Brazilian Journal of Probability and Statistics
2017, Vol. 31, No. 3, 561–568
DOI: 10.1214/16-BJPS324

A note on curvature influence diagnostics in elliptical regression models

Mauricio Zevallos and Luiz Koodi Hotta

University of Campinas

Abstract. In this paper, we derive analytical expressions for the curvature influence statistic proposed by Cook [*J. Roy. Statist. Soc. Ser. B* **48** (1986) 133–169] in elliptical regression models under a data perturbation scheme. A relationship between the curvature statistics and the residuals is established and the effects of the shape parameter are assessed. The results reveal the role of the shape parameter in applying the curvature influence diagnostics technique.

1 Introduction

The local influence method introduced by Cook (1986) is a useful paradigm for detecting observations that have a strong effect on a model; see Zhu et al. (2007) and references therein. Taking advantage of the concepts of differential geometry, Cook (1986) assessed the effect (influence) of minor perturbations in the model through the curvature of the influence graph.

Galea et al. (1997, 2000) and Liu (2000), among others, have investigated the use of curvature influence diagnostics in linear regression models with elliptical errors. The class of elliptical errors includes the normal distribution, Student-*t* distribution, logistic distribution, and exponential power distribution.

Following the strategy of Schwarzmann (1991), here we derive analytic expressions for curvature influence diagnostics in elliptical regression models under a data perturbation scheme. The results are important to understand the role of the shape parameter in finding influential points through appropriate curvature statistics. In addition, practical consequences in terms of data analysis are discussed. Previous works on this topic have calculated the curvature numerically, mainly focusing on a subset of parameters, say the regression or scale parameters.

The remainder of this paper is organized as follows. In Section 2, we present the main results, and some conclusions and the final remarks are given in Section 3.

2 Curvature influence statistic and residuals

Let $y = (y_1, \dots, y_n)^\top$ be observations generated by the (postulated) model

$$y = X\beta + \varepsilon, \tag{2.1}$$

Key words and phrases. Data perturbation, local influence, influential observations.
Received February 2015; accepted June 2016.

where X is a known $n \times p$ matrix of rank p, β is a $p \times 1$ vector of unknown parameters and ε is a random $n \times 1$ perturbation vector following an elliptical symmetric distribution with zero mean and covariance ϕI, with I being the identity matrix. Thus, the density of y is

$$f(y) = \frac{n}{\sqrt{\phi}} g(s, \eta), \tag{2.2}$$

$$s = \frac{1}{\phi}(y - X\beta)^\top (y - X\beta), \qquad s \geq 0, \tag{2.3}$$

where $\phi > 0$, g is an elliptical density generator and η is the shape parameter. This model is called the elliptical linear regression model; see Fang and Anderson (1990) and Galea et al. (1997) for details. In Table 1, we show the expressions for the function g associated with some very well-known elliptical distributions; see Galea et al. (2000).

The density of y depends on three parameters: β, ϕ and η. However, in this paper we consider the shape parameter η as fixed and known. Therefore, we only estimate the parameter vector $\boldsymbol{\theta} = (\beta, \phi)$, maximizing the log-likelihood denoted by $l(\boldsymbol{\theta})$. The maximum likelihood estimators satisfy

$$\hat{\beta} = (X^\top X)^{-1} X^\top y, \tag{2.4}$$

$$\hat{\phi} = -2\hat{W} \|e\|^2 / n, \tag{2.5}$$

where $\hat{W}$ is the expression of

$$W = \frac{1}{g}\left(\frac{\partial g}{\partial s}\right) \tag{2.6}$$

evaluated at the maximum likelihood estimates of $\boldsymbol{\theta}$ and e is the vector of residuals,

$$e = y - X\hat{\beta}. \tag{2.7}$$

Suppose y is perturbed according to a data perturbation scheme

$$\tilde{y} = y + \omega \tag{2.8}$$

Table 1 *Some multivariate elliptical distributions. Expression A is defined in (2.15) and for each distribution c is a normalizing constant.*

Density	$g(s, \eta)$	A
Normal	$c \exp(-s/2)$	2
Logistic	$c \exp(-s)/[1 + \exp(-s)]^2$	$2 + 2n \exp(\hat{s})/[1 - \exp(\hat{s})]^2$
Student-t	$c(1 + s/\eta)^{-(\eta+n)/2}$	$2 - 2n/(n + \eta)$
Power Exponential	$c \exp(-s^\eta/2)$	2η

with perturbation vector $\omega = (\omega_1, \dots, \omega_n)^\top$. As a result, we obtain the *perturbed log-likelihood*

$$l(\boldsymbol{\theta}|\omega) = -\frac{n}{2}\ln\phi + \ln g\left(\frac{1}{\phi}(y+\omega - X\beta)^\top(y+\omega-X\beta), \eta\right). \tag{2.9}$$

Let $\omega_0 = (0, \dots, 0)^\top$ be the point of null perturbation, the point which satisfies $l(\boldsymbol{\theta}|\omega_0) = l(\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_\omega$ be the maximum likelihood estimates under $l(\boldsymbol{\theta})$ and $l(\boldsymbol{\theta}|\omega)$, respectively, and note that $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\omega_0}$.

To assess the influence of minor perturbations ω on the postulated model, Cook (1986) suggested using the likelihood displacement $\mathrm{LD}(\omega) = 2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_\omega)]$. The objective is to find the vector of the maximum normal curvature evaluated at ω_0 and $\hat{\boldsymbol{\theta}}$, denoted by $C = (c_1, \dots, c_n)^\top$. This is done by solving

$$|F - \lambda I| = 0, \tag{2.10}$$

where

$$F = \frac{\partial^2 \mathrm{LD}(\omega)}{\partial\omega\partial\omega^\top} = \Delta^\top(-G^{-1})\Delta, \tag{2.11}$$

I is the identity matrix,

$$\Delta = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\omega^\top} \tag{2.12}$$

is a matrix of order $q \times n$ and $-G$ is the observed information matrix:

$$G = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}. \tag{2.13}$$

The curvature vector C is the normalized eigenvector associated with the largest absolute eigenvalue of F, λ_c, and the curvature is $C_{\max} = 2\lambda_c$.

For elliptical regression models, Galea et al. (1997) and Liu (2000) derived the expressions for the observed information matrix and Δ, respectively. In this paper we rewrite these expressions in a convenient way. Thus, the components of $\Delta^\top = [\Delta_1^\top, \Delta_2^\top]$ evaluated at ω_0 and $\hat{\boldsymbol{\theta}}$ are

$$\Delta_1^\top = -\frac{2}{\hat{\phi}}\hat{W}X, \qquad \Delta_2^\top = -\frac{\hat{W}}{\hat{\phi}^2}Ae, \tag{2.14}$$

where

$$A = 2 - \frac{n}{\hat{W}^2}\left(\frac{\partial\hat{W}}{\partial s}\right), \tag{2.15}$$

and A for some multivariate elliptical distribution is given in Table 1. The observed information matrix evaluated at ω_0 and $\hat{\boldsymbol{\theta}}$ is

$$-G^{-1} = \operatorname{diag}\left\{-\frac{\hat{\phi}}{2\hat{W}}(X^\top X)^{-1}, \frac{4\hat{\phi}^2}{n}A^{-1}\right\}. \tag{2.16}$$

Then, substituting (2.14) and (2.16) in (2.11) yields

$$F = -\frac{2\hat{W}}{\hat{\phi}} X(X^\top X)^{-1} X^\top + \frac{2\hat{W}A}{\hat{\phi}}\left(\frac{2\hat{W}\|e\|^2}{n\hat{\phi}}\right) hh^\top, \tag{2.17}$$

where $h = e/\|e\|$ is the normalized residual vector. By replacing (2.5) in the last expression, we obtain

$$F = \frac{n}{\|e\|^2} H + \frac{An}{\|e\|^2} hh^\top,$$

where $H = X(X^\top X)^{-1}X^\top$ and $hh^\top$ are projection matrices with $Hh = 0$. Therefore the F matrix has eigenvalues $\lambda_1 = n/\|e\|^2$ (with multiplicity p) and $\lambda_2 = An/\|e\|^2$ (with multiplicity 1). This result generalizes the findings of Schwarzmann (1991) for normal errors. The next theorem summarizes these findings.

Theorem 2.1. *Let a elliptical regression model* (2.1)–(2.2) *and assume the data perturbation scheme* (2.8). *Then, matrix F defined in* (2.11) *can be writen as*

$$F = \frac{n}{\|e\|^2} H + \frac{An}{\|e\|^2} hh^\top, \tag{2.18}$$

where $H = X(X^\top X)^{-1}X^\top$ and $h = e/\|e\|$ is the normalized least squares residual vector. In addition, $Hh = 0$ and F has eigenvalues $\lambda_1 = n/\|e\|^2$ (*with multiplicity* p) *and $\lambda_2 = An/\|e\|^2$* (*with multiplicity* 1).

In the curvature influence diagnostics, we usually choose the eigenvector associated with the largest eigenvalue. The eigenvector associated with λ_2 is the normalized residual vector h. This is the appropriate curvature statistic for assessing influence under data perturbation schemes because the eigenvectors associated with λ_1 (which are the eigenvectors of H) do not consider the response variable y. Therefore, we are interested in finding the conditions where the appropriate curvature statistic is chosen, that is, when $\lambda_2 > \lambda_1$ holds. From Theorem 2.1, this occurs if and only if the quantity A defined in (2.15) satisfies $A > 1$. We establish this result as a corollary.

Corollary 2.1. *Under the conditions of Theorem* 2.1, *the eigenvector associated with the largest eigenvalue of F is h if and only if $A > 1$.*

Next, we discuss when the condition $A > 1$ holds, for some well-known elliptical densities reported in Table 1. For instance, the condition is always satisfied by the normal and logistic densities. For the exponential power density, the condition is $\eta > 1/2$, which includes heavy-tailed case ($\eta > 1$). However, for the Student-t density, the normalized residuals are chosen if and only if the degree of freedom parameter is larger than the sample size ($\eta > n$). This means that for the Cauchy distribution, which is a particular case of the Student-t distribution when $\eta = 1$, the residual vector is never chosen as the curvature vector. Moreover, if, for example, the degree of freedom parameter of the Student-t distribution is small (in order to reproduce heavy tails), the appropriate curvature statistic only might be chosen for, accordingly, very small data sets.

In order to highlight the consequences of the main findings of the paper in terms of local influence analysis, we present next two examples.

Example 2.1. Suppose that data were generated by a Student-t model with $\eta < n$. To find the curvature influential points a researcher who is not aware of the results given in this paper calculates the curvature numerically. From (2.18), this means the researcher will find eigenvectors of F associated with $\lambda_1 = n/\|e\|^2$, which are non-unique, and are the eigenvectors associated with the eigenvalue 1 of $H = X(X^\top X)^{-1}X^\top$ (because H is a projection matrix, it has eigenvalues equal to 0 or 1). For instance, consider a linear regression where the columns of matrix X are $\mathbf{1} = (1, 1, \dots, 1)$ and $\mathbf{x} = (x_1, x_2, \dots, x_n)$. The eigenvector v associated with the eigenvalue 1 must satisfy $Hv = v$. Since $HX = X$, the columns of X are eigenvectors. The eigenvectors corresponding to the eigenvalue 1 are not unique; they are elements of the eigenspace generated by $\mathbf{1}$ and $\mathbf{x}$. For example, $v = \mathbf{x} - \bar{x}\mathbf{1}$, where $\bar{x} = \sum_{i=1}^n x_i/n$, is also an eigenvector. Therefore, according to the curvature influence diagnostics, we could choose, for example, the following normalized eigenvectors:

$$v_1 = (1, \dots, 1)/\sqrt{n}, \tag{2.19}$$

$$v_2 = (x_1 - \bar{x}, \dots, x_n - \bar{x})/\sqrt{\lambda}, \tag{2.20}$$

where $\lambda = \sum_{i=1}^n (x_i - \bar{x})^2$.

If the researcher chooses v_1, the curvature is constant and then there are no influential points, regardless of the data, which is nonsense. When v_2 is chosen, because the leverage is equal to

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad i = 1, \dots, n, \tag{2.21}$$

that is, $h_{ii} = n^{-1} + v_{2i}^2$, where v_{2i} is the ith element of v_2, the influential points are the points with high leverage. However, the researcher can choose other eigenvectors and accordingly identify other influential points.

Therefore, a researcher who is not aware of the results given in this paper does not obtain a unique answer in terms of the identification of influential points. In fact, the answers ranged from the non-existence of influential points to the identification of the x's points with high leverage. Even worse, those solutions are independent of the response values y which should be considered for data perturbation scheme analysis.

Example 2.2. In this example, we present an empirical illustration. We analized the water salinity data of Ruppert and Carroll (1980) where the response variable is biweekly salinity (y) and the explanatory variables are the salinity lagged two weeks (x_1), a dummy variable which indicates the time period (x_2) and river discharge (x_3). Regression diagnostics is performed by these authors as well as by Atkinson (1985), Carroll and Ruppert (1985) and Davison and Tsai (1992). Besides, curvature influence diagnostics using elliptical models was performed by Galea et al. (1997) assuming model perturbation and by Liu (2000) assuming data perturbation scheme.

We perform local influence via Cook's curvature diagnostics assuming data perturbation scheme on an elliptical model with regressors $\mathbf{1} = (1, 1, \ldots, 1)$, x_1, x_2, x_3. First, we calculated the eigenvector associated with λ_2, h, which is the normalized least square residual vector. Since eigenvectors associated with matrix H are non-unique we decided to calculate an orthonormal basis via the Gram–Schmidt process on $\{\mathbf{1}, x_1, x_2, x_3\}$. The five curvature vectors are depicted in Figure 1, being h the black solid line and the four vectors associated with H in dashed lines. Here we can see that the curvature vector h indicates no influential point. Then, assuming that errors follow Normal and Logistic distributions we do not identify influential points. The same applies assuming Power Exponential

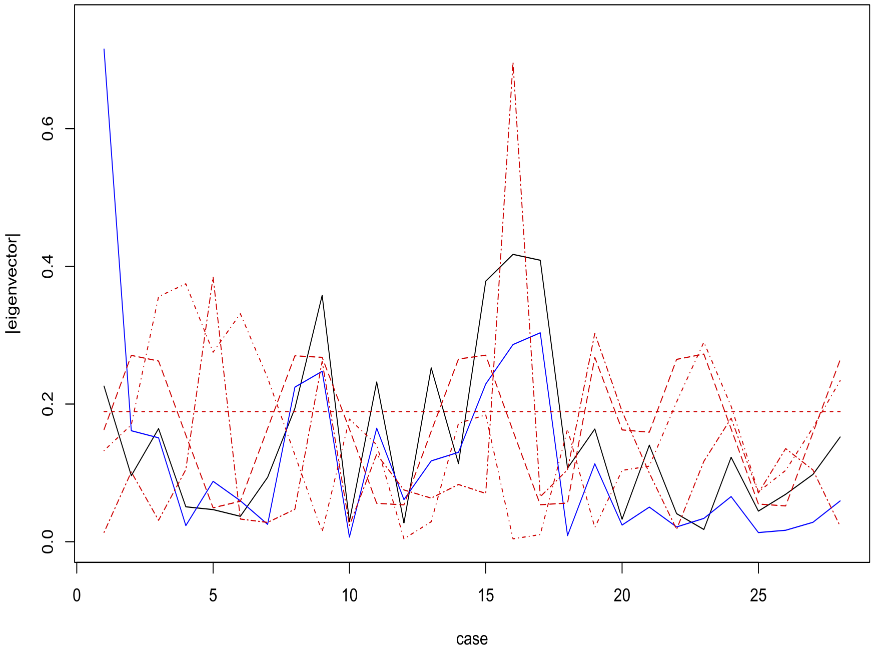

Figure 1 *Absolute values of eigenvectors: of H in red dashed lines, in solid black line for the residuals in the unperturbed case and in solid blue line for the residuals in the perturbed case.*

with $\eta > 1/2$. However, if we assume a Student-t distribution with $\eta = 3$ since $n = 28 > 3$ then the curvature vector is any of the four vectors associated with H.

In addition, we perturbed the first response observation y_1 by $y_1(\omega) = y_1 + 9$ (as a reference, the standard deviation of y is 3.01 and the standard deviation of the residuals in the OLS fit with the unperturbed data is 1.25) and perform again the curvature influence analysis. The new residual vector is the solid blue line in Figure 1. We can clearly see that this vector identify, correctly, the first observation as the perturbed observation. However, if we assume a Student-t distribution with $\eta < 28$ then the curvature vectors associated with H indicates no influential point.

3 Conclusions and further research

In this paper, we derived analytic expressions for curvature influence diagnostics in elliptical regression models under a data perturbation scheme. The results reveal the role of the shape parameter in applying the curvature influence diagnostics technique. Since the proper statistic for a data perturbation scheme, that is, the residuals, is only chosen under some conditions on the shape parameter, the specification of the distribution of the errors is crucial when dealing with data. Therefore, the application of local influence approach to regression diagnostics in the case of perturbation of the response values should not be done automatically, as illustrated by the examples in Section 2.

In practice, the shape parameter has to be estimated. Thus, a topic of further research is to derive the analytic expression of the curvature statistic in this situation.[1] An additional topic of further research is the analysis of the *independent case* where the errors are defined as following univariate elliptical distributions.

Acknowledgments

The authors thank the editor and the anonymous referees for their valuable comments and suggestions, which led to a substantial improvement of the paper. This research was partially supported by São Paulo Research Foundation (FAPESP), grants 2013/00506-1 and 2013/22930-0.

References

Atkinson, A. C. (1985). *Plots, Transformations and Regression*. Oxford: Clarendon.

Carroll, R. J. and Ruppert, D. (1985). Transformations in regression: A robust analysis. *Technometrics* **27**, 1–12. MR0772893

[1] However, it is worth to mention that Zellner (1976) has pointed out the nonexistence of the ML estimates for β, ϕ and η when considering the Student-t distribution.

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society B* **48**, 133–169. MR0867994

Davison, A. C. and Tsai, C. L. (1992). Regression model diagnostics. *International Statistical Review* **60**, 337–353.

Fang, K. T. and Anderson, T. W. (1990). *Statistical Inferences in Elliptical Contoured and Related Distributions*. New York: Allerton Press. MR1066887

Galea, M., Paula, G. A. and Bolfarine, H. (1997). Local influence in elliptical linear regression models. *The Statistician* **46**, 71–79.

Galea, M., Riquelme, M. and Paula, G. A. (2000). Diagnostic methods in elliptical linear regression models. *Brazilian Journal of Probability and Statistics* **14**, 167–184. MR1860055

Liu, S. (2000). On local influence for elliptical linear models. *Statistical Papers* **41**, 211–224. MR1769062

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association* **75**, 828–838. MR0600964

Schwarzmann, B. (1991). A connection between local-influence analysis and residual diagnostics. *Technometrics* **33**, 103–104.

Zhu, H., Ibrahim, J., Lee, S. and Zhang, H. (2007). Perturbation selection and influence measures in local influence analysis. *The Annals of Statistics* **35**, 2565–2588. MR2382658

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate student-t error terms. *Journal of the American Statistical Association* **71**, 400–405. MR0405699

Department of Statistics
IMECC-UNICAMP
University of Campinas
Rua Sérgio Buarque de Holanda 651
Campinas, São Paulo
Brasil
E-mail: amadeus@ime.unicamp.br
hotta@ime.unicamp.br