Safe density ratio modeling

Kjell Konis, Konstantinos Fokianos

Safe Density Ratio Modeling

Kjell Konis[1] Konstantinos Fokianos[2]

[1]Institute of Mathematics, EPFL
e-mail: kjell.konis@epfl.ch

[2]Department of Mathematics & Statistics, University of Cyprus
e-mail: fokianos@ucy.ac.cy

First Version: April 2009

Abstract

An important problem in logistic regression modeling is the existence of the maximum likelihood estimators. Especially when the sample size is small, the maximum likelihood estimator of the regression parameters does not exist if the data are completely, or quasi–completely separated. Recognizing that this phenomenon has a serious impact on the fitting of the density ratio model–which is a semiparametric model whose profile empirical log-likelihood has the logistic form because of the equivalence between prospective and retrospective sampling–we suggest a linear programming methodology for examining whether the maximum likelihood estimators of the finite dimensional parameter vector of the model exist. It is shown that the methodology can be effectively utilized in the analysis of case control gene expression data by identifying cases where the density ratio model cannot be applied. It is demonstrated that naive application of the density ratio model yields to erroneous conclusions.

Keywords: biased sampling, differential expression, empirical likelihood, linear programming, separation

1 Motivation

Consider two independent samples of observations such that

$$\begin{array}{lll} X_{11},\ldots,X_{1n_1} & \text{is a random sample from} & g_1(x) \quad = \exp\left(\alpha+\beta^T h(x)\right) g_2(x), \\ X_{21},\ldots,X_{2n_2} & \text{is a random sample from} & g_2(x). \end{array} \tag{1}$$

In the above $g_i(x)$, $i=1,2$ are unknown probability density functions, α is an unknown scalar parameter, β is a d-dimensional vector of parameters and $h(x)$ is a d-dimensional vector which consists of known functions of X. Model (1) is called density ratio model because it specifies that the log–ratio of two unknown probability density functions is linear in some parameters, see Anderson (1972), Qin and Zhang (1997). The model is motivated by means of the standard logistic regression and the equivalence between prospective and retrospective sampling, Prentice and Pyke (1979). It should be mentioned that the class of distributions for which (1) holds is rather general and it includes the exponential family of distributions. Hence a vast collection of data types can be modeled by means of model (1).

An important observation is that when model (1) holds, and if $\beta=0$, then the two samples are identically distributed. We conclude that model (1) is useful to the semiparametric comparison of two samples in the sense that the densities $g_i(.)$, $i=1,2$ are left completely unspecified but the weight function $\exp\left(\alpha+\beta^T h(x)\right)$ depends on some finite dimensional parameter. Hence (1) provides a sound framework for addressing the problem of comparing two independent samples. In addition, it provides a compromise between the fully parametric and non–parametric approaches to the problem of testing equality of two distribution, see Qin et al. (2002), Kedem et al. (2004) and Fokianos et al. (2005), among others, for applications of the density ratio model to real data problems.

Inference regarding both finite and infinite dimensional parameters of model (1) has been studied by various authors assuming that the sample sizes tend to infinity in a suitable way. Following the empirical likelihood methodology, as advocated by Owen (2001), a parametric likelihood function for the finite dimensional parameters is obtained after profiling out the infinite dimensional parameter of the model. However, there are applications where the sample sizes are small and therefore direct application of the aforementioned techniques might suffer from non-existence or even non-convergence problems. To make this point clear, we follow Qin and Zhang (1997) and Fokianos et al. (2001) who show that the empirical log-likelihood is given by

$$l(\theta)\equiv l(\alpha,\beta) \quad = \quad -\sum_{i=1}^{2}\sum_{j=1}^{n_i}\log\left[1+\rho_1\exp(\alpha+\beta^T h(x_{ij}))\right]+\sum_{j=1}^{n_1}\left(\alpha+\beta^T h(x_{1j})\right) \tag{2}$$

where $\rho_1=n_1/n_2$. Furthermore, if $\hat{\theta}=(\hat{\alpha},\hat{\beta})'$ denotes the maximum likelihood estimator of θ, assuming that it exists, then it can be shown that

$$\hat{p}_{ij}=\frac{1}{n_2}\frac{1}{1+\rho_1\exp(\hat{\alpha}+\hat{\beta}^T h(x_{ij}))}. \tag{3}$$

where $p_{ij} = dG_2(x_{ij})$, the size of the jump of $G_2(\cdot)$ at the observed datum $X_{ij} = x_{ij}$. A consistent estimator for both of $G_1(.)$ and $G_2(.)$ can be constructed provided that the total sample size $n = n_1 + n_2$ tends to infinity such that $n_1/n_2 \to \rho_1$–see Qin and Zhang (1997) and Fokianos et al. (2001) for more.

Note that (2), after reparametrization, is equivalent to the standard logistic regression likelihood–a direct consequence of the equivalence between retrospective and prospective sampling as it was mentioned before. Hence, the finite dimensional vector of parameters $\theta = (\alpha, \beta^T)^T$ can be estimated by any of the numerous statistical programs which include logistic regression modeling. A standard approach for computing the maximum likelihood estimate of θ is to use the Fisher scoring method which, under some regularity assumptions, yields a sequence of approximations that converge to the maximum likelihood estimators of α and β. Occasionally this sequence of approximations does not converge to a finite value. Therefore application of the density ratio model is questionable, see Fokianos (2008) who proposes penalization for the resolution of this problem and Davidov and Iliopoulos (2009) who give necessary and sufficient conditions for the existence of maximum likelihood estimators for both the finite and infinite dimensional parameter. The non existence issue of the maximum likelihood estimators for the logistic regression model, sometimes, referred to as *monotone likelihood* or *infinite parameters*, occurs when a condition among the sample points, known as *separation*, prevails.

More specifically, the concept of separation in the logistic regression context was introduced by Albert and Anderson (1984) who showed that the sample points can be classified into one of three mutually exclusive configurations: *complete separation*, *quasicomplete separation* or *overlap*. To rephrase the concept of separation in terms of the density ratio model, define the sample points by the $(d+1)$-dimensional vector $u_{ij} = (1, h^T(x_{ij}))^T$ for $j = 1, 2, \ldots, n_i$, $i = 1, 2$. Then, there exists complete separation among the sample points if there is a nonzero vector γ such that

$$\gamma^T u_{ij} < 0 \quad \text{when} \quad i = 1 \quad \text{and} \quad \gamma^T u_{ij} > 0 \quad \text{when} \quad i = 2.$$

In other words, there is complete separation when there exits a hyperplane H such that all the values of the first sample lie strictly on one side of H and all the values of the second sample lie strictly on the other side of H. If complete separation is not present among the sample points but there exists a nonzero vector γ satisfying

$$\gamma^T u_{ij} \leq 0 \quad \text{when} \quad i = 1 \quad \text{and} \quad \gamma^T u_{ij} \geq 0 \quad \text{when} \quad i = 2,$$

then there is quasicomplete separation among the sample points. Finally, if there is no nonzero vector γ satisfying the last set of equations, then the sample points overlap. Albert and Anderson (1984) and Santner and Duffy (1986) show that there is a finite and unique maximum likelihood estimate of the logistic regression model parameter θ provided that there exists overlap among the sample points. Therefore, the density ratio model (1) is applicable only when there is overlap among the sample points according to the above conventions. The aim of this contribution is to suggest a test for checking whether there exists separation among sample points. By providing such a test, we answer the question of the applicability of model (1).

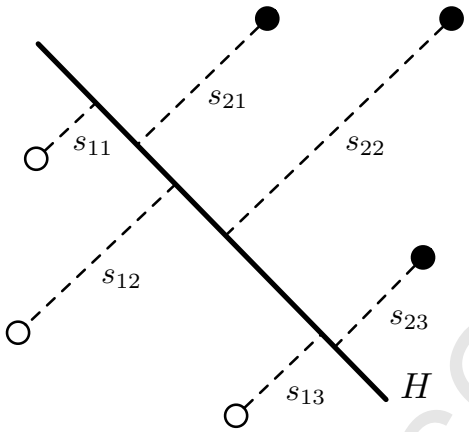

Figure 1: Six sample points completely separated by the hyperplane H. The inner products $s_{ij} = u_{ij} \cdot \gamma$ are strictly positive when $i = 2$ and strictly negative when $i = 1$.

2 Methodology

The previous discussion shows that before fitting the density ratio model, we need to test for separation among the sample points. We outline a method based on linear programming. In the case that separation is found among the sample points we report that the density ratio model is not applicable. Otherwise (i.e., in the case of an overlap configuration), we fit the underlying logistic regression and proceed with inference, as in Fokianos et al. (2001), for instance.

The use of linear programming to check for separated configurations of the sample points was first suggested in Albert and Anderson (1984) where the authors specified the necessary constraints but not an objective function. Furthermore, Santner and Duffy (1986) described a mixed integer linear program capable of distinguishing between the three configurations while Silvapulle and Burridge (1986) and Clarkson and Jennrich (1991) used linear programming to check for the existence of a finite maximum likelihood estimate for the logistic regression model. We test for separation based on recent work by Konis (2007) because it is relatively simpler to apply. In addition, the test is implemented by means of the `safeBinaryRegression` R package and therefore it is easily accessible to data analysts, Konis (2009). In what follows, we assume that the $n \times (d + 1)$ design matrix U with rows given by u_{ij}^T is of full rank so that the associated logistic regression model is well-defined.

Recalling the notation from the previous section, let s_{ij} be the inner product of u_{ij} and a vector $\gamma = (\gamma_1, \ldots, \gamma_{d+1})^T$ perpendicular to H, as shown in Figure 1. The linear program proposed in Konis (2007) seeks to maximize the sum of the absolute values of the s_{ij} subject to the constraint that $s_{1j} \leq 0$ and $s_{2j} \geq 0$. In the notation of linear

programming (see Konis (2007)) this problem is stated by seeking the vector γ such that

$$\begin{array}{ll} \text{maximize:} & \sum_{j=1}^{n_2} u_{2j} \cdot \gamma - \sum_{j=1}^{n_1} u_{1j} \cdot \gamma, \\ \text{subject to:} & u_{2j} \cdot \gamma \geq 0, \quad j = 1, 2, \ldots, n_2, \\ & u_{1j} \cdot \gamma \leq 0, \quad j = 1, 2, \ldots, n_1, \\ & \gamma_l \in R, \quad l = 1, \ldots, d+1. \end{array} \tag{4}$$

If there is overlap among the sample points then there is no nonzero vector γ that satisfies the constraints imposed by (4). In this case $\gamma \equiv 0$ is the only feasible point and the optimal value of the objective function is zero. On the other hand, if there is either complete or quasicomplete separation among the sample points then there is γ such that at least one of the inner products is nonzero. However, if γ satisfies the constraints of (4) then so does $k\gamma$ for real $k > 1$. Hence, in the separated case, the linear program is unbounded.

The `safeBinaryRegression` package uses `lp_solve` (Berkelaar et al., 2009) via the `lpSolveAPI` (Konis and lp_solve, 2009) R package to solve (4). If the linear program is bounded and the optimal value of the objective function is zero then there is overlap among the sample points and the density ratio model is appropriate. However, if the linear program is unbounded then the density ratio model cannot be applied to two sample problems.

3 Examples

Consider data from the Affymetrix Spike-In study (Irizarry et al., 2003). This experiment was used by Affymetrix to develop the MAS 5.0 preprocessing algorithms. The data set consists of measurements from 12626 human genes. We focus on the two array groups among the 14 array groups that contain 12 replicates each, leading to a case-control design with a total of 24 Affymetrix HGU95 biochips. For the data analysis, probe level summaries are computed using the RMA method (Irizarry et al., 2003), yielding intensity levels on a $\log 2$ scale. We test whether the density ratio model can be employed for inference for these data in the sense of comparing the distributions of cases and controls. Table 1 shows the number of separated cases for the whole data. It turns out that when $h(\cdot)$ is chosen to be univariate we obtain 15 separated samples and therefore the density ratio model is not applicable for these particular genes. The results are consistent for all choices of $h(\cdot)$ since both $h(x) = \log x$ and $h(x) = \sqrt{x}$ are monotone functions of x when $x > 0$. Note that the choices for $h(\cdot)$ are motivated by considering the log ratio of some standard distributions. For instance, choosing $h(x) = x$ can be motivated by calculating the log ratio of two normal densities with identical variances. The non–applicability of the density ratio model turns out to be a more serious issue when $h(\cdot)$ is chosen to be a two dimensional function. In this case the number of separated samples increases considerably. Figure 2 demonstrates clearly the concept of separation for some selected genes from the Affymetrix Spike-In study.

As a second example we considered the data set described in the breast cancer study by Hedenfalk et al.

$h(x)$	x	$\log x$	$\sqrt{x}$	$(x, \log x)^T$	$(x, x^2)^T$
#	15	15	15	556	1210

Table 1: The number of separated cases for Affymetrix Spike-In study for different choices of $h(\cdot)$.

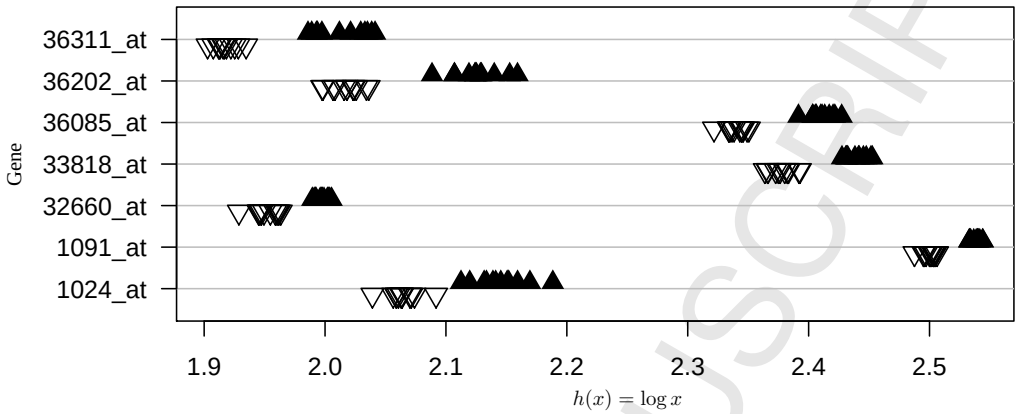

Figure 2: Illustrating separation for seven genes in the Affymetrix Spike-In data. The filled triangles represent one sample and the up-side-down open triangles represent the other.

(2001). These data were produced using cDNA technology and they describe intensity ratios of 3226 genes across 22 breast cancer samples. One of the aims of the experiment conducted by Hedenfalk et al. (2001) was to detect differences in the genetic profiles between the BRCA1 and BRCA2 mutations. For these two conditions there are 7 and 8 arrays available, respectively. Notice that these data were log_{10} transformed and therefore they assume both positive and negative values. We apply the density ratio model with $h(x) = x$ (respectively, $h(x) = x^2$) and we obtain 19 (10, respectively) separated cases. In addition, the choice $h(x) = (x, x^2)^T$ yields 28 separated cases.

Some interesting observations follow next. Figure 3 shows data from a specific gene from the breast cancer study. For this particular case, it is worthwhile to notice that the density ratio model (1) is applicable when $h(x) = x$ or $h(x) = x^2$–a projection of the data onto either the x or the y axis produces an overlapped configuration among the sample points. However, the plot shows that when $h(x) = (x, x^2)^T$ then there is complete separation among the sample points and therefore model (1) is not applicable with this choice of $h(\cdot)$.

Furthermore, Figure 4 demonstrates the consequences of naively applying the density ratio model without testing for separation. It shows plots of the estimated distribution functions of genes 1 and 521 from the breast cancer study data by fitting the density ratio model using $h(x) = x^2$ and without testing for separation. The cdf of

$g_2(.)$, say $G_2(.)$, is estimated by using (3) and

$$\hat{G}_2(x) = \sum_{i=1}^{2} \sum_{j=1}^{n_i} \hat{p}_{ij} I(X_{ij} \leq x),$$

while

$$\hat{G}_1(x) = \sum_{i=1}^{2} \sum_{j=1}^{n_i} \hat{p}_{ij} \exp(\hat{\alpha}_1 + \hat{\beta}^T h(x_{ij})) I(X_{ij} \leq x),$$

where $I(\cdot)$ denotes the indicator function. The associated logistic regression model in the upper plot (gene 1) has an overlap among the sample points. The lower panel (gene 521) appears to indicate that some of the data points overlap. However, for this particular gene, the corresponding sample points are completely separated. Therefore model (1) is not even applicable. Hence, testing for separation among the sample points is essential prior to fitting the density ratio model.

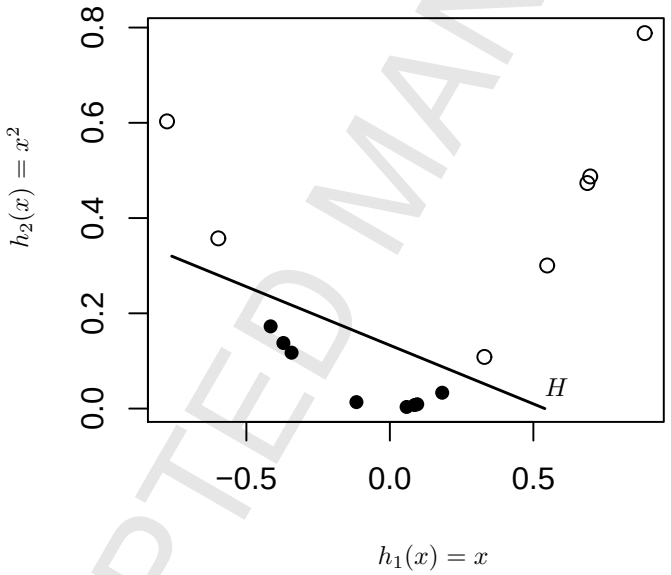

Figure 3: The open circles correspond to gene 3094 of condition BRCA1 and the filled circles correspond to condition BRCA2 for the same gene. H is the separating hyperplane when $h(x) = (x, x^2)^T$ in (1).

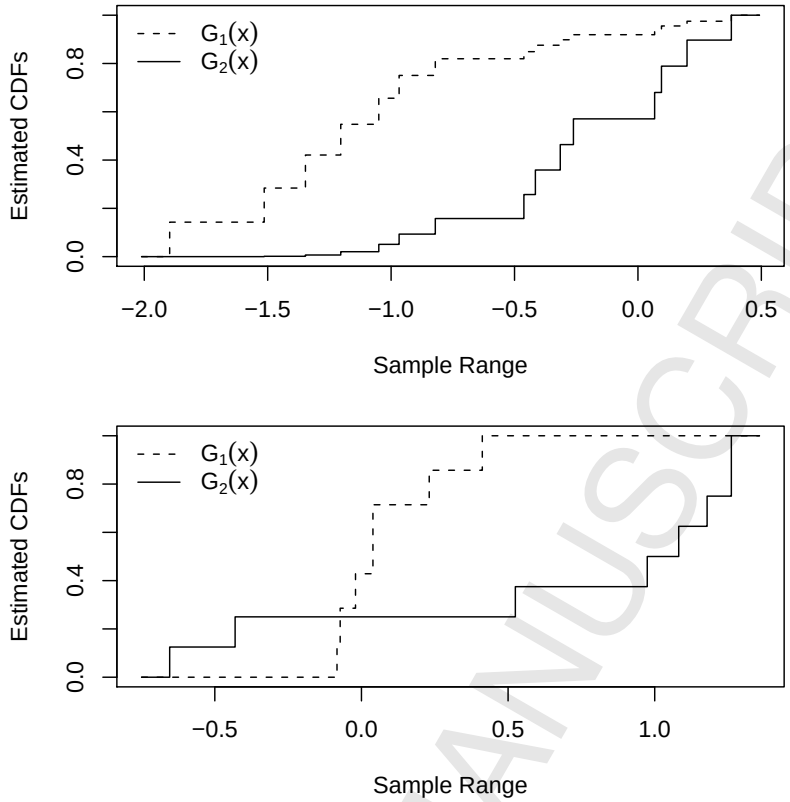

Figure 4: Two comparisons of the estimated cumulative distributions. There is overlap among the sample points in the upper plot and complete separation in the lower plot. However, there are no features in the second plot that suggest anything is amiss.

4 Conclusions

We have outlined a methodology for testing for separation among the sample points prior to fitting the density ratio model. It has been shown, using real data examples, that the proposed methodology effectively identifies cases where the model is not applicable. Some misleading results unfold if the technique is not applied properly.

Acknowledgements

The authors thank A. Davison for constructive comments. This work was carried out while K. Fokianos was visiting the Institute of Mathematics, EPFL. The hospitality of all faculty members is greatly acknowledged.

References

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

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika 59*, 19–35.

Berkelaar, M., K. Eikland, and P. Notebaert (2009). *lp_solve*.

Clarkson, D. B. and R. I. Jennrich (1991). Computing extended maximum likelihood estimates for linear parameter models. *Journal of the Royal Statistical Society, Series B 53*, 417–426.

Davidov, O. and G. Iliopoulos (2009). On the existence and uniqueness of the npmle in biased sampling models. *Journal of Statistical Planning and Inference 139*, 176–183.

Fokianos, K. (2008). Comparing two samples by penalized logistic regression. *Electronic Journal of Statistics 2*, 564–580.

Fokianos, K., B. Kedem, J. Qin, and D. A. Short (2001). A semiparametric approach to the one–way layout. *Technometrics 43*, 56–64.

Fokianos, K., I. Sarrou, and I. Pashalidis (2005). A two-sample model for the comparison of radiation doses. *Chemometrics and Intelligent Laboratory Systems 79*, 1–9.

Hedenfalk, I., D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent (2001). Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med. 344*, 539–48.

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed (2003). Summaries of Affymetrix GeneChip probe level data. *31*, e15.

Kedem, B., D. Wolff, and K. Fokianos (2004). Statistical comparisons of algorithms. *IEEE Transactions on Instrumentation and Measurement 53*, 770–776.

Konis, K. (2007). *Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models*. DPhil in Computational Statistics, Worcester College, University of Oxford, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, United Kingdom.

Konis, K. (2009). *safeBinaryRegression: Safe Binary Regression*. R package version 0.1.

Konis, K. and lp_solve (2009). *lpSolveAPI: Interface to lp_solve 5.5.0.14*. R package version 5.5.0.14.

Owen, A. B. (2001). *Empirical Likelihood*. Boca Raton, Florida: Chapman and Hall/CRC.

Prentice, R. L. and R. Pyke (1979). Logistic disease incidence models and case–control studies. *Biometrika 66*, 403–411.

Qin, J., M. Barwick, R. Ashbolt, and T. Dwyer (2002). Quantifying the change of melanoma incidence by Breslow thickness. *Biometrics 58*, 665–670.

Qin, J. and B. Zhang (1997). A goodness of fit test for the logistic regression model based on case–control data. *Biometrika 84*, 609–618.

Santner, T. J. and D. E. Duffy (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika 73*, 755–758.

Silvapulle, M. J. and J. Burridge (1986). Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *Journal of the Royal Statistical Society, Series B 48*, 100–106.