Statistics and Probability Letters 128 (2017) 84–88

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Point separation in logistic regression on Hilbert space-valued variables

Agne Kazakevičiūtė [a,*], Malini Olivo [b,c]

[a] Department of Statistical Science, University College, London, UK
[b] Agency for Science, Technology and Research (A*STAR), Singapore
[c] School of Physics, National University of Ireland, Galway, Ireland

ARTICLE INFO

Article history:
Received 8 October 2016
Received in revised form 6 March 2017
Accepted 20 April 2017
Available online 4 May 2017

Keywords:
Point separation
Functional data analysis
Logistic regression

ABSTRACT

We study point separation for the logistic regression model for Hilbert space-valued variables. We prove that the separating hyperplane can be found from a finite set of candidates and give an upper bound for the probability of point separation.

1. Introduction

The problem of point separation in logistic regression has been studied since as early as in Albert and Anderson (1984), and more than 700 papers have appeared in this research area since then. In Albert and Anderson (1984), the authors established the conditions on the maximum-likelihood estimate of the parameter vector in logistic regression model to exist in the case, where data come from the $\mathbb{R}^k$ space. Three scenarios of the arrangement of the data points were introduced: complete separation, quasi-complete separation and overlap. The authors proved that in the first two scenarios, the maximum-likelihood estimate of parameter vector does not exist or exists but is not unique, while in the third (overlap) scenario the maximum-likelihood estimate exists and is unique. The authors also suggested an iterative algorithm to be used when checking, whether or not the data points are in quasi-complete separation. Other methods on detecting overlap have been established as well (see, e.g. Christmann and Rousseeuw, 2001).

The majority of papers in this research area are devoted to proposing new parameter estimates that would exist and would have good theoretical properties in the case, where the data is already known to be in complete or quasi-complete separation. For example, the penalized maximum-likelihood estimator was introduced by Firth (1993) and asymptotically investigated by Gao and Shen (2007), while (Rousseeuw and Christmann, 2001) proposed a hidden logistic regression model to overcome the problem of point separation. Based on the recent activity in the field (see, e.g. Fu et al., 2015 or Sauter and Held, 2016, where they investigated which methods work well in quasi-complete separation, or Held and Sauter, 2016, where they proposed adaptive prior weighting to avoid complete separation), we believe that various results on the problem of point separation in logistic regression in the $\mathbb{R}^k$ setting are still of a great interest.

Moreover, with the recent expansion of Functional Data Analysis (FDA) (see Ramsay and Silverman, 2002, 2005 for an overview of the topic), the functional logistic regression models have been widely studied. The logistic estimate in

[*] Corresponding author.
E-mail address: a.kazakeviciute.12@ucl.ac.uk (A. Kazakevičiūtė).

http://dx.doi.org/10.1016/j.spl.2017.04.019
0167-7152/

abstract Hilbert spaces can be called a Naïve approach because the dimensionality reduction is achieved by simply 'cutting' the infinite-dimensional observation after some $k_n < n$ time point, where n is the number of sample points. In such a way, the first k_n parameter values are estimated via maximum-likelihood and the rest are set to zero. This approach is avoided in the literature for various reasons. For example, Escabias et al. (2007) argued that the Naïve approach in the context of functional data introduces multicollinearity (strong dependence among predictors) which in turn causes inaccurate parameter estimates and increases their variance. Therefore, the standard approaches include dimensionality reduction based on Principal Component Analysis (PCA) or Partial Least Squares (PLS) (see, e.g. Escabias et al., 2004; Aguilera et al., 2008; Denhere and Billor, 2016; James, 2002) or by basis expansion with some added penalty (see e.g. Aguilera-Morillo et al., 2013 or Müller, 2005). In none of these cases, consistency of functional logistic regression model parameter was established, mainly because the optimal rule for selecting the number of principal components or basis functions has not been established. The closest attempt to provide the theoretical justification of such a rule was done in Müller and Stadtmüller (2005). However, in the latter work, the authors approximated infinite-dimensional model by a finite-dimensional one without proving that the error of such an approximation tends to 0.

There are two theoretical contributions of this work. First is that we provide a theorem which transforms the problem of finding the separating hyperplane from the set of infinitely many elements into the feasible problem of finding it from the finite set of candidate hyperplanes and we describe how to construct such a set. We believe this theorem could speed up various established algorithms used by practitioners for determining, whether or not a maximum-likelihood estimate exists for the given datasets. Second contribution is that we provide an upper bound of the probability of the event that a sample is in quasi-complete separation by giving its upper bound. As a corollary of the latter result, we derive the minimal requirements on the selection of the dimension k_n for projection the data so that the consistency of the resulting functional logistic estimate could be expected. Such result could advance the study of the consistency of logistic classifier in abstract Hilbert spaces, for example, where weaker assumptions than those in Müller and Stadtmüller (2005) could be achieved.

2. Logistic estimate in abstract Hilbert spaces

Let E be a separable Hilbert space with the inner product $\langle\cdot,\cdot\rangle$. Let $X \in E$ be a Hilbert space-valued random variable and Y a random variable, gaining values -1 and 1, with conditional probabilities (w.r.t. X), $1 - p_{\theta_0}(X)$ and $p_{\theta_0}(X)$, respectively. Here, $\theta_0 \in E$ is an unknown parameter and

$$p_\theta(x) = \frac{1}{1+\mathrm{e}^{-\langle\theta,x\rangle}}, \quad \theta, x \in E.$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from the distribution of (X, Y). For $\theta, x \in E$ and $y \in \{-1, 1\}$ define

$$m_\theta(x, y) = \log(1 + \mathrm{e}^{-y\langle\theta,x\rangle})$$

and denote

$$M_n(\theta) = \overline{m_\theta(X, Y)} = \frac{m_\theta(X_1, Y_1) + \cdots + m_\theta(X_n, Y_n)}{n}, \quad M(\theta) = \mathrm{E}m_\theta(X, Y).$$

Obviously,

$$m_\theta(x, 1) = -\log p_\theta(x), \quad \text{and} \quad m_\theta(x, -1) = -\log(1 - p_\theta(x)).$$

Therefore, $M_n(\theta)$ might be interpreted as the logarithm of the quasi-likelihood function, multiplied by $-1/n$. Naturally, for various practical tasks, it is of great interest to provide an estimate of p_θ.

Let (E_k) be some fixed sequence of the linear subspaces of the space E such that the following conditions are satisfied: (1) $\dim E_k = k$ for all k, (2) $E_k \subset E_{k+1}$ for all k, and (3) $\overline{\bigcup_k E_k} = E$. For any k and n define

$$\hat{\theta}_{kn} = \arg\min_{\theta \in E_k} M_n(\theta). \tag{1}$$

Then, fix some sequence (k_n) and set

$$\hat{\theta} = \hat{\theta}_{k_n n} \quad \text{and} \quad \hat{p} = p_{\hat{\theta}}. \tag{2}$$

We will call $\hat{p}$ the *logistic estimate* of the conditional probability p_{θ_0}. For example, let $E = L^2(T)$ with the usual inner product

$$\langle\theta, x\rangle = \int_T \theta(t)x(t)\mathrm{d}t,$$

where $T \subset \mathbb{R}$ is an interval. The standard method for obtaining logistic estimate from a given sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is expanding X and θ via selected basis functions $\{e_j\}$

$$X_i(t) = \sum_{j=1}^{\infty} X_{ij} e_j(t), \qquad \theta(t) = \sum_{j=1}^{\infty} \theta_j e_j(t),$$

choosing $k = k_n$ and then using (1), where $E_k = \left\{ \sum_{j=1}^{k} c_j e_j \mid c_1, \dots, c_k \in \mathbb{R} \right\}$. The number k_n of basis functions to be used is usually selected less than n so that the parameter vector could be estimable. However, there are two open problems. First is that (as discussed before) the estimate (1) does not exist, if sample points are separable. This results in convergence to a false estimate which causes biased results. Second problem is that it is not clear how to select k_n with respect to n so that the resulting estimate would be consistent, for example. In Section 4, we solve the first problem, where we describe how separation of points can be checked against in practice. In Section 5, we partially solve the second problem, where we give the minimal requirements for k_n so that consistency of the resulting estimate (1) could be expected.

Remark 1. If $\theta \in E_k$, then $\langle \theta, X \rangle = \langle \theta, X^{(k)} \rangle$, where $X^{(k)}$ is the orthogonal projection of X on the space E_k. Therefore, $\hat{\theta}_{kn}$ is obtained only from $X_i^{(k)}$, $i = 1, \dots, n$. One could get a wrong idea that then the data are from $\mathbb{R}^k$ and we do not need to consider the general case when calculating the probability of point separation. However, the situation is more difficult than this. While the conditional probability of $Y = 1$, w.r.t. X, is denoted by $p_\theta(X)$ and has a nice expression, the same conditional probability w.r.t. $X^{(k)}$ is not $p_\theta(X^{(k)})$ but $\mathsf{E}^{X^{(k)}} p_\theta(X)$, where $\mathsf{E}^{X^{(k)}}$ is the conditional expectation w.r.t. $X^{(k)}$.

3. Separability of sample points

Let $(x_1, y_1), \dots, (x_n, y_n)$ be n vectors from $E_k \times \{-1, 1\}$. We will call them *sample points*. Let $a \neq 0$ be another vector from E_k. We will say that a vector *a separates sample points* if, for all i,

$$y_i \langle a, x_i \rangle \geq 0.$$

We say that sample points are *separable*, if there exists some $a \neq 0$ that separates them. Note that this definition is equivalent to the definition of quasi-complete separation in the $\mathbb{R}^k$ case, established by Albert and Anderson (1984).

Obviously, if some vector a separates sample points, then vector ca with any $c > 0$ also separates them. However, $-ca$ with any $c > 0$ does not separate them. The separability of sample points has also a geometric interpretation. Any nonzero vector a corresponds to a hyperplane H_a which is defined by the equation $\langle a, x \rangle = 0$ (note that 0 is used in this equation due to the fact that in this work we consider the logistic model without an intercept term). The vector a is then a normal of a hyperplane H_a. The subsets of E, defined by inequalities $\langle a, x \rangle \geq 0$ and $\langle a, x \rangle \leq 0$, are then called *half-spaces* of E. If we change a to ca with $c > 0$, the associated hyperplane as well as the associated half-spaces will not change. If we change a to $-ca$ with $c > 0$, the associated hyperplane will not change but the associated half-spaces will have the reversed order. If a' is not proportional to a, the associated hyperplanes differ. Therefore, a hyperplane defines a normal to a precision up to a constant c. Moreover, a hyperplane uniquely defines the pair of half-spaces, rather than individual half-spaces. If we want a hyperplane to define a normal to a precision up to a positive constant c, we have to introduce an *oriented hyperplane*. Formally speaking, an oriented hyperplane is a hyperplane with a fixed unit length normal. An oriented hyperplane uniquely defines individual half-spaces, and we can call one of the two half-spaces *an upper half-space*, and another one *a lower half-space*. For example, the upper half-space is defined by the equation $\langle a, x \rangle \geq 0$, where a is that fixed normal. If a separates sample points and H is the corresponding hyperplane, we can say that points from different groups fall into different half-spaces. Of course, one has to keep in mind that those half-spaces overlap, that is, points on the hyperplane belong to both half-spaces. If H is an oriented hyperplane and $a/\|a\|$ is its fixed normal, then points from the group $y = 1$ belong to the upper half-space, while the rest belong to the lower half-space.

Denote by $X_i^{(k)}$ the projection of the point X_i on the space E_k. We will say that the sample $(X_1, Y_1), \dots, (X_n, Y_n)$ is *k-separable*, if the random sample points $(X_1^{(k)}, Y_1), \dots, (X_n^{(k)}, Y_n)$ are separable. The latter definition defines some subset of the event space Ω that consists of $\omega \in \Omega$ for which the sample points

$$(X_1^{(k)}(\omega), Y_1(\omega)), \dots, (X_n^{(k)}(\omega), Y_n(\omega)) \tag{3}$$

are separable. It is well-known that if the sample is k-separable, then the maximum quasi-likelihood estimate of θ does not exist or is not unique (Albert and Anderson, 1984).

When searching for a separating hyperplane, there are infinitely many candidate hyperplanes to consider. This fact makes the theoretical investigation of the probability that the sample is separable harder since the sums of infinitely many possible separating hyperplanes are involved in the calculations. In practice, the search area of an algorithm for finding the possible separating hyperplane is restricted to some set of finite number of candidate hyperplanes that is guaranteed to contain the true separating hyperplane. However, this fact has not been proved yet. In the following section, we give a proof for this.

4. Criteria for separability

Let $(e_1, \dots, e_k)$ be the orthonormal basis in E_k and let x^i denote the coordinates of a vector $x \in E_k$ in that basis system, that is,

$$x = x^1 e_1 + \dots + x^k e_k.$$

For any $x_1, \dots, x_k \in E_k$, we will denote

$$\det[x_1, \dots, x_k] = \begin{vmatrix} x_1^1 & \cdots & x_k^1 \\ \vdots & \ddots & \vdots \\ x_1^k & \cdots & x_k^k \end{vmatrix}.$$

Obviously, det is a k-linear antisymmetric form.

Since $\det[x_1, \dots, x_{k-1}, x]$ is a linear function w.r.t. x, it is of the form $\langle a, x\rangle$ with some a. In other words, there exists a unique a such that, for all x, $\det[x_1, \dots, x_{k-1}, x] = \langle a, x\rangle$. Obviously, a is a function of $x_1, \dots, x_{k-1}$.

If $x_1, \dots, x_{k-1}$ are linearly dependent, the determinant is equal to 0 for all x, that is, $a = 0$. Conversely, if $a = 0$, then $x_1, \dots, x_{k-1}$ are linearly dependent (otherwise we could find x_k for which $x_1, \dots, x_k$ are linearly independent which would imply that the determinant is nonzero, that is, $a \neq 0$).

There is an intrinsic relationship between a determinant and a hyperplane. If $x_1, \dots, x_{k-1}$ are linearly independent, then $a \neq 0$ defines some hyperplane H_a. This hyperplane has the special property that points $x_1, \dots, x_{k-1}$ belong to it (because determinant is equal to 0 when any two columns in it are equal). In fact, it is the unique hyperplane that contains these points because all a that are perpendicular to all $x_1, \dots, x_{k-1}$ are proportional.

Suppose $n \geq k$. We will prove that when checking the separability of sample points it is enough to sort out the finite number of potential vectors a that possibly separate the sample. Note that the set of such possible vectors is random. For any family of distinct indices $(i_1, \dots, i_{k-1}) \subset \{1, \dots, n\}$ denote by $Z_{i_1 \dots i_{k-1}}$ a random vector from E_k such that, for all $x \in E_k$,

$$\det[X_{i_1}^{(k)}, \dots, X_{i_{k-1}}^{(k)}, x] = \langle Z_{i_1 \dots i_{k-1}}, x\rangle.$$

Let

$$S = \{\pm Z_{i_1 \dots i_{k-1}} \mid (i_1, \dots, i_{k-1}) \subset \{1, \dots, n\}\}.$$

Note that the set S is finite and the number of elements in it is

$$|S| = 2\binom{n}{k-1}.$$

Theorem 1. *If $n \geq k$, then the sample is k-separable if and only if the points $X_1^{(k)}, \dots, X_n^{(k)}$ can be separated by some vector from the set S.*

Remark 2. If $n \leq k$, the points are always k-separable. If $n = k$, any properly oriented hyperplane passing through $k - 1$ point separates the sample points. If $n = k - 1$, there is only one hyperplane passing through all the sample points, and it separates the sample points, regardless of its orientation. If $n < k - 1$, then there are infinitely many hyperplanes passing through the sample points, and all of them separates the sample points, regardless of their orientation.

5. Probability that sample is separable

Theorem 1 implies that the sample is k-separable if and only if, for some distinct $i_1, \dots, i_{k-1}$,

$$\forall i \quad Y_i \det[X_{i_1}^{(k)}, \dots, X_{i_{k-1}}^{(k)}, X_i^{(k)}] \geq 0 \tag{4}$$

or

$$\forall i \quad Y_i \det[X_{i_1}^{(k)}, \dots, X_{i_{k-1}}^{(k)}, X_i^{(k)}] \leq 0. \tag{5}$$

Let q_{kn} be the probability of such event. We will need the following assumption on the distribution of X:

(FR) We will say that the distribution of X is *of full rank*, if $\mathsf{P}(\langle\theta, X\rangle = 0) = 0$, for all $\theta \neq 0$.

Theorem 2. *If (FR) holds and $n \geq k$, then with some $q < 1$ that does not depend neither on n nor on k,*

$$q_{kn} \leq 2\binom{n}{k-1} q^{n-k+1}.$$

Theorem 2 gives an upper bound of the probability that sample points are k-separable. It may not be the lowest upper bound but it gives a good understanding about what sequence (k_n) should be chosen for projecting X so that we could expect estimate (1) to be consistent. The following Corollary summarizes this.

Corollary 2.1. *If $k_n/n \to 0$, then $q_{k_n n} \to 0$.*

For example, if we take $k_n = \lfloor\sqrt{n}\rfloor$, the probability that the logistic estimate exists is close to 1, for n large enough.

6. Discussion

The results presented in this work can be directly used for the theoretical investigations of the properties of logistic classifier in abstract Hilbert spaces, such as consistency, for example. When working with functional data, an infinitely-dimensional parameter vector cannot be uniquely estimated only from the finite number of observations. Therefore, a common practice is to 'cut' the parameter vector θ after, say, the kth coordinate, and set the remaining coordinates to zero. However, this approach is avoided in literature, mainly due to the fact that the quantitative rule of selecting such k in a way that the resulting estimate would have desirable theoretical properties has not been established yet. Theorem 2 contributes to the understanding of what a good rule for selecting k could possibly be. Corollary 2.1 tells us that at least $k_n/n \to 0$ should be required so that we could expect a maximum quasi-likelihood estimate in logistic regression models in abstract Hilbert spaces to have desirable theoretical properties.

Acknowledgments

This work was supported by the Department of Statistical Science, University College London, United Kingdom, and Singapore Bioimaging Consortium, Agency for Science, Technology and Research, Singapore.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.spl.2017.04.019.

References

Aguilera, A.M., Escabias, M., Valderrama, M.J., 2008. Discussion of different logistic models with functional data. Application to systemic lupus erythematosus. Comput. Statist. Data Anal. 53 (1), 151–163.

Aguilera-Morillo, M.C., Aguilera, A.M., Escabias, M., Valderrama, M.J., 2013. Penalized spline approaches for functional logit regression. TEST 22 (2), 251–277.

Albert, A., Anderson, J.A., 1984. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71 (1), 1–10.

Christmann, A., Rousseeuw, P.J., 2001. Measuring overlap in binary regression. Comput. Statist. Data Anal. 37 (1), 65–75.

Denhere, M., Billor, N., 2016. Robust principal component functional logistic regression. Comm. Statist. Simulation Comput. 45 (1), 264–281.

Escabias, M., Aguilera, A., Valderrama, M.J., 2004. Principal component estimation of functional logistic regression: discussion of two different approaches. J. Nonparametr. Stat. 16 (3–4), 365–384.

Escabias, M., Aguilera, A.M., Valderrama, M.J., 2007. Functional PLS logit regression model. Comput. Statist. Data Anal. 51 (10), 4891–4902.

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80 (1), 27–38.

Fu, P., Panneerselvam, A., Clifford, B., Dowlati, A., Ma, P.C., Zheng, G., Halmos, B., Leidner, R.S., 2015. Simpsons paradoxaggregating and partitioning populations in health disparities of lung cancer patients. Stat. Methods Med. Res. 24 (6), 937–948.

Gao, S., Shen, J., 2007. Asymptotic properties of a double penalized maximum likelihood estimator in logistic regression. Statist. Probab. Lett. 77 (9), 925–930.

Held, L., Sauter, R., 2016. Adaptive prior weighting in generalized regression. Biometrics http://dx.doi.org/10.1111/biom.12541.

James, G.M., 2002. Generalized linear models with functional predictors. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (3), 411–432.

Müller, H.-G., 2005. Functional modelling and classification of longitudinal data. Scand. J. Stat. 32 (2), 223–240.

Müller, H.-G., Stadtmüller, U., 2005. Generalized functional linear models. Ann. Statist. 33 (2), 774–805.

Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis: Methods and Case Studies. Springer.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer.

Rousseeuw, P.J., Christmann, A., 2001. Robustness against separation and outliers in logistic regression. Comput. Statist. Data Anal. 43 (3), 315–332.

Sauter, R., Held, L., 2016. Quasi-complete separation in random effects of binary response mixed models. J. Stat. Comput. Simul. 86 (14), 2781–2796.