DOCUMENT DE TREBALL

XREAP2012-11

Local Distance-Based Generalized Linear Models using the dbstats package for R

Eva Boj (CREB, XREAP)
Pedro Delicado
Josep Fortiana
Anna Esteve
Adrià Caballé

XARXA DE REFERÈNCIA EN ECONOMIA APLICADA

Local Distance-Based Generalized Linear Models using the `dbstats` package for `R`

Eva Boj[1,4] Pedro Delicado[2] Josep Fortiana[1]
Anna Esteve[3] Adrià Caballé[2]

May 21, 2012

This paper introduces local distance-based generalized linear models. These models extend (weighted) distance-based linear models firstly with the generalized linear model concept, then by localizing. Distances between individuals are the only predictor information needed to fit these models. Therefore they are applicable to mixed (qualitative and quantitative) explanatory variables or when the regressor is of functional type. Models can be fitted and analysed with the `R` package `dbstats`, which implements several distance-based prediction methods.

Keywords: Distance-based prediction, Generalized Linear Model, Local Likelihood, Iteratively Weighted Least Squares, `R`

1. Introduction

Boj, Delicado, and Fortiana (2010) introduced local Distance-Based Linear Model (DB-LM), a nonparametric prediction technique extending (weighted) DB-LM. In the present paper we introduce further extensions. In general, any statistical technique based on

[1]Universitat de Barcelona
[2]Universitat Politècnica de Catalunya
[3]CEEISCAT
[4]**Corresponding author:** Eva Boj. Dept. de Matemàtica Econòmica, Financera i Actuarial, Univ. de Barcelona, Diagonal 690, 08034 Barcelona, Spain. Tel: +34 934035744. Fax: +34-934034892. E-mail: `evaboj@ub.edu`

Weighted Least Squares (WLS) can be adapted to data presented as an inter-individual distances matrix by just replacing each WLS step by the corresponding weighted DB-LM. This procedure is easily extended to Iterative Weighted Least Squares (IWLS), as applied in many statistical methods, ranging from Generalized Linear Models (GLM) McCullagh and Nelder (1989) to Robust Regression (see, for instance, Green (1984), Street, Carroll, and Ruppert (1988)). Here we develop in detail Distance-Based Generalized Linear Models (DB-GLM), then we construct its local version.

The `dbstats` `R` package (Boj, Caballé, Delicado, and Fortiana 2012) contains classes and functions implementing distance-based prediction methods such as DB-LM, local DB-LM, DB-GLM, local DB-GLM and Distance-Based Partial Least Squares Regression (DB-PLSR) Boj, Claramunt, Grané, and Fortiana (2007).

The paper is structured as follows: In Section 2.1 we review the main features of DB-LM; in Section 2.2 we develop DB-GLM as an extension of DB-LM; in Section 2.3 we introduce local DB-GLM. In Section 3 we describe the `dbstats` package for `R`. Finally, in Section 4, we illustrate the use of `dbstats` to fit DB-GLM and local DB-GLM with several examples.

2. Distance-Based Prediction

In this section, after recalling the main characteristics of DB-LM, we present DB-GLM and then we show how to construct its local version.

2.1. Distance-Based Linear Model: Definition and results

DB-LM was introduced by Cuadras (1989) and has been developed in Cuadras and Arenas (1990), Cuadras, Arenas, and Fortiana (1996), Boj, Claramunt, and Fortiana (2007), Esteve, Boj, and Fortiana (2009) and Boj, Delicado, and Fortiana (2010). Here we recall its main concepts, as given in these articles, where the reader is referred to for more details and proofs.

A sketchy description of DB-LM is as follows: Let y_i be a real-valued observation for each i-th individual Ω_i in a given set $\Omega = \{\Omega_1, \ldots, \Omega_n\}$, randomly drawn from a population, and let $w_i \in (0, 1)$ be the constant positive weight of Ω_i. The $n \times 1$ weight vector $\boldsymbol{w} = (w_1, \ldots, w_n)'$ is standardized to unit sum, i.e., $\mathbf{1}' \cdot \boldsymbol{w} = 1$, where $\mathbf{1}$ is the $n \times 1$ vector of ones. We assume that the $n \times 1$ response vector $\boldsymbol{y} = (y_i)$ is $\boldsymbol{w}$-centered, i.e., $\boldsymbol{w}' \cdot \boldsymbol{y} = 0$.

Individuals in Ω are described by a set $\boldsymbol{Z}$ of variables, henceforth *observed predictors*, possibly including both quantitative and qualitative measurements or, possibly, other nonstandard quantities, such as character strings or functions. A distance (metric or semi-metric) $\delta(\,\cdot\,,\,\cdot\,)$ is defined in Ω, as a function of the $\boldsymbol{Z}$ variables. We denote by $\boldsymbol{\Delta}$ the $n \times n$ matrix, whose entries are the squared distances $\delta^2(\Omega_i, \Omega_j)$.

We define the $n \times n$ *inner-products matrix* as:

$$\boldsymbol{G_w} = -\frac{1}{2}\, \boldsymbol{J_w} \cdot \boldsymbol{\Delta} \cdot \boldsymbol{J_w}',$$

where $\boldsymbol{J_w}$ is the $\boldsymbol{w}$-centering matrix, defined as $\boldsymbol{J_w} = \boldsymbol{I} - \boldsymbol{1} \cdot \boldsymbol{w}'$. We denote by $\boldsymbol{g_w}$ a $1 \times n$ row vectorcontaining the (necessarily nonnegative) diagonal entries of $\boldsymbol{G_w}$.

Any $n \times k$ matrix $\boldsymbol{X_w}$ such that $\boldsymbol{G_w} = \boldsymbol{X_w} \cdot \boldsymbol{X_w}'$ is called a *Euclidean configuration of* $\boldsymbol{\Delta}$. $k \geq r \equiv rank\, \boldsymbol{G_w}$ and $\boldsymbol{w}' \cdot \boldsymbol{X_w} = \boldsymbol{0}$.

The DB-LM of response $\boldsymbol{y}$ with weights $\boldsymbol{w}$ and predictor matrix $\boldsymbol{\Delta}$, an $n \times n$ square distances matrix, is defined as the WLS regression of $\boldsymbol{y}$ on a $\boldsymbol{w}$-centered Euclidean configuration of $\boldsymbol{\Delta}$, $\boldsymbol{X_w}$, a *latent Euclidean configuration*.

Assume a new case Ω_{n+1} is available, and we are given the $1 \times n$ vector $\boldsymbol{\delta}_{n+1}$ of squared distances from Ω_{n+1} to the n previously known individuals. Ω_{n+1} can be represented as a k-vector $\boldsymbol{x}_{n+1}$ in the row space of $\boldsymbol{X_w}$. Then, the predicted Y for Ω_{n+1} is $\boldsymbol{x}_{n+1} \cdot \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the vector of estimated regression coefficients.

DB-LM does not depend on a specific $\boldsymbol{X_w}$, since the final quantities are obtained directly from the distances. Usually such a configuration needs not be made explicit, and neither do $\hat{\boldsymbol{\beta}}$ or $\boldsymbol{x}_{n+1}$. In DB-LM the hat matrix is:

$$\boldsymbol{H_w} = \boldsymbol{G_w} \cdot \left(\boldsymbol{D_w}^{1/2} \cdot \boldsymbol{F_w}^{+} \cdot \boldsymbol{D_w}^{1/2}\right), \tag{1}$$

where $\boldsymbol{D_w} = \mathrm{diag}(\boldsymbol{w})$ is the diagonal matrix whose diagonal entries are the weights $\boldsymbol{w}$,

$$\boldsymbol{F_w} = \boldsymbol{D_w}^{1/2} \cdot \boldsymbol{G_w} \cdot \boldsymbol{D_w}^{1/2},$$

and $\boldsymbol{F_w}^{+}$ is the Moore-Penrose pseudo-inverse of $\boldsymbol{F_w}$. Thus, $\boldsymbol{H_w}$ is an intrinsic quantity, meaning that it can be expressed directly as a function of the distances or, equivalently, the inner products.

The predicted Y for a new case Ω_{n+1}, given its $\boldsymbol{\delta}_{n+1}$ vector is:

$$\hat{y}_{n+1} = \frac{1}{2}\left(\boldsymbol{g_w} - \boldsymbol{\delta}_{n+1}\right) \cdot \left(\boldsymbol{D_w}^{1/2} \cdot \boldsymbol{F_w}^{+} \cdot \boldsymbol{D_w}^{1/2}\right) \cdot \boldsymbol{y}. \tag{2}$$

In DB-LM the rank r of the hat-matrix in (1), as in an ordinary linear regression, is equal to the number of linearly independent linear predictors. Since for n cases,

depending on the chosen metric, r can be as high as $n-1$, giving an overparametrized model with unstable predictions, a sensible procedure is to replace the pseudo-inverse $\boldsymbol{F}_{\boldsymbol{w}}^{+}$ with a lower-rank approximation. This can be easily implemented by the Singular Value Decomposition which, by the Schmidt-Eckart-Young Theorem (see, e.g., Stewart (1993)), gives the best ℓ^2 approximation of any given rank k, $1 \leq k \leq r$. Cross-validation can then be used to select a suitable k.

DB-LM contains WLS as a particular instance: if we start from a $n \times r$ $\boldsymbol{w}$-centered matrix $\boldsymbol{X}_{\boldsymbol{w}}$ of r continuous predictors corresponding to n individuals and we define $\boldsymbol{\Delta}$ as the matrix of squared Euclidean distances between rows of $\boldsymbol{X}_{\boldsymbol{w}}$, then $\boldsymbol{X}_{\boldsymbol{w}}$ is trivially a Euclidean configuration of $\boldsymbol{\Delta}$, hence the DB-LM hat matrix, response and predictions coincide with the corresponding WLS quantities of ordinary Linear Model (LM).

2.2. Distance-Based Generalized Linear Model

In this section we review the basic concepts and notations of GLM, for the sake of an easy reference. As it is well-known (see, eg., McCullagh and Nelder (1989)), in a GLM we have a linear predictor $\boldsymbol{\eta} = \boldsymbol{X} \cdot \boldsymbol{\beta}$, which is related to the response variable Y by means of a link function $g(\cdot)$, $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, then,

$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}). \tag{3}$$

In a GLM it is assumed that each component of the response has a distribution in the exponential family, taking the form:

$$f_Y(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \exp\{(\boldsymbol{y} \cdot \boldsymbol{\theta} - b(\boldsymbol{\theta}))/a(\boldsymbol{\phi}) + c(\boldsymbol{y}, \boldsymbol{\phi})\}, \tag{4}$$

for some specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. If $\boldsymbol{\phi}$ is known, this is an exponential family model with canonical parameter $\boldsymbol{\theta}$.

The log-likelihood function for a GLM is $l(\boldsymbol{\theta}; \boldsymbol{y}) = (\boldsymbol{y} \cdot \boldsymbol{\theta} - b(\boldsymbol{\theta}))/a(\boldsymbol{\phi}) + c(\boldsymbol{y}, \boldsymbol{\phi})$ and the mean and the variance of Y can be derived easily from the relations $E\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right) = 0$ and $E\left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}^2}\right) + E\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)^2 = 0$. From (4) we have that $\frac{\partial l}{\partial \boldsymbol{\theta}} = \{y - b'(\boldsymbol{\theta})\}/a(\boldsymbol{\phi})$ and $\frac{\partial^2 l}{\partial \boldsymbol{\theta}^2} = -b''(\boldsymbol{\theta})/a(\boldsymbol{\phi})$ and then,

$$E(Y) = \boldsymbol{\mu} = b'(\boldsymbol{\theta}), \tag{5}$$

and

$$var\left(Y\right) = b''\left(\boldsymbol{\theta}\right) \cdot a\left(\boldsymbol{\phi}\right). \tag{6}$$

The variance of Y is the product of two functions; one, $b''\left(\boldsymbol{\theta}\right)$, depends on the canonical parameter only (and hence on the mean (5)) and will be called the *variance function*, while the other is independent of θ and depends only on $\boldsymbol{\phi}$. The variance function as a function of $\boldsymbol{\mu}$ will be written $V\left(\boldsymbol{\mu}\right) = b''\left(\boldsymbol{\theta}\right)$. Commonly $a\left(\boldsymbol{\phi}\right)$ is of the form $a\left(\boldsymbol{\phi}\right) = \boldsymbol{\phi}/\boldsymbol{w}$ and $\boldsymbol{\phi}$, the *dispersion parameter*, is constant over observations. Respect $\boldsymbol{w}$ it is a known *prior weight* that varies from observations to observation. If we have n independent readings $\boldsymbol{w} = n$. Finally, we can write (6) as

$$var\left(Y\right) = V\left(\boldsymbol{\mu}\right) \cdot \frac{\boldsymbol{\phi}}{\boldsymbol{w}}. \tag{7}$$

2.2.1. Construction of the Distance-Based Generalized Linear Model

In a GLM the maximum-likelihood estimates of the parameters $\boldsymbol{\beta}$ in the linear predictor $\boldsymbol{\eta}$ can be obtained by IWLS (see, e.g., McCullagh and Nelder (1989) pp. 40-43 or Wood (2006) pp. 63-66 for a more detailed description and justification of the algorithm). In the IWLS the dependent variable of the regression is not $\boldsymbol{y}$ but $\boldsymbol{z}$, a linearized form of the link function applied to $\boldsymbol{y}$, and the weights are functions of the fitted values $\hat{\boldsymbol{\mu}}$. The process is iterative because both the adjusted dependent variable $\boldsymbol{z}$ and the weight $\boldsymbol{W}$ depend on the fitted values, for which only current estimates are available. The procedure underlying the iteration is as follows. Let $\hat{\boldsymbol{\eta}}_0$ be the current estimate of the linear predictor, with corresponding fitted value $\hat{\boldsymbol{\mu}}_0$ derived from the link function $\boldsymbol{\eta} = g\left(\boldsymbol{\mu}\right)$. Form the adjusted dependent variate with typical value

$$\boldsymbol{z}_0 = \hat{\boldsymbol{\eta}}_0 + \left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0\right) \cdot \left(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\right)_0, \tag{8}$$

where the link derivative is evaluated at $\hat{\boldsymbol{\mu}}_0$. The quadratic weight is defined by:

$$\boldsymbol{W}_0^{-1} = \left(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\right)_0^2 \cdot \frac{\boldsymbol{V}_0}{\boldsymbol{w}}, \tag{9}$$

where $\boldsymbol{V}_0$ is the variance function evaluated at $\hat{\boldsymbol{\mu}}_0$. Now regress $\boldsymbol{z}_0$ on the covariates with weight $\boldsymbol{W}_0$ to give new estimates of $\hat{\boldsymbol{\beta}}_1$ of the parameters; from these form a new

estimate $\hat{\boldsymbol{\eta}}_1$ of the linear predictor. Repeat until changes are sufficiently small.

Note that $\boldsymbol{z}$ is just a linearized form of the link function applied to the data, for, to first order, $g(\boldsymbol{y}) \simeq g(\boldsymbol{\mu}) + (\boldsymbol{y} - \boldsymbol{\mu}) \cdot g'(\boldsymbol{\mu})$. The variance of $\boldsymbol{z}$ is $\boldsymbol{W}^{-1}$ (ignoring the dispersion parameter), assuming that $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ are fixed and known.

Both DB-GLM and DB-LM share the same elements: a set of n individuals with associated weights vector $\boldsymbol{w}$ (standardized to unit sum), for which we have observed the $\boldsymbol{w}$-centered response vector $\boldsymbol{y}$ and a set of predictors $\boldsymbol{Z}$. From the latter we calculate the $n \times n$ distances matrix $\boldsymbol{\Delta}$. Just as GLM with respect to LM, DB-GLM differs from DB-LM in two aspects:

1. We assume the responses distribution is in an exponential dispersion family (4), as in any GLM.

2. The relation between the linear predictor $\boldsymbol{\eta} = \boldsymbol{X_w} \cdot \boldsymbol{\beta}$, obtained from the latent Euclidean configuration $\boldsymbol{X_w}$, and the response $\boldsymbol{y}$ is given by a link function $g(\cdot)$ as in (3).

Then we have an underlying GLM, with link function $g : C(Supp(Y)) \to \mathbb{R}$,

$$g(\mu_i) = \eta_i, \quad \text{where } \mu_i = E\{Y_i\}, \quad \eta_i = \boldsymbol{x}_i \cdot \boldsymbol{\beta}, \tag{10}$$

where $\boldsymbol{\beta} \in \mathbb{R}^r$ is an $r \times 1$ parameter vector. Model (10) relates each response Y_i to the Euclidean coordinates of Ω_i. That is, the linear predictor η_i is a linear combination of the Euclidean coordinates $\boldsymbol{x}_i$ of Ω_i, the i-th row of the $n \times r$ matrix $\boldsymbol{X_w}$ of a Euclidean configuration of $\boldsymbol{\Delta}$.

The DB-GLM model definition is sound since it does not depend of the particular choice of the Euclidean configuration. Indeed, as stated, the model consists of random vectors $(Y_1, \ldots, Y_n)'$ whose expectation, $(\mu_1, \ldots \mu_n)'$, transformed by the link function, is a vector in the column space $\mathscr{G}$ of $\boldsymbol{X_w}$. Since $\mathscr{G}$ is also the column space of $\boldsymbol{X_w} \cdot \boldsymbol{X_w}' = \boldsymbol{G_w}$, (Rao 1973, p. 27), we conclude the model depends only on $\boldsymbol{\Delta}$.

To fit DB-GLM we use the IWLS algorithm described above, where DB-LM substitutes LM in formulas (8) and (9) to regress $\boldsymbol{z}_0$ on the covariates with weight $\boldsymbol{W}_0$, in order to obtain the new estimation $\hat{\boldsymbol{\eta}}_1$. Observe that the IWLS estimation process for DB-GLM does not depend on a specific $\boldsymbol{X_w}$ since the final quantities are obtained directly from distances. In the first step we need and initial $\hat{\boldsymbol{\mu}}_0$. Then we calculate $\hat{\boldsymbol{\eta}}_0$ and $\left(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\right)_0$. These two elements only depend on the link function. Finally, we calculate $\boldsymbol{V}_0$, the function (7) evaluated at $\hat{\boldsymbol{\mu}}_0$, which only depends on the fitted vales $\hat{\boldsymbol{\mu}}$ at each step.

Prediction for new observations is also independent of the choice of $\boldsymbol{X}_{\boldsymbol{w}}$. Given a new case Ω_{n+1}, described by the $1 \times n$ vector $\boldsymbol{\delta}_{n+1}$ of squared distances from Ω_{n+1} to the n previously known individuals, the predicted η_{n+1} for Ω_{n+1} is $\boldsymbol{x}_{n+1} \cdot \hat{\boldsymbol{\beta}}$, which is calculated with formula (2) with the quantities of the last IWLS step. Then we can calculate $\mu_{n+1} = g^{-1}(\eta_{n+1})$.

DB-GLM contains GLM as a particular case: if we start from a $n \times r$ $\boldsymbol{w}$-centered matrix $\boldsymbol{X}_{\boldsymbol{w}}$ of r continuous predictors corresponding to n individuals and we define $\boldsymbol{\Delta}$ as the matrix of squared Euclidean distances between rows of $\boldsymbol{X}_{\boldsymbol{w}}$, then $\boldsymbol{X}_{\boldsymbol{w}}$ is trivially a Euclidean configuration of $\boldsymbol{\Delta}$, hence the DB-GLM hat matrix, response and predictions coincide with the corresponding IWLS quantities of ordinary GLM.

2.3. Local Distance-Based Generalized Linear Model

We consider again the framework stated in Section 2.2 when DB-GLM was introduced. Our objective is now to fit a *local DB-GLM*, where *local* refers to the fact that when the DB-GLM is used to predict the value of the response variable for an object Ω_{n+1}, we use only the information provided by observed objects Ω_i, $i = 1, \ldots, n$, that are *close* to Ω_{n+1}, giving to Ω_i a weight that is a decreasing function of the distance between Ω_i and Ω_{n+1}. The idea is to translate to the DB-GLM context the principles of *local likelihood*, as stated in Loader (1999) (see also Section 3.4 in Bowman and Azzalini (1997), Section 6.5 in Hastie, Tibshirani, and Friedman (2009) or Section 5.10 in Wasserman (2006)). Our approach parallels that used in Boj, Delicado, and Fortiana (2010) when local DB-LM is defined.

Let $m(\Omega_{n+1})$ be the expected value of the response y corresponding to the object Ω_{n+1}. This is the value we want to estimate and we do that by using DB-GLM. We assume that two distance functions, δ_1 and δ_2, are defined between the elements of Ω (the set of observable objects). We consider the weights

$$w_i(\Omega_{n+1}) = K(\delta_1(\Omega_{n+1}, \Omega_i)/h) / \sum_{j=1}^{n} K(\delta_1(\Omega_{n+1}, \Omega_j)/h),$$

where h is an smoothing parameter (depending on n). Let $\boldsymbol{\Delta}_2$ be the matrix of squared distances between functions defined from δ_2. We fit a DB-GLM starting from the initial elements

$$\boldsymbol{\Delta}_2 = (\delta_2(\Omega_i, \Omega_j)^2)_{i=1\ldots n, j=1\ldots n},\ \boldsymbol{y} = (y_i)_{i=1\ldots n},\ \text{ and } \boldsymbol{w} = (w_i(\Omega_{n+1}))_{i=1\ldots n}.$$

We consider the new individual Ω_{n+1} and we compute the squared distances from object Ω_{n+1} to other individuals Ω_i:

$$\boldsymbol{\delta}_{2,n+1} = (\delta_2(\Omega_{n+1}, \Omega_1)^2, \cdots, \delta_2(\Omega, \Omega_n)^2).$$

Then we run the IWLS algorithm for DB-GLM to obtain the local DB-GLM estimator of $m(\Omega_{n+1})$:

$$\hat{m}_{localDB-GLM}(\Omega_{n+1}) = \hat{y}_{n+1}.$$

There are two distance functions involved in the local DB-GLM: one of them, δ_1, is used to compute the weight of observed objects Ω_i around the object Ω_{n+1} where the response function is estimated, and the other, δ_2, defines the distances between observations for computing DB-GLM. The distances δ_1 and δ_2 can coincide or not. In the context of local DB-LM, Boj, Delicado, and Fortiana (2010) show that using two distance functions provides more flexibility than using only one (that is, $\delta_1 = \delta_2$).

3. The dbstats Package

The `dbstats` package for `R` (Boj, Caballé, Delicado, and Fortiana 2012) implements several distance-based prediction methods. Currently the response is univariate. Distances can either be directly input as an interdistances matrix, a squared interdistances matrix, an inner-products matrix or computed from observed explanatory variables. We distinguish *observed explanatory variables*, denoted by $\boldsymbol{Z}$, from *Euclidean coordinates* $\boldsymbol{X_w}$. Observed explanatory variables $\boldsymbol{Z}$ are possibly a mixture of continuous, qualitative or more general quantities. `dbstats` does not provide specific methods for computing distances, depending instead on other available functions and packages, such as:

`dist` in the `stats` package.

`daisy` in the `cluster` package (Maechler 2012). Compared to `dist` above whose input must be numeric variables, the main feature of `daisy` is its ability to handle other variable types as well (e.g. nominal, ordinal, (a)symmetric binary) even when different types occur in the same data set.

`dist` in the `proxy` package (Meyer and Buchta 2012). Supersedes the one in the `stats` package. It allows a user-provided function, entered as a parameter, for evaluating distances between observations, hence it can deal with any type of data.

Distance-related classes in `dbstats` are `dist` and `dissimilarity` (as in `stats`), `D2`, for squared distances matrices; and `Gram`, for doubly centered inner product matrices. Utility functions such as `as.D2`, `as.Gram`, `D2toDist`, `D2toG`, `distoD2` and `GtoD2` allow their mutual interconversions (see (Boj, Caballé, Delicado, and Fortiana 2012) for details).

The main functions of `dbstats` are:

Linear and local linear models with a continuous response:

- `dblm` for DB-LM.
- `ldblm` for local DB-LM.
- `dbplsr` for DB-PLSR.

Generalized linear and local generalized linear models with a univariate response:

- `dbglm` for DB-GLM.
- `ldbglm` for local DB-GLM.

In the next subsections we describe the usage of `dblm`, `dbglm` and `ldbglm`. For `ldblm` and `plsr` we refer to (Boj, Caballé, Delicado, and Fortiana 2012).

3.1. dblm

The usage of `dblm` depends on the input information. There are two ways to incorporate predictors information: either as a `formula` or as a distance-type object of any of the four classes: `dist`, `dissimilarity`, `D2` or `Gram`.

The usage of `dblm` is:

For class `formula`

```
dblm(formula, data, ... , metric = "euclidean", method = "OCV",
  full_search=FALSE, weights, rel.gvar = 0.95, eff.rank)
```

For class `dist` or `dissimilary`

```
dblm(distance, y, ... , method = "OCV", full_search = FALSE,
  weights, rel.gvar = 0.95, eff.rank)
```

For class `D2`

```
dblm(D2, y, ... , method = "OCV", full_search = FALSE, weights,
  rel.gvar = 0.95, eff.rank)
```

For class `Gram`

```
dblm(G, y, ... , method = "OCV", full_search = FALSE, weights,
  rel.gvar = 0.95, eff.rank)
```

The arguments in `dblm` are:

formula an object of class `formula`. A formula of the form $\boldsymbol{y}$~$\boldsymbol{Z}$. This argument is a remnant of the `lm` function, kept for compatibility.

data an optional data frame containing the variables in the model (both response and explanatory variables, either the observed ones, $\boldsymbol{Z}$, or a Euclidean configuration $\boldsymbol{X}_{\boldsymbol{w}}$).

metric metric function to be used when computing distances from observed explanatory variables. One of `"euclidean"` (default), `"manhattan"`, or `"gower"`.

distance a `dist` or `dissimilarity` class object. See functions `dist` in the package `stats` and `daisy` in the package `cluster`.

D2 a `D2` class object. Squared distances matrix between individuals $\boldsymbol{\Delta}$. See details above to learn the usage of `dblm`.

G a `Gram` class object. Doubly centered inner product matrix of the squared distances matrix `D2`, i.e., $\boldsymbol{G}_{\boldsymbol{w}}$. See details above to learn the usage of `dblm.Gram`.

y (required if no formula is given as the principal argument). Response (dependent variable) must be numeric, matrix or data.frame.

method sets the method to be used in deciding the *effective rank*, which is defined as the number of linearly independent Euclidean coordinates used in prediction. There are six different methods: `"AIC"`, `"BIC"`, `"OCV"` (default), `"GCV"`, `"eff.rank"` and `"rel.gvar"`. `OCV` and `GCV` take the effective rank minimizing a cross-validatory quantity (either ordinary `ocv` or generalized `gcv`). `AIC` and `BIC` take the effective rank minimizing, respectively, the Akaike or Bayesian Information Criterion (see the `R` function `AIC` for more details). The optimizacion procedure to be used in the above four methods can be set with the `full_search` optional parameter.

When method is `eff.rank`, the effective rank is explicitly set by the user through the `eff.rank` optional parameter which, in this case, becomes mandatory.

When method is `rel.gvar`, the fraction of the data *geometric variability* for model fitting is explicitly set by the user through the `rel.gvar` optional parameter which, in this case, becomes mandatory.

full_search sets which optimization procedure will be used to minimize the modelling

criterion specified in `method`. Needs to be specified only if `method` is `"AIC"`, `"BIC"`, `"OCV"` or `"GCV"`. If `full_search=TRUE`, *effective rank* is set to its global best value, after evaluating the criterion for all possible ranks. Potentially too computationally expensive. If `full_search=FALSE`, the `R` function `optimize` is called. Then computation time is shorter, but the result may be found a local minimum.

weights an optional numeric vector of weights to be used in the fitting process. By default all individuals have the same weight.

rel.gvar relative geometric variability (real between 0 and 1). Take the lowest effective rank with a relative geometric variability higher or equal to `rel.gvar`. Default value (`rel.gvar=0.95`) uses a 95% of the total variability. Applies only `rel.gvar` if `method = "rel.gvar"`.

eff.rank integer between 1 and the number of observations minus one. Number of Euclidean coordinates used for model fitting. Applies only if `method="eff.rank"`.

... arguments passed to or from other methods to the low level.

When using method `method="eff.rank"` or `method="rel.gvar"`, a compromise between possible consequences of a bad choice has to be reached. If the rank is too large, the model can be overfitted, possibly leading to an increased prediction error for new cases (even though `R2`, the determination coefficient, is high). On the other hand, a small rank suggests a model inadequacy (`R2` is small). The other four methods are less error prone (but still they do not guarantee good predictions).

The function returns a list of class `dblm` containing the following components:

residuals the residuals (response minus fitted values).

fitted.values the fitted mean values.

df.residuals the residual degrees of freedom.

weights the specified weights.

y the response used to fit the model.

H the hat matrix projector.

call the matched call.

rel.gvar the relative geometric variabiliy, used to fit the model.

eff.rank the dimensions chosen to estimate the model.

ocv the ordinary cross-validation estimate of the prediction error.

gcv the generalized cross-validation estimate of the prediction error.

aic the Akaike Value Criterium of the model (only if `method="AIC"`).

bic the Bayesian Value Criterium of the model (only if `method="BIC"`).

3.2. dbglm

`dbglm` is a variety of GLM where explanatory information is coded as distances between individuals. These distances can either be computed from observed explanatory variables or directly input as a squared inter-distances matrix. Response and link function could be as in the `glm` function of `stats` for ordinary GLM.

The usage of `dbglm` is:

For class `formula`

```
dbglm(formula, data, family = gaussian, ... , metric = "euclidean",
        method = "OCV", full_search = FALSE, weights, maxiter = 100,
        eps1 = 1e-10, eps2 = 1e-10, rel.gvar = 0.95, eff.rank = NULL,
        offset, mustart = NULL)
```

For class `dist` or `dissimilary`

```
dbglm(distance, y, family = gaussian, method = "OCV", full_search
  = FALSE, weights, maxiter = 100, eps1 = 1e-10, eps2 = 1e-10,
        rel.gvar = 0.95, eff.rank = NULL, offset, mustart = NULL, ...)
```

For class `D2`

```
dbglm(D2, y, ... , family = gaussian, method = "OCV", full_search =
        FALSE, weights, maxiter=100, eps1 = 1e-10, eps2 = 1e-10,
        rel.gvar = 0.95, eff.rank = NULL, offset, mustart = NULL)
```

For class `Gram`

```
dbglm(G, y, ... , family = gaussian, method = "OCV", full_search =
        FALSE, weights, maxiter = 100, eps1 = 1e-10, eps2 = 1e-10,
```

rel.gvar = 0.95, eff.rank = NULL, offset, mustart = NULL)

The arguments in `dbglm` are:

formula an object of class `formula`. A formula of the form $\boldsymbol{y \tilde{} Z}$. This argument is a remnant of the `glm` function, kept for compatibility.

data an optional data frame containing the variables in the model (both response and explanatory variables, either the observed ones, $\boldsymbol{Z}$, or a Euclidean configuration $\boldsymbol{X_w}$).

metric metric function to be used when computing distances from observed explanatory variables. One of `"euclidean"` (the default), `"manhattan"`, or `"gower"`.

y (required if no formula is given as the principal argument). Response (dependent variable) must be numeric, factor, matrix or data.frame.

distance a `dist` or `dissimilarity` class object. See functions `dist` in the package `stats` and `daisy` in the package `cluster`.

D2 a `D2` class object. Squared distances matrix between individuals $\boldsymbol{\Delta}$. See details in `dblm`.

G a `Gram` class object. Doubly centered inner product matrix of the squared distances matrix `D2`, i.e., $\boldsymbol{G_w}$. See details in `dblm`.

family a description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function. (See the `R` function `family` for details of family functions.)

method sets the method to be used in deciding the *effective rank*, used in prediction. There are six different methods: `"AIC"`, `"BIC"`, `"OCV"` (default), `"GCV"`, `"eff.rank"` and `"rel.gvar"`. `OCV` and `GCV` take the effective rank minimizing a cross-validatory quantity (either ordinary `ocv` or generalized `gcv`). `AIC` and `BIC` take the effective rank minimizing, respectively, the Akaike or Bayesian Information Criterion (see the `R` function `AIC` for more details). The optimizacion procedure to be used in the above four methods can be set with the `full_search` optional parameter.

When method is `eff.rank`, the effective rank is explicitly set by the user through the `eff.rank` optional parameter which, in this case, becomes mandatory.

When method is `rel.gvar`, the fraction of the data *geometric variability* for model

fitting is explicitly set by the user through the `rel.gvar` optional parameter which, in this case, becomes mandatory.

full_search sets which optimization procedure will be used to minimize the modelling criterion specified in `method`. Needs to be specified only if `method` is `"AIC"`, `"BIC"`, `"OCV"` or `"GCV"`. If `full_search=TRUE`, *effective rank* is set to its global best value, after evaluating the criterion for all possible ranks. Potentially too computationally expensive. If `full_search=FALSE`, the `R` function `optimize` is called. Then computation time is shorter, but the result may be found a local minimum.

weights an optional numeric vector of prior weights to be used in the fitting process. By default all individuals have the same weight.

maxiter maximum number of iterations in the iterated `dblm` algorithm. (Default = 100)

eps1 stopping criterion 1, `"DevStat"`: convergence tolerance `eps1`, a positive (small) number; the iterations converge when `|dev - dev_old|/(|dev|) < eps1`. Stationarity of deviance has been attained.

eps2 stopping criterion 2, `"mustat"`: convergence tolerance `eps2`, a positive (small) number; the iterations converge when `|mu - mu_old|/(|mu|) < eps2`. Stationarity of fitted.values `mu` has been attained.

rel.gvar relative geometric variability (a real number between 0 and 1). At each `dblm` iteration, take the lowest effective rank, with a relative geometric variability higher or equal to `rel.gvar`. Default value (`rel.gvar=0.95`) uses the 95% of the total variability.

eff.rank integer between 1 and the number of observations minus one. Number of Euclidean coordinates used for model fitting in each `dblm` iteration. If specified its value overrides `rel.gvar`. When `eff.rank=NULL` (default), calls to `dblm` are made with `method=rel.gvar`.

offset this can be used to specify an a priori known component to be included in the linear predictor during fitting. This should be NULL or a numeric vector of length equal to the number of cases.

mustart starting values for the vector of means.

... arguments passed to or from other methods to the low level.

For Gamma-distributed responses, the domain of the canonical link function is not the same as the permitted range of the mean. In particular, the linear predictor might be

negative, obtaining an impossible negative mean. Should that event occur, `dbglm` stops with an error message. Proposed alternative is to use a non-canonical link function.

The function returns a list of class `dbglm` containing the following components:

residuals the `working` residuals, that is the `dblm` residuals in the last iteration of `dblm` fit.

fitted.values the fitted mean values, results of final `dblm` iteration.

family the `family` object used.

deviance measure of discrepancy or badness of fit. Proportional to twice the difference between the maximum achievable log-likelihood and that achieved by the current model.

aic.model A version of Akaike's Information Criterion. Equal to minus twice the maximized log-likelihood plus twice the number of parameters. Computed by the aic component of the family. For binomial and Poison families the dispersion is fixed at one and the number of parameters is the number of coefficients. For Gaussian, Gamma and Inverse Gaussian families the dispersion is estimated from the residual deviance, and the number of parameters is the number of coefficients plus one. For a Gaussian family the MLE of the dispersion is used so this is a valid value of AIC, but for Gamma and Inverse Gaussian families it is not. For families fitted by quasi-likelihood the value is NA.

null.deviance the deviance for the null model. The null model will include the offset, and an intercept if there is one in the model. Note that this will be incorrect if the link function depends on the data other than through the fitted mean: specify a zero offset to force a correct calculation.

iter number of Fisher scoring (`dblm`) iterations.

prior.weights the original weights.

weights the `working` weights, that are the weights in the last iteration of `dblm` fit.

df.residual the residual degrees of freedom.

df.null the residual degrees of freedom for the null model.

y the response vector used.

convcrit convergence criterion. One of: `"DevStat"` (stopping criterion 1), `"muStat"` (stopping criterion 2), `"maxiter"` (maximum allowed number of iterations has been exceeded).

H hat matrix projector of the last `dblm` iteration.

rel.gvar the relative geometric variabiliy in the last `dblm` iteration.

eff.rank the `working` effective rank, that is the `eff.rank` in the last `dblm` iteration.

3.3. ldbglm

`ldbglm` is a localized version of a DB-GLM. As in the global model `dbglm`, explanatory information is coded as distances between individuals. Neighborhood definition for localizing is done by the (semi)metric `dist1` whereas a second (semi)metric `dist2` (which may coincide with `dist1`) is used for distance-based prediction. Both `dist1` and `dist2` can either be computed from observed explanatory variables or directly input as a squared interdistances matrix or as a `Gram` matrix. Response and link function as in the `dbglm` function for ordinary generalized linear models. The model allows for a mixture of continuous and qualitative explanatory variables or, in fact, from more general quantities such as functional data.

The usage of `ldbglm` is:

For class `formula`

```
ldbglm(formula, data, ..., family = gaussian(), kind.of.kernel = 1,
        metric1 = "euclidean", metric2 = metric1, method = "GCV",
        weights, user_h = NULL, h.range = NULL, noh = 10, k.knn = 3,
        rel.gvar = 0.95, eff.rank = NULL, maxiter = 100, eps1 = 1e-10,
        eps2 = 1e-10)
```

For class `dist` or `dissimilary`

```
ldbglm(dist1, dist2 = dist1, y, family = gaussian(), kind.of.kernel
        = 1, method = "GCV", weights, user_h = quantile(dist1, .25) ^ .5,
        h.range = quantile(as.matrix(dist1), c(.05,.25)) ^ .5, noh = 10,
        k.knn = 3, rel.gvar = 0.95, eff.rank = NULL, maxiter = 100,
        eps1 = 1e-10, eps2 = 1e-10, ...)
```

For class `D2`

```
ldbglm(D2_1, D2_2 = D2_1, y, family = gaussian(), kind.of.kernel = 1,
        method = "GCV", weights, user_h = NULL, h.range = NULL, noh = 10,
        k.knn = 3, rel.gvar = 0.95, eff.rank = NULL, maxiter = 100,
        eps1 = 1e-10, eps2 = 1e-10, ...)
```

For class `Gram`

```
ldbglm(G1, G2 = G1, y, kind.of.kernel = 1, user_h = NULL,
        family = gaussian(), method = "GCV", weights, h.range = NULL,
        noh = 10, k.knn = 3, rel.gvar = 0.95, eff.rank = NULL, maxiter
        = 100, eps1 = 1e-10, eps2 = 1e-10, ...)
```

The arguments in `ldbglm` are:

formula an object of class `formula`. A formula of the form $\boldsymbol{y}$~$\boldsymbol{Z}$. This argument is a remnant of the `loess` function, kept for compatibility.

data an optional data frame containing the variables in the model (both response and explanatory variables, either the observed ones, $\boldsymbol{Z}$, or a Euclidean configuration $\boldsymbol{X}_{\boldsymbol{w}}$).

y (required if no formula is given as the principal argument). Response (dependent variable) must be numeric, matrix or data.frame.

dist1 a `dist` or `dissimilarity` class object. Distances between observations, used for neighborhood localizing definition. Weights for observations are computed as a decreasing function of their `dist1` distances to the neighborhood center, e.g. a new observation whose reoponse has to be predicted. These weights are then entered to a `dbglm`, where distances are evaluated with `dist2`.

dist2 a `dist` or `dissimilarity` class object. Distances between observations, used for fitting `dbglm`. Default `dist2=dist1`.

D2_1 a `D2` class object. Squared distances matrix between individuals. One of the alternative ways of entering distance information to a function. See the Details section in `dblm`. See above `dist1` for explanation of its role in this function.

D2_2 a `D2` class object. Squared distances between observations. One of the alternative ways of entering distance information to a function. See the Details section in `dblm`. See above `dist2` for explanation of its role in this function. Default `D2_2=D2_1`.

G1 a `Gram` class object. Doubly centered inner product matrix associated with the squared distances matrix `D2_1`.

G2 a `Gram` class object. Doubly centered inner product matrix associated with the squared distances matrix `D2_2`. Default `G2=G1`

family a description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function. (See the `R` function `family` for details of family functions.)

kind.of.kernel integer number between 1 and 6 which determines the user's choice of smoothing kernel. (1) Epanechnikov (Default), (2) Biweight, (3) Triweight, (4) Normal, (5) Triangular, (6) Uniform.

metric1 metric function to be used when computing `dist1` from observed explanatory variables. One of `"euclidean"` (default), `"manhattan"`, or `"gower"`.

metric2 metric function to be used when computing `dist2` from observed explanatory variables. One of `"euclidean"` (default), `"manhattan"`, or `"gower"`.

method sets the method to be used in deciding the *optimal bandwidth h*. There are five different methods, `AIC`, `BIC`, `OCV`, `GCV` (default) and `user_h`. `OCV` and `GCV` take the optimal bandwidth minimizing a cross-validatory quantity (either `ocv` or `gcv`). `AIC` and `BIC` take the optimal bandwidth minimizing, respectively, the Akaike or Bayesian Information Criterion (see the `R` function `AIC` for more details). When method is `user_h`, the bandwidth is explicitly set by the user through the `user_h` optional parameter which, in this case, becomes mandatory.

user_h global bandwidth `user_h`, set by the user, controlling the size of the local neighborhood of Z. Smoothing parameter (Default: 1st quartile of all the distances d(i,j) in `dist1`). Applies only if `method="user_h"`.

h.range a vector of length 2 giving the range for automatic bandwidth choice. (Default: quantiles 0.05 and 0.5 of d(i,j) in `dist1`).

noh number of bandwidth `h` values within `h.range` for automatic bandwidth choice (if `method!="user_h"`).

k.knn minimum number of observations with positive weight in neighborhood localizing. To avoid runtime errors due to a too small bandwidth originating neighborhoods with only one observation. By default `k.nn=3`.

rel.gvar relative geometric variability (a real number between 0 and 1). At each `dblm` iteration, take the lowest effective rank, with a relative geometric variability higher or equal to `rel.gvar`. Default value (`rel.gvar=0.95`) uses the 95% of the total variability.

eff.rank integer between 1 and the number of observations minus one. Number of Euclidean coordinates used for model fitting in each `dblm` iteration. If specified its value overrides `rel.gvar`. When `eff.rank=NULL` (default), calls to `dblm` are made with `method=rel.gvar`.

weights an optional numeric vector of weights to be used in the fitting process. By default all individuals have the same weight.

maxiter maximum number of iterations in the iterated `dblm` algorithm. (Default = 100)

eps1 stopping criterion 1, `"DevStat"`: convergence tolerance `eps1`, a positive (small) number; the iterations converge when `|dev - dev_old|/(|dev|) < eps1`. Stationarity of deviance has been attained.

eps2 stopping criterion 2, `"mustat"`: convergence tolerance `eps2`, a positive (small) number; the iterations converge when `|mu - mu_old|/(|mu|) < eps2`. Stationarity of fitted.values `mu` has been attained.

`...` arguments passed to or from other methods to the low level.

The set of bandwidth `h` values checked in automatic bandwidth choice is defined by `h.range` and `noh`, together with `k.knn`. For each `h` in it a local generalized linear model is fitted and the optimal `h` is decided according to the statistic specified in `method`.

And that `kind.of.kernel` designates which kernel function is to be used in determining individual weights from `dist1` values. See the `R` function `density` for more information.

The function returns a list of class `ldbglm` containing the following components:

residuals the residuals (response minus fitted values).

fitted.values the fitted mean values.

family the `family` object used.

y the response variable used.

S the Smoothing hat projector.

weights the specified weights.

call the matched call.

dist1 the distance matrix (object of class `"D2"` or `"dist"`) used to calculate the weights of the observations.

dist2 the distance matrix (object of class `"D2"` or `"dist"`) used to fit the `dbglm`.

4. Examples

4.1. An example of DB-GLM

We fit DB-GLM to the data set on Swedish third-party motor insurance in 1977 described in Hallin and Ingenbleek (1983). The file is included in `faraway` package with the name *motorins* (Faraway 2012). Data for factor Zone =1 can be found too in Andrews and Herzberg (1985, pp. 413-421). These data correspond to the cities of Stockholm, Gteburg and Malmo, and were obtained from a committee study of risk premiums in motor insurance. The total number of observations (for Zone=1) is $n = 295$ corresponding to different non-empty risk groups. For each group, Y is the number of claims suffered by the automobile insured in the exposure $\boldsymbol{w}$, which is the number of insured in policy-years. The factors thought to be important in modeling the occurrence of claims are three: Distance (Kilometers Travelled), Bonus (No-claims bonus) and Make (specified car makes). The number of levels of each factor are 5, 7 and 9 respectively. Distance and Bonus are continuous numerical predictors and we have coded numerically versions of them as follows:

We have represented each state of Distance by a class mark. Central classes are represented by the interval average, whereas class marks for the extreme classes are reasonably representative values. The codes are:

$<$ 1000 *Km per year*: 750 Kilometers travelled per year

1000 − 15000 Km per year: 8000 Kilometers travelled per year

15000 − 20000 Km per year: 17500 Kilometers travelled per year

20000 − 25000 Km per year: 22500 Kilometers travelled per year

$>$ 25000 *Km per year*: 40000 Kilometers travelled per year

Bonus is represented by the (arbitrary) numerical codes, 1 to 7. Insured starts in the class 1 and is moved up one class (to a maximum of 7) each year there is no claim.

Make will be considered as a nominal categorical variable in Gower's formula (11). It is coded numerically (as 1 to 9) just as a programming convenience. It represents 9 specified car makes.

```
R> library(dbstats)
R> require(faraway)
R> data(motorins)
R> Motor1 <- subset(motorins, Zone == 1)
R> Motor1$frequency <- Motor1$Claims / Motor1$Insured
R> y <- Motor1$frequency
R> w <- Motor1$Insured
R> Motor1$KmC <- rep(0,nrow(Motor1))
R> Motor1$KmC[Motor1$Kilometres == "1"] <-   750
R> Motor1$KmC[Motor1$Kilometres == "2"] <-  8000
R> Motor1$KmC[Motor1$Kilometres == "3"] <- 17500
R> Motor1$KmC[Motor1$Kilometres == "4"] <- 22500
R> Motor1$KmC[Motor1$Kilometres == "5"] <- 40000
R> Motor1$BonC  <- as.numeric(Motor1$Bonus)
R> Motor1$MakeC <- as.numeric(Motor1$Make)
```

The first step in the treatment of these data by DB-GLM is the choice of a suitable metric. In principle it is possible to tailor a metric to reflect specific information on predictors and on how their proximity relates to the particular prediction under study. Here it is sufficient to utilize an omnibus metric function which satisfies the Euclidean condition. One very popular such metric for mixtures of numerical continuous, categorical and binary predictor variables is the one based on Gower's general similarity coefficient (see Gower (1971) for further details):

$$s_{ij} = \frac{\sum_{h=1}^{p_1} (1 - |x_{ih} - x_{jh}|/G_h) + a + \alpha_{ij}}{p_1 + (p_2 - d) + p_3} \tag{11}$$

where p_1 is the number of continuous variables, a and d are the number of positive and negative matches, respectively, for the p_2 binary variables, and α_{ij} is the number of matches for the p_3 multi-state categorical variables. G_h is the range of the h-th continuous variable. The squared distance is computed as: $\delta_{ij}^2 = 1 - s_{ij}$. Gower (1971) proves that this distance satisfies the Euclidean condition. In our example, $p_1 = 2$, $p_2 = 0$ and $p_3 = 1$ in (11).

For GLM, we use 11 parameters: 2 for Distance and Bonus with the class marks defined above and 9 binary variables for each nominal class of Make, taking into account an intercept term. Both in GLM and DB-GLM, we assume both Poisson and Binomial

distributions for claim frequency, combined with its associated canonical links. The weights of regressions are the exposures $\boldsymbol{w}$.

We fit three cases for the DB-GLM:

1. `rel.gvar = 1`, i.e., we take into account for the model all the dimensions of the latent Euclidean configurations;

2. `method = "GCV"`, i.e., we choose the effective rank which minimizes a generalized cross-validation (leave-one-out) statistic;

3. `rel.gvar = 0.90`, i.e., we take into account for the model the 90% of explained geometric variability. In both cases (Poisson and Binomial) coincide with the choice of taking into account an effective rank of 10, which coincide with the number of parameters used in the fitted GLM (without counting the intercept term).

We obtain in both cases (the Poisson and the Binomial ones), lower residual deviances with the distance-based treatment of the GLM than those obtained with the classical GLM, see Tables 1 and 2. The detailed instructions used to elaborate these tables for the function `dbglm` can be found in Annex A.

To illustrate the `summary` command, we choose a DB-GLM, the one with Poisson response and Logarithmic link, using Gower's distance and fitted taking into account the `"GCV"` method. In Annex A it is the one named *dbglm2*. We show the results too for *dbglm4* which corresponds to the case when we fit the Poisson with Logarithmic link, using the Euclidean distance and taking into account the complete geometric variability. That case is relevant because it is the particular case in which the results coincide with the classical GLM assuming Poisson and Logarithmic link (named *glm1* in Annex A). Similarly, the `summary` of *dbglm8* coincides with the `summary` of *glm2* for Binomial response and Logit link (see Annex A)

If we compare the output of the `summary` command of *dbglm4* and *glm1* below, we can observe that both are similarly programmed. The main difference is that in `dbglm` we have not estimations of coefficients, because DB-GLM does not assign a coefficient to each explanatory variable.

```
R> summary(dbglm2)

Call:  dbglm.formula(formula = y ~ KmC + BonC + factor(MakeC),
    data = Motor1, family = poisson(link = "log"), ... =
    list(method = "GCV"), metric = "gower", weights = w)
```

```
Deviance Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-6.39000 -0.71300  0.05910  0.02042  0.83250  6.72400

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6977.53  on 294 degrees of freedom
Residual deviance: 485.72  on 281 degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 11
Convergence criterion: muStat
```

R> summary(dbglm4)

```
Call:  dbglm.formula(formula = y ~ KmC + BonC + factor(MakeC),
    data = Motor1, family = poisson(link = "log"), metric =
    "euclidean", weights = w, rel.gvar = 1)

Deviance Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-6.51300 -0.89800 -0.06430 -0.06486  0.80760 10.09000

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6977.53  on 294 degrees of freedom
Residual deviance: 779.36  on 284 degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 100
Convergence criterion: MaxIter
```

R> summary(glm1)

```
Call:
glm(formula = y ~ KmC + BonC + factor(MakeC), family =
    poisson(link = "log"), data = Motor1, weights = w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.5134  -0.8980  -0.0643   0.8076  10.0902

Coefficients:
```

```
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.640e+00  2.637e-02 -62.181  < 2e-16 ***
KmC             1.431e-05  6.381e-07  22.424  < 2e-16 ***
BonC           -2.165e-01  2.762e-03 -78.387  < 2e-16 ***
factor(MakeC)2  1.282e-01  4.598e-02   2.788  0.00531 **
factor(MakeC)3 -2.140e-01  5.162e-02  -4.146 3.38e-05 ***
factor(MakeC)4 -5.162e-01  4.987e-02 -10.352  < 2e-16 ***
factor(MakeC)5  1.270e-01  4.850e-02   2.618  0.00883 **
factor(MakeC)6 -3.976e-01  4.467e-02  -8.900  < 2e-16 ***
factor(MakeC)7 -1.320e-01  5.891e-02  -2.240  0.02508 *
factor(MakeC)8  1.396e-01  8.673e-02   1.609  0.10762
factor(MakeC)9 -3.079e-02  2.276e-02  -1.353  0.17618
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6977.53  on 294  degrees of freedom
Residual deviance:  779.36  on 284  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 5
```

We calculate the mean difference of `fitted.values` of *glm1* and *glm4*:

```
R> mean(glm1$fitted.values - dbglm4$fitted.values)

[1] -2.521631e-09
```

obtaininig that they coincide, up to a negligeable value.

Respect to the `plot` command, its usage is:

```
plot(x, which = c (1:3, 5), id.n = 3, main = "", cook.levels =
  c(0.5, 1), cex.id = 0.75, type_glm = c("link", "response"), ...)
```

The arguments are:

x an object of class `dblm` or `dbglm`.

which if a subset of the plots is required, specify a subset of the numbers 1:6.

id.n number of points to be labelled in each plot, starting with the most extreme.

main an overall title for the plot. Only if one of the six plots is selected.

cook.levels levels of Cook's distance at which to draw contours.

cex.id magnification of point labels.

type_glm the type of prediction (required only for a `dbglm` class object). Like `predict.dbglm`, the default `"link"` is on the scale of the linear predictors; the alternative `"response"` is on the scale of the response variable.

The first five plots are useful for residual analysis and are the same as in `plot.lm`. The last plot allows us to view the `"OCV"`, `"GCV"`, `"AIC"` or `"BIC"` criterion according to which the rank used `dblm` function has been chosen. It applies only if the parameter `full_search` in `dblm` its `TRUE`.

It is easy to get the predicted mean values, as these are calculated by the inverse link function on the linear predictors. We refer to the `R` function `family` to view how to insert user-defined `linkfun` and `linkinv` in `dbstats`.

To illustrate the `plot` command, we exhibit for model *dbglm2* the five possible plots: `Residuals vs Fitted`, `Normal Q-Q`, `Scale-location`, `Cook's distance` and `Residuals vs Leverage`, which can be found in Figures 1, 2, 3, 4 and 5 respectively.

```
R> plot(dbglm2, type_glm = "response", which = 1:5)
```

Values predicted with `predict` may be the expected mean values of the response for the new data (`type="response"`), or the linear predictors evaluated at the estimated `dblm` of the last iteration, as is in the `plot` command above. Additionally, we can choose the type of the new data, which can be: `"Z"` if `newdata` contains the values of the explanatory variables, `"D2"` if contains the squared distances matrix or `"G"` if contains the inner products matrix. Its usage is:

```
predict(object, newdata, type = c("link","response"), type_var = "Z",...)
```

With model *dbglm2*, for data in the original set, such as insureds with 750 kilometers travelled per year, with Bonus and Make both in class 1, we can execute:

```
R> newdata <- data.frame(KmC = 750, BonC = 1, MakeC = 1)
R> pr <- predict(dbglm2, newdata, type = "response", type_var = "Z"); pr
```

```
          [,1]
[1,] 0.1711442
```

Poisson / Logarithmic	Residual Deviance	Eff.rank
DB-GLM (`rel.gvar = 1`)	454.05	18
DB-GLM (`method = "GCV"`)	485.72	13
DB-GLM (`rel.gvar = 0.90`)	539.55	10
GLM	779.36	10

Table 1: Results of the fitting for the Poisson model with the Logarithmic link

Binomial / Logit	Residual Deviance	Eff.rank
DB-GLM (`rel.gvar = 1`)	498.42	18
DB-GLM (`method = "GCV"`)	538.05	13
DB-GLM (`rel.gvar = 0.90`)	595.74	10
GLM	889.07	10

Table 2: Results of the fitting for the Binomial model with the Logit link

Comparing this prediction with the corresponding one from `fitted.values`, we confirm that both values agree (with a precission of 1e-15).

```
R> dbglm2$fitted.values[1] - pr

              [,1]
[1,] 4.746203e-15
```

Now, we compute the prediction for a new insured, e.g., with 900 kilometers travelled per year, with Bonus and Make both in class 1, and we obtain the prediction:

```
R> newdata <- data.frame(KmC = 900, BonC = 1, MakeC = 1)
R> pr <- predict(dbglm2, newdata, type = "response", type_var = "Z"); pr

          [,1]
[1,] 0.1717889
```

4.2. An example of local DB-GLM

In this section we make an example using local DB-GLM with functional data as explanatory variable and a binary response. We say that an observed variable is functional when a whole function is registered for each individual in the sample (see Ramsay and Silverman (2005) for a general perspective on Functional Data Analysis and Ferraty and Vieu (2006a) for a nonparametric approach).

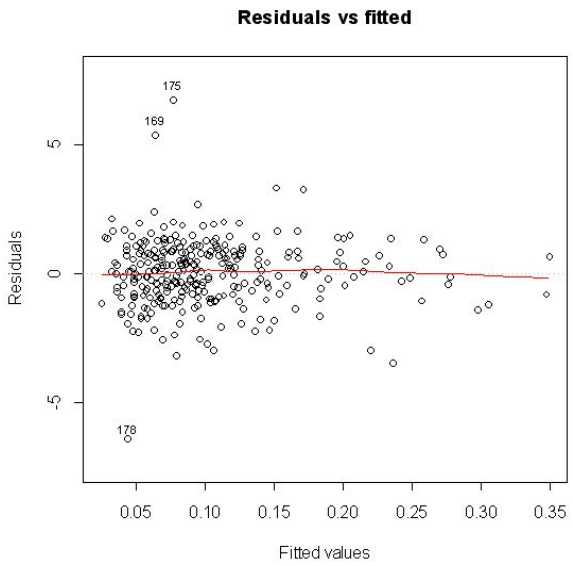

Figure 1: `Residuals vs Fitted` plot for DB-GLM with Poisson response and Logarithmic link, using Gower's distance and fitted taking into account the `GCV` method

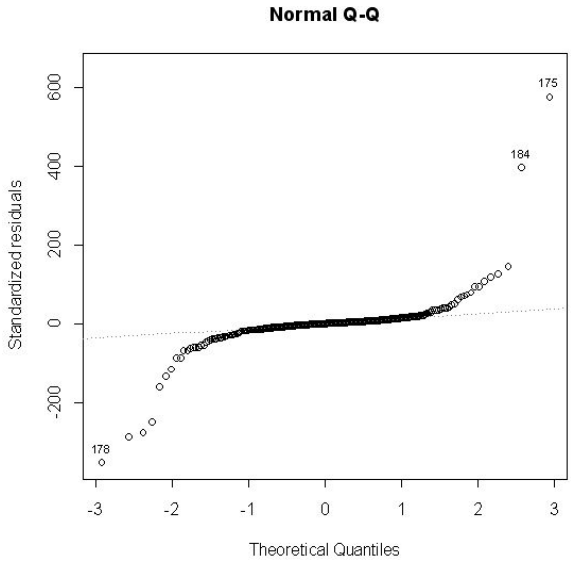

Figure 2: `Normal Q-Q` plot plot for DB-GLM with Poisson response and Logarithmic link, using Gower's distance and fitted taking into account the `GCV` method

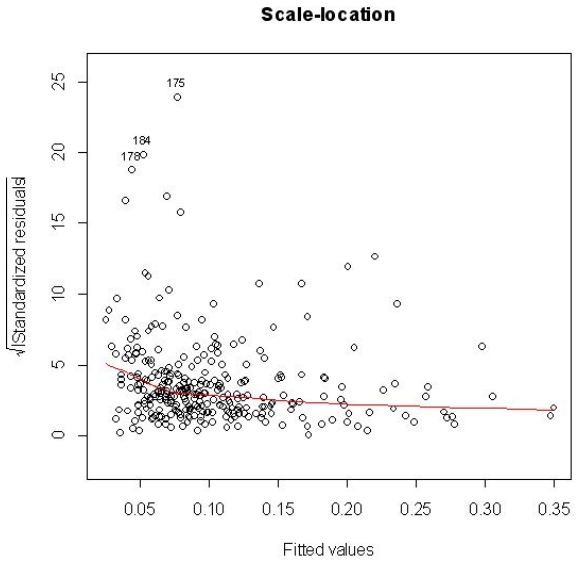

Figure 3: `Scale-location` plot for DB-GLM with Poisson response and Logarithmic link, using Gower's distance and fitted taking into account the `GCV` method

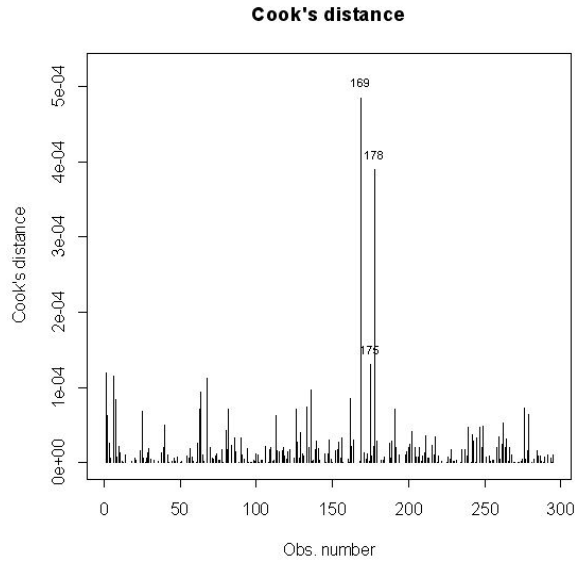

Figure 4: `Cook's distance` plot for DB-GLM with Poisson response and Logarithmic link, using Gower's distance and fitted taking into account the `GCV` method

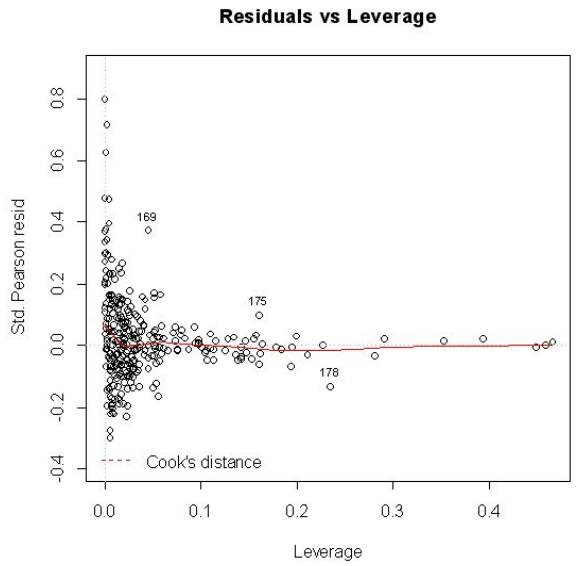

Figure 5: `Residuals vs Leverage` plot for DB-GLM with Poisson response and Logarithmic link, using Gower's distance and fitted taking into account the `GCV` method

We consider the near infrared (NIR) spectral data set contending with wheat samples that was described in Kalivas (1997). This data set contains data from 100 wheat samples. The available information for each sample consists of two scalar measures (protein and moisture contents; only protein content are used here) and a functional variable, the NIR spectra: samples were measured using diffuse reflection in units of log inverse reflectance $\log(1/R)$ at wavelengths going from 1100 to 2500 nm in 2 nm intervals (reflectance refers to the fraction of incident electromagnetic power that is reflected by the sample; see Brenchley, Hörchner, and Kalivas (1997) for more details about NIR measurements). The protein and spectrum data are available at `ftp://ftp.clarkson.edu/users/h/o/hopkepk/chemdata/kalivas/`, at files `protein.asc` and `whtspec.asc`, respectively.

Let us define the binary variable `y` indicating for each wheat sample in the data set whether its protein content is over the median value or not:

```
R> whtspec <- read.table("whtspec.asc")
R> protein <- read.table("protein.asc")
R> wave.length <- seq(1100, 2500, by = 2)
R> y <- as.numeric(protein > median(protein[ , 1]))
```

Our goal is to predict the variable `y` using the NIR spectra function (`whtspec`) as predictor.

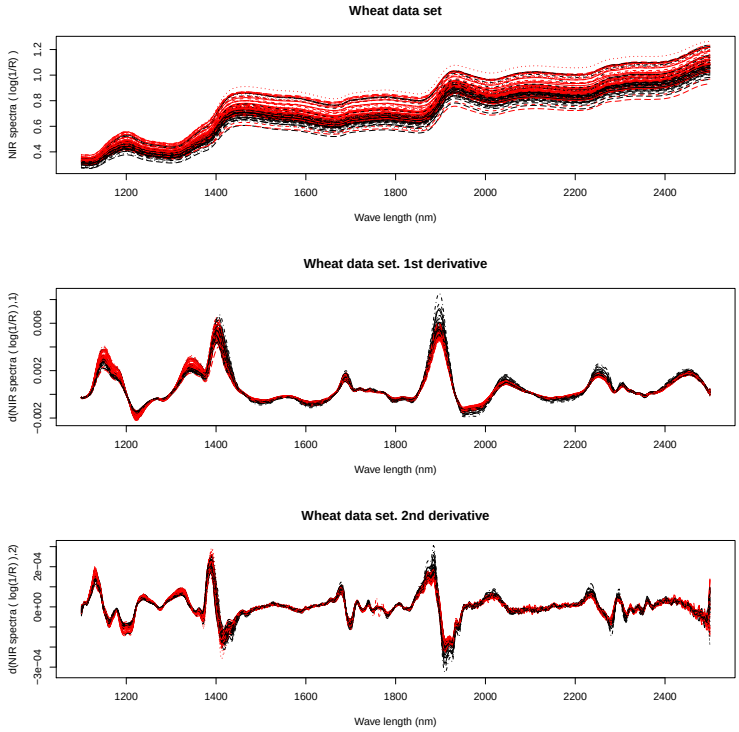

Figure 6: Wheat data set. NIR spectra functions, jointly with their first and second derivatives. Functions in red corresponds to wheat samples with protein content over the median.

We use the `R` package `fda.usc` (Febrero-Bande and Oviedo 2012) to deal with NIR spectra data as functional data:

```
R> library(fda.usc)
R> whtspec.fdata <- fdata(mdata = whtspec, argvals = wave.length, names
   = list(main = "Wheat data set", xlab = "Wave length (nm)", ylab =
   "NIR spectra ( log(1/R) )"))
R> plot(whtspec.fdata, col = y+1)
R> plot(fdata.deriv(whtspec.fdata, nderiv = 1), col = y+1, main =
   "Wheat data set. 1st derivative")
R> plot(fdata.deriv(whtspec.fdata, nderiv = 2), col = y+1, main=
   "Wheat data set. 2nd derivative")
```

This way the functions, as well as their first and second derivatives, are plotes as Figure 6 shows. Wheat samples have been colored according to the value of the binary variable `y` (red when `y == 1`). From the figure it is not obvious how NIR spectra functions or their derivatives could allows us to predict the value of `y`, the indicator of high protein content.

In order to measure the prediction ability of a given prediction rule we randomly divide the data set into a training set (with 60 wheat samples) and a validation set (with the remainig 40 samples):

```
R> set.seed(2)
R> trai <- sample(1:100)[1:60]
```

We will compare the performance of the following binary prediction tools, all of them using functional predictors:

- Generalized linear model with functional predictor using basis representation, as it is implemented in function `fregre.glm`, form package `fda.usc`.
- DB-GLM: Distance based generalized linear model, developed in Section 2.2.
- Local DB-GLM: Local distance based generalized linear model, developed in Section 2.3.

For each of these three prediction tools we have used three different functional predictors: NIR spectra functions, their first derivatives and their second derivatives. The measure of prediction quality that we are using for comparing the 9 procedures under consideration will be the number of bad classified wheat samples among the 40 in the validation set.

The following code was used to prepare the functional data to fit the functional GLM with `fregre.glm`:

```
R> yt <- y[trai]
R> yt.df <- as.data.frame(yt); names(yt.df) <- "protein"
R> yv <- y[-trai]
R> yv.df <- as.data.frame(yv); names(yv.df) <- "protein"
R> wst.d0 <- whtspec.fdata[trai,]
R> wsv.d0 <- whtspec.fdata[-trai, ]
R> wst.d1 <- fdata.deriv(wst.d0, nderiv = 1)
R> wsv.d1 <- fdata.deriv(wsv.d0, nderiv = 1)
R> wst.d2 <- fdata.deriv(wst.d0, nderiv = 2)
R> wsv.d2 <- fdata.deriv(wsv.d0, nderiv = 2)
R> ldata <- list("df" = yt.df, "ws.d0" = wst.d0, "ws.d1" = wst.d1,
   "ws.d2" = wst.d2)
R> newldata <- list("df" = yv.df, "ws.d0" = wsv.d0, "ws.d1" = wsv.d1,
   "ws.d2" = wsv.d2)
R> basis1 <- create.bspline.basis(rangeval = range(wave.length),
   nbasis = 18)
R> basis2 <- create.bspline.basis(rangeval = range(wave.length),
   nbasis = 18)
```

	Prediction tool		
	Functional glm	DB-GLM	local DB-GLM
Functional predictor:			
NIR spectra functions	9	14	8
First derivative	12	10	8
Second derivative	14	12	11

Table 3: Number of bad classified wheat samples among the 40 in the validation set.

```
R> basis.ws <- list("x" = basis1); basis.b <- list("x" = basis2)
```

The choice of a basis of B-splines with 18 elements (`nbasis = 18`) is arbitrary and it could be improved by using a choice based on leave-one-out prediction error criterium. These code lines allows us to call the function `fregre.glm` when the functional predictor used is observed NIR spectra function:

```
R> f0 <- protein~ws.d0
R> res.glm.0 <- fregre.glm(f0, ldata, family = binomial, basis.x =
   basis.ws, basis.b = basis.b)
R> pred.glm.0 <- predict.fregre.glm(res.glm.0, newldata)
R> glm.err.0 <- (abs(pred.glm.0 - yv) > .5)
R> print(sum(glm.err.0))
```

In order to use the first or the second derivatives the definition of formula `f0` must be modified as folows:

```
R> f1 <- protein~ws.d1
R> f2 <- protein~ws.d2
```

Then the call to `fregre.glm` must be changed accordingly (see Annex B). The first column of Table 3 shows the number of wheat samples in the validation set that are bad clssified when using the functional GLM. It can be seen that the best results are obtained when using the observed functions.

In order to fit a DB-GLM with the function `dbglm` the first step is to compute the inter-individual distance matrix. When dealing with functional data our choice is to use one of the semimetrics defined in Ferraty and Vieu (2006a) as they are implemented in their own `R` library NPFDA (Ferraty and Vieu (2006b); free access on line at `http://www.lsp.ups-tlse.fr/staph/npfda/`), also available at the package `fda.usc` (see functions `metric.lp`, `semimetric.basis` and `semimetric.NPFDA` in this package). In particular here we use L^2 distances between NIR spectra functions or their derivatives calculated after representing functions in a B-spline basis.

The following code was used to fit the functional DB-GLM with `dbglm`. NIR spectra functions are used as predictors:

```
R> ws.d0 <- whtspec.fdata
R> D2.0 <- semimetric.basis(ws.d0, ws.d0, nderiv = 0, nbasis1 = 18,
   nbasis2 = 18) ^2
R> D2.0.trai <- D2.0[trai,trai]
R> class(D2.0) <- "D2"; class(D2.0.trai) <- "D2"
R> res.dbglm.0 <- dbglm(D2.0.trai, y = yt, family = "binomial",
   maxiter = 25, rel.gvar = 0.9)
R> pred.dbglm.0 <- predict(res.dbglm.0, D2.0[-trai,trai],
   type_var = "D2", type = "response")
R> dbglm.err.0 <- (abs(pred.dbglm.0 - yv) > .5)
R> print(sum(dbglm.err.0))
```

In order to use the first or the second derivatives, the parameter `nderiv` at the second sentence must be set equal to 1 or 2, respectively, to obtain the corresponding distance matrices `D2.1`, `D2.1.trai`, `D2.2` and `D2.0.trai`. The rest of the code must be changed accordingly (see Annex B). The results of DB-GLM fits, in term of the number of wheat samples in the validation set that are bad classified, are shown at the second column of Table 3. In this case the use of derivatives improves the results. The performances of functional GLM and DB-GLM are comparable.

We are fitting now local distance based generalized linear models (local DB-GLM) with the function `ldbglm`. We are using again L^2 distances between NIR spectra functions or their derivatives. The automatic choice of the smoothing parameter h will be done with the Generalized Cross Validation criterium (`method="GCV"`; results using different methods are similar).

We show the code used for fitting the local DB-GLM when NIR spectra functions are used as predictors.

```
R> res.ldbglm.0 <- ldbglm(D2.0.trai, y = yt, family = "binomial",
   maxiter = 25, method = "GCV", h.range = c(2,4), rel.gvar = 0.9)
```

Observe that it has been necessary to modify the default range of candidate values for h (`h.range = c(2,4)`). By default the range for h was $[.5, 2]$ and the optimum value for h was the upper limit of this interval. Plotting the fitted DB-GLM model (using `plot(res.ldbglm.0, which = 3)`) it can be seen that the range $[2, 4]$ is addequate for h. The optimal value is attained at $h = 2.3331$, as it can be seen when doing the summary of the fitted model:

```
R > summary(res.ldbglm.0)
```

```
call:   ldbglm.D2(D2_1 = D2.0.trai, y = yt, family = "binomial", method = "GCV",
    h.range = c(2, 4), rel.gvar = 0.9, maxiter = 25)

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.7100 -0.2875 -0.0550 -0.0150  0.2150  0.8400

Number of Observations: 60
R-squared : 0.3852
Trace of smoother matrix: 4.74
family: binomial


kind of kernel= (1) Epanechnikov
optimal bandwidth h : 2.333058
GCV value criterion : 3.323117e-03
```

Now the prediction for the validation set is done:

```
R> pred.ldbglm.0 <- predict(res.ldbglm.0, D2.0[-trai,trai],
   type_var = "D2", type = "response")$fit
R> ldbglm.err.0 <- (abs(pred.ldbglm.0 - yv) > .5)
R> print(sum(ldbglm.err.0))
```

The number of bad classified wheat samples is 8, as it can be seen at the third column of Table 3. In order to use the first or the second derivatives of NIR spectra functions as predictors, distance matrices `D2.0` and `D2.0.trai` must be replaced by `D2.1` and `D2.1.trai`, or by `D2.2` and `D2.0.trai`. The rest of the code must be changed accordingly (see Annex B). Care must be taken when choosing a sensible range where bandwidth h must be, because the range of values for distances between the observed functions is quite different to that corresponding to their first or second derivatives. In this case the range used h when using first derivatives whas `h.range=c(0.003,0.007)` and it was `h.range=c(0.0004,0.001)` when using second derivatives. Look at the third column of Table 3 to see the number of bad classified samples in the validation set. It follows from these number that the three local DB-GLM fits performs similarly and that they do a little better job than functional GLM or global DB-GLM.

Acknowlegments

Work supported in part by the Spanish Ministerio de Educación y Ciencia and FEDER, grants MTM2010-17323 and MTM2010-14887, and by Generalitat de Catalunya, AGAUR

grant 2009SGR970.

A. Code excerpts

```
R> ###### Poisson (Logarithmic link) and Gower's distance
R> ## Case: rel.gvar = 1
R> dbglm1 <- dbglm(y ~ KmC + BonC + factor(MakeC), Motor1,
 family = poisson (link = "log"), metric = "gower", weights
 = w, rel.gvar = 1)
R> ## Case: method = "GCV"
R> dbglm2 <- dbglm(y ~ KmC + BonC + factor(MakeC), Motor1,
 family = poisson (link = "log"), metric = "gower", weights
 = w, method= "GCV", full_search=TRUE)
R> ## Case: rel.gvar = 0.90
R> dbglm3 <- dbglm(y ~ KmC + BonC + factor(MakeC), Motor1,
 family = poisson(link = "log"), metric = "gower", weights
 = w, rel.gvar = 0.90)

R> ###### Poisson (Logarithmic link) and Euclidean distance
R> ## With dbglm, case: rel.gvar = 1
R> dbglm4 <- dbglm(y ~ KmC + BonC + factor(MakeC), Motor1,
 family = poisson(link = "log"), metric = "euclidean", weights
 = w, rel.gvar = 1)
R> ## With glm
R> glm1 <- glm(y ~ KmC + BonC + factor(MakeC), family =
 poisson(link = "log"), data = Motor1, weights = w)

R> ###### Binomial (Logit link) and Gower's distance
R> ## rel.gvar = 1
R> dbglm5 <- dbglm(y ~ KmC + BonC + factor(MakeC), Motor1,
 family = binomial(link = "logit"), metric = "gower", weights
 = w, rel.gvar = 1, full_search=TRUE)
R> ## method = "GCV"
R> dbglm6 <- dbglm(y ~ KmC + BonC + factor(MakeC), Motor1,
 family = binomial(link = "logit"), metric = "gower", weights
 = w, method = "GCV")
R> ## rel.gvar = 0.90
R> dbglm7 <- dbglm(y ~ KmC + BonC + factor(MakeC), Motor1,
 family = binomial(link = "logit"), metric = "gower", weights
 = w, rel.gvar = 0.90)
```

```
R> ###### Binomial (Logit link) and Euclidean distance
R> ## With dbglm, case: rel.gvar = 1
R> dbglm8 <- dbglm(y ~ KmC + BonC + factor(MakeC), Motor1,
 family = binomial(link = "logit"), metric = "euclidean", weights
 = w, rel.gvar = 1)
R> ## With glm
R> glm2 <- glm(y ~ KmC + BonC + factor(MakeC), family =
 binomial(link = "logit"), data = Motor1, weights = w)
```

B. Code excerpts

```
## fregre.glm. Functional predictor:
## 1st derivative of NIR spectra function
R> f1 <- protein~ws.d1
R> res.glm.1 <- fregre.glm(f1, ldata, family = binomial,
 basis.x = basis.ws, basis.b = basis.b)
R> pred.glm.1 <- predict.fregre.glm(res.glm.1, newldata)
R> glm.err.1 <- (abs(pred.glm.1 - yv) > .5)
R> print(sum(glm.err.1))

## fregre.glm. Functional predictor:
## 2nd derivative of NIR spectra function
R> f2 <- protein~ws.d2
R> res.glm.2 <- fregre.glm(f2, ldata, family = binomial,
 basis.x = basis.ws, basis.b = basis.b)
R> pred.glm.2 <- predict.fregre.glm(res.glm.2, newldata)
R> glm.err.2 <- (abs(pred.glm.2 - yv) > .5)
R> print(sum(glm.err.2))

## dbglm. Functional predictor:
## 1st derivative of NIR spectra function
R> ws.d1 <- fdata.deriv(whtspec.fdata)
R> D2.1 <- semimetric.basis(ws.d1, ws.d1, nbasis1 = 18,
 nbasis2 = 18) ^2
R> class(D2.1) <- "D2"
R> D2.1.trai <- D2.1[trai,trai]
R> class(D2.1.trai) <- "D2"
R> res.dbglm.1 <- dbglm(D2.1.trai, y = yt, family = "binomial",
 maxiter = 25, rel.gvar = 0.9)
R> pred.dbglm.1 <- predict(res.dbglm.1, D2.1[-trai,trai],
 type_var = "D2", type = "response")
```

```
R> dbglm.err.1 <- (abs(pred.dbglm.1 - yv) > .5)
R> print(sum(dbglm.err.1))

## dbglm. Functional predictor:
## 2nd derivative of NIR spectra function
R> ws.d2 <- fdata.deriv(ws.d1)
R> D2.2 <- semimetric.basis(ws.d2, ws.d2, nbasis1 = 18,
 nbasis2 = 18) ^2
R> class(D2.2) <- "D2"
R> D2.2.trai <- D2.2[trai,trai]
R> class(D2.2.trai) <- "D2"
R> res.dbglm.2 <- dbglm(D2.2.trai, y = yt, family = "binomial",
 maxiter = 25, rel.gvar = 0.9)
R> pred.dbglm.2 <- predict(res.dbglm.2, D2.2[-trai,trai],
 type_var = "D2", type = "response")
R> dbglm.err.2 <- (abs(pred.dbglm.2 - yv) > .5)
R> print(sum(dbglm.err.2))

## ldbglm. Functional predictor:
## 1st derivative of NIR spectra function
R> res.ldbglm.1 <- ldbglm(D2.1.trai, y = yt, family = "binomial",
 maxiter = 25, method = "GCV", h.range = c(0.003,0.007),
 rel.gvar = 0.9)
R> pred.ldbglm.1 <- predict(res.ldbglm.1, D2.1[-trai,trai],
 type_var = "D2", type = "response")$fit
R> ldbglm.err.1 <- (abs(pred.ldbglm.1 - yv) > .5)
R> print(sum(ldbglm.err.1))

## ldbglm. Functional predictor:
## 2nd derivative of NIR spectra function
R> res.ldbglm.2 <- ldbglm(D2.2.trai, y = yt, family = "binomial",
 maxiter = 25, method = "GCV", h.range = c(0.0004,0.001),
 rel.gvar = 0.9)
R> pred.ldbglm.2 <- predict(res.ldbglm.2, D2.2[-trai,trai],
 type_var = "D2", type = "response")$fit
R> ldbglm.err.2 <-  (abs(pred.ldbglm.2 - yv) > .5)
R> print(sum(ldbglm.err.2))
```

References

Andrews, D. F. and A. M. Herzberg (1985). *Data. A collection of problems from many fields for the student and research worker*. New York, NY, USA: Springer.

Boj, E., A. Caballé, P. Delicado, and J. Fortiana (2012). *dbstats: Distance-based statistics (dbstats)*. R package version 1.0.2.

Boj, E., M. M. Claramunt, and J. Fortiana (2007). Selection of predictors in distance-based regression. *Communications in Statistics A. Theory and Methods 36*, 87–98.

Boj, E., M. M. Claramunt, A. Grané, and J. Fortiana (2007). Implementing pls for distance-based regression: computational issues. *Computational Statistics 22*, 237–248.

Boj, E., P. Delicado, and J. Fortiana (2010). Local linear functional regression based on weighted distance-based regression. *Computational Statistics and Data Analysis 54*, 429–437.

Bowman, A. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford University Press.

Brenchley, J. M., U. Hörchner, and J. H. Kalivas (1997). Wavelength selection characterization for nir spectra. *Applied Spectroscopy 51*, 689–699.

Cuadras, C. and C. Arenas (1990). A distance-based regression model for prediction with mixed data. *Communications in Statistics A. Theory and Methods 19*, 2261–2279.

Cuadras, C. M. (1989). Distance analysis in discrimination and classification using both continuous and categorical variables. In Y. Dodge (Ed.), *Statistical Data Analysis and Inference*, Amsterdam, The Netherlands, pp. 459–473. North-Holland Publishing Co.

Cuadras, C. M., C. Arenas, and J. Fortiana (1996). Some computational aspects of a distance-based model for prediction. *Communications in Statistics B. Simulation and Computation 25*, 593–609.

Esteve, A., E. Boj, and J. Fortiana (2009). Interaction terms in distance-based regression. *Communications in Statistics A. Theory and Methods 38*, 3498–3509.

Faraway, J. (2012). *faraway: Functions and datasets for books by Julian Faraway*. R package version 1.0.5.

Febrero-Bande, M. and M. Oviedo (2012). *fda.usc: Functional Data Analysis and Utilities for Statistical Computing (fda.usc)*. R package version 0.9.7.

Ferraty, F. and P. Vieu (2006a). *Non parametric functional data analysis. Theory and practice*. Springer.

Ferraty, F. and P. Vieu (2006b). *Reference manual for implementing NonParametric Functional Data Analysis (NPFDA)*. Companion manual of the book: Non-Parametric Functional Data Analysis: Theory and Practice, Springer-Verlag (New York), 2006.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics 27*, 857–874.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical*

Society. Series B (Methodological) 46(2), 149–192.

Hallin, M. and J. F. Ingenbleek (1983). The Swedish automobile portfolio in 1977. a statistical study. *Skandinavisk Aktuarietidskrift (Scandinavian Actuarial Journal) 83*, 49–64.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction (2nd ed.)*. Springer.

Kalivas, J. H. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems 37*, 255–259.

Loader, C. (1999). *Local regression and likelihood*. New York: Springer.

Maechler, M. (2012). `cluster`*: Cluster Analysis Extended Rousseeuw et al.* `R` package version 1.14.2.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models (2nd ed)*. London: Chapman and Hall.

Meyer, D. and C. Buchta (2012). `proxy`*: Distance and Similarity Measures*. `R` package version 0.4-7.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis (2nd ed)*. New York: Springer.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. New York, NY, USA: John Wiley & Sons.

Stewart, G. W. (1993). On the early history of the singular values decomposition. *SIAM Review 35*, 551–566.

Street, J. O., R. J. Carroll, and D. Ruppert (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician 42*(2), 152–154.

Wasserman, L. (2006). *All of Nonparametric Statistics*. New York: Springer.

Wood, S. N. (2006). *Generalized Additive models: An Introduction with R*. Boca Raton, FL, USA: Chapman & Hall/CRC.

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

2006

CREAP2006-01
Matas, A. (GEAP)**; Raymond, J.Ll.** (GEAP)
"Economic development and changes in car ownership patterns"
(Juny 2006)

CREAP2006-02
Trillas, F. (IEB)**; Montolio, D.** (IEB)**; Duch, N.** (IEB)
"Productive efficiency and regulatory reform: The case of Vehicle Inspection Services"
(Setembre 2006)

CREAP2006-03
Bel, G. (PPRE-IREA)**; Fageda, X.** (PPRE-IREA)
"Factors explaining local privatization: A meta-regression analysis"
(Octubre 2006)

CREAP2006-04
Fernàndez-Villadangos, L. (PPRE-IREA)
"Are two-part tariffs efficient when consumers plan ahead?: An empirical study"
(Octubre 2006)

CREAP2006-05
Artís, M. (AQR-IREA)**; Ramos, R.** (AQR-IREA)**; Suriñach, J.** (AQR-IREA)
"Job losses, outsourcing and relocation: Empirical evidence using microdata"
(Octubre 2006)

CREAP2006-06
Alcañiz, M. (RISC-IREA)**; Costa, A.; Guillén, M.** (RISC-IREA)**; Luna, C.; Rovira, C.**
"Calculation of the variance in surveys of the economic climate"
(Novembre 2006)

CREAP2006-07
Albalate, D. (PPRE-IREA)
"Lowering blood alcohol content levels to save lives: The European Experience"
(Desembre 2006)

CREAP2006-08
Garrido, A. (IEB)**; Arqué, P.** (IEB)
"The choice of banking firm: Are the interest rate a significant criteria?"
(Desembre 2006)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

CREAP2006-09
Segarra, A. (GRIT)**; Teruel-Carrizosa, M.** (GRIT)
"Productivity growth and competition in spanish manufacturing firms:
What has happened in recent years?"
(Desembre 2006)

CREAP2006-10
Andonova, V.; Díaz-Serrano, Luis. (CREB)
"Political institutions and the development of telecommunications"
(Desembre 2006)

CREAP2006-11
Raymond, J.L.(GEAP)**; Roig, J.L..** (GEAP)
"Capital humano: un análisis comparativo Catalunya-España"
(Desembre 2006)

CREAP2006-12
Rodríguez, M.(CREB)**; Stoyanova, A.** (CREB)
"Changes in the demand for private medical insurance following a shift in tax incentives"
(Desembre 2006)

CREAP2006-13
Royuela, V. (AQR-IREA)**; Lambiri, D.**; **Biagi**, **B.**
"Economía urbana y calidad de vida. Una revisión del estado del conocimiento en España"
(Desembre 2006)

CREAP2006-14
Camarero, M.; Carrion-i-Silvestre, J.LL. (AQR-IREA)**.**;**Tamarit**, **C.**
"New evidence of the real interest rate parity for OECD countries using panel unit root tests with breaks"
(Desembre 2006)

CREAP2006-15
Karanassou, M.; Sala, H. (GEAP)**.**;**Snower** , **D. J.**
"The macroeconomics of the labor market: Three fundamental views"
(Desembre 2006)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

2007

XREAP2007-01
Castany, L (AQR-IREA)**; López-Bazo, E.** (AQR-IREA).;**Moreno** , **R.** (AQR-IREA)
"Decomposing differences in total factor productivity across firm size"
(Març 2007)

XREAP2007-02
Raymond, J. Ll. (GEAP)**; Roig, J. Ll.** (GEAP)
"Una propuesta de evaluación de las externalidades de capital humano en la empresa"
(Abril 2007)

XREAP2007-03
Durán, J. M. (IEB)**; Esteller, A.** (IEB)
"An empirical analysis of wealth taxation: Equity vs. Tax compliance"
(Juny 2007)

XREAP2007-04
Matas, A. (GEAP)**; Raymond, J.Ll.** (GEAP)
"Cross-section data, disequilibrium situations and estimated coefficients: evidence from car ownership demand"
(Juny 2007)

XREAP2007-05
Jofre-Montseny, J. (IEB)**; Solé-Ollé, A.** (IEB)
"Tax differentials and agglomeration economies in intraregional firm location"
(Juny 2007)

XREAP2007-06
Álvarez-Albelo, C. (CREB)**; Hernández-Martín, R.**
"Explaining high economic growth in small tourism countries with a dynamic general equilibrium model"
(Juliol 2007)

XREAP2007-07
Duch, N. (IEB)**; Montolio, D.** (IEB); **Mediavilla, M.**
"Evaluating the impact of public subsidies on a firm's performance: a quasi-experimental approach"
(Juliol 2007)

XREAP2007-08
Segarra-Blasco, A. (GRIT)
"Innovation sources and productivity: a quantile regression analysis"
(Octubre 2007)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

XREAP2007-09
Albalate, D. (PPRE-IREA)
"Shifting death to their Alternatives: The case of Toll Motorways"
(Octubre 2007)

XREAP2007-10
Segarra-Blasco, A. (GRIT); **Garcia-Quevedo, J.** (IEB); **Teruel-Carrizosa, M.** (GRIT)
"Barriers to innovation and public policy in catalonia"
(Novembre 2007)

XREAP2007-11
Bel, G. (PPRE-IREA); **Foote, J.**
"Comparison of recent toll road concession transactions in the United States and France"
(Novembre 2007)

XREAP2007-12
Segarra-Blasco, A. (GRIT);
"Innovation, R&D spillovers and productivity: the role of knowledge-intensive services"
(Novembre 2007)

XREAP2007-13
Bermúdez Morata, Ll. (RFA-IREA); **Guillén Estany, M.** (RFA-IREA), **Solé Auró, A**. (RFA-IREA)
"Impacto de la inmigración sobre la esperanza de vida en salud y en discapacidad de la población española"
(Novembre 2007)

XREAP2007-14
Calaeys, P. (AQR-IREA); **Ramos, R.** (AQR-IREA), **Suriñach, J**. (AQR-IREA)
"Fiscal sustainability across government tiers"
(Desembre 2007)

XREAP2007-15
Sánchez Hugalbe, A. (IEB)
"Influencia de la inmigración en la elección escolar"
(Desembre 2007)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

2008

XREAP2008-01
Durán Weitkamp, C. (GRIT)**; Martín Bofarull, M.** (GRIT) ; **Pablo Martí, F.**
"Economic effects of road accessibility in the Pyrenees: User perspective"
(Gener 2008)

XREAP2008-02
Díaz-Serrano, L.; Stoyanova, A. P. (CREB)
"The Causal Relationship between Individual's Choice Behavior and Self-Reported Satisfaction: the Case of Residential Mobility in the EU"
(Març 2008)

XREAP2008-03
Matas, A. (GEAP)**; Raymond, J. L.** (GEAP)**; Roig, J. L.** (GEAP)
"Car ownership and access to jobs in Spain"
(Abril 2008)

XREAP2008-04
Bel, G. (PPRE-IREA) **; Fageda, X.** (PPRE-IREA)
"Privatization and competition in the delivery of local services: An empirical examination of the dual market hypothesis"
(Abril 2008)

XREAP2008-05
Matas, A. (GEAP); **Raymond, J. L.** (GEAP); **Roig, J. L.** (GEAP)
"**Job accessibility and employment probability**"
(Maig 2008)

XREAP2008-06
Basher, S. A.; Carrión, J. Ll. (AQR-IREA)
Deconstructing Shocks and Persistence in OECD Real Exchange Rates
(Juny 2008)

XREAP2008-07
Sanromá, E. (IEB)**; Ramos, R.** (AQR-IREA); Simón, H.
Portabilidad del capital humano y asimilación de los inmigrantes. Evidencia para España
(Juliol 2008)

XREAP2008-08
Basher, S. A.; Carrión, J. Ll. (AQR-IREA)
Price level convergence, purchasing power parity and multiple structural breaks: An application to US cities
(Juliol 2008)

XREAP2008-09
Bermúdez, Ll. (RFA-IREA)
A priori ratemaking using bivariate poisson regression models
(Juliol 2008)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

XREAP2008-10
Solé-Ollé, A. (IEB), **Hortas Rico, M.** (IEB)
Does urban sprawl increase the costs of providing local public services? Evidence from Spanish municipalities
(Novembre 2008)

XREAP2008-11
Teruel-Carrizosa, M. (GRIT), **Segarra-Blasco, A.** (GRIT)
Immigration and Firm Growth: Evidence from Spanish cities
(Novembre 2008)

XREAP2008-12
Duch-Brown, N. (IEB), **García-Quevedo, J.** (IEB), **Montolio, D.** (IEB)
Assessing the assignation of public subsidies: Do the experts choose the most efficient R&D projects?
(Novembre 2008)

XREAP2008-13
Bilotkach, V., **Fageda, X.** (PPRE-IREA), **Flores-Fillol, R.**
Scheduled service versus personal transportation: the role of distance
(Desembre 2008)

XREAP2008-14
Albalate, D. (PPRE-IREA), **Gel, G.** (PPRE-IREA)
Tourism and urban transport: Holding demand pressure under supply constraints
(Desembre 2008)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

2009

XREAP2009-01
Calonge, S. (CREB)**; Tejada, O.**
"A theoretical and practical study on linear reforms of dual taxes"
(Febrer 2009)

XREAP2009-02
Albalate, D. (PPRE-IREA)**; Fernández-Villadangos, L.** (PPRE-IREA)
"Exploring Determinants of Urban Motorcycle Accident Severity: The Case of Barcelona"
(Març 2009)

XREAP2009-03
Borrell, J. R. (PPRE-IREA)**; Fernández-Villadangos, L.** (PPRE-IREA)
"Assessing excess profits from different entry regulations"
(Abril 2009)

XREAP2009-04
Sanromá, E. (IEB)**; Ramos, R.** (AQR-IREA), **Simon, H.**
"Los salarios de los inmigrantes en el mercado de trabajo español. ¿Importa el origen del capital humano?"
(Abril 2009)

XREAP2009-05
Jiménez, J. L.; Perdiguero, J. (PPRE-IREA)
"(No)competition in the Spanish retailing gasoline market: a variance filter approach"
(Maig 2009)

XREAP2009-06
Álvarez-Albelo,C. D. (CREB)**, Manresa, A.** (CREB)**, Pigem-Vigo, M.** (CREB)
"International trade as the sole engine of growth for an economy"
(Juny 2009)

XREAP2009-07
Callejón, M. (PPRE-IREA)**, Ortún V, M.**
"The Black Box of Business Dynamics"
(Setembre 2009)

XREAP2009-08
Lucena, A. (CREB)
"The antecedents and innovation consequences of organizational search: empirical evidence for Spain"
(Octubre 2009)

XREAP2009-09
Domènech Campmajó, L. (PPRE-IREA)
"Competition between TV Platforms"
(Octubre 2009)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

XREAP2009-10
Solé-Auró, A. (RFA-IREA)**,Guillén, M.** (RFA-IREA)**, Crimmins, E. M.**
"Health care utilization among immigrants and native-born populations in 11 European countries. Results from the Survey of Health, Ageing and Retirement in Europe"
(Octubre 2009)

XREAP2009-11
Segarra, A. (GRIT)**, Teruel, M.** (GRIT)
"Small firms, growth and financial constraints"
(Octubre 2009)

XREAP2009-12
Matas, A. (GEAP)**, Raymond, J.Ll.** (GEAP), **Ruiz, A.** (GEAP)
"Traffic forecasts under uncertainty and capacity constraints"
(Novembre 2009)

XREAP2009-13
Sole-Ollé, A. (IEB)
"Inter-regional redistribution through infrastructure investment: tactical or programmatic?"
(Novembre 2009)

XREAP2009-14
Del Barrio-Castro, T., **García-Quevedo, J.** (IEB)
"The determinants of university patenting: Do incentives matter?"
(Novembre 2009)

XREAP2009-15
Ramos, R. (AQR-IREA), **Suriñach, J.** (AQR-IREA)**, Artís, M.** (AQR-IREA)
"Human capital spillovers, productivity and regional convergence in Spain"
(Novembre 2009)

XREAP2009-16
Álvarez-Albelo, C. D. (CREB), **Hernández-Martín, R.**
"The commons and anti-commons problems in the tourism economy"
(Desembre 2009)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

2010

XREAP2010-01
García-López, M. A. (GEAP)
"The Accessibility City. When Transport Infrastructure Matters in Urban Spatial Structure"
(Febrer 2010)

XREAP2010-02
García-Quevedo, J. (IEB), **Mas-Verdú, F.** (IEB), **Polo-Otero, J.** (IEB)
"Which firms want PhDs? The effect of the university-industry relationship on the PhD labour market"
(Març 2010)

XREAP2010-03
Pitt, D., **Guillén, M.** (RFA-IREA)
"An introduction to parametric and non-parametric models for bivariate positive insurance claim severity distributions"
(Març 2010)

XREAP2010-04
Bermúdez, Ll. (RFA-IREA), **Karlis, D.**
"Modelling dependence in a ratemaking procedure with multivariate Poisson regression models"
(Abril 2010)

XREAP2010-05
Di Paolo, A. (IEB)
"Parental education and family characteristics: educational opportunities across cohorts in Italy and Spain"
(Maig 2010)

XREAP2010-06
Simón, H. (IEB), **Ramos, R.** (AQR-IREA), **Sanromá, E.** (IEB)
"Movilidad ocupacional de los inmigrantes en una economía de bajas cualificaciones. El caso de España"
(Juny 2010)

XREAP2010-07
Di Paolo, A. (GEAP & IEB), **Raymond, J. Ll.** (GEAP & IEB)
"Language knowledge and earnings in Catalonia"
(Juliol 2010)

XREAP2010-08
Bolancé, C. (RFA-IREA), **Alemany, R.** (RFA-IREA), **Guillén, M.** (RFA-IREA)
"Prediction of the economic cost of individual long-term care in the Spanish population"
(Setembre 2010)

XREAP2010-09
Di Paolo, A. (GEAP & IEB)
"Knowledge of catalan, public/private sector choice and earnings: Evidence from a double sample selection model"
(Setembre 2010)

XREAP2010-10
Coad, A., **Segarra, A.** (GRIT), **Teruel, M.** (GRIT)
"Like milk or wine: Does firm performance improve with age?"
(Setembre 2010)

XREAP2010-11
Di Paolo, A. (GEAP & IEB), **Raymond, J. Ll.** (GEAP & IEB), **Calero, J.** (IEB)
"Exploring educational mobility in Europe"
(Octubre 2010)

XREAP2010-12
Borrell, A. (GiM-IREA), **Fernández-Villadangos, L.** (GiM-IREA)
"Clustering or scattering: the underlying reason for regulating distance among retail outlets"
(Desembre 2010)

XREAP2010-13
Di Paolo, A. (GEAP & IEB)
"School composition effects in Spain"
(Desembre 2010)

XREAP2010-14
Fageda, X. (GiM-IREA), **Flores-Fillol, R.**
"Technology, Business Models and Network Structure in the Airline Industry"
(Desembre 2010)

XREAP2010-15
Albalate, D. (GiM-IREA), **Bel, G.** (GiM-IREA), **Fageda, X.** (GiM-IREA)
"Is it Redistribution or Centralization? On the Determinants of Government Investment in Infrastructure"
(Desembre 2010)

XREAP2010-16
Oppedisano, V., **Turati, G.**
"What are the causes of educational inequalities and of their evolution over time in Europe? Evidence from PISA"
(Desembre 2010)

XREAP2010-17
Canova, L., **Vaglio, A.**
"Why do educated mothers matter? A model of parental help"
(Desembre 2010)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

2011

XREAP2011-01
Fageda, X. (GiM-IREA), **Perdiguero, J.** (GiM-IREA)
"An empirical analysis of a merger between a network and low-cost airlines"
(Maig 2011)

XREAP2011-02
Moreno-Torres, I. (ACCO, CRES & GiM-IREA)
"What if there was a stronger pharmaceutical price competition in Spain? When regulation has a similar effect to collusion"
(Maig 2011)

XREAP2011-03
Miguélez, E. (AQR-IREA); **Gómez-Miguélez, I.**
"Singling out individual inventors from patent data"
(Maig 2011)

XREAP2011-04
Moreno-Torres, I. (ACCO, CRES & GiM-IREA)
"Generic drugs in Spain: price competition vs. moral hazard"
(Maig 2011)

XREAP2011-05
Nieto, S. (AQR-IREA)**, Ramos, R.** (AQR-IREA)
"¿Afecta la sobreeducación de los padres al rendimiento académico de sus hijos?"
(Maig 2011)

XREAP2011-06
Pitt, D., **Guillén, M.** (RFA-IREA), **Bolancé, C.** (RFA-IREA)
"Estimation of Parametric and Nonparametric Models for Univariate Claim Severity Distributions - an approach using R"
(Juny 2011)

XREAP2011-07
Guillén, M. (RFA-IREA), **Comas-Herrera, A.**
"How much risk is mitigated by LTC Insurance? A case study of the public system in Spain"
(Juny 2011)

XREAP2011-08
Ayuso, M. (RFA-IREA), **Guillén, M.** (RFA-IREA), **Bolancé, C.** (RFA-IREA)
"Loss risk through fraud in car insurance"
(Juny 2011)

XREAP2011-09
Duch-Brown, N. (IEB), **García-Quevedo, J.** (IEB), **Montolio, D.** (IEB)
"The link between public support and private R&D effort: What is the optimal subsidy?"
(Juny 2011)

XREAP2011-10
Bermúdez, Ll. (RFA-IREA), **Karlis, D.**
"Mixture of bivariate Poisson regression models with an application to insurance"
(Juliol 2011)

XREAP2011-11
Varela-Irimia, X-L. (GRIT)
"Age effects, unobserved characteristics and hedonic price indexes: The Spanish car market in the 1990s"
(Agost 2011)

XREAP2011-12
Bermúdez, Ll. (RFA-IREA), **Ferri, A.** (RFA-IREA), **Guillén, M.** (RFA-IREA)
"A correlation sensitivity analysis of non-life underwriting risk in solvency capital requirement estimation"
(Setembre 2011)

XREAP2011-13
Guillén, M. (RFA-IREA), **Pérez-Marín, A.** (RFA-IREA), **Alcañiz, M.** (RFA-IREA)
"A logistic regression approach to estimating customer profit loss due to lapses in insurance"
(Octubre 2011)

XREAP2011-14
Jiménez, J. L., **Perdiguero, J.** (GiM-IREA), **García, C.**
"Evaluation of subsidies programs to sell green cars: Impact on prices, quantities and efficiency"
(Octubre 2011)

XREAP2011-15
Arespa, M. (CREB)
"A New Open Economy Macroeconomic Model with Endogenous Portfolio Diversification and Firms Entry"
(Octubre 2011)

XREAP2011-16
Matas, A. (GEAP), **Raymond, J. L.** (GEAP), **Roig, J.L.** (GEAP)
"The impact of agglomeration effects and accessibility on wages"
(Novembre 2011)

XREAP2011-17
Segarra, A. (GRIT)
"R&D cooperation between Spanish firms and scientific partners: what is the role of tertiary education?"
(Novembre 2011)

XREAP2011-18
García-Pérez, J. I.; **Hidalgo-Hidalgo, M.**; **Robles-Zurita, J. A.**
"Does grade retention affect achievement? Some evidence from PISA"
(Novembre 2011)

XREAP2011-19
Arespa, M. (CREB)
"Macroeconomics of extensive margins: a simple model"
(Novembre 2011)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

XREAP2011-20
García-Quevedo, J. (IEB), **Pellegrino, G.** (IEB), **Vivarelli, M.**
"The determinants of YICs' R&D activity"
(Desembre 2011)

XREAP2011-21
González-Val, R. (IEB), **Olmo, J.**
"Growth in a Cross-Section of Cities: Location, Increasing Returns or Random Growth?"
(Desembre 2011)

XREAP2011-22
Gombau, V. (GRIT), **Segarra, A.** (GRIT)
"The Innovation and Imitation Dichotomy in Spanish firms: do absorptive capacity and the technological frontier matter?"
(Desembre 2011)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

2012

XREAP2012-01
Borrell, J. R. (GiM-IREA), **Jiménez, J. L.**, **García, C.**
"Evaluating Antitrust Leniency Programs"
(Gener 2012)

XREAP2012-02
Ferri, A. (RFA-IREA), **Guillén, M.** (RFA-IREA), **Bermúdez, Ll.** (RFA-IREA)
"Solvency capital estimation and risk measures"
(Gener 2012)

XREAP2012-03
Ferri, A. (RFA-IREA), **Bermúdez, Ll.** (RFA-IREA), **Guillén, M.** (RFA-IREA)
"How to use the standard model with own data"
(Febrer 2012)

XREAP2012-04
Perdiguero, J. (GiM-IREA), **Borrell, J.R.** (GiM-IREA)
"Driving competition in local gasoline markets"
(Març 2012)

XREAP2012-05
D'Amico, G., **Guillen, M.** (RFA-IREA), Manca, R.
"Discrete time Non-homogeneous Semi-Markov Processes applied to Models for Disability Insurance"
(Març 2012)

XREAP2012-06
Bové-Sans, M. A. (GRIT), Laguado-Ramírez, R.
"Quantitative analysis of image factors in a cultural heritage tourist destination"
(Abril 2012)

XREAP2012-07
Tello, C. (AQR-IREA), **Ramos, R.** (AQR-IREA), **Artís, M.** (AQR-IREA)
"Changes in wage structure in Mexico going beyond the mean: An analysis of differences in distribution, 1987-2008"
(Maig 2012)

XREAP2012-08
Jofre-Monseny, J. (IEB), **Marín-López, R.** (IEB), **Viladecans-Marsal, E.** (IEB)
"What underlies localization and urbanization economies? Evidence from the location of new firms"
(Maig 2012)

XREAP2012-09
Muñiz, I. (GEAP), **Calatayud, D.**, **Dobaño, R.**
"Los límites de la compacidad urbana como instrumento a favor de la sostenibilidad. La hipótesis de la compensación en Barcelona medida a través de la huella ecológica de la movilidad y la vivienda"
(Maig 2012)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

XREAP2012-10
Arqué-Castells, P. (GEAP), **Mohnen, P.**
"Sunk costs, extensive R&D subsidies and permanent inducement effects"
(Maig 2012)

XREAP2012-11
Boj, E. (CREB), **Delicado, P.**, **Fortiana, J**., **Esteve, A.**, **Caballé, A.**
"Local Distance-Based Generalized Linear Models using the dbstats package for R"
(Maig 2012)

xreap@pcb.ub.es