IMUB Institut de Matemàtica

STATISTICAL ASPECTS OF GOWER'S INTERPOLATION: ERROR TERM AND ITS INFLUENCE ON PREDICTION

Eva Boj, Maria Mercè Claramunt, Aurea Grané and Josep Fortiana

No. 378
October 2005

Mathematics Preprint Series

UNIVERSITAT DE BARCELONA

IMUB, Institut de Matemàtica Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain

Statistical Aspects of Gower's Interpolation: Error Term and its Influence on Prediction

Eva Boj[1] Maria Mercè Claramunt[1]
Aurea Grané[2] Josep Fortiana[3]

July 22, 2005

New points can be superimposed on a Euclidean configuration obtained as a result of a metric Multidimensional Scaling at coordinates given by Gower's interpolation formula. The procedure amounts to discarding a, possibly non-null, coordinate along an additional dimension. We compute this error term, assessing its influence on distance-based predictions.

Keywords: Distance-based prediction, Metric multidimensional scaling, Gower interpolation, Typicality in discrimination.

AMS Subject Classification: 62H25, 62H30, 62G99, 62H99.

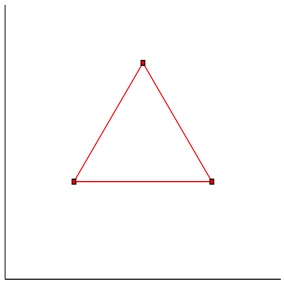
1. Introduction

Gower's *interpolation formula* (Gower 1968, see also Gower and Hand 1996, pp. 246*ff*) is a well-known device for inserting new points in a given Euclidean map obtained from a set of interdistances by means of metric Multidimensional Scaling (MMDS). General context references are Borg and Groenen (1997), Cox and Cox (1994), and Krzanowski and Marriott (1994), as well as Gower and Hand (1996). In general, a faithful representation of the augmented set should require a new dimension for each new point. For a quick illustration,

[1]Departament de Matemàtica Econòmica, Financera i Actuarial, Facultat de Ciències Econòmiques i Empresarials. Universitat de Barcelona, Avinguda Diagonal, 690, 08034 Barcelona, Spain.
Tel: +34 934 035 744. Fax: +34 934 034 892. Email: evaboj@ub.edu
[2]Departamento de Estadística, Universidad Carlos III, Calle Madrid, 126, 28903 Getafe (Madrid), Spain.
[3]Departament de Probabilitat, Lògica i Estadística, Facultat de Matemàtiques, Universitat de Barcelona, Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain.

Figure 1: Interdistances matrix for a set of three objects admitting a two-dimensional Euclidean representation.

$$D_3 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$



let us assume a set of three objects, with interdistances shown in Figure 1, to which a fourth object is added, equidistant at distance δ from the former three, resulting:

$$D_4 = \begin{pmatrix} 0 & 1 & 1 & \delta \\ 1 & 0 & 1 & \delta \\ 1 & 1 & 0 & \delta \\ \delta & \delta & \delta & 0 \end{pmatrix}.$$

Depending on δ, there are four possible situations.
1. If $\delta < \frac{1}{2}$, the dissimilarity defined by D_4 is nonmetric, as it fails to satisfy the triangular inequality.
2. If $\frac{1}{2} \leq \delta < \frac{\sqrt{3}}{3}$, D_4 defines a non Euclidean metric (Figure 2, left hand side).
3. If $\delta = \frac{\sqrt{3}}{3}$, the dissimilarity defined by D_4 is a Euclidean metric admitting a two-dimensional representation (Figure 2, right hand side).
4. If $\delta > \frac{\sqrt{3}}{3}$, the dissimilarity defined by D_4 is a Euclidean metric admitting a three-dimensional representation (Figure 3).

Figure 2: A non Euclidean metric configuration (left hand side) and a two-dimensional Euclidean metric configuration (right hand side).

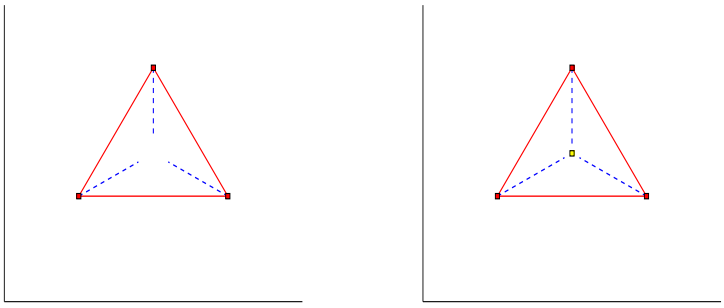


Figure 3: A three-dimensional Euclidean metric configuration.

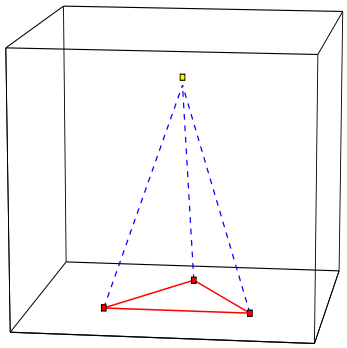


Figure 4: A new observation can be assigned to one of the two known groups (left hand side) or it is far from both (right hand side).

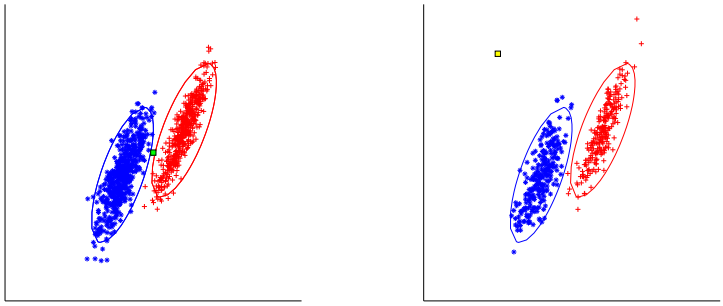


Figure 5: Existence of a new group may be inferred.

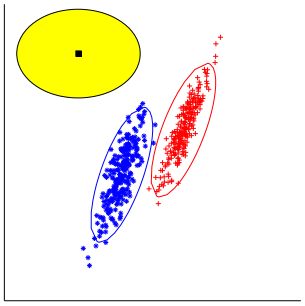
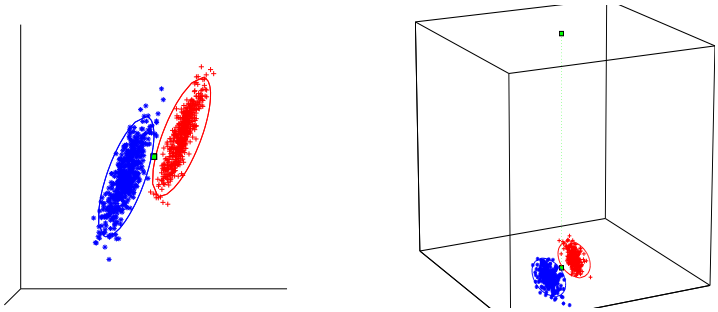


Figure 6: A new observation appears close to the known groups due to the projection effect.



If the exact representation is projected onto the preexisting lower dimensional subspace, which is what Gower's interpolation does, this additional coordinate is lost. The *squared error term*, the square of the additional coordinate, lost in projection, is a measure of interpolation quality. Its significance may lead to incorrect results in distance-based statistical prediction methods. For instance, the *typicality problem* in Discriminant Analysis (Rao 1973) amounts to deciding whether a new observation to be classified belongs to one of the already known groups or, instead, it falls far apart from all of them (Figure 4), so much so as to lead us to postulate the existence of a third, so far unobserved, group (Figure 5). In (Cuadras and Fortiana 1998) the problem of typicality is considered in the context of Distance-Based Discrimination (Cuadras et al. 1997). A shortcoming of the proposed method is its implicit use of Gower's interpolation, which may cause that a new observation appears as assignable to the known groups, when it actually lies very far along the additional dimension which is lost under projection (Figure 6). Similarly, predictions in Distance Based Regression (see, e.g., Cuadras 1989, Cuadras et al. 1996) implicitly rely upon Gower's interpolation and will be systematically biased for observations with a significant projection error.
This paper is organized as follows: In Section 2 we establish notations and assumptions. In Section 3 we evaluate this squared error term, rederiving Gower's own formula. In Section 4 we address the problem of learning the probability distribution of the observed error term in a given statistical context, showing a particular instance thereof for an ideal gaussian configuration. For actual data, in Section 5 we suggest a method for estimating the squared error term distribution under the null hypothesis that the underlying geometrical quantity is zero. Concluding remarks are in Section 6 and Appendix A contains accessory derivations.

2. Notations and assumptions

We will denote by $\Omega_n = \{\omega_1, \ldots, \omega_n\}$ a set of n points or observational units, and by $\boldsymbol{D}_n = (d_{ij})$ an $n \times n$ matrix of squared interdistances between them. We assume throughout that the Euclidean property is satisfied, thus for some $k > 0$ there exists a set of n vectors, $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, in a k-dimensional Euclidean space such that

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = d_{ij}, \quad 1 \leq i, j \leq n.$$

Such a set of vectors is called a *Euclidean configuration of* $\boldsymbol{D}_n$ or *of* Ω_n, and the ambient Euclidean space will be denoted by $\mathscr{F}_n \subset \mathbb{R}^k$. The usual assumption that such configurations are centered entails no loss of generality. Let

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}$$

be an $n \times k$ matrix containing the $\boldsymbol{x}_i$ vectors as rows. The matrix equation $\mathbf{1}' \cdot \boldsymbol{X} = \mathbf{0}$, where $\mathbf{1}_n$ is the $n \times 1$ column of ones, expresses the centering condition. Clearly there are many centered Euclidean configurations of a given $\boldsymbol{D}_n$. More precisely, a centered matrix with n rows, $\boldsymbol{X}$, is a centered Euclidean configuration of $\boldsymbol{D}_n$ if its *inner products matrix* $\boldsymbol{G}_n = \boldsymbol{X} \cdot \boldsymbol{X}'$ satisfies the equality

$$\boldsymbol{G}_n = -\frac{1}{2}\, \boldsymbol{J}_n \cdot \boldsymbol{D}_n \cdot \boldsymbol{J}_n, \tag{1}$$

where $\boldsymbol{J}_n = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n \cdot \mathbf{1}_n{}'$ is the *centering matrix*. This same equation (1), interpreted now as a definition of $\boldsymbol{G}_n$, leads to the statement of Schoenberg's theorem (Schoenberg 1935), that $\boldsymbol{D}_n$ is a Euclidean distance matrix if $\boldsymbol{G}_n$ is positive semidefinite. Expansion of (1) gives:

$$\boldsymbol{G}_n = -\frac{1}{2}\left(\boldsymbol{D}_n - \overline{\boldsymbol{D}}_n \cdot \mathbf{1}_n{}' - \mathbf{1}_n \cdot \overline{\boldsymbol{D}}_n{}' + \overline{\overline{\boldsymbol{D}}}_n\, \mathbf{1}_n \cdot \mathbf{1}_n{}'\right),$$

where

$$\overline{\boldsymbol{D}}_n = \frac{1}{n}\boldsymbol{D}_n \cdot \mathbf{1}_n, \qquad \overline{\overline{\boldsymbol{D}}}_n = \frac{1}{n}\mathbf{1}_n{}' \cdot \overline{\boldsymbol{D}}_n = \frac{1}{n^2}\mathbf{1}_n{}' \cdot \boldsymbol{D}_n \cdot \mathbf{1}_n.$$

Additional notations, also used in (Cuadras et al. 1997), are the *geometric variability of* $\boldsymbol{D}_n$, defined as:

$$v_n = \frac{1}{2}\overline{\overline{\boldsymbol{D}}}_n, \tag{2}$$

and the column $\boldsymbol{g}_n$, containing the diagonal entries in $\boldsymbol{G}_n$, the n *proximity functions* of the elements of Ω_n, i.e., the squared norms of their Euclidean representantives in $\mathscr{F}_n$,

$$\boldsymbol{g}_n = \operatorname{diag}(\boldsymbol{G}_n) = \overline{\boldsymbol{D}}_n - \frac{1}{2}\overline{\overline{\boldsymbol{D}}}_n\, \mathbf{1}_n = \overline{\boldsymbol{D}}_n - v_n\, \mathbf{1}_n = \begin{pmatrix} \|\boldsymbol{x}_1\|^2 \\ \vdots \\ \|\boldsymbol{x}_n\|^2 \end{pmatrix}. \tag{3}$$

Note that these squared norms depend only on the distances themselves; in particular they are invariant across all centered Euclidean configurations of $\boldsymbol{D}_n$. The mean squared norm is equal to the geometric variability, $\frac{1}{n}\mathbf{1}_n{}' \cdot \boldsymbol{g}_n = \frac{1}{n}\operatorname{tr}(\boldsymbol{G}_n) = v_n$.

3. The error term

By adding a new ω_{n+1} to Ω_n, we have to consider an $(n+1) \times (n+1)$ matrix of squared interdistances of the augmented set $\Omega_{n+1} = \Omega_n \cup \{\omega_{n+1}\}$, which may be represented as a

2×2 block matrix,

$$\boldsymbol{D}_{n+1} = \left(\begin{array}{c|c} \boldsymbol{D}_n & \boldsymbol{d}_n \\ \hline \boldsymbol{d}_n{}' & 0 \end{array}\right), \tag{4}$$

where the column $\boldsymbol{d}_n$ contains the n squared distances from ω_{n+1} to $\omega_i \in \Omega_n$. As in (1),

$$\begin{aligned} \boldsymbol{G}_{n+1} &= -\tfrac{1}{2}\, \boldsymbol{J}_{n+1} \cdot \boldsymbol{D}_{n+1} \cdot \boldsymbol{J}_{n+1} \\ &= -\tfrac{1}{2}\, (\boldsymbol{D}_{n+1} - \overline{\boldsymbol{D}}_{n+1} \cdot \mathbf{1}_{n+1}{}' - \mathbf{1}_{n+1} \cdot \overline{\boldsymbol{D}}_{n+1}{}' + \overline{\overline{\boldsymbol{D}}}_{n+1}\, \mathbf{1}_{n+1} \cdot \mathbf{1}_{n+1}{}'). \end{aligned} \tag{5}$$

Computing in terms of previously defined quantities and of

$$\begin{aligned} \overline{\boldsymbol{d}}_n &= \frac{1}{n}\, \mathbf{1}_n{}' \cdot \boldsymbol{d}_n, \\ \phi^2_{n+1} &= \overline{\boldsymbol{d}}_n - v_n, \end{aligned} \tag{6}$$

we obtain:

$$\overline{\boldsymbol{D}}_{n+1} = \frac{1}{n+1}\, \boldsymbol{D}_{n+1} \cdot \begin{pmatrix} \mathbf{1}_n \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{n}{n+1}\, \overline{\boldsymbol{D}}_n + \frac{1}{n+1}\, \boldsymbol{d}_n \\ \frac{n}{n+1}\, \overline{\boldsymbol{d}}_n \end{pmatrix} \tag{7}$$

and

$$\overline{\overline{\boldsymbol{D}}}_{n+1} = \frac{1}{n+1}\, (\mathbf{1}_n{}', 1) \cdot \overline{\boldsymbol{D}}_{n+1} = \frac{1}{n+1} \left(\frac{n^2}{n+1} \overline{\overline{\boldsymbol{D}}}_n + 2\, \frac{n}{n+1} \overline{\boldsymbol{d}}_n \right). \tag{8}$$

In particular, the geometric variability of the augmented set is:

$$v_{n+1} = \left(\frac{n}{n+1}\right)^2 (v_n + \frac{1}{n} \overline{\boldsymbol{d}}_n). \tag{9}$$

Similarly, the vector of proximities $\boldsymbol{g}_{n+1} = \mathrm{diag}(\boldsymbol{G}_{n+1})$ is:

$$\begin{aligned} \boldsymbol{g}_{n+1} &= \overline{\boldsymbol{D}}_{n+1} - v_{n+1}\, \mathbf{1}_{n+1} = \begin{pmatrix} \frac{n}{n+1} \overline{\boldsymbol{D}}_n + \frac{1}{n+1}\, \boldsymbol{d}_n \\ \\ \frac{n}{n+1} \overline{\boldsymbol{d}}_n \end{pmatrix} - \left(\frac{n}{n+1}\right)^2 (v_n + \frac{1}{n} \overline{\boldsymbol{d}}_n) \begin{pmatrix} \mathbf{1}_n \\ \\ 1 \end{pmatrix} \\ \\ &= \begin{pmatrix} \frac{n}{n+1}\, \boldsymbol{g}_n + \frac{1}{n+1}\, \boldsymbol{d}_n - \frac{n}{(n+1)^2}\, \phi^2_{n+1}\, \mathbf{1}_n \\ \\ (\frac{n}{n+1})^2\, \phi^2_{n+1} \end{pmatrix}, \end{aligned}$$

thus, denoting by $\boldsymbol{x}_{n+1}$ the vector representing ω_{n+1} in a centered Euclidean configuration of Ω_{n+1}, we have

$$\|\boldsymbol{x}_{n+1}\|^2 = \left(\frac{n}{n+1}\right)^2 \phi^2_{n+1}. \tag{10}$$

Note that the form of (10) makes it clear that the equality holds for *any* centered Euclidean configuration of Ω_{n+1}, a result which will become useful below.

On the other hand, we have the *superimposed coordinates* of ω_{n+1}, $\widehat{\boldsymbol{x}}_{n+1} \in \mathscr{F}_n$, result of Gower's interpolation, or *add-a-point* formula (Gower 1968, see also Gower and Hand 1996, pp. 246*ff*). In order to derive it, we assume a Euclidean configuration for Ω_{n+1},

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_{n+1} \end{pmatrix}, \tag{11}$$

in a subspace $\mathscr{F}_{n+1} \subset \mathbb{R}^{k+1}$. $\boldsymbol{Y}$ consists of n vectors from $\boldsymbol{X}$,

$$\boldsymbol{y}_i = (\boldsymbol{x}_i, 0) \in \mathbb{R}^{k+1}, \quad \text{for } 1 \leq i \leq n,$$

plus another $\boldsymbol{y}_{n+1} = (\widehat{\boldsymbol{x}}_{n+1}, r_{n+1}) \in \mathbb{R}^{k+1}$ with a possibly nonnull $(k+1)$-th coordinate. Below it will become clear that such a configuration is always possible provided that $\boldsymbol{D}_{n+1}$ satisfies the Euclidean condition. Since $\boldsymbol{Y}$ is obviously noncentered, $\boldsymbol{y}_{n+1} \neq \boldsymbol{x}_{n+1}$ in (10). From the n entries in $\boldsymbol{d}_n = (d_1, \ldots, d_n)'$, defined in (4) as the squared distances between ω_{n+1} and the elements of Ω_n, we have the following set of equations,

$$\begin{aligned} d_i &= (\boldsymbol{y}_{n+1} - \boldsymbol{y}_i) \cdot (\boldsymbol{y}_{n+1} - \boldsymbol{y}_i)' \\ &= \|\boldsymbol{y}_{n+1}\|^2 + \|\boldsymbol{y}_i\|^2 - 2\,\boldsymbol{y}_{n+1} \cdot \boldsymbol{y}_i{}' \\ &= \|\boldsymbol{y}_{n+1}\|^2 + \|\boldsymbol{x}_i\|^2 - 2\,\widehat{\boldsymbol{x}}_{n+1} \cdot \boldsymbol{x}_i{}', \end{aligned}$$

which, taking into account (3), can be written as a single matricial equation as

$$(\boldsymbol{d}_n - \boldsymbol{g}_n)' = \|\boldsymbol{y}_{n+1}\|^2\, \boldsymbol{1}_n - 2\,\widehat{\boldsymbol{x}}_{n+1} \cdot \boldsymbol{X}',$$

Posmultiplication by $\boldsymbol{X}$ gives

$$(\boldsymbol{d}_n - \boldsymbol{g}_n)' \cdot \boldsymbol{X} = -2\,\widehat{\boldsymbol{x}}_{n+1} \cdot (\boldsymbol{X}' \cdot \boldsymbol{X}),$$

since $\boldsymbol{X}$ is centered. If $\boldsymbol{X}' \cdot \boldsymbol{X}$ is nonsingular, $\widehat{\boldsymbol{x}}_{n+1}$ is given by the usual *add-a-point* formula

$$\widehat{\boldsymbol{x}}_{n+1} = \frac{1}{2}\,(\boldsymbol{g}_n - \boldsymbol{d}_n)' \cdot \boldsymbol{X} \cdot (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1}, \tag{12}$$

otherwise the inverse must be replaced by a g-inverse, resulting in a nonunique $\widehat{\boldsymbol{x}}_{n+1}$. Nonetheless, its squared norm

$$\|\widehat{\boldsymbol{x}}_{n+1}\|^2 = \frac{1}{4}\,(\boldsymbol{g}_n - \boldsymbol{d}_n)' \cdot \boldsymbol{X} \cdot (\boldsymbol{X}' \cdot \boldsymbol{X})^{-2} \cdot \boldsymbol{X}' \cdot (\boldsymbol{g}_n - \boldsymbol{d}_n),$$

is invariant under different choices of a g-inverse, since $\boldsymbol{X} \cdot (\boldsymbol{X}' \cdot \boldsymbol{X})^{-2} \cdot \boldsymbol{X}' = \boldsymbol{G}_n{}^+$, the Moore-Penrose pseudo-inverse of $\boldsymbol{G}_n$. Hence

$$\|\widehat{\boldsymbol{x}}_{n+1}\|^2 = \frac{1}{4}\,(\boldsymbol{g}_n - \boldsymbol{d}_n)' \cdot \boldsymbol{G}_n{}^+ \cdot (\boldsymbol{g}_n - \boldsymbol{d}_n). \tag{13}$$

The Euclidean configuration $\boldsymbol{Y}$, defined in (11), can easily be centered by subtracting from each row their overall average, giving, in an obvious block matrix notation,

$$\boldsymbol{Y}_0 = \left(\begin{array}{c|c} \boldsymbol{X} - \frac{1}{n+1}\, \mathbf{1}_n \cdot \widehat{\boldsymbol{x}}_{n+1} & -\frac{1}{n+1}\, \mathbf{1}_n\, r_{n+1} \\ \hline \frac{n}{n+1}\, \widehat{\boldsymbol{x}}_{n+1} & \frac{n}{n+1}\, r_{n+1} \end{array} \right). \tag{14}$$

Equating the squared norm of its $(n+1)$-th vector,

$$\left(\frac{n}{n+1}\right)^2 \left(\|\widehat{\boldsymbol{x}}_{n+1}\|^2 + r_{n+1}^2 \right)$$

with $\|\boldsymbol{x}_{n+1}\|^2$ in (10), we have an expression for the squared error term, the squared norm of the residual of the orthogonal projection $\mathscr{F}_{n+1} \to \mathscr{F}_n$,

$$r_{n+1}^2 = \phi_{n+1}^2 - \frac{1}{4}\, (\boldsymbol{g}_n - \boldsymbol{d}_n)' \cdot \boldsymbol{G}_n{}^+ \cdot (\boldsymbol{g}_n - \boldsymbol{d}_n). \tag{15}$$

An alternative derivation of Gower's interpolation formula, and a further insight, can be obtained by explicitly writing the relationship between $\boldsymbol{G}_n$ and $\boldsymbol{G}_{n+1}$ (see Appendix A).

4. Distribution of the error term

In a statistical setting, for instance in the problem of typicality referred to in the Introduction, we would like to decide whether the residual term is significantly non-null or else it can be safely ignored. To this end we need to know the probability distribution of the (observed) residual under the null hypothesis that the geometrical residual is equal to zero. Analytical derivation of such a distribution is an exceedingly complex task for any realistic situation, hence for actual data resampling methods provide sensible estimations —see Section 5. Meanwhile, and in order to gain some theoretical insight on the qualitative properties of error term distributions we pose an ideal scenario, in which we assume that normality holds *for the Euclidean configuration*: A random $(n+1)$-th point is added to n points in $\mathscr{F}_n = \mathbb{R}^p$, realized as the hyperplane of all vectors in $\mathscr{F}_{n+1} = \mathbb{R}^{p+1}$ with a null $(p+1)$-th coordinate. Let us denote them by

$$\boldsymbol{\mu}_i = (\boldsymbol{\theta}_i, 0) \in \mathbb{R}^{p+1}, \quad \boldsymbol{\theta}_i \in \mathbb{R}^p, \quad 1 \leq i \leq n.$$

In order to evaluate the projection we do not need the actual coordinates $\boldsymbol{\theta}_i$, only the subspace they generate. For $n \geq p$ we can, with probability 1, take them to be a set of p constant vectors generating $\mathscr{F}_n$. The additional random point will be

$$\boldsymbol{y}_{p+1} = (\boldsymbol{x}_{p+1}, t) \sim \mathbf{N}(\boldsymbol{\mu}_{p+1}, \boldsymbol{\Sigma}_{p+1}),$$

where $\boldsymbol{\mu}_{p+1} = (\boldsymbol{\theta}_{p+1}, \tau) \in \mathbb{R}^{p+1}$ and $\boldsymbol{\Sigma}_{p+1}$, which can be assumed nonsingular, has the block structure:

$$\boldsymbol{\Sigma}_{p+1} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_p & \boldsymbol{\Sigma}_{xt} \\ \hline \boldsymbol{\Sigma}'_{xt} & \sigma_t^2 \end{array} \right), \tag{16}$$

where $\boldsymbol{\Sigma}_p$ is $p \times p$, $\boldsymbol{\Sigma}_{xt}$ is $p \times 1$ and σ_t^2 is a scalar. The projection of $\boldsymbol{y}_{p+1}$ on the subspace generated by the rows of

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_p \end{pmatrix}$$

can be expressed as:

$$\left\langle \boldsymbol{y}_{p+1}, \boldsymbol{M} \right\rangle \left\langle \boldsymbol{M}, \boldsymbol{M} \right\rangle^{-1} \boldsymbol{M},$$

where the inner products are to be computed with the Mahalanobis metric:

$$\left\langle \boldsymbol{y}_{p+1}, \boldsymbol{M} \right\rangle = \boldsymbol{y}_{p+1} \cdot \boldsymbol{\Sigma}_{p+1}^{-1} \cdot \boldsymbol{M}', \quad \left\langle \boldsymbol{M}, \boldsymbol{M} \right\rangle = \boldsymbol{M} \cdot \boldsymbol{\Sigma}_{p+1}^{-1} \cdot \boldsymbol{M}'.$$

Using a well-known formula to invert a block nonsingular symmetric matrix (see, e.g., Seber 1984, Appendix),

$$\boldsymbol{\Sigma}_{p+1}{}^{-1} = \left(\begin{array}{c|c} \boldsymbol{A} & \boldsymbol{b} \\ \hline \boldsymbol{b}' & a \end{array} \right),$$

where

$$\boldsymbol{A} = \left[\Sigma_p - \tfrac{1}{\sigma_t^2} \boldsymbol{\Sigma}_{xt} \cdot \boldsymbol{\Sigma}'_{xt} \right]^{-1}, \quad \boldsymbol{b} = -\tfrac{1}{\sigma_t^2} \boldsymbol{A} \cdot \boldsymbol{\Sigma}_{xt}, \quad a = \tfrac{1}{\sigma_t^2} + \tfrac{1}{\sigma_t^4} \boldsymbol{\Sigma}_{xt} \cdot \boldsymbol{A} \cdot \boldsymbol{\Sigma}_{xt}.$$

If we assume, as we are entitled to do, that $\{\theta_i\}$ are the $\mathbb{R}^p$ standard basis vectors,

$$\theta_i = \boldsymbol{e}_i, \quad 1 \leq i \leq n = p,$$

then $\boldsymbol{M} = (\boldsymbol{I}_p, \boldsymbol{0})$, $\left\langle \boldsymbol{M}, \boldsymbol{M} \right\rangle^{-1} = \boldsymbol{A}^{-1}$ and

$$\left\langle \boldsymbol{y}_{p+1}, \boldsymbol{M} \right\rangle = (\boldsymbol{x}_{p+1}, t) \cdot \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{b}' \end{pmatrix} = \boldsymbol{x}_{p+1} \cdot \boldsymbol{A} + t\, \boldsymbol{b}' = \left(\boldsymbol{x}_{p+1} - \tfrac{t}{\sigma_t^2} \boldsymbol{\Sigma}'_{xt} \right) \cdot \boldsymbol{A}.$$

Thus, the projected vector and its residual are

$$\begin{aligned} \widehat{y}_{p+1} &= (\boldsymbol{x}_{p+1} - \tfrac{t}{\sigma_t^2} \boldsymbol{\Sigma}'_{xt}, 0), \\ \widetilde{y}_{p+1} &= y_{p+1} - \widehat{y}_{p+1} = \left(\tfrac{t}{\sigma_t^2} \boldsymbol{\Sigma}'_{xt},\, t \right). \end{aligned}$$

The error term is the squared norm

$$Q = \|\widetilde{y}_{p+1}\|^2 = t^2 \left(1 + \frac{\|\boldsymbol{\Sigma}_{xt}\|^2}{\sigma_t^4} \right).$$

Since t, the $(p+1)$-th coordinate of y_{p+1}, is normally distributed with mean τ and variance σ_t^2, under the null hypothesis that τ is null we have $t = Z\,\sigma_t$, where $Z \sim \mathrm{N}(0,1)$. Hence Q is a multiple of a χ_1^2 variate, with the proportionality factor

$$\kappa = \sigma_t^2 + \frac{\|\boldsymbol{\Sigma}_{xt}\|^2}{\sigma_t^2}.$$

Under $H_1:\ \tau \neq 0$ we can write

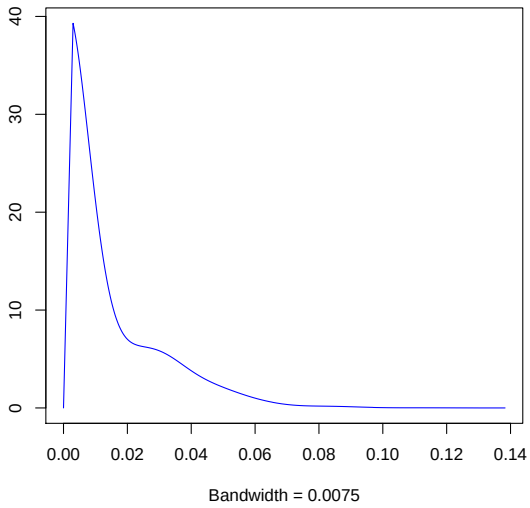
$$t = \tau + \sigma_t\, Z = \sigma_t\,(\delta + Z), \quad \text{where } \delta = \tau/\sigma_t,$$

then $t^2 = \sigma_t^2\, W$ and $Q = \kappa\, W$, where $W = (\delta + Z)^2$ is a noncentral χ_1^2 with δ^2 as its noncentrality parameter.

5. Computations with real data

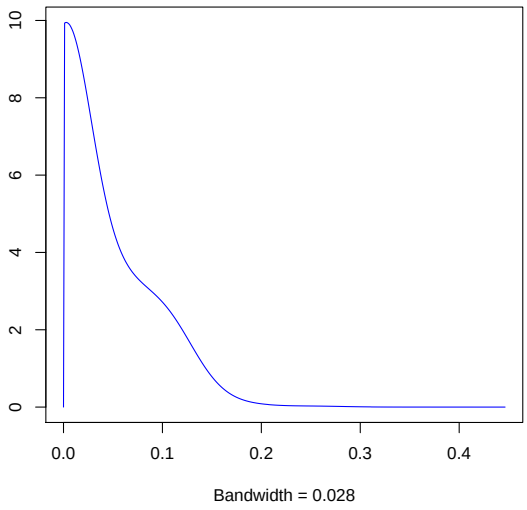
A dataset consisting of observations on an n-set $\Omega_n = \{\omega_1, \ldots, \omega_n\}$ is understood as an n-sample drawn from a population Ω. Within this framework the Euclidean space $\mathscr{F}_n$ derived from Ω_n as explained above is an "estimation" of the Euclidean space $\mathscr{F}$ likewise associated with Ω. By hypothesis each ω_i in Ω_n satisfies the null hypothesis that it can be exactly represented by a vector in $\mathscr{F}$ or, equivalently, that when projected on $\mathscr{F}$ the error

Figure 7: Density estimation for Data set 1



term is zero. Thus we can implement a cross-validatory scheme by randomly splitting Ω_n in two disjoint subsets, Ω_A and Ω_B, with n_A and n_B observations, respectively. We take

Figure 8: Density estimation for Data set 2



$\mathscr{F}_A$, the space generated by the Euclidean configuration of the n_A observations in Ω_A, as an estimation of $\mathscr{F}$, and the n_B residuals obtained by projecting on $\mathscr{F}_A$ the observations in Ω_B, i.e., by applying (15), as samples of the observed squared error term under H_0. We perform a suitable number $\nu \leq \binom{n}{n_A}$ of such splittings so to yield $N = \nu\, n_B$ values of r^2 to estimate its probability density. In particular, an expedient choice is the jacknife-like, or leave-one-out, computation, with $n_A = n - 1$ and $N = n$.

To illustrate this procedure we provide two examples, in the contexts of classification (Data set 1) and regression (Data set 2). Data set 1, taken from Krzanowski (1975), has observations of six continuous and three qualitative variables on 186 subjects from two groups, of 99 and 87 cases, respectively. We shall not need to delve into specifics on meaning of variables and groups. Let us only point out that for such data, i.e., intended for a classification problem, the subsets Ω_A and Ω_B should be formed with proportional numbers of cases in each group, as representing the whole sample. Data set 2 is Gasoline Mileage Data, from Henderson and Velleman (1981), where a continuous numerical response is regressed on 8 continuous numerical and two categorical predictors, observed for 32 cars. In both examples the distance matrix is derived from the mixed predictor variables using the widely used metric derived from Gower's omnibus similarity coefficient (Gower 1971). Figures 7 and 8 show kernel-smoothed r^2 density estimations for both data sets, derived from $\nu = 2500$ resamplings with $n_B \approx 30\%$ of the total sample size n. Each resampling is realized as a boolean n-vector, randomly drawn from a Bernoulli distribution with parameter 0.30.

6. Concluding remarks

In the first place, the present work provides some additional insight on the mechanics of distance-based prediction methods, especially on the role played by Gower's interpolation and the dangers of implicitly and uncritically assuming the equal Euclidean dimension hypothesis. Figure 6 shows an example, with an artificial data-set, where consideration of the error term is highly relevant to avoid misleading results. As a tangible result of our investigation we propose an auxiliary statistical tool, useful as a cautionary device to verify the validity of the aforementioned hypothesis.

Appendices

A. Expressing G_{n+1} in terms of G_n and Gower's interpolation

Direct computation from the equality

$$-2\,\boldsymbol{G}_{n+1} = \boldsymbol{D}_{n+1} - \boldsymbol{g}_{n+1} \cdot (\mathbf{1}_n{}', 1) - \begin{pmatrix} \mathbf{1}_n \\ 1 \end{pmatrix} \cdot \boldsymbol{g}_{n+1}{}',$$

gives

$$\boldsymbol{G}_{n+1} = \begin{pmatrix} \dfrac{\boldsymbol{G}_n - \frac{1}{2\,(n+1)}\ [(\boldsymbol{g}_n - \boldsymbol{d}_n) \cdot \mathbf{1}_n{}' + \mathbf{1}_n \cdot (\boldsymbol{g}_n - \boldsymbol{d}_n)'] - \frac{n}{(n+1)^2}\ \phi_{n+1}^2\, \mathbf{1}_n \cdot \mathbf{1}_n{}'}{\frac{1}{2}\,(\frac{n}{n+1})\ \left[(\boldsymbol{g}_n - \boldsymbol{d}_n)' + \frac{(n-1)}{(n+1)}\ \phi_{n+1}^2\, \mathbf{1}_n{}'\right]} \\ \dfrac{\frac{1}{2}\,(\frac{n}{n+1})\ \left[(\boldsymbol{g}_n - \boldsymbol{d}_n) + \frac{(n-1)}{(n+1)}\ \phi_{n+1}^2\, \mathbf{1}_n\right]}{(\frac{n}{n+1})^2\ \phi_{n+1}^2} \end{pmatrix}, \tag{17}$$

where notations are as in Section 3. A somewhat clearer expression is obtained by defining

$$\boldsymbol{k}_n = \frac{1}{2}\ \left[(\boldsymbol{g}_n - \boldsymbol{d}_n) + \phi_{n+1}^2\, \mathbf{1}_n\right]. \tag{18}$$

Note that $\mathbf{1}_n \cdot \boldsymbol{k}_n = 0$, taking into account (3) and (6). Thus,

$$\boldsymbol{G}_{n+1} = \begin{pmatrix} \dfrac{\boldsymbol{G}_n - \frac{1}{n+1}\ (\boldsymbol{k}_n \cdot \mathbf{1}_n{}' + \mathbf{1}_n \cdot \boldsymbol{k}_n{}') + \frac{1}{(n+1)^2}\ \phi_{n+1}^2\, \mathbf{1}_n \cdot \mathbf{1}_n{}'}{\frac{n}{n+1}\ \boldsymbol{k}_n{}' - \frac{n}{(n+1)^2}\ \phi_{n+1}^2\, \mathbf{1}_n{}'} \\ \dfrac{\frac{n}{n+1}\ \boldsymbol{k}_n - \frac{n}{(n+1)^2}\ \phi_{n+1}^2\, \mathbf{1}_n}{(\frac{n}{n+1})^2\ \phi_{n+1}^2} \end{pmatrix}. \tag{19}$$

Since also $\boldsymbol{G}_{n+1} = \boldsymbol{Y}_0 \cdot \boldsymbol{Y}_0{}'$, comparing with

$$\boldsymbol{G}_{n+1} = \left(\begin{array}{c|c} \boldsymbol{X} \cdot \boldsymbol{X}' - \frac{1}{n+1} \left[\boldsymbol{X} \cdot \widehat{\boldsymbol{x}}_{n+1}{}' \cdot \boldsymbol{1}_n{}' + \boldsymbol{1}_n \cdot \widehat{\boldsymbol{x}}_{n+1} \cdot \boldsymbol{X}' \right] + \frac{1}{(n+1)^2} \phi^2_{n+1} \boldsymbol{1}_n \cdot \boldsymbol{1}_n{}' & \\ \hline \frac{n}{n+1} \widehat{\boldsymbol{x}}_{n+1} \cdot \boldsymbol{X}' - \frac{n}{(n+1)^2} \phi^2_{n+1} \boldsymbol{1}_n{}' & \end{array} \right| \left| \begin{array}{c} \\ \hline \frac{n}{n+1} \boldsymbol{X} \cdot \widehat{\boldsymbol{x}}_{n+1}{}' - \frac{n}{(n+1)^2} \phi^2_{n+1} \boldsymbol{1}_n \\ \hline \left(\frac{n}{n+1}\right)^2 \phi^2_{n+1} \end{array} \right), \tag{20}$$

we recover the interpolation formula (12) equating $\boldsymbol{k}_n$ in (18) with $\boldsymbol{X} \cdot \widehat{\boldsymbol{x}}_{n+1}$ in (20):

$$\boldsymbol{X} \cdot \widehat{\boldsymbol{x}}_{n+1} = \frac{1}{2} \left[(\boldsymbol{g}_n - \boldsymbol{d}_n) + \phi^2_{n+1} \boldsymbol{1}_n \right].$$

References

Borg, I. and P. Groenen (1997). *Modern Multidimensional Scaling*. New York: Springer-Verlag.

Cox, T. F. and M. A. A. Cox (1994). *Multidimensional Scaling*. London: Chapman & Hall.

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.

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. and J. Fortiana (1998). Typicality in discriminant analysis with mixed variables. In A. Rizzi, M. Vichi, and H. H. Bock (Eds.), *Advances in Data Science and Classification. Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98). Rome, Italy, July 21-24, 1998*, Heidelberg, pp. 82–85. Springer-Verlag.

Cuadras, C. M., J. Fortiana, and F. Oliva (1997). The proximity of an individual to a population with applications in discriminant analysis. *Journal of Classification 14*, 117–136.

Gower, J. C. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika 55*, 582–585.

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

Gower, J. C. and D. J. Hand (1996). *Biplots*. London, UK: Chapman & Hall.

Henderson, H. V. and P. F. Velleman (1981). Building multiple regression models interactively. *Biometrics 37*, 391–411.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association 70*, 782–790.

Krzanowski, W. J. and F. H. C. Marriott (1994). *Multivariate Analysis. Part 1, Distributions, ordination and inference*. London: Edward Arnold.

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

Schoenberg, I. J. (1935). Remarks to Maurice Fréchet's article "Sur la Définition Axiomatique d'une Classe d'Espaces Distanciés Vectoriellement Applicables sur l'Espace de Hilbert". *The Annals of Mathematics 36*, 724–732.

Seber, G. A. F. (1984). *Multivariate observations*. New York, NY, USA: John Wiley & Sons.

Relació dels últims Preprints publicats:

– **361** *Strict modules and homotopy modules in stable homotopy.* Javier J. Gutiérrez. MS Subject Classification: 55P43, 18E30, 55P42. October 2004.

– **362** *Lattice approximation for a stochastic wave equation.* Lluís Quer-Sardanyons and Marta Sanz-Solé. MS Subject Classification: 60H35, 60H15. October 2004.

– **363** *An adaptive goodness-of-fit test.* Aurea Grané and Josep Fortiana. AMS Subject Classification: 62G10, 62G30, 62E15. November 2004.

– **364** *A stabilization phenomenon for a class of stochastic partial differential equations.* David Nualart and Pierre-A. Vuillermot. AMS Subject Classification: 60H15, 35R60. November 2004.

– **365** *Homotopy localization of groupoids.* Carles Casacuberta, Marek Golasiński, and Andrew Tonks. AMS Subject Classification (2000): Primary 18B40, 55P60; Secondary 18E35, 18A40. November 2004.

– **366** *Bifurcations of three-dimensional diffeomorphisms with non-simple quadratic homoclinic tangencies and generalized Hénon maps.* S.V. Gonchenko, V.S. Gonchenko, and J.C. Tatjer. AMS Subject Classification: 37G25, 37C29, 37G05. December 2004.

– **367** *Bootstrapping repeated measures data in a nonlinear mixed-models context.* Jordi Ocaña, Rachid El Halimi, M. Carme Ruiz de Villa, and Josep A. Sánchez. AMS Subject Classification: 62F35, 62F40, 62G09, 62P10. February 2005.

– **368** *The orthogonal subcategory problem in homotopy theory.* Carles Casacuberta and Boris Chorny. AMS Subject Classification: 55U35, 55P60, 18G55. February 2005.

– **369** *On two fragments with negation and without implication of the logic of residuated lattices.* Félix Bou, Àngel García-Cerdaña, and Ventura Verdú. AMS Subject Classification: 03B47, 03B50, 03F99, 06D15, 06B99. February 2005.

– **370** *Variational solutions for partial differential equations driven by a fractional noise.* David Nualart and Pierre-A. Vuillermot. AMS Subject Classification: 60H15, 35R60. March 2005.

– **371** *Malliavin calculus for stochastic differential equations driven by a fractional Brownian motion.* David Nualart and Bruno Saussereau. AMS Subject Classification: 60H05, 60H07. March 2005.

– **372** *Glivenko's Theorem in algebraizable logics.* Antoni Torrens. AMS Subject Classification: 03G, 08C15. March 2005.

– **373** *Power variation of some integral long-memory processes.* José Manuel Corcuera, David Nualart, and Jeannette H.C. Woerner. AMS Subject Classification: 60F05, 60G15, 60G18, 62M99. April 2005.

– **374** *Intersection local time for two independent fractional Brownian motions.* David Nualart and Salvador Ortiz. AMS Subject Classification (2000): 60G15, 60F05, 60F25, 60G18, 60J55. July 2005.

– **375** *On the infinite-valued Łukasiewicz logic that preserves degrees of truth.* Josep Maria Font, Àngel J. Gil, Antoni Torrens, and Ventura Verdú. AMS Subject Classification (2000): 03B50, 03G20, 06D35, 03B22. July 2005.

– **376** *A lattice scheme for stochastic partial differential equations of elliptic type in dimension $d \geq 4$.* Teresa Martínez and Marta Sanz-Solé. AMS Subject Classification (2000): 60H15, 60H35, 35J05. September 2005.

– **377** *Approximation of rough paths of fractional Brownian motion.* Annie Millet and Marta Sanz-Solé. AMS Subject Classification (2000): Primary 60G15; Secondary 60H05, 60H07. September 2005.

Editat pel Servei de Textos Matemàtics
E-mail: stm@mat.ub.es