que calcular las matrices de covarianzas y de correlaciones utilizaremos la función de Matlab *princomp.* En el ejemplo siguiente aplicamos dicha función a los datos que presentamos en el apartado 2.1.4.2 y en concreto a los datos que representaban una relación estocástica lineal no perfecta, obteniendo los siguientes resultados.

```
>> [eigenVectors,observaciones_en_espacio_ACP,eigenValues] = princomp(zscore([x1;x2]'))

eigenVectors =

    0.7071    0.7071
    0.7071   -0.7071

eigenValues =

    1.8413
    0.1587
```

Fig. 11. ACP a partir de las variables originales.

Que corresponden a las ecuaciones transformadas:

$$C_1 = 0.7071\tilde{X}_1 + 0.7071\tilde{X}_2$$
$$C_2 = 0.7071\tilde{X}_1 - 0.7071\tilde{X}_2$$

y a la matriz de correlación:

$$\begin{pmatrix} 1.8413 & 0 \\ 0 & 0.1587 \end{pmatrix}$$

En dicha matriz podemos observar como $\lambda_2 = 0.16$ es un valor muy próximo a 0 y muy distante del primer autovalor, lo cual denota que existe colinealidad como ya sabíamos previamente.
Además en la matriz *observaciones_en_espacio_ACP* obtenemos los coeficientes principales (C_1,C_2) para cada punto correspondiente con el de las variables originales (X_1,X_2). Si generamos un diagrama de dispersión tanto para las variables originales como para los coeficientes principales obtenemos la siguiente figura:

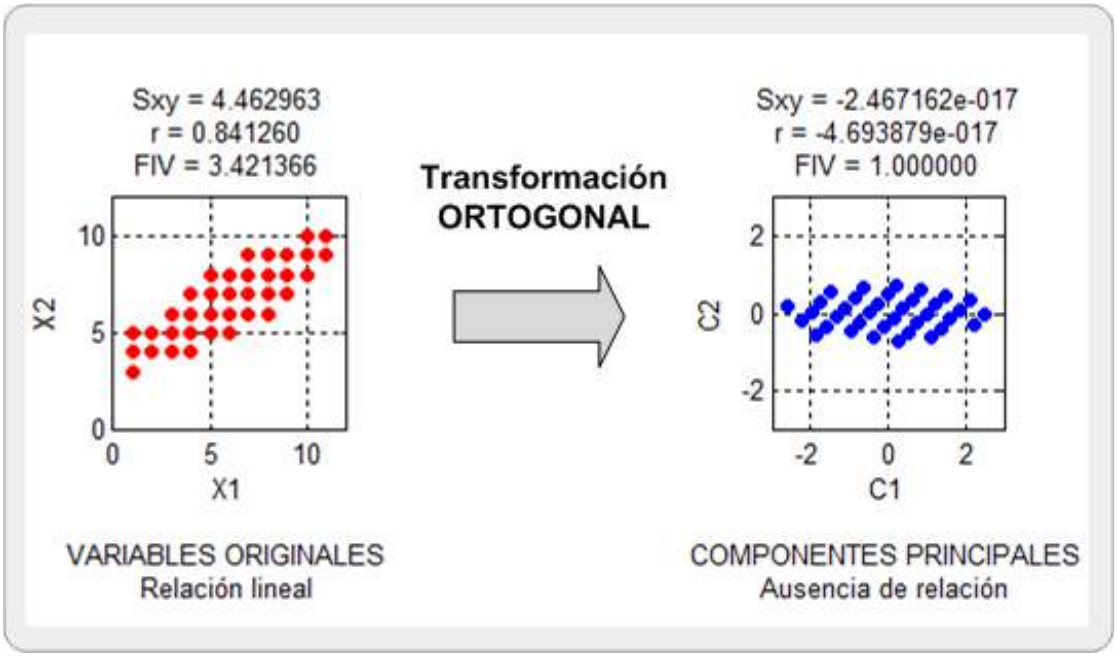

Fig. 12. Transformación ortogonal de datos originales.

En la Fig. 12 podemos observar como a partir de unas variables con una relación bastante lineal las podemos transformar en otras variables con ausencia de toda relación entre ellas, reflejado en el valor del factor de inflación de la varianza que es igual a la unidad.

[Belsley, 1980] propuso un índice denominado número de condición η, el cual está basado en la relación entre el máximo autovalor de la matriz de correlación y el mínimo, tal como se indica a continuación:

$$\eta = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

El número de condición siempre será más grande de 1. Para valores de $\eta < 2.26$ puede ser ignorado, para valores $2.26 < \eta < 3.16$ existe una colinealidad débil. Para valores $3.16 < \eta < 5.48$ se califica como moderada, para $5.48 < \eta < 10$ se considera fuerte y para $\eta > 10$ se considera muy fuerte.

Si calculamos el número de condición para los dos últimos ejemplos que hemos mostrado en el estudio de componentes principales obtenemos $\eta = 1.83$ y $\eta = 3.41$. Lo cual indica en el primer caso que la colinealidad existente puede ser despreciable y que para el segundo caso tenemos una colinealidad moderada.

2.3.2 Técnicas de corrección

Se han planteado técnicas y algoritmos para corregir la colinealidad en los datos; sin embargo, algunos procedimientos funcionan en un modelo, mientras que en otros no.

2.3.2.1 Eliminación de variables del análisis

Es la solución más cómoda ya que únicamente hay que eliminar aquellos predictores correlacionados con otros a partir de una detección previa de ellos. Los estimadores que resultan tienen una varianza de error menor. Este enfoque es aceptado por ser reduccionista y simplificar el modelo, sin embargo reduce el rango de la matriz de información de variables independientes y esto lo puede convertir en una técnica que genere un modelo con menor poder explicativo ante nuevas entradas.

2.3.2.2 Componentes principales

El análisis de componentes principales visto anteriormente no solo sirve como método para conocer si una variable independiente está correlacionada con otra u otras variables independientes. El espacio ortogonal de variables transformadas cumple la condición de que son independientes entre sí y por tanto carecen de colinealidad entre ellas. Por tanto se puede trabajar en este espacio con dichas variables utilizando MCO con total seguridad de que no observaremos problemas de inestabilidad en los coeficientes obtenidos, signos incorrectos en dichos coeficientes, ni elevados errores estándar en el ajuste.

2.3.2.3 La técnica *"Ridge Regression"*

Cuando las variables predictoras están muy correlacionadas, los coeficientes de regresión resultantes de un ajuste por MCO pueden llegar a ser muy erráticos e imprecisos, debido a los efectos desastrosos que la multicolinealidad tiene sobre su varianza. Estos coeficientes originan predicciones erróneas a la hora de vaticinar nuevas respuestas correspondientes a entradas similares que deberían pronosticar salidas similares. Esto es así, como hemos visto, debido a la inversión de la matriz singular X^TX (singular debido a las colinealidades). Afortunadamente, la técnica *Ridge Regression* (RR) **[Hoerl y Kennard, 1970]**, es un método que trata estas colinealidades minimizando el problema al contraer los coeficientes *w* de MCO, logrando coeficientes ajustados con menor varianza, dando estabilidad así a la predicción del modelo y solucionando dicho problema. La matriz X^TX es reemplazada por otra matriz numéricamente más estable debido a la agregación (suma) de un sesgo con la finalidad de reducir el error estándar de éstos (Fig. 13) **[Shawe-Taylor, 2004]**.

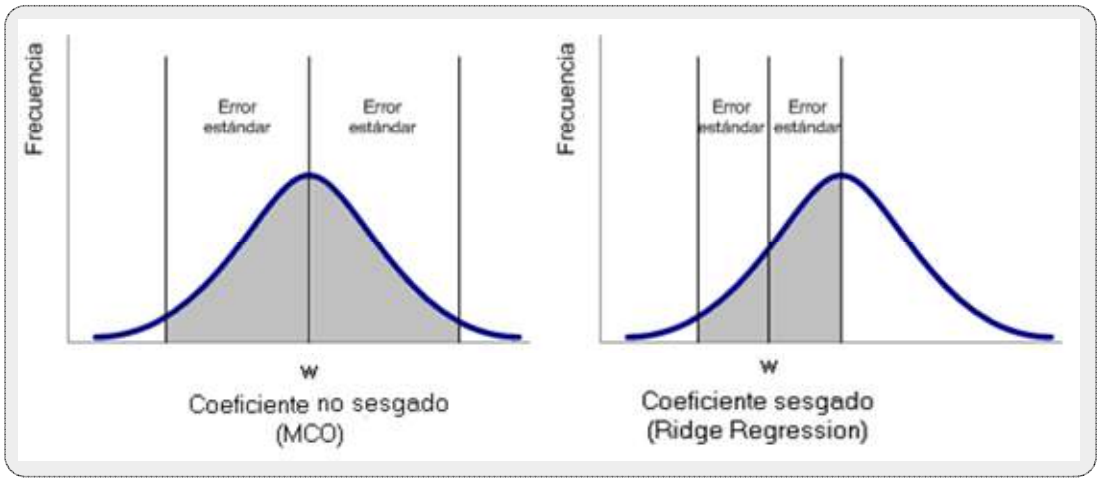

Fig. 13. Agregación de un sesgo a MCO.

Si procedemos de esa forma a partir de la fórmula (1.1) que define el método de MCO, el procedimiento RR no es más que una ligera modificación (adicción de un término constante a cada coeficiente o factor de regularización *k*) de dicha ecuación:

$$F(w) = k\|w\|^2 + \sum_i (y_i - X_i w)^2 \qquad (1.1)$$

2.4 Exploración de regresión sesgada

A la técnica RR también se le conoce como 'regresión de cresta' o 'regresión sesgada'. Veamos las dos modalidades de cómputo con las que contamos para poder realizar este tipo de regresión.

2.4.1 Primera solución

Encontrar la función en la cual la suma de los cuadrados de las diferencias junto con el sesgo para los valores observados y esperados sea menor, corresponderá a encontrar los coeficientes de regresión *w* para los cuales la función por la cual determinamos dicho error, sea un error mínimo, o dicho de otra forma, corresponde a diferenciar la ecuación (1.2) en *w*.

$$\frac{\partial F}{\partial w} = 0 \quad \Rightarrow \quad \frac{\partial}{\partial w}(k\|w\|^2) + \sum_i \frac{\partial}{\partial w}(y_i - X_i w)^2 = 0$$

$$\Rightarrow \quad 2kw + \sum_i 2X_i^T (y_i - X_i w) = 0$$

$$\Rightarrow \quad \left(\sum_i X_i^T X_i\right) w \; + \; kw = \sum_i X_i^T y_i$$

$$\Rightarrow \quad \left(X^T X + kI_n\right) w = X^T y \quad \Rightarrow$$

$$\boxed{w = \left(X^T X + kI_n\right)^{-1} X^T y}$$

I_n corresponde a la matriz identidad de dimensiones (*n* x *n*) y como podemos observar la matriz $\left(X^T X + k\,\mathrm{I}_n\right)^{-1}$ es siempre invertible si $k > 0$. Como veremos más adelante, sabemos que existe un k (de hecho, un intervalo de valores de *k*), mejorando el error del estimador MCO. El inconveniente reside en la elección de *k* que no debe ser de modo intuitivo, ya que si este valor es muy grande, se produce una *sobre-regularización* **[Ramos, 2007]**, la cual puede originar pérdida de información importante, y si *k* resulta pequeño, se produce una *sub-regularización*, que puede provocar que la solución no sea robusta, es decir, que sea sensible a errores en los datos (*k*=0 supone volver a un estimador MCO). Los procedimientos o técnicas de elección de este factor de regularización se discutirán más adelante.
Al igual que ocurría con el método de MCO donde *w* es función lineal del vector de la variable respuesta dependiente (*y*), solucionar la ecuación anterior para los coeficientes *w* implica entonces solucionar un sistema de ecuaciones lineales con *n* ecuaciones y *n* incógnitas. Por tanto, la complejidad computacional de esta tarea resulta (n^3) operaciones. Una vez que tenemos los coeficientes de regresión *w*, la función de predicción de un nuevo vector de entrada *x* será,

$$\hat{y}(x) = xw = \sum_{i=1}^{n} w_i (x)_i$$

con complejidad computacional (*n*) operaciones.

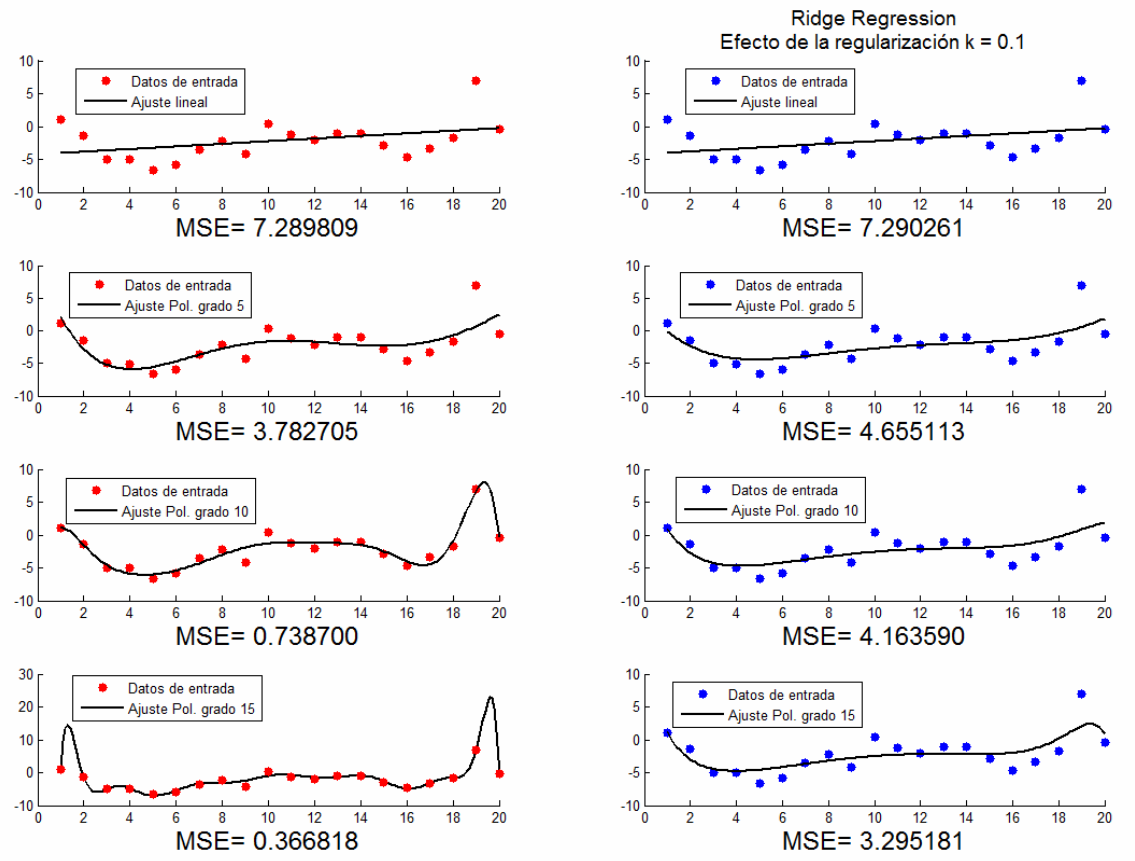

Fig. 14. Efecto de la regularización.

En la (Fig. 14) se puede observar el efecto de regularización que provoca la regresión sesgada sobre la regresión clásica, independientemente de utilizar un ajuste lineal o no lineal. Efectivamente, el ajuste mediante regresión por MCO consigue el menor error de ajuste frente a RR, lógico debido a que solamente tenemos una única variable independiente para la variable explicativa y por lo tanto no existen colinealidades y como comentamos en apartados anteriores, RR mejora en términos del error del ajuste cuando existen variables independientes correlacionadas entre sí, cuando esto no es así, MCO es el mejor ajuste con el mínimo error que se puede realizar.
No obstante e independientemente de que en este ejemplo no se pueda distinguir perfectamente toda la fortaleza de RR, sí podemos observar como la varianza global del error que se produce para los dos tipos de regresión es menor en el ejemplo de RR que en el ejemplo de la regresión clásica, independientemente incluso del orden del polinomio que utilicemos para hacer el ajuste. Esto quiere decir que RR juega un papel muy importante a la hora de regularizar y homogeneizar el ajuste final, haciéndolo más robusto, menos variante y por tanto más sensible a errores en los datos y posibles *outliers* que se pudieran presentar.

Los efectos de una mala elección del factor *k*, se discuten en los ejemplos siguientes. Cuando escogemos un factor de *k* muy grande producimos una sobre-regularización con una varianza global del error casi inapreciable porque prácticamente e

independientemente del ajuste que realicemos, los datos siempre se ajustarán a una línea horizontal.

Cuando seleccionamos un factor de *k* muy pequeño, perdemos robustez, el error obtenido para cada tipo de ajuste es más variable, pero nos acercamos otra vez al ajuste de MCO y por tanto con sensibilidad a errores en los datos y a posibles efectos perjudiciales si las variables independientes están correlacionadas.

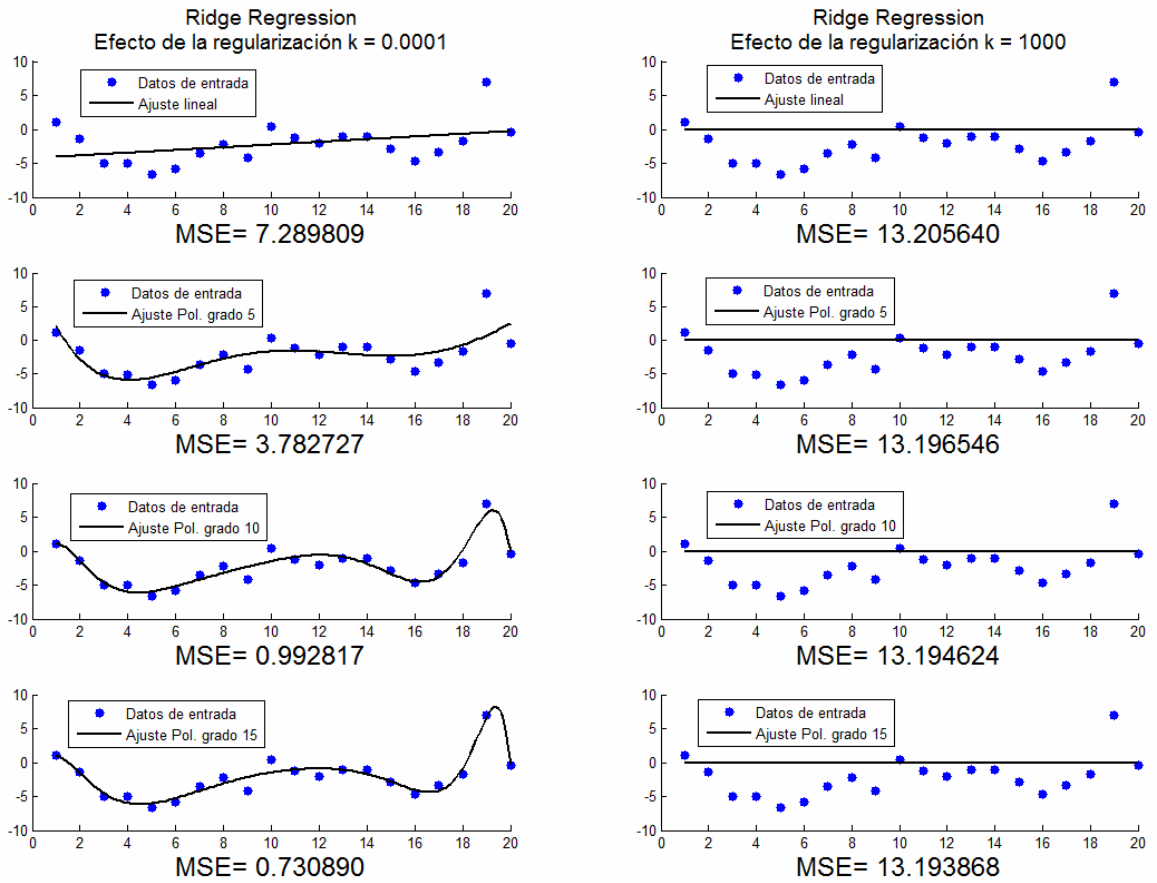

Fig. 15. Sub-regularización y sobre-regularización.

2.4.2 Solución dual

A partir de la solución anterior para los coeficientes de regresión *w*, podemos deducir lo siguiente:

$$w = \left(X^T X + kI_n \right)^{-1} X^T y$$

$$\left(X^T X + kI_n \right) w = X^T y$$

$$X^T Xw + kw = X^T y$$

$$kw = X^T y - X^T Xw = X^T (y - Xw)$$

$$w = k^{-1} X^T (y - Xw) = X^T \alpha$$

Donde el término α matemáticamente significa:

$$\alpha = k^{-1}(y - Xw)$$
$$\alpha k = y - Xw$$
$$\alpha k = y - XX^T\alpha$$
$$\alpha k + XX^T\alpha = y$$
$$y = \alpha(kI_l + XX^T)$$
$$\alpha = (XX^T + kI_l)^{-1}y$$
$$\alpha = (G + kI_l)^{-1}y$$

La matriz $G = XX^T$ se le conoce como "*Gram matrix*". Esta matriz G y la matriz $(G + kI_l)$ tiene dimensiones (l x l). Los parámetros α son conocidos como "*dual variables*" o variables duales y resolver α implica resolver l ecuaciones lineales con l incógnitas, una tarea de complejidad (l^3), como se muestra en la función de predicción a partir de estas variables, que viene dada por:

$$\hat{y} = Xw = XX^T\alpha = XX^T(G + kI_l)^{-1}y$$

Para predecir un nuevo punto o vector x, implica complejidad computacional (nl), ya que los coeficientes w son una combinación lineal de los puntos de entrenamiento X^T.

$$w = X^T\alpha$$
$$w = \sum_{i=1}^{l} \alpha_i x_i$$
$$\hat{y}(x) = x\ w = x\ \sum_{i=1}^{l} \alpha_i x_i = \sum_{i=1}^{l} \alpha_i \left(\sum_{j=1}^{n} (x_i)_j (x)_j \right)$$

Si la dimensión n del espacio de características es mayor que el número l de ejemplos de entrenamiento, es mejor y más eficiente resolver el sistema por este segundo método (*dual*) en vez del primer método (*primal*) ya que éste último implica resolver la matriz $(X^TX + kI_n)$, que es de dimensiones (n x n). La evaluación de la función predictiva es, sin embargo, siempre más costosa la solución *dual*, debido a que comporta (nl) operaciones, frente a (n) operaciones que conlleva la primera solución.

2.4.3 La técnica *"Kernel Ridge Regression"*

Si los datos de entrenamiento (las variables independientes) muestran relaciones no lineales, las técnicas de regresión anteriores serán incapaces de modelarlas adecuadamente con un error mínimo aceptable (el sesgo introducido en RR ayuda pero a veces también resulta insuficiente). Sin embargo, una solución no lineal puede ser tratada y formulada moviéndonos a un espacio de características lineales a partir del espacio de entrada no lineal. *Kernel Ridge Regression* (KRR) es una técnica que

encuentra y realiza un mapeo de los datos de entrada (considerados no lineales) en un espacio de características de más alta dimensión (donde corresponden a un modelo aproximadamente lineal) obteniendo errores de ajuste mucho menores que los conseguidos en el espacio de entrada inicial, y conservando la eficiencia del factor de regularización *k* utilizada en la técnica RR.
La idea básica de KRR consiste en realizar un mapeo de los datos de entrenamiento x $\in$ X, a un espacio de mayor dimensión F a través de un mapeo no lineal $\Phi(x) : X \rightarrow F$, donde podemos realizar una regresión lineal.

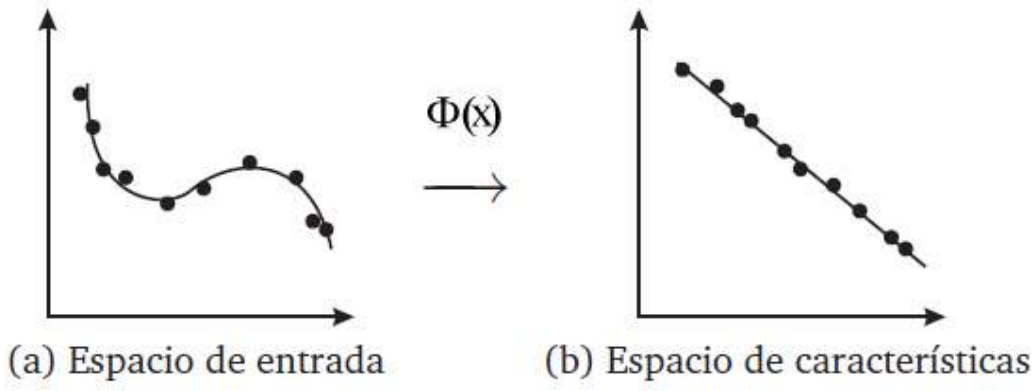

Fig. 16. Idea básica de los métodos Kernel.

A partir de esto, la matriz *G* utilizada en la versión dual de la técnica RR se transforma en una matriz o kernel *K* de productos escalares para valores transformados de *X*. Esta es la ventaja de la aproximación dual de RR, se puede reemplazar la matriz *G* mediante cualquier matriz kernelizada *K*, para el caso que nos ocupa de un kernel lineal:

$$G = XX^T \rightarrow K = \phi(X)\phi(X)^T$$

Dicho kernel K sigue manteniendo dimensiones (*l* x *l*), y por tanto complejidad operacional (l^3).

Para el cálculo de los coeficientes de regresión *w* se procederá como sigue:

$$w = \phi(X)^T \alpha$$
$$w = \phi(X)^T (K + kI_l)^{-1} y$$

Y la función de predicción resultante a partir de estos coeficientes *w* quedaría:

$$\hat{y} = \phi(X)w = \phi(X)\phi(X)^T (K + kI_l)^{-1} y$$
$$\hat{y} = z(K + kI_l)^{-1} y$$

Es de señalar que si utilizamos un kernel lineal, entonces *z* = *K*, por lo que esto correspondería a utilizar una solución dual de RR, no obstante podemos probar y jugar con diferentes kernels *K* (polinomial, función de base radial, tangente hiperbólica, etc.) junto con diferentes parámetros de regularización *k* con el objetivo de encontrar el

mejor modelo explicativo en ese espacio de características y poder aplicarlo posteriormente a las aproximaciones a realizar para nuevos ejemplos de entrada.

En la predicción de un nuevo punto ϕ(x) se sigue conservando la misma complejidad de cómputo (*nl*) que el conseguido mediante la técnica RR, como se muestra a continuación:

$$\hat{y}(\phi(x)) = \phi(x)w = \phi(x)\sum_{i=1}^{l}\alpha_i\phi(x_i) = \sum_{i=1}^{l}\alpha_i\left(\sum_{j=1}^{n}(\phi(x_i))_j(\phi(x))_j\right)$$

En la figura siguiente (Fig. 17) se puede observar un ajuste de regresión utilizando un kernel RBF (Radial Basis Function) de tipo Gaussiano, de forma que:

$$K = \exp\left(-\frac{\|x-u\|^2}{\sigma}\right)$$

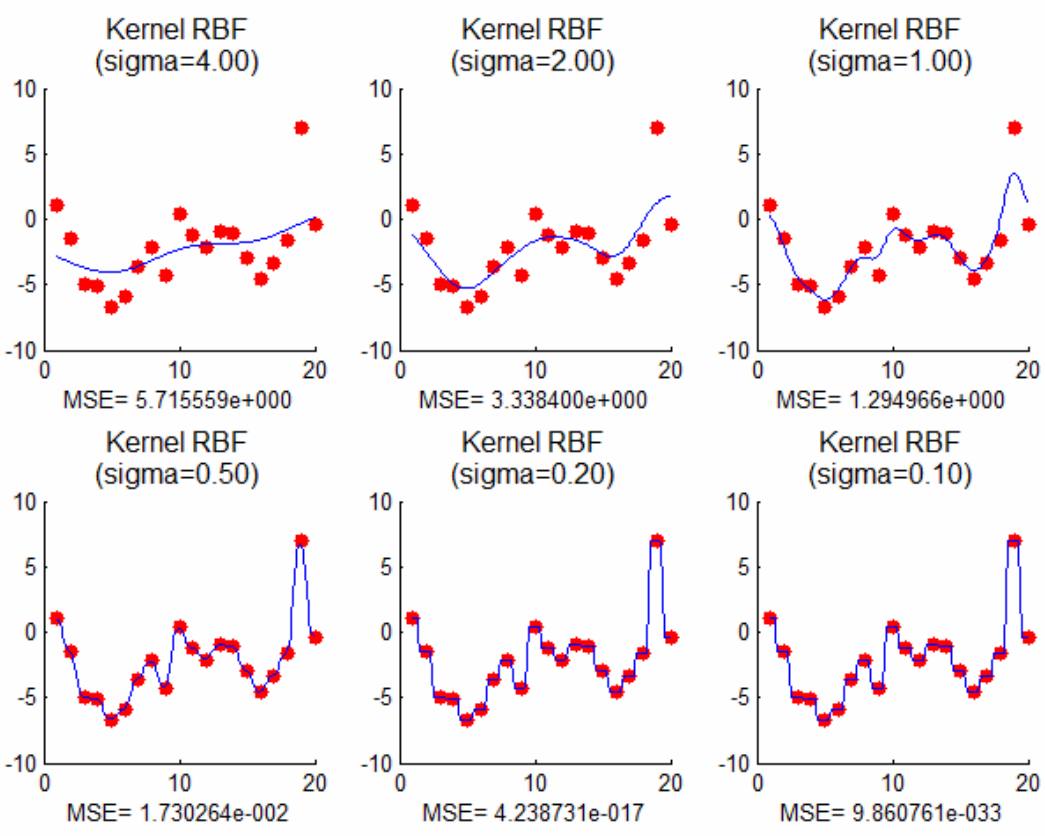

Fig. 17. Regresión con kernel RBF-Gaussiano para diferentes valores de sigma.

Modificando el valor de la dispersión σ en la función Gaussiana, se puede observar como podemos alcanzar un ajuste casi perfecto (MSE ≈ 0) sobre los datos de entrenamiento (para valores de σ inferiores a 0.2).

2.4.4 Estandarización de datos para la regresión sesgada.

Si ajustamos un modelo del tipo:

$$Y = w_0 + w_1 X_1 + \ldots + w_n X_n$$

Necesitaremos centrar (ya que aparece un término constante) y/o escalar las variables que integran dicha ecuación. Una variable centrada se obtiene restando a cada observación la media de todas las observaciones para cada variable. Por ejemplo la variable respuesta centrada $(Y - \overline{y})$ y la variable predictora j-ésima centrada $(X_j - \overline{x}_j)$.

Las variables centradas también pueden ser escaladas, existiendo dos tipos principales de escalado en los datos, el *escalado de longitud unidad* y la *estandarización*. Generalmente, tanto el escalado de longitud unidad como la estandarización (escalado mediante la desviación estándar) se utiliza, como veremos en el apartado siguiente, para poder comparar los coeficientes *w* entre sí (en la misma escala) para diferentes valores de *k*. El centrado, ayuda a agrupar los datos y por ello disminuir la dispersión de los mismos con efectos beneficiosos en la reducción del error del ajuste para aproximar nuevos datos de prueba. No está demasiado claro y por tanto es una fuente de controversia, según **[Pasha, 2004]**, que sea necesario estandarizar las variables a la hora de realizar ajustes de regresión. No se trata que las variables sean esencialmente similares en sus rangos (da igual que un conjunto de variables de temperatura estén en ^{o}C o en ^{o}F) sino más bien que sean independientes, no correlacionadas y con bastante poder explicativo.

Un modelo de ecuación de regresión en términos de variables estandarizadas es del tipo:

$$\tilde{Y} = \theta_1 \tilde{X}_1 + \ldots + \theta_n \tilde{X}_n$$

De tal forma que a cada variable original de datos $\tilde{X}_j$, $\tilde{Y}$ le corresponde una transformación por estandarización de media cero y desviación estándar la unidad:

$$X_j \rightarrow \tilde{X}_j = \frac{X_j - \overline{x}_j}{\sigma_j}$$

$$Y \rightarrow \tilde{Y} = \frac{Y - \overline{y}}{\sigma_Y}$$

y donde σ_j y σ_Y son respectivamente:

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}{n-1}} \qquad \sigma_Y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}}$$

Procediendo a despejar dichas transformaciones en la ecuación de variables transformadas tendremos:

$$\tilde{Y} = \theta_1 \tilde{X}_1 + \ldots + \theta_n \tilde{X}_n$$

$$\frac{Y - \overline{y}}{\sigma_Y} = \theta_1 \frac{X_1 - \overline{x}_1}{\sigma_1} + \ldots + \theta_n \frac{X_n - \overline{x}_n}{\sigma_n}$$

$$Y = \overline{y} + \theta_1 \frac{\sigma_Y (X_1 - \overline{x}_1)}{\sigma_1} + \ldots + \theta_n \frac{\sigma_Y (X_n - \overline{x}_n)}{\sigma_n}$$

$$Y = \overline{y} + \frac{\theta_1 \sigma_Y X_1 - \theta_1 \sigma_Y \overline{x}_1}{\sigma_1} + \ldots + \frac{\theta_n \sigma_Y X_n - \theta_n \sigma_Y \overline{x}_n}{\sigma_n}$$

$$Y = \overline{y} + \frac{\theta_1 \sigma_Y X_1}{\sigma_1} - \frac{\theta_1 \sigma_Y \overline{x}_1}{\sigma_1} + \ldots + \frac{\theta_n \sigma_Y X_n}{\sigma_n} - \frac{\theta_n \sigma_Y \overline{x}_n}{\sigma_n}$$

$$Y = \overline{y} - \sum_{j=1}^{n} \frac{\theta_j \sigma_Y \overline{x}_j}{\sigma_j} + \frac{\theta_1 \sigma_Y X_1}{\sigma_1} + \ldots + \frac{\theta_n \sigma_Y X_n}{\sigma_n}$$

resultando,

$$Y = w_0 + w_1 X_1 + \ldots + w_n X_n$$

para todo,

$$w_j = \left(\sigma_Y \Big/ \sigma_j \right) \theta_j$$

$$w_0 = \overline{y} - \sum_{j=1}^{n} w_j \overline{x}_j$$

Si la normalización que utilizamos es el escalado de longitud unidad, el modelo de ecuación de regresión en términos de estas variables transformadas será del tipo:

$$\tilde{Z}_y = \theta_1 \tilde{Z}_1 + \ldots + \theta_n \tilde{Z}_n$$

De tal forma que a cada variable original de datos $\tilde{Z}_j$, $\tilde{Z}_y$ le corresponde una transformación de media cero y longitudes la unidad según:

$$X_j \rightarrow \tilde{Z}_j = \frac{X_j - \overline{x}_j}{L_j}$$

$$Y \rightarrow \tilde{Z}_y = \frac{Y - \overline{y}}{L_y}$$

y donde L_j y L_y son respectivamente:

$$L_j = \sqrt{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2} \qquad L_y = \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Como se indica en la formulación anterior, la cantidad L_y se refiere a la longitud de la variable centrada $Y - \overline{y}$. Similarmente, L_j mide la longitud de la variable $X_j - \overline{x}_j$.

Procediendo a despejar dichas transformaciones en la ecuación de variables transformadas de la misma forma que hicimos con la estandarización, tendremos:

$$\tilde{Z}_y = \theta_1 \tilde{Z}_1 + \ldots + \theta_n \tilde{Z}_n$$

$$\frac{Y - \overline{y}}{L_y} = \theta_1 \frac{X_1 - \overline{x}_1}{L_1} + \ldots + \theta_n \frac{X_n - \overline{x}_n}{L_n}$$

...

resultando,

$$Y = w_0 + w_1 X_1 + \ldots + w_n X_n$$

para todo,

$$w_j = \left(L_y \Big/ L_j \right) \theta_j$$

$$w_0 = \overline{y} - \sum_{j=1}^{n} w_j \overline{x}_j$$

Es obvio que si solamente deseamos centrar los datos, nuestras variables originales quedarían de la siguiente manera:

$$Y = w_0 + w_1 X_1 + \ldots + w_n X_n \text{ para todo,}$$

$$w_j = \theta_j$$

$$w_0 = \overline{y} - \sum_{j=1}^{n} w_j \overline{x}_j$$

2.4.5 Ejemplo de aplicación mediante regresión múltiple

Veamos algún ejemplo donde pongamos en práctica la formulación anterior. Para ello, hacemos uso de una base de datos llamada Aqua-all.txt obtenida desde la dirección web: http://www.rpi.edu/~bennek/class/mds/Aqua-all.txt, que es una versión reducida de variables (solamente 525 variables independientes), a su vez extraída de la dirección web: http://www.pharmacy.arizona.edu/outreach/aquasol/ y donde se almacena una extensa recopilación y un gran repositorio de datos con información que tratan temas farmacológicos de solubilidad en agua para compuestos orgánicos.
Nuestra matriz de datos original se compone de 525 variables descriptoras independientes que definen una variable respuesta dependiente, para un total de 197 registros u observaciones. Es de señalar el elevado número de dimensiones con los que se va a trabajar, a priori no sabemos si esas variables tienen alguna correlación entre ellas, no obstante como vamos a utilizar la técnica RR mitigamos cualquier efecto perjudicial que estas correlaciones pudieran tener sobre los resultados.
Separaremos 100 registros para definir el conjunto de entrenamiento y el resto (97 registros) para definir el conjunto de prueba o de test.

Veamos que sucede si no realizamos ninguna transformación de los datos originales.

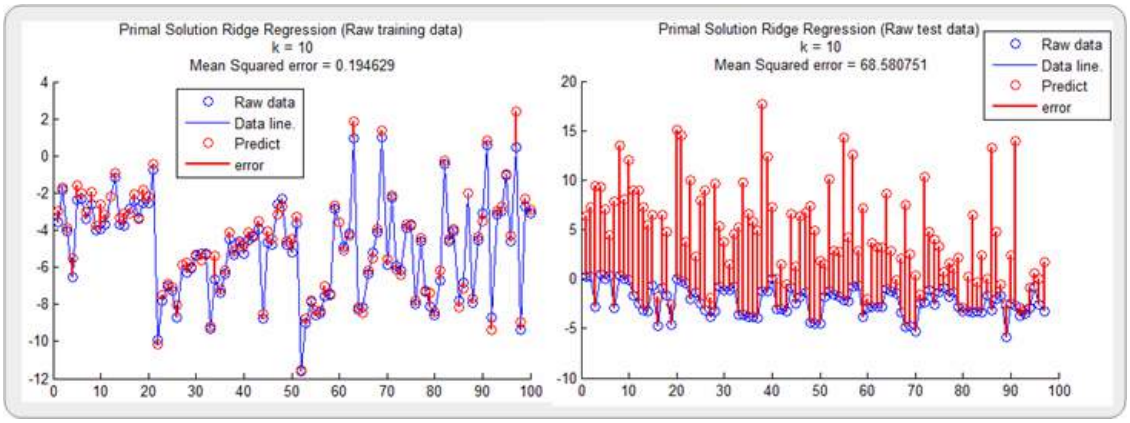

Fig. 18. *Ridge Regression* (Primera solución) con datos sin normalizar.

Como observamos en la Fig. 18, el ajuste para los datos de entrenamiento mediante la primera solución expuesta en pasos anteriores para RR (coeficientes $w = \left(X^T X + kI_n \right)^{-1} X^T y$ y función predictiva $\hat{Y} = Xw$) parece comportarse bastante bien, pero no es tan óptimo cuando intentamos aproximar las 97 observaciones del conjunto de validación, obteniendo aquí en términos de MSE un valor muy alto.
Procediendo como lo discutido en el apartado de la normalización de datos, el modelo puede ser mejorado añadiéndole un término independiente a la ecuación y por tanto transformando *X* e *Y* en otras variables, resultado de substraer el valor de sus medias.

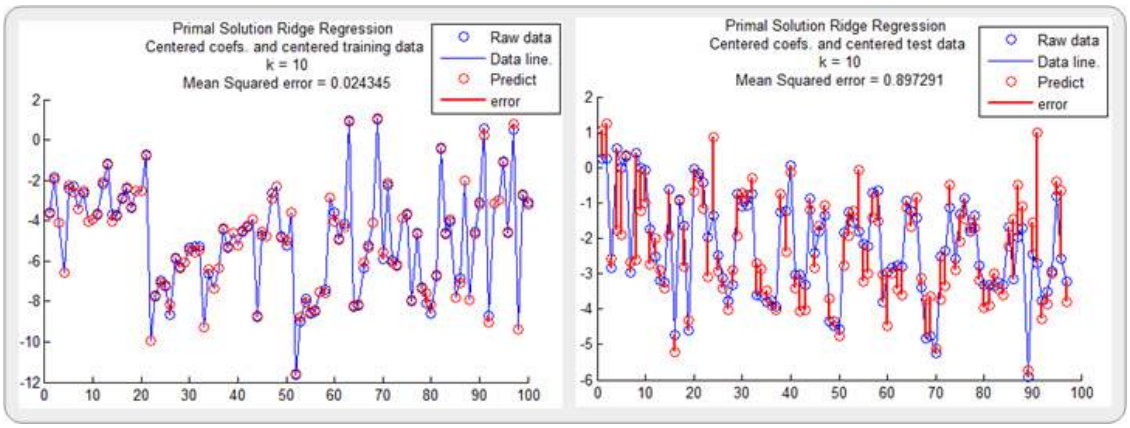

Fig. 19. *Ridge Regression* (Primera solución) con datos centrados.

Observamos en la Fig. 19, como tanto para los datos de entrenamiento como para el conjunto de validación se ha conseguido reducir drásticamente el valor del error en el ajuste, incluso manteniendo el mismo factor de regularización *k*. La centralización de los datos origina una agrupación de los mismos en torno a su media con lo que disminuye su dispersión mejorando el ajuste de mínimos cuadrados. Cuando trabajamos con coeficientes normalizados podemos definir nuevas predicciones trabajando con estos coeficientes, pero los datos también tienen que estar procesados (centrados sobre su media) de la forma `Ypred2=Xtest2*w2 + b;` donde b resulta el

término independiente (en este caso es igual a la media de la variable dependiente original *y*). Estos mismos resultados pueden ser obtenidos si de-normalizamos los coeficientes procesados y trabajamos con las variables originales, de la forma `Ypred3=Xtest*w2 + (b-mean(X)*w2)`.

Otra comprobación importante que podemos realizar es el cálculo del tiempo de cómputo al utilizar la primera solución de RR frente a la versión dual de dicha técnica.
En el primer caso, calculamos el tiempo empleado en obtener la matriz de coeficientes *w* a partir de la matriz `Xtrain2'*Xtrain2` de dimensiones (525x525) de la siguiente manera:

```
% Model: Primal solution with bias
time1 = cputime;
w2 = inv(Xtrain2'*Xtrain2+ L*I)*(Xtrain2'*Ytrain2);
elapsedTime1 = cputime - time1
```

Procedemos de la misma manera para el cálculo de los alfas y w's mediante la versión dual (*G* tiene dimensiones 100 x 100):

```
% Model: Dual solution with bias
time2 = cputime;
% Gram matrix
G = Xtrain2*Xtrain2';
% Dual variables
alpha = inv(G+L*I2)*Ytrain2;
w3 = Xtrain2'*alpha;
elapsedTime2 = cputime - time2
```

La tabla de tiempos en 2 ordenadores diferentes es la siguiente:

	Laptop Medion Akoya Intel Atom 1.6 GHz 1Gb RAM	PC Lenovo ThinkStation Intel Core i5 3.33 GHz 8 Gb RAM
`elapsedTime1`	1.0156	0.0936
`elapsedTime2`	4.687500e-002	0

Como se puede observar, al ser el número de dimensiones mucho mayor que el número de observaciones (n >> l), resulta más eficiente computacionalmente hablando utilizar la versión dual de RR para el cálculo de los coeficientes de regresión.

Pongamos ahora algún ejemplo con la técnica KRR. La fortaleza de esta técnica de regresión es la posibilidad de utilizar funciones Kernel que nos permiten construir una función de regresión lineal en un espacio de características de más alta dimensión (lo que equivale a una regresión no lineal en el espacio de entrada).
Utilicemos un kernel polinomial de grado 2 de la forma:

$$K(x, y) = ((x \bullet y) + 1)^2$$

Como se puede observar en la Fig. 20, el uso de un kernel polinomial mejora notablemente el ajuste sobre los datos de entrenamiento. Pero este sobreajuste impide

generalizar bien sobre los datos de validación, obteniendo peores resultados que los conseguidos con RR para el mismo factor de regularización k = 10.

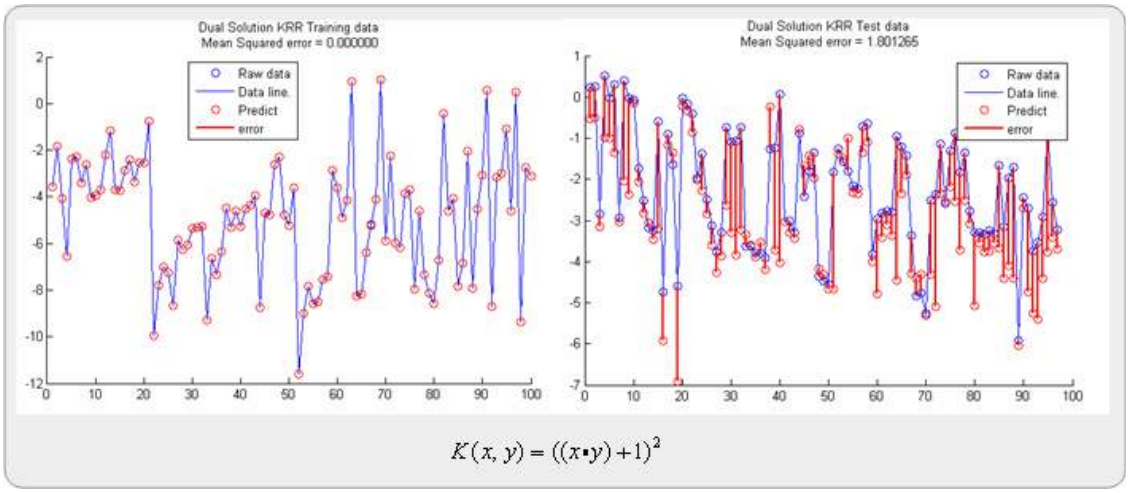

Fig. 20. *Kernel Ridge Regression* (polinomial grado 2) con datos centrados.

Probemos ahora con un kernel de tipo sigmoide o también conocido como tangente hiperbólica.

$$K(x, y) = \tanh(\eta(x \bullet y) + c)$$

El kernel mediante la tangente hiperbólica se conoce también como 'kernel sigmoide' o como 'kernel perceptron multicapa' y procede del campo de las redes neuronales.
Hay dos parámetros que son ajustables en esta función, el término η y la constante c. El valor que se le suele asignar a η es $1/n$, siendo n la dimensión de los datos que se están tratando **[Souza, 2010]**.

El resultado de ajustar mediante un kernel de tipo sigmoide, el ejemplo que estamos tratando con parámetros $\eta = 1/525$, c = 1 es el siguiente:

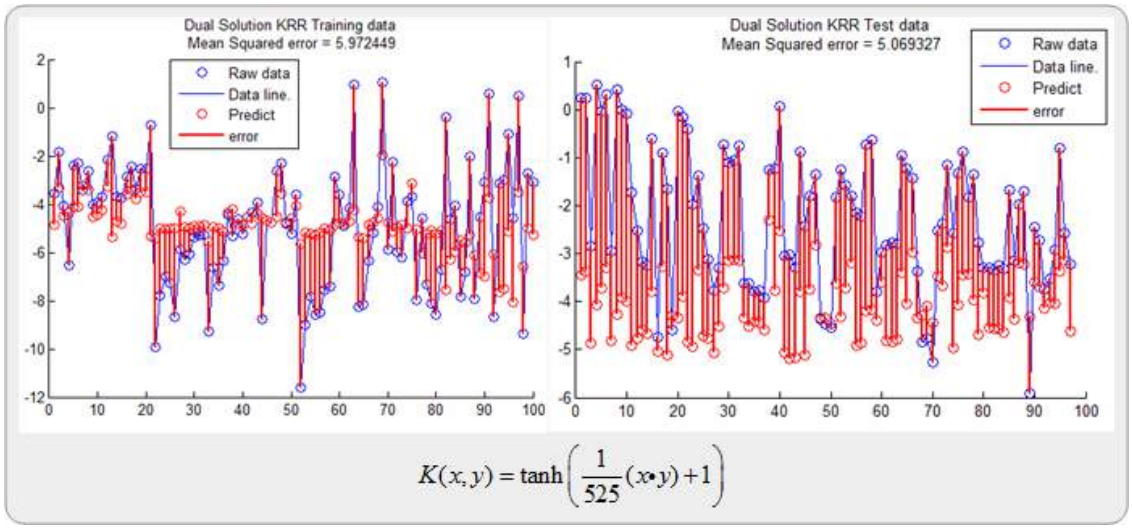

Fig. 21. *Kernel Ridge Regression* (sigmoide) con datos centrados.

Donde comprobamos (Fig. 21) que no generaliza bien ni para el conjunto de entrenamiento, ni para el conjunto de validación.

Todavía no hemos aprovechado toda la potencia que nos brindan las técnicas RR. Todos estos ejemplos los hemos calculado fijando el término del factor de regularización *k* a un valor arbitrario de 10. Evidentemente, si modificamos este valor, los resultados también se verán modificados.

2.4.6 Elección del factor de regularización

Sabemos que existe un factor de regularización *k* (de hecho, un intervalo de valores de *k*) mejorando el MSE del estimador MCO; pero nada en la discusión anterior nos permite decidir cuál es su valor. Al ser *k* un parámetro que introduce un sesgo en los estimadores, es deseable seleccionar el valor más pequeño de *k* por el cual se estabilizan los coeficientes de regresión. En la práctica, se recurre a alguna o varias de las siguientes soluciones **[Núñez, 2005]**:

2.4.6.1 Uso de trazas de regresión sesgada

Es una aproximación gráfica y por lo tanto debe ser vista como una técnica exploratoria de datos visual. Se prueban diversos valores de *k* representándose las diferentes estimaciones del vector de coeficientes *w* (trazas RR); se retiene entonces aquel valor de *k* a partir del cual se estabilizan las estimaciones. La idea es intuitivamente atrayente: pequeños incrementos de *k* partiendo de cero (MCO) tienen habitualmente un efecto drástico sobre *w*, al coste de introducir algún sesgo. Incrementaremos *k* por tanto hasta que parezca que su influencia sobre *w* se atenúa (hasta que las trazas RR sean casi horizontales). El decidir dónde ocurre esto es, no obstante, bastante subjetivo.

Siguiendo las recomendaciones del "statistics toolbox" de Matlab para su función *ridge*, a la hora de realizar trazas RR es conveniente utilizar los coeficientes de regresión normalizados o transformados θ en lugar de los correspondientes originales *w*, para que aparezcan gráficamente en la misma escala. No obstante, dependiendo de que normalización utilicemos, obtendremos unas trazas u otras, eso sí, todas para valores de *k* entre cero y uno, (0<=k<=1).

YEAR	IMPORT	DOPROD	STOCK	CONSUM
49	15.9	149.3	4.2	108.1
50	16.4	161.2	4.1	114.8
51	19.0	171.5	3.1	123.2
52	19.1	175.5	3.1	126.9
53	18.8	180.8	1.1	132.1
54	20.4	190.7	2.2	137.7
55	22.7	202.1	2.1	146.0
56	26.5	212.4	5.6	154.1
57	28.1	226.1	5.0	162.3
58	27.6	231.9	5.1	164.3
59	26.3	239.0	0.7	167.6
60	31.1	258.0	5.6	176.8
61	33.3	269.8	3.9	186.6
62	37.0	288.4	3.1	199.7
63	43.3	304.5	4.6	213.9
64	49.0	323.4	7.0	223.8
65	50.3	336.8	1.2	232.0
66	56.6	353.9	4.5	242.9

Fig. 22. Datos sobre la economía francesa.

Para mostrar el ejemplo de trazas RR haremos uso de un conjunto de datos extraídos desde **[Chatterjee, 2006]** sobre variables de producción y consumo de la economía francesa (Fig. 22), éstas son por orden, el año, las importaciones, la producción doméstica, los productos almacenados y el consumo doméstico.

Si realizamos las trazas RR con la variable IMPORT como variable dependiente, para los dos tipos de normalización explicados en apartados anteriores obtenemos las dos gráficas de la Fig. 23. Como se puede comprobar en la primera gráfica (los datos tienen media cero y desviación típica la unidad), las variables DOPROD y CONSUM mantienen una correlación entre ellas. Dicha correlación se estabiliza a medida que aumentamos el sesgo por medio del parámetro *k*. Visualmente podemos establecer dicha estabilización a partir de un valor de k = 0.2. Para valores superiores a 0.2 los coeficientes parecen mantener ya una constante bastante lineal y su varianza disminuye.
En la gráfica de la derecha, representamos los coeficientes normalizados utilizando el escalado de longitud unidad explicado también en apartados anteriores. Como se puede observar, los coeficientes se muestran estables para valores de *k* a partir de 0.04 - 0.05.

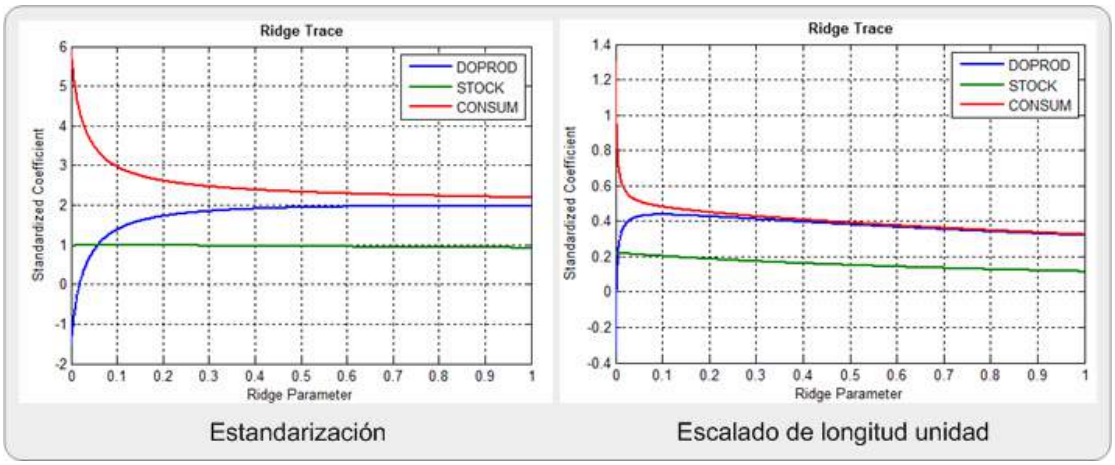

Fig. 23. Trazas RR para diferentes escalas.

2.4.6.2 Método del punto fijo

En el trabajo **[Hoerl, Kennard y Baldwin, 1975]**, se sugirió calcular matemáticamente el parámetro k de la forma:

$$k = \frac{n\sigma^2(0)}{\sum_{j=1}^{n}\left[\theta_j(0)\right]^2}$$

donde $\theta_1(0),\ldots,\theta_n(0)$ son los coeficientes de regresión transformados cuando k=0 (estimadores de MCO) y $\sigma^2_{(i)} = \frac{SSE_{(i)}}{l-n-2}$, la varianza de los residuales (errores), siendo $\sigma^2(0)$, la correspondiente varianza cuando el parámetro de regularización k=0.

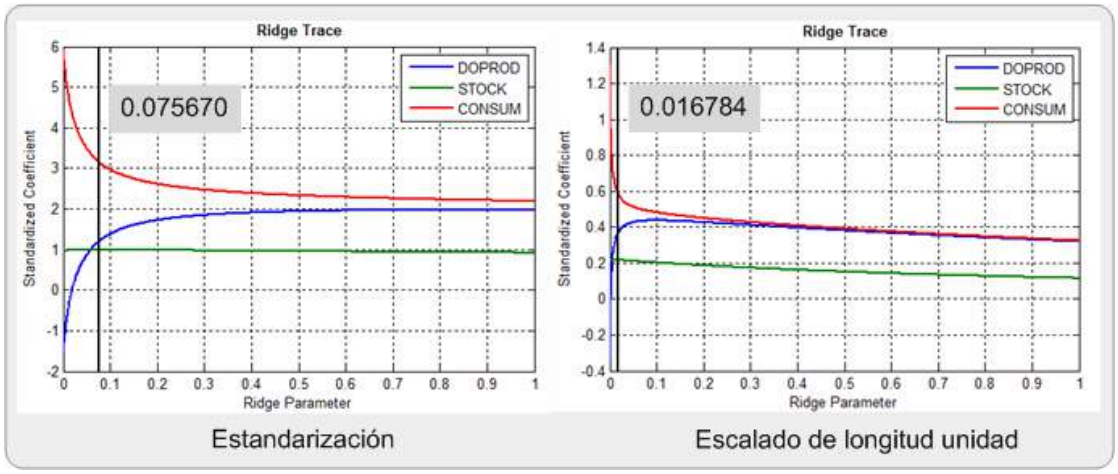

Fig. 24. Elección de *k* (método del punto fijo).

En la Fig. 24 podemos observar el punto de corte de la estimación del parámetro *k* con la traza de los coeficientes y en la tabla de la Fig. 25 observamos el valor de esos coeficientes para dichos puntos de corte junto con sus valores originales (de-normalizados), después de aplicar las correspondientes operaciones siguiendo las fórmulas del apartado 2.4.4.

	k=0.07567	k=0.016784
Variables normalizadas	Coeficientes Estandarizados	Coeficientes E. Long. Unidad
Constante	0	0
DOPROD	1.1980	0.3633
STOCK	0.9881	0.2166
CONSUM	3.1619	0.5919

Variables originales	Coeficientes originales	Coeficientes originales
Constante	-9.2697	-8.9983
DOPROD	0.0399	0.0550
STOCK	0.5991	0.5967
CONSUM	0.1532	0.1303

Fig. 25. Coeficientes de regresión para la variable IMPORT (método del punto fijo).

2.4.6.3 Método iterativo

Hoerl y Kennard un año después (1976) **[Hoerl y Kennard, 1976]**, propusieron un procedimiento repetitivo y más complejo para seleccionar el valor de *k*.

- ✓ Comenzar calculando k_0, siendo este valor el parámetro *k* que se obtiene haciendo uso del método anterior (método del punto fijo).
- ✓ Posteriormente, utilizar k_0 para calcular $k_1 = \dfrac{n\sigma^2(0)}{\sum_{j=1}^{n}\left[\theta_j(k_0)\right]^2}$
- ✓ Entonces, usar k_1 para calcular $k_2 = \dfrac{n\sigma^2(0)}{\sum_{j=1}^{n}\left[\theta_j(k_1)\right]^2}$
- ✓ Repetir este proceso hasta que $k_{j+1} \simeq k_j$, o sea, hasta que las diferencias encontradas para valores de *k* sucesivos sean casi despreciables.

Nuevamente en esta aproximación aparece la subjetividad de lo que se considera despreciable para las diferencias de *k* consecutivos, además se supone que a partir del *k* obtenido por el método del punto fijo, los valores de *k* serán muy parecidos debido a la influencia de la varianza en los residuales, que va a ser también muy similar.
Si aplicamos estos cálculos, tomando como condición de parada $k_{j+1} - k_j \leq 0.0001$, obtenemos los siguientes parámetros y coeficientes:

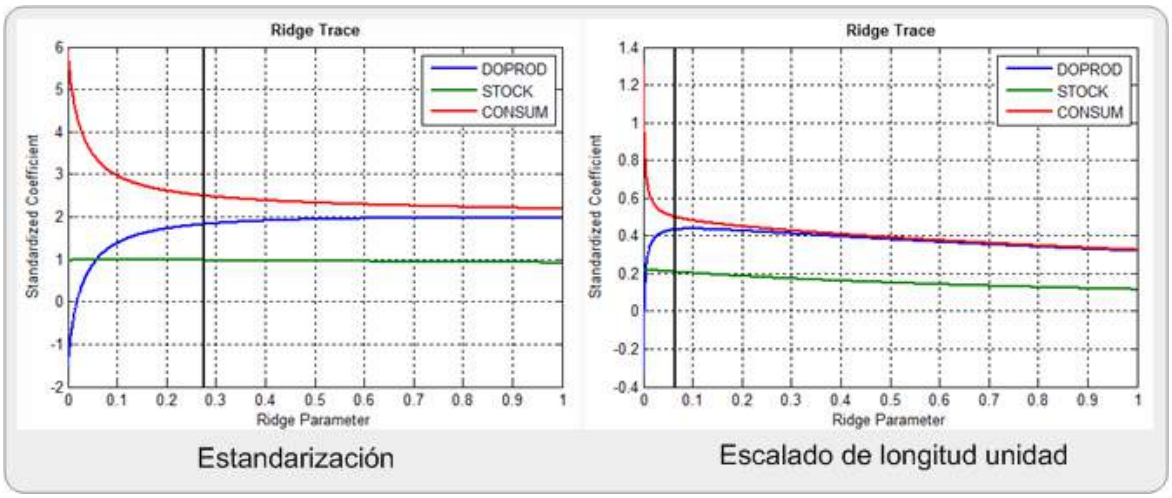

Fig. 26. Elección de *k* (método iterativo).

	k=0.276453	k=0.064777
Variables normalizadas	Coeficientes Estandarizados	Coeficientes E. Long. Unidad
Constante	0	0
DOPROD	1.8236	0.4329
STOCK	0.9762	0.2082
CONSUM	2.4935	0.5003

Variables originales	Coeficientes originales	Coeficientes originales
Constante	-8.7771	-8.1554
DOPROD	0.0608	0.0656
STOCK	0.5919	0.5736
CONSUM	0.1208	0.1102

Fig. 27. Coeficientes de regresión para la variable IMPORT (método iterativo).

2.4.6.4 Validación cruzada

La idea es también muy simple, aunque computacionalmente algo laborioso. Se estima el error de predicción dividiendo al azar el conjunto de datos en varias partes. En cada paso una de las partes se convierte en una muestra de prueba que sirve para validar el modelo y las restantes partes constituyen lo que es llamado una muestra de entrenamiento que sirve para construir el modelo.

Si por ejemplo, se usasen 10 partes, se llamaría una “10 fold cross-validation” , por lo general se usa 1 parte y en ese caso es llamado el método “leave-one-out” (dejar uno afuera).
Sea $\hat{y}_j^{[-i]}$ el valor predicho (la predicción que hacemos de la observación y_j) para la j-ésima observación usando una línea de regresión que ha sido estimada sin haber usado las observaciones de dicha parte.

El cálculo del error por validación cruzada usando *p* partes es:

$$CV_{(t)} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{l}(y_j - \hat{y}_j^{[-i]})^2}{p} \quad \text{para } t \text{ valores de } k$$

Entonces el mejor modelo (el mejor factor de regularización *k* por validación cruzada) es aquel *k* que tiene el error de validación cruzada promedio más pequeño:

$$k = \arg mín\, CV$$

En principio, calcular $CV_{(k)}$ para un valor de *k* requeriría llevar a cabo *l* regresiones, excluyendo cada vez una observación distinta.

3. PREDICCION DE SERIES TEMPORALES NO LINEALES

Denominamos predicción a la estimación de valores futuros de una variable en función del comportamiento pasado de la serie. Se trata de seguir la evolución de una variable con el fin de regular su resultado. La predicción en series temporales es una línea de investigación fundamental en la estadística. El hecho de poder reproducir el comportamiento de un sistema dinámico no lineal a partir de medidas discretas (series temporales) de sus variables posibilita la aplicación de los modelos de predicción basados en series temporales a innumerables campos del conocimiento, complementando la modelización física.

3.1 Precisión en la predicción de series temporales sometidas a ruidos en los datos

La estimación de mínimos cuadrados para modelos lineales es notoria por su falta de robustez frente a valores atípicos (*outliers*), como hemos comprobado en apartados anteriores. Si la distribución de los atípicos es asimétrica, los estimadores pueden estar sesgados y aunque las técnicas RR ayudan a corregir el error del ajuste, precisamente por la introducción de un sesgo, si los atípicos son muy pronunciados, en presencia de cualquier valor de estos atípicos, los estimadores mínimos cuadráticos son ineficientes y pueden serlo en extremo. No obstante, ¿qué ocurre si las variables a estudiar están sometidas a ruidos continuos en todo su recorrido temporal?. En la práctica, cuando se trabajan con datos reales suministrados por los sistemas de adquisición de datos, que a su vez son suministrados por los diagnósticos de medidas, conllevan errores implícitos no sólo en sus sistemas físicos de medida (que tienen una precisión o resolución mínima) sino en las interferencias externas a las que están expuestos dichos sistemas. Veamos que ocurre en estos casos.

3.2. Analítica predictiva en series temporales sometidas a ruido gaussiano continuo

Retomando el repositorio de datos analizado en el apartado 2.4.5, integrado por 197 observaciones y 525 variables descriptoras independientes que definen una variable única dependiente, se pretende analizar la precisión del error en el ajuste de esa variable continua dependiente en presencia de ruido gaussiano añadido a todas y cada una de las variables independientes que modelan dicha variable respuesta. Para ello se compararán los resultados obtenidos en presencia de dos tipos de intensidades de ruido gaussiano añadido con los resultados a obtener en ausencia de ruido (datos brutos originales).

3.2.1 Supuestos de partida para el análisis

- Ante el gran número de variables descriptoras independientes, el estudio de la colinealidad entre dichas variables se hace intratable una a una con todas las demás. Por

ello utilizaremos la técnica RR para obviar si existen relaciones lineales entre las 525 variables independientes. Como hemos demostrado en apartados anteriores, dicha técnica mitiga los efectos perjudiciales de las colinealidades mediante la introducción de un sesgo o factor de regularización.

- Al ser el número de dimensiones mucho más elevado que el número de observaciones (525>>197), utilizaremos la versión dual de RR para el cálculo de los coeficientes de regresión y para obtener el error del ajuste final, como hemos demostrado en apartados anteriores que es mucho más eficiente en términos de cálculo y de computación.

- Utilizaremos KRR porque no sabemos si los datos de entrenamiento muestran relaciones no lineales entre sus variables independientes. Dicha técnica, como también hemos visto, obtiene una solución más óptima al movernos a un espacio de características lineal a partir del espacio de entrada no lineal. Además utilizaremos diferentes funciones kernel (lineal, polinomial grado 2 y tangente hiperbólica), para comparar cual obtiene mejores resultados en la precisión del ajuste.

- En la elección del factor de regularización descartaremos las trazas RR debido también al elevado número de dimensiones del problema a tratar. Resultaría muy engorroso pintar 525 trazas de las variables para un intervalo de factores de regresión. Por ello utilizaremos la validación cruzada para obtener el factor de regularización más óptimo. En este caso el que obtenga la serie temporal más similar a una de referencia (el error del ajuste promedio más pequeño para un rango de factores de regularización).

3.2.2 Resultados finales obtenidos

En la tabla siguiente se adjuntan los resultados finales obtenidos. Como se puede comprobar el kernel lineal consigue mejores resultados para los tres conjuntos de datos (datos brutos, adicción de ruido gaussiano débil y adicción de ruido gaussiano más elevado).

		Datos originales		Adicción ruido gaussiano débil		Adicción ruido gaussiano elevado	
		MSE (datos de entrenamiento)	**MSE (datos de prueba)**	MSE (datos de entrenamiento)	**MSE (datos de prueba)**	MSE (datos de entrenamiento)	**MSE (datos de prueba)**
Kernel	Lineal	0.149	**0.509**	1.594	**1.970**	17.610	**12.156**
	Polinomial grado2	0.0	**1.801**	0.0	**3.7860**	2.550	**16.967**
	Tangente hiperb.	3.465	**3.219**	4.645	**3.416**	35.616	**12.631**

Hay que recalcar que aunque se consiguen errores en los ajustes casi nulos en los datos de entrenamiento utilizando el kernel polinomial tanto en los datos originales como en los datos con ruido gaussiano débil, al utilizar ese mismo modelo para la predicción de los datos de prueba, obtenemos peores resultados que con el kernel lineal. Esto es debido a que el kernel polinomial sobre ajusta excesivamente los datos de entrenamiento y el modelo obtenido no es capaz de generalizar bien para los datos de prueba.
El kernel mediante la tangente hiperbólica obtiene peores resultados, no obstante se observa que en los datos añadiendo elevado ruido gaussiano, se acercan los resultados a los obtenidos mediante el kernel lineal, siendo mejores y superando los conseguidos por el kernel polinomial para dicho caso.

En la Fig. 28 podemos observar como la predicción de la serie temporal del conjunto de prueba en los datos brutos originales, el error en el ajuste es casi mínimo, reproduciendo casi en su conjunto la serie temporal observada original de dicho conjunto.

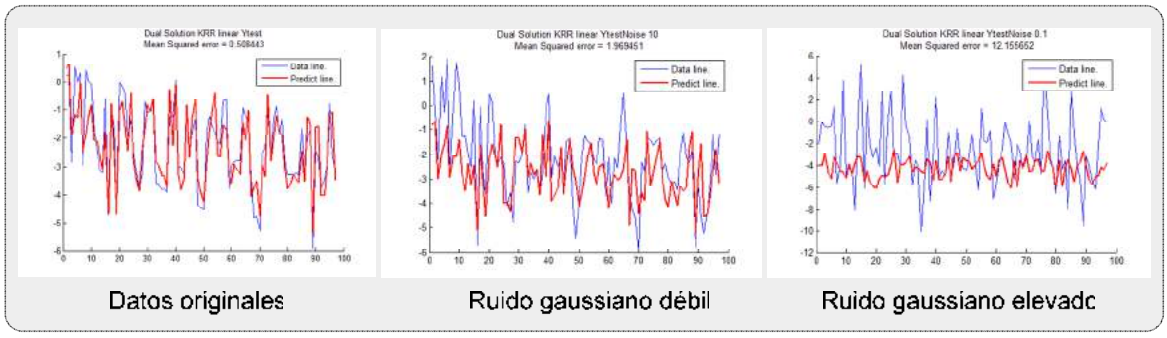

Fig. 28. Precisión en la predicción de la serie temporal dependiente para el conjunto de prueba utilizando un kernel lineal.

En el caso de un kernel polinomial de grado 2 (Fig. 29), la reproducción de la serie temporal final para el conjunto de datos de prueba es más imprecisa, resultando un aumento del error en el ajuste con respecto al conseguido por el kernel lineal.

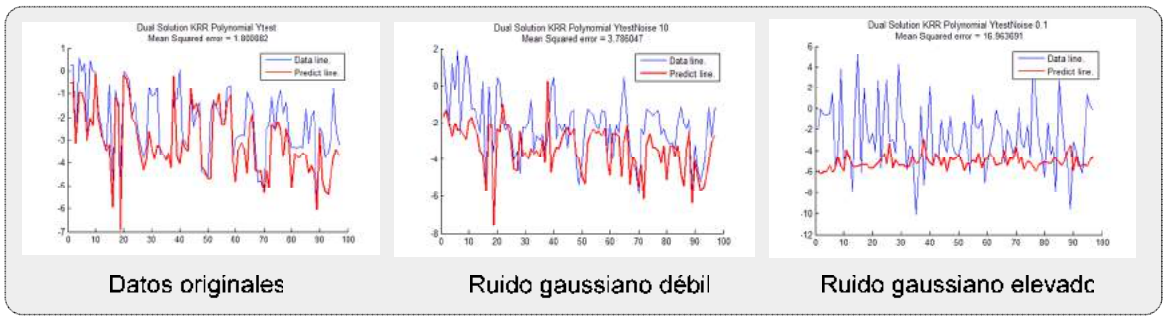

Fig. 29. Precisión en la predicción de la serie temporal dependiente para el conjunto de prueba utilizando un kernel polinomial de grado 2.

Finalmente para el caso del kernel mediante la tangente hiperbólica (Fig. 30), comprobamos como los resultados son aún peores, siendo más imprecisa la reproducción en la predicción de la serie temporal observada.

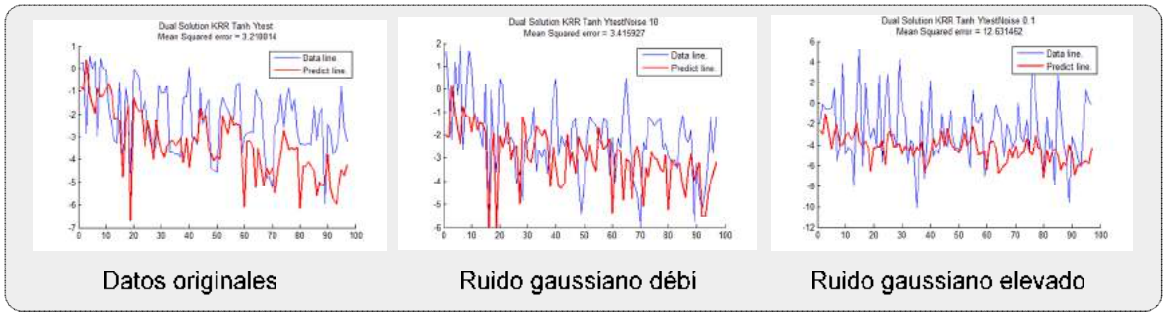

Fig. 30. Precisión en la predicción de la serie temporal dependiente para el conjunto de prueba utilizando un kernel mediante la tangente hiperbólica.

No obstante, es de señalar que utilizando el kernel de la tangente hiperbólica se consiguen mejores resultados que utilizando el kernel polinomial para el conjunto de prueba (3.416 vs. 3.786), cuando nos fijamos en la segunda serie temporal (adicción de ruido gaussiano moderado).

4. CONCLUSIONES

En este trabajo se han descrito diferentes técnicas de regresión que se engloban dentro de la analítica predictiva de datos. El análisis predictivo de datos es muy útil para estudiar y ajustar de manera eficiente el comportamiento de un sistema dinámico lineal o no lineal a partir de las medidas discretas de sus variables. Por tanto, el objetivo principal de un modelo de regresión generado a partir de un análisis predictivo es obtener una ecuación matemática que nos permita "predecir" con el mínimo error posible el valor de una variable dependiente Y una vez conocidos los valores de X_1, X_2 .. X_n o variables independientes predictoras. Dicha ecuación servirá como modelo o función de aproximación para la predicción de futuras observaciones.

Cuando las variables predictoras están muy correlacionadas, los coeficientes de regresión resultantes de un ajuste por mínimos cuadrados ordinarios (MCO) pueden llegar a ser muy erráticos e imprecisos, debido a los efectos desastrosos que la multicolinealidad tiene sobre su varianza. Estos coeficientes originan predicciones erróneas a la hora de vaticinar nuevas respuestas correspondientes a entradas similares que deberían pronosticar salidas similares. La técnica *Ridge Regression* (RR) trata estas colinealidades minimizando el problema al contraer los coeficientes de regresión de MCO mediante la introducción de un sesgo, logrando coeficientes ajustados con menor varianza, dando estabilidad así a la predicción del modelo y solucionando dicho problema. En este trabajo se ha estudiado también las diferentes modalidades que existen para obtener y elegir un sesgo o factor de regularización óptimo (aquel que obtenga predicciones con el mínimo error posible entre lo observado y lo esperado).

Los métodos kernel (ampliamente utilizadas en las máquinas de aprendizaje supervisado) han demostrado ser técnicas muy eficaces en la resolución de problemas no lineales. Si los datos de entrenamiento (las variables independientes) muestran relaciones no lineales, la técnica RR será incapaz de modelarlas adecuadamente con un error mínimo aceptable (el sesgo introducido en RR ayuda pero a veces también resulta insuficiente). Sin embargo, una solución no lineal puede ser tratada y formulada moviéndonos a un espacio de características lineal a partir del espacio de entrada no lineal. *Kernel Ridge Regression* (KRR) es una técnica que encuentra y realiza un mapeo de los datos de entrada (considerados no lineales) en un espacio de características de más alta dimensión (donde corresponden a un modelo aproximadamente lineal) obteniendo errores en el ajuste mucho menores con un gasto computacional razonable, lo cual puede posibilitar su implementación en tareas de tiempo real. No obstante, se ha visto como estas funciones conllevan algunos inconvenientes. Como se ha podido comprobar, el sobre-ajuste a un conjunto de entrenamiento puede provocar no saber generalizar adecuadamente cuando se usa el modelo para un conjunto de prueba diferente.

Finalmente, se han aplicado estas técnicas predictivas a diferentes series temporales no linales. Para ello se compararon los resultados en presencia de dos tipos de intensidades de ruido gaussiano añadido, con los resultados obtenidos en ausencia de ruido (datos brutos originales), concluyendo que el uso del kernel lineal mediante la solución dual de RR es el que mejor rendimiento proporciona en términos de mínimo error en el

ajuste. Además la forma estructural de la serie temporal esperada seguía conservándose, incluso en presencia de ruido gaussiano moderado.

5. MOTIVACION Y TRABAJOS FUTUROS

Los inicios de este proyecto comienzan en la Unidad de Adquisición de Datos del Laboratorio de Fusión del CIEMAT. En dicha unidad, se vienen desarrollando durante años, una serie de técnicas y procedimientos analíticos aplicados a la base de datos del stellerator español TJ-II **[TJ-II]** (Fig. 31), y donde se almacenan más de tres millones de señales de evolución temporal.

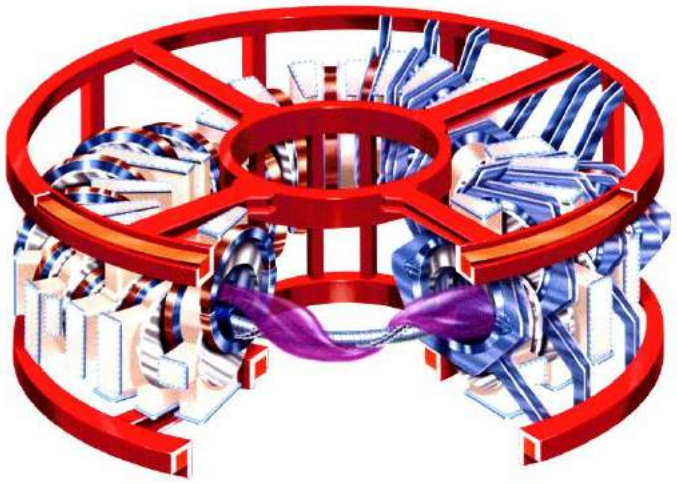

Fig. 31. Stellerator TJ-II (CIEMAT).

La extracción del conocimiento oculto en bases de datos masivas, requiere el uso de herramientas y técnicas automáticas de minería de datos que faciliten la generación de modelos predictivos eficientes y con elevado poder explicativo.
Los modelos teóricos resultantes de este estudio supervisado, tienen que ser capaces de generalizar con el mínimo error posible frente a nuevas entradas de datos, con el fin de poder ayudar no solo en diferentes tareas críticas de control, sino también en la aportación y anticipación de resultados, incluso antes de que se produzcan éstos.
Por tanto, el objetivo principal de este trabajo es el estudio, comprensión y análisis de técnicas predictivas de datos que hagan uso de la regresión estadística como paradigma de aprendizaje válido para poder ser aplicado en el futuro a las diferentes señales que integran la base de datos del stellerator TJ-II.

Básicamente, este trabajo se subdividió en dos partes, la primera parte expone exhaustivamente la analítica predictiva basada en la regresión, y en la segunda parte del proyecto se hace uso de dicho análisis, aplicando diferentes técnicas a unos datos experimentales multidimensionales en presencia de ruido gaussiano añadido, comparando estos resultados con los obtenidos en ausencia de ruido. Con esto, se ha tratado de reflejar cómo diferentes técnicas de regresión actúan en presencia de *outliers* generalizados y contínuos, con el objetivo de poder simular y acercarse así al comportamiento de un sistema de adquisición de datos real, sometido a elevadas interferencias externas.

Futuros trabajos que complementen este proyecto sería comprobar si los resultados aquí obtenidos son reproducibles mediante los datos y señales de evolución temporal que se encuentran en la base de datos del TJ-II y además, sería interesante también comparar estas técnicas de regresión con el poder predictivo de otros sistemas de aprendizaje como son las redes neuronales ó las máquinas de vectores soporte.

REFERENCIAS Y BIBLIOGRAFIA

[Akdeniz, 2001] Akdeniz, F. 2001. "The examination and analysis of residuals for some biased estimators in linear regression". *Communications in Statistics: Theory and Methods*. 30. 1171-1183.

[Belsley, 1980] Belsley, D. Kuth, E. Welsh, R. 1980. ·Regression diagnostics. Identifying influential data and sources of collinearity". New York, John Wiley & Sons, Inc.

[Chatterjee, 2006] Chatterjee, S. Hadi, A. 2006. "Regression analysis by example". Fourth Edition. *Wiley-Interscience*. ISBN: 100471746967

[Galton, 1886] Galton, Francis. 1886. "Regression towards mediocrity in hereditary stature". *Journal of the Anthropological Institute*. 15. 246-263.

[García, 2006] García, Jorge. et al. 2006. "Efectos de la colinealidad en el modelado de la regresión y su solución". *Cultura Científica y Tecnológica*. 16. 23-34

[GEA, 2006] Grupo de Estadística Aplicada. 2006. Universidad de Salamanca. "Regresión y correlación". *Introducción a la Estadística*. http://biplot.usal.es/problemas/libro/index.html

[Hoerl y Kennard, 1970] Hoerl, Arthur E. Kennard, Robert W." Ridge Regression: Applications to Nonorthogonal Problems". *Technometrics*, Vol. 12, No. 1. (Feb., 1970), pp. 69-82.

[Hoerl y Kennard, 1976] Hoerl, A. E.,R. W. Kennard. 1976. "Ridge Regression Iterative Estimation of the Biased Parameter". *Communication in statistics*, A5(1), 77-88.

[Hoerl, Kennard y Baldwin, 1975] Hoerl, A.E., Kennard, R.W., and Baldwin, K.F. (1975), "Ridge regression: some simulations," *Communications in Statistics*, 4, 105-123.

[Mason, 1991] Mason, C. Perreault, W. 1991. "Collinearity, power and interpretation of multiple regression analysis". *Journal of marketing Research*. 28. 268-220.

[NIST, 2003] NIST/SEMATECH 2003. *e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/

[Núñez, 2005] Núñez, V. Tussell, F. 2005. "Regresión y Análisis de Varianza". http://www.et.bs.ehu.es/~etptupaf/nuevo/ficheros/estad3/reg.pdf

[Pasha, 2004] Pasha, G.R. Shah, Ali. 2004. "Application of ridge regression to multicollinear data". *Journal of Research (Science)*, Bahauddin Zakariya University, Multan, Pakistan. 15. 97-106. ISSN 1021-1012.

[Ramos, 2007] Ramos, C. Martínez, M. Sanchís, J. Salcedo, J.V. 2007. "LQR Robusto mendiante incertidumbre acotada en los datos". *Revista Iberoamericana de Automática e Informática Industrial*. 4. 61-72.

[Shawe-Taylor, 2004] Shawe-Taylor, J. Cristianini, N. 2004. "Kernel Methods for Pattern Analysis". *Cambridge University Press*. ISBN: 978-0-521-81397-6.

[Souza, 2010] Souza, Cesar. 2010. Kernel Functions for Machine Learning Applications. http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html

[Thibaux, 2008] Thibaux, Romain. 2008. "Regression". *Computer Science 294, Practical Machine Learning*. http://www.cs.berkeley.edu/~pliang/cs294-spring08/lectures/regression/

[TJ-II] http://fudaqs2.ciemat.es/TJ2WEB/indexGAD.jsp

[Wang, 1994] Wang, S. and Akabay, C. 1994. "Autocorrelation: problems and solution in regression analysis". *The Journal of Business and Forecasting Methods and Systems*. 13. 18-26.

[Zhang, 2009] Zhang, Jian. 2009. “Risk Minimization”. *Statistical Learning Theory*. http://www.stat.purdue.edu/~jianzhan/STAT598Y/NOTES/slt02.pdf

Autorización de difusión.

El abajo firmante, matriculado en el Master en Investigación en Informática de la Facultad de Informática, autoriza a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente Trabajo Fin de Máster: "ANALISIS PREDICTIVO DE DATOS MEDIANTE TECNICAS DE REGRESION ESTADISTICA", realizado durante el curso académico 2009-2010 bajo la dirección de MATILDE SANTOS PEÑAS [y con la colaboración externa de dirección de JESUS A. VEGA SANCHEZ] en el Departamento de Arquitectura de Computadores y Automática, y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo.

Firmado: Augusto Pereira González