*J. Appl. Math. & Informatics Vol.* **26***(2008), No. 5 - 6, pp. 1133 - 1138*
Website: http://www.kcam.biz

# OUTLIER DETECTION BASED ON A CHANGE OF LIKELIHOOD

MYUNG GEUN KIM

ABSTRACT. A general method of detecting outliers based on a change of likelihood by using the influence function is suggested. It can be applied to all kinds of distributions that are specified by parameters. For the multivariate normal case, specific computations are made to get the corresponding conditional influence function. A numerical example is provided for illustration.

AMS Mathematics Subject Classification : 62H99, 62J20.
*Key words and phrases* : Conditional influence function, likelihood, outliers.

## 1. Introduction

Methods of detecting outliers and influential observations have been studied in wide areas and some of them can be found in Barnett and Lewis [1], Cook and Weisberg [2], and Kim [7]. As one of the methods, the influence function method introduced by Hampel [3] has been developed in various fields of statistics. However, it is confined to a class of parameters which can be regarded as statistical functionals of the parent distributions. The influence function method is performed by perturbing the parent distribution towards a distribution having unit mass at a point.

In this work the influence function method is adapted to the likelihood function. The result in this work can be used for detecting outliers when the underlying distribution is specified by parameters. In Section 2 the conditional influence function is defined, which can be used for detecting observations that have a large influence on the likelihood. In Section 3 the conditional influence function is derived when the underlying distribution is a multivariate normal. In Section 4 a numerical example is provided for illustration.

## 2. Conditional influence function

Received December 12, 2007. Accepted January 14, 2008.

We denote a vector in $R^p$ by $x_0$. For a given distribution function $F$ defined on $R^p$, the perturbation of $F$ at $x_0$ is defined by $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_{x_0}$ for $0 \leq \varepsilon \leq 1$, where $\delta_{x_0}$ denotes the distribution having unit mass at $x_0$. Let $\theta(F)$ be a parameter which is a functional of $F$ and $\theta(F_\varepsilon)$ be the perturbation of $\theta(F)$ at $x_0$. The influence function for $\theta(F)$ at $x_0$ (Hampel [3]) is defined by

$$\lim_{\varepsilon \to 0} \frac{\theta(F_\varepsilon) - \theta(F)}{\varepsilon}$$

and it measures the instantaneous rate of change of $\theta(F)$ as $F$ moves infinitesimally towards $\delta_{x_0}$. It is used for identifying observations that have a large influence on the estimate of $\theta(F)$. If $\theta(F_\varepsilon)$ can be expanded in a Taylor series of $\varepsilon$, then the influence function for $\theta(F)$ at $x_0$ is the coefficient of the first order $\varepsilon$-term in a series expansion of $\theta(F_\varepsilon)$.

Let $x_1,...,x_n$ be a random sample from a $p$-variate distribution $F$ specified by a parameter vector $\theta(F)$. We denote by $L(\theta(F))$ the likelihood function of $\theta(F)$ given $x_1,...,x_n$. When the parent distribution $F$ is perturbed at $x_0$, the perturbation of $L(\theta(F))$ is defined by $L(\theta(F_\varepsilon))$ which is the likelihood function of the perturbation $\theta(F_\varepsilon)$ given $x_1,...,x_n$. In view of the above definition of the influence function, we define the conditional influence function for $L(\theta(F))$ given $x_1,...,x_n$ by

$$\lim_{\varepsilon \to 0} \frac{L(\theta(F_\varepsilon)) - L(\theta(F))}{\varepsilon}.$$

Thus the conditional influence function can be used for detecting outliers when the underlying distribution is $F$.

## 3. A multivariate normal case

We consider a case in which $F$ denotes a multivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$. Then the likelihood function $L(\mu, \Sigma)$ given the random sample $x_1,...,x_n$ is

$$L(\mu, \Sigma) = (2\pi)^{-np/2} \mid \Sigma \mid^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu)\right\}.$$

The mean vector and covariance matrix for the perturbed distribution $F_\varepsilon$ are

$$\mu_\varepsilon = (1-\varepsilon)\mu + \varepsilon x_0$$

$$\Sigma_\varepsilon = (1-\varepsilon)\Sigma + \varepsilon(1-\varepsilon)(x_0 - \mu)(x_0 - \mu)^T.$$

Let

$$\alpha_{ij} = (x_i - \mu)^T\Sigma^{-1}(x_j - \mu)$$

for $i, j = 0, 1, ..., n$. We can easily get

$$\mid \Sigma_\varepsilon \mid = \mid \Sigma \mid (1-\varepsilon)^p(1 + \alpha_{00}\varepsilon).$$

Further it is easy to show that

$$\begin{aligned}(1-\varepsilon)^{-np/2} &= 1+\frac{np}{2}\varepsilon+\frac{np}{4}\Big(\frac{np}{2}+1\Big)\varepsilon^2+O(\varepsilon^3)\\ (1+\alpha_{00}\varepsilon)^{-n/2} &= 1-\frac{n}{2}\alpha_{00}\varepsilon+\frac{n}{4}\Big(\frac{n}{2}+1\Big)\alpha_{00}^2\varepsilon^2+O(\varepsilon^3).\end{aligned}$$

Then Taylor series expansion of $|\,\Sigma_\varepsilon\,|^{-n/2}$ with respect to $\varepsilon$ is computed as

$$\begin{aligned}|\,\Sigma_\varepsilon\,|^{-n/2} &= |\,\Sigma\,|^{-n/2}\Big[1+\Big\{\frac{n}{2}(p-\alpha_{00})\Big\}\varepsilon\\ &\quad+\Big\{\frac{np}{4}\Big(\frac{np}{2}+1\Big)-\frac{n^2p}{4}\alpha_{00}+\frac{n}{4}\Big(\frac{n}{2}+1\Big)\alpha_{00}^2\Big\}\varepsilon^2\Big]\\ &\quad+O(\varepsilon^3).\end{aligned}$$

The inverse of $\Sigma_\varepsilon$ is easily computed as

$$\begin{aligned}\Sigma_\varepsilon^{-1} &= \Sigma^{-1}+\Big\{\Sigma^{-1}-\Sigma^{-1}(x_0-\mu)(x_0-\mu)^T\Sigma^{-1}\Big\}\varepsilon\\ &\quad+\Big\{\Sigma^{-1}+(\alpha_{00}-1)\Sigma^{-1}(x_0-\mu)(x_0-\mu)^T\Sigma^{-1}\Big\}\varepsilon^2+O(\varepsilon^3)\end{aligned}$$

so that Taylor series expansion of $(x_i-\mu_\varepsilon)^T\Sigma_\varepsilon^{-1}(x_i-\mu_\varepsilon)$ is given by

$$\alpha_{ii}+(\alpha_{ii}-\alpha_{i0}^2-2\alpha_{i0})\varepsilon+\Big\{\alpha_{ii}+1+(\alpha_{00}-1)(\alpha_{i0}+1)^2\Big\}\varepsilon^2+O(\varepsilon^3).$$

Thus the likelihood function $L(\mu_\varepsilon,\Sigma_\varepsilon)$ of $\mu_\varepsilon$ and $\Sigma_\varepsilon$ given $x_1,\ldots,x_n$ is expanded as

$$\begin{aligned}L(\mu_\varepsilon,\Sigma_\varepsilon) &= L(\mu,\Sigma)\Big[1+\Big(\frac{np}{2}-\frac{n}{2}\alpha_{00}+a_1\Big)\varepsilon\\ &\quad+\Big\{\frac{1}{2}\Big(\frac{np}{2}-\frac{n}{2}\alpha_{00}+a_1\Big)^2+\frac{n}{4}(\alpha_{00}^2+p)+a_2\Big\}\varepsilon^2\Big]+O(\varepsilon^3),\end{aligned}$$

where

$$\begin{aligned}a_1 &= -\frac{1}{2}\sum_{i=1}^{n}(\alpha_{ii}-\alpha_{i0}^2-2\alpha_{i0})\\ a_2 &= -\frac{1}{2}\sum_{i=1}^{n}\Big\{\alpha_{ii}+1+(\alpha_{00}-1)(\alpha_{i0}+1)^2\Big\}\\ &= (1-\alpha_{00})a_1-\frac{\alpha_{00}}{2}\sum_{i=1}^{n}(\alpha_{ii}+1).\end{aligned}$$

Hence the conditional influence function for the likelihood function becomes

$$\Big\{\frac{np}{2}-\frac{n}{2}\alpha_{00}-\frac{1}{2}\sum_{i=1}^{n}(\alpha_{ii}-\alpha_{i0}^2-2\alpha_{i0})\Big\}L(\mu,\Sigma). \tag{1}$$

This cannot be directly used because it involves unknown parameters. Three sample versions of the conditional influence function are possible as in the influence analysis (see Kim [5] for more details): the sample conditional influence function (SCIF), the empirical conditional influence function (ECIF) and the

delected empirical conditional influence function (DCIF). It is well known that in general these three sample versions provide very similar results.

Let $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$ and $S = (1/n)\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$. Then $\hat{\mu} = \bar{x}$ and $\hat{\Sigma} = S$ are the maximum likelihood estimators for $\mu$ and $\Sigma$ respectively. The likelihood function $L(\mu, \Sigma)$ attains its maximum value at $\mu = \bar{x}$ and $\Sigma = S$ and the maximum value is

$$L(\bar{x}, S) = (2\pi)^{-np/2} \mid S \mid^{-n/2} \exp\left(-\frac{np}{2}\right).$$

The maximum likelihood estimator for $\alpha_{ij}$ is given by $\hat{\alpha}_{ij} = (x_i - \bar{x})^T S^{-1}(x_j - \bar{x})$. It is then always true that $L(\hat{\mu}_\varepsilon, \hat{\Sigma}_\varepsilon) \leq L(\hat{\mu}, \hat{\Sigma})$.

Let $\bar{x}_{-r}$ be the sample mean based on the random sample of size $n-1$ with the $r$th observation $x_r$ deleted and $S_{-r}$ be the corresponding sample covariance matrix. Then the SCIF is defined by

$$SCIF = (n-1)\Big\{L(\bar{x}, S) - L(\bar{x}_{-r}, S_{-r})\Big\}.$$

Since

$$\begin{aligned}
\bar{x}_{-r} &= \frac{n}{n-1}\bar{x} - \frac{1}{n-1}x_r \\
S_{-r} &= \frac{n}{n-1}S - \frac{n}{(n-1)^2}(x_r - \bar{x})(x_r - \bar{x})^T \\
&= \frac{n}{n-1}S\left(I_p - \frac{1}{n-1}S^{-1}(x_r - \bar{x})(x_r - \bar{x})^T\right),
\end{aligned}$$

a little computation yields

$$\begin{aligned}
L(\bar{x}_{-r}, S_{-r}) &= (2\pi)^{-np/2} \mid S_{-r} \mid^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x}_{-r})^T S_{-r}^{-1}(x_i - \bar{x}_{-r})\right\} \\
&= e^{p/2}(\mid S_{-r} \mid / \mid S \mid)^{-n/2} \\
&\quad \times \exp\left\{-(x_r - \bar{x}_{-r})^T S_{-r}^{-1}(x_r - \bar{x}_{-r})/2\right\} L(\bar{x}, S).
\end{aligned}$$

Thus the SCIF is computed as

$$\begin{aligned}
SCIF &= (n-1)\Big[1 - e^{p/2}(\mid S_{-r} \mid / \mid S \mid)^{-n/2} \\
&\qquad \times \exp\{-(x_r - \bar{x}_{-r})^T S_{-r}^{-1}(x_r - \bar{x}_{-r})/2\}\Big] L(\bar{x}, S).
\end{aligned}$$

Let $\hat{F}$ be the empirical distribution function based on $x_1,\ldots,x_n$. The ECIC is obtained by replacing $F$ with $\hat{F}$ in (1) and it is given by

$$ECIF = \left\{-\frac{n}{2}\hat{\alpha}_{00} + \frac{1}{2}\sum_{i=1}^{n}(\hat{\alpha}_{i0}^2 + 2\hat{\alpha}_{i0})\right\} L(\bar{x}, S)$$

since $\sum_{i=1}^{n}\hat{\alpha}_{ii} = np$.

We define $\hat{F}_{-r} = \{1 + (n-1)^{-1}\}\hat{F} - (n-1)^{-1}\delta_{x_r}$. Then $\hat{F}_{-r}$ is the deleted version of $\hat{F}$ with the $r$th observation deleted. The mean vector and covariance matrix for $\hat{F}_{-r}$ are given by

$$\mu(\hat{F}_{-r}) = \bar{x}_{-r}$$

$$\Sigma(\hat{F}_{-r}) = S_{-r}.$$

The DCIF is obtained by replacing F with $\hat{F}_{-r}$ in (1)and it has a form similar to the ECIF.

## 4. A numerical example

For illustration, we will consider the cost data consisting of 36 measurements on the per mile cost of three variables - fuel ($x_1$), repair ($x_2$) and capital ($x_3$)- which is taken from p.276 of Johnson and Wichern [4]. The cost data was analyzed by Kim [6] and by some authors in the references therein. By their results, it is reasonable to conclude that observations 9 and 21 are possible outliers.

Table 1. SCIF for the cost data

| # | SCIF | # | SCIF | # | SCIF | # | SCIF | # | SCIF | # | SCIF |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.032 | 7 | 0.023 | 13 | 0.025 | 19 | 0.022 | 25 | 0.303 | 31 | 0.096 |
| 2 | 0.105 | 8 | 0.105 | 14 | 0.022 | 20 | 0.351 | 26 | 0.064 | 32 | 0.069 |
| 3 | 0.116 | 9 | 0.998 | 15 | 0.169 | 21 | 0.775 | 27 | 0.129 | 33 | 0.034 |
| 4 | 0.110 | 10 | 0.021 | 16 | 0.064 | 22 | 0.023 | 28 | 0.036 | 34 | 0.030 |
| 5 | 0.028 | 11 | 0.029 | 17 | 0.029 | 23 | 0.237 | 29 | 0.068 | 35 | 0.056 |
| 6 | 0.035 | 12 | 0.033 | 18 | 0.097 | 24 | 0.050 | 30 | 0.046 | 36 | 0.174 |

We will provide only the first sample version SCIF for illustration since in general the three sample versions of the influence function yield similar results. The term $(n-1)L(\bar{x}, S)$ in SCIF is common to all observations and therefore it is redundant for investigating the influence of each observation. Thus we will compute the values $SCIF/[(n-1)L(\bar{x}, S)]$ which are included in the column with the heading SCIF of Table 1. Numbers following # in each column represent the observations.

The stem-and-leaf display or the index plot of the sample values for SCIF (not provided here) can be useful for getting information about possible outliers and they show that observation 9 has the largest influence on the likelihood and observation 21 is the next. The influence of the others is not relatively severe compared with those of observations 9 and 21. Thus observations 9 and 21 are possible outliers.

## REFERENCES

1. V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed., Wiley, New York, 1994.
2. R. D. Cook and S. Weisberg, *Residuals and Influence in Regression*, Chapman and Hall, New York, 1982.
3. F. R. Hampel, *The Influence curve and its role in robust estimation.* J. Amer. Statist. Assoc., **60** (1974), 383–393.
4. A. J. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis.* Englewood Cliffs: Prentice-Hall, 1992.
5. M. G. Kim, *Influence curve for the Cholesky root of a covariance matrix*, Communications in Statistics: Theory and Methods, **23** (1994), 1399–1412.
6. M. G. Kim, *Multivariate outliers and decompositions of Mahalanobis distance*, Communications in Statistics: Theory and Methods, **29** (2000), 1511–1526.
7. M. G. Kim, *Influence analysis for a linear hypothesis in multivariate regression model*, J. Appl. Math. & Computing, **13** (2003), 479–485.

**M.G. Kim** received his Ph.D from Ohio State University. He is now a professor of Mathematics Education Department at Seowon University. His research interest centers on diagnostics in multivariate analysis and linear model.

Department of Mathematics Education, Seowon University, 231 Mochung-Dong, Heungduk-Gu, Cheongju, Chung-Buk, 361-742, Korea
e-mail : mgkim@seowon.ac.kr