МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2008
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2008
Proceedings of the Thirty Seventh Spring Conference of the Union of Bulgarian Mathematicians
Borovetz, April 2–6, 2008

CONSISTENCY OF THE k-th NEAREST NEIGHBOR ESTIMATOR OF THE RELATIVE RISK AND ITS APPLICATION TO INJURY SURVEILLANCE*

Svetla Slavova, Richard Kryscio, Terry Bunn

The relative risk function is defined as the ratio of two probability density functions, usually cases to controls, at a fixed point. The nonparametric k-th nearest neighbor (kNN) approach is used for the density estimation. The kNN relative risk estimator at a fixed location is shown to be asymptotically consistent. An application of the kNN relative risk estimator to injury surveillance is discussed.

1. Introduction. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a set of n independent and identically distributed (iid) random variables with values in R^2 and continuous density function $f(X)$. The bold face type indicates random variables, and the capitals are used for vectors and matrices. The kNN density estimator of the unknown density $f(X)$ at a point $X \in R^2$ is defined as $\hat{f}_n(X) = \dfrac{k}{n\boldsymbol{v}(X)}$, where k is the pre-specified number of nearest neighbors of $X \in R^2$; $\boldsymbol{v}(X)$ is the volume of minimal sphere $\boldsymbol{S}(X)$ centered at X and containing at least k of the observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$. The volume of the region $\boldsymbol{S}(X)$ is $\boldsymbol{v}(X) \equiv \int\limits_{\boldsymbol{S}(X)} dY$. The coverage of $\boldsymbol{S}(X)$ is defined as $\boldsymbol{u}(X) \equiv \int\limits_{S(X)} f(Y)dY = \Pr\{\mathbf{X} \in \boldsymbol{S}(X)\}$ and it is known to have a Beta distribution $Beta(k, n-k+1)$ with parameters k and $n-k+1$, and to be independent of the underlying distribution [1].

2. Moments of the coverage and the volume. Beta function $B(a,b)$, $a>0$, $b>0$ is defined by $B(a,b) = \int\limits_{(0,1)} z^{a-1}(1-z)^{b-1}dz$. For a,b positive integers Beta function can be calculated as $B(a,b) = (a-1)!(b-1)!/(a+b-1)!$ Suppose $\boldsymbol{x}$ has the Beta distribution with parameters a and b. Then it can be shown that $E[\boldsymbol{x}^m] = B(a+m,b)/B(a,b)$. Therefore, some of the moments of $\boldsymbol{u}(X) \sim Beta(k, n-k+1)$ are:

$$E[\boldsymbol{u}(X)] = \frac{k}{n+1}, E\left[\boldsymbol{u}^2(X)\right] = \frac{k(k+1)}{(n+1)(n+2)},$$

*2000 Mathematics Subject Classification: 66G05, 62G20.
Key words: k-th nearest neighbor, relative risk, occupational surveillance.
This work was partially supported by Grant/Cooperative Agreement Number 5U60OHH008483-03 from NIOSH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIOSH. The work of RK was partially supported by a University of Kentucky Research Professorship award.

$$E\left[\frac{1}{\boldsymbol{u}(X)}\right] = \frac{n}{k-1}, E\left[\frac{1}{\boldsymbol{u}^2(X)}\right] = \frac{n(n-1)}{(k-1)(k-2)}.$$

Fukunaga and Hostetler [1] expressed the volume $\boldsymbol{v}(X)$ as a function of the coverage of the region $\boldsymbol{u}(X)$ in order to take advantage of the Beta distribution. By definition, the kNN estimator deals with a small local region around the point X. Therefore, the local density $f(X)$ can be approximated with a truncated Taylor series about X. Under the assumption that $f(X)$ has continuous partial derivatives of sufficient order in a neighborhood of X,

$$\boldsymbol{u}(X) \equiv \int_{S(X)} f(Y)dY \cong \int_{S(X)} f(X)dY + \int_{S(X)} \left[\frac{\partial f(X)}{\partial X}\right]^T (Y-X)dY$$
$$+\frac{1}{2}\int_{S(X)} (Y-X)^T \left[\frac{\partial^2 f(X)}{\partial X^2}\right](Y-X)dY,$$

where $\left[\frac{\partial f(X)}{\partial X}\right]^T = \left(\frac{\partial f(X)}{\partial x_1}, \frac{\partial f(X)}{\partial x_2}\right)$ and $\left[\frac{\partial^2 f(X)}{\partial X^2}\right]_{ij} = \frac{\partial^2 f(X)}{\partial x_i \partial x_j}$.

Due to the symmetry of the region, $\int_{S(X)} (Y-X)dY = 0$. Therefore,

$$\boldsymbol{u}(X) \cong f(X)\boldsymbol{v}(X) + \frac{1}{2}\int_{S(X)} (Y-X)^T \left[\frac{\partial^2 f(X)}{\partial X^2}\right](Y-X)dY.$$

Using matrix properties a) $Z^T AZ = tr\, AZZ^T$, b) $tr\, AZ = tr\, ZA$, c) linearity of the trace, we have

$$\boldsymbol{u}(X) \cong f(X)\boldsymbol{v}(X) + \frac{1}{2}tr\left\{\left(\int_{S(X)} (Y-X)(Y-X)^T dY\right)\left[\frac{\partial^2 f(X)}{\partial X^2}\right]\right\},$$

Fukunaga and Hostetler [1] showed that $\int_{S(X)} (Y-X)(Y-X)^T dY = c(X)\boldsymbol{v}(X)$, where $c(X)$ is a function of the second partial derivatives of $f(X)$ in a neighborhood of X. Then,

$$\boldsymbol{u}(X) \cong f(X)\boldsymbol{v}(X) + c(X)\boldsymbol{v}^2(X) \tag{1}$$

and

$$\frac{1}{\boldsymbol{v}(X)} \cong \frac{f(X)}{\boldsymbol{u}(X)} + \frac{c(X)\boldsymbol{v}(X)}{\boldsymbol{u}(X)}. \tag{2}$$

Using only the first term in (2.1) as an approximation for $\boldsymbol{u}(X)$ we get that $\boldsymbol{u}(X) \cong f(X)\boldsymbol{v}(X)$, and subsequently $\boldsymbol{v}(X) \cong \frac{\boldsymbol{u}(X)}{f(X)}$. After substituting $\boldsymbol{v}(X)$ in the second term

of (2.2), we have:

$$\frac{1}{\boldsymbol{v}(X)} \cong \frac{f(X)}{\boldsymbol{u}(X)} + \frac{c(X)\boldsymbol{u}(X)}{\boldsymbol{u}(X)f(X)} = \frac{f(X)}{\boldsymbol{u}(X)} + \frac{c(X)}{f(X)}.$$

Therefore,

$$E\left[\frac{1}{\boldsymbol{v}(X)}\right] \cong f(X)E\left[\frac{1}{\boldsymbol{u}(X)}\right] + \frac{c(X)}{f(X)} = f(X)\frac{n}{k-1} + \frac{c(X)}{f(X)},$$

$$E\left[\frac{1}{\boldsymbol{v}^2(X)}\right] \cong f^2(X)\frac{n(n-1)}{(k-1)(k-2)} + 2c(X)\frac{n}{k-1} + \frac{c^2(X)}{f^2(X)},$$

$$E\left[\boldsymbol{v}(X)\right] \cong E\left[\frac{\boldsymbol{u}(X)}{f(X)}\right] = \frac{1}{f(X)}\frac{k}{n+1},$$

$$E\left[\boldsymbol{v}^2(X)\right] \cong E\left[\frac{\boldsymbol{u}^2(X)}{f^2(X)}\right] = \frac{1}{f^2(X)}\frac{k(k+1)}{(n+1)(n+2)}.$$

3. Properties of the kNN estimator of the relative risk. Let $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ be iid random vectors with values in R^2 and a probability density function $f_1(X)$. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$ be iid random vectors with values in R^2 and a probability density function $f_2(X)$. The kNN estimate of the relative risk $\gamma(X) = f_1(X)/f_2(X)$ at a point $X \in R^2$ is defined as

$$\hat{\gamma}(X) = \frac{\hat{f}_1(X)}{\hat{f}_2(X)} = \left[\frac{k_1}{n_1\boldsymbol{v}_1(X)}\right] \Big/ \left[\frac{k_2}{n_2\boldsymbol{v}_2(X)}\right].$$

The Mean Square Error (MSE) of the relative risk estimator can be expressed as

$$MSE[\hat{\gamma}(X)] = \frac{k_1^2 n_2^2}{k_2^2 n_1^2} E\left[\boldsymbol{v}_2^2(X)\right] E\left[\frac{1}{\boldsymbol{v}_1^2(X)}\right] - 2\gamma(X)\frac{k_1 n_2}{k_2 n_1} E\left[\boldsymbol{v}_2(X)\right] E\left[\frac{1}{\boldsymbol{v}_1(X)}\right] + \gamma^2(X)$$

$$\cong \frac{k_1^2 n_2^2}{k_2^2 n_1^2}\left[\frac{1}{f_2^2(X)}\frac{k_2(k_2+1)}{(n_2+1)(n_2+2)}\right]\left[f_1^2(X)\frac{n_1(n_1-1)}{(k_1-1)(k_1-2)} + 2c_1(X)\frac{n_1}{k_1-1} + \frac{c_1^2(X)}{f_1^2(X)}\right]$$

$$-2\gamma(X)\frac{k_1 n_2}{k_2 n_1}\left[\frac{1}{f_2(X)}\frac{k_2}{n_2+1}\right]\left[f_1(X)\frac{n_1}{k_1-1} + \frac{c_1(X)}{f_1(X)}\right] + \gamma^2(X)$$

$$= \underbrace{\frac{f_1^2(X)}{f_2^2(X)}\frac{k_1^2(k_2+1)n_2^2(n_1-1)}{k_2(k_1-1)(k_1-2)n_1(n_2+1)(n_2+2)}}_{A}$$

$$+ \underbrace{2c_1(X)\frac{1}{f_2^2(X)}\frac{k_1^2(k_2+1)n_2^2}{k_2\,(k_1-1)\,n_1(n_2+1)(n_2+2)}}_{B}$$

$$+ \underbrace{\frac{c_1^2(X)}{f_2^2(X)f_1^2(X)}\frac{k_1^2 n_2^2(k_2+1)}{k_2 n_1^2(n_2+1)(n_2+2)}}_{C} - \underbrace{2\gamma(X)\frac{f_1(X)}{f_2(X)}\frac{k_1 n_2}{(k_1-1)\,(n_2+1)}}_{D}$$

$$\underbrace{- 2\gamma(X)\frac{c_1(X)}{f_1(X)f_2(X)}\frac{k_1 n_2}{n_1\,(n_2+1)}}_{F} + \gamma^2(X).$$

When $n_1 \to \infty$, $n_2 \to \infty$, the terms B, C and F in the expression above tend to zero. Assuming $n_1/(n_1-1) \cong 1$, $n_2/(n_2+1) \cong 1$, and $n_2/(n_2+2) \cong 1$, we get the following expression for the MSE:

$$MSE[\hat{\gamma}(X)] \xrightarrow[n_1,n_2\to\infty]{} \gamma^2(X)\left[\frac{k_1^2(k_2+1)}{k_2(k_1-1)(k_1-2)} - 2\frac{k_1}{(k_1-1)} + 1\right].$$

This result shows that when nearest neighbor parameters k_1 and k_2 tend to infinity, and $\dfrac{k_1}{n_1} \to 0$, $\dfrac{k_2}{n_2} \to 0$ as $n_1 \to \infty$, and $n_2 \to \infty$, the kNN relative risk estimator is asymptotically unbiased and consistent.

4. Asymptotic distribution. Let $F_{n_1,n_2}(t|k_1,k_2)$ be the distribution function of $\hat{\gamma}(X)$.

Theorem. *Let the probability density functions $f_1(X)$, $f_2(X)$ be strictly positive and two times differentiable with bounded derivatives in a neighborhood of X. If $n_1 \to \infty$, $n_2 \to \infty$ such that $n_1/\,n_2 \to const > 0$, then for fixed k_1, k_2 and $t \in (0,\infty)$*

$$F_{n_1,n_2}(t|k_1,k_2) = W(t|\gamma(X),k_1,k_2) + O\left(\frac{1}{n_1}\right) + O\left(\frac{1}{n_2}\right).$$

Here $W(t|\gamma(X),k_1,k_2) = I_{[1-\phi\{t,\gamma(x)\}]}(k_2,k_1)$, where $\phi\{t,\gamma(X)\} = \left\{1 + \dfrac{k_2}{k_1}\dfrac{t}{\gamma(X)}\right\}^{-1}$, and $I_a(i,j) = \dfrac{1}{B(i,j)}\int\limits_0^a y^{i-1}(1-y)^{j-1}dy$, the incomplete beta-function.

The proof of the theorem for densities defined in R^1 appears in the Appendix to [2]. The proof in the higher-dimensional cases is similar, but requires minor changes in the definition of the volume and the coverage function (details can be found in [3]). The theorem states that for sufficiently large samples, $\hat{\gamma}(X)$ has approximately the distribution $W(t|\gamma(X),k_1,k_2)$. The large sample distribution, obtained under fairly general conditions, provides easily computable critical regions, level of significance, power of the test and confidence bounds, without relying on computationally dependent algorithms.

Let $f_1(X)$ be the probability density function of cases (diseases or injuries) given by their geographical coordinates. Let $f_2(X)$ be the probability density function of the controls (or the population at risk), represented by their geographical locations. It is of practical interest to test a hypothesis for excessive risk at a particular point (geographical location) X. Consider testing $H_0 : \gamma(X) = 1$ *vs.* $H_1 : \gamma(X) > 1$. This is equivalent to testing if the two unknown densities $f_1(X)$, $f_2(X)$ are equal at the point X, versus the alternative hypothesis that the case group density is larger than the population group density, indicating high risk. Let α be a fixed level of significance and t_α be the lower boundary of the critical region of the right-tailed test at a level α. Then, $\Pr\{\hat{\gamma}(X) \geq t_\alpha|H_0\} = \alpha = 1 - I_{[1-\phi\{t_\alpha,1\}]}(k_2,k_1)$. For a given confidence level c and numbers of nearest neighbors $k_1,\ k_2$, let $\eta_c = \eta_c(k_2,k_1)$ be the c-th quantile of the Beta-distribution

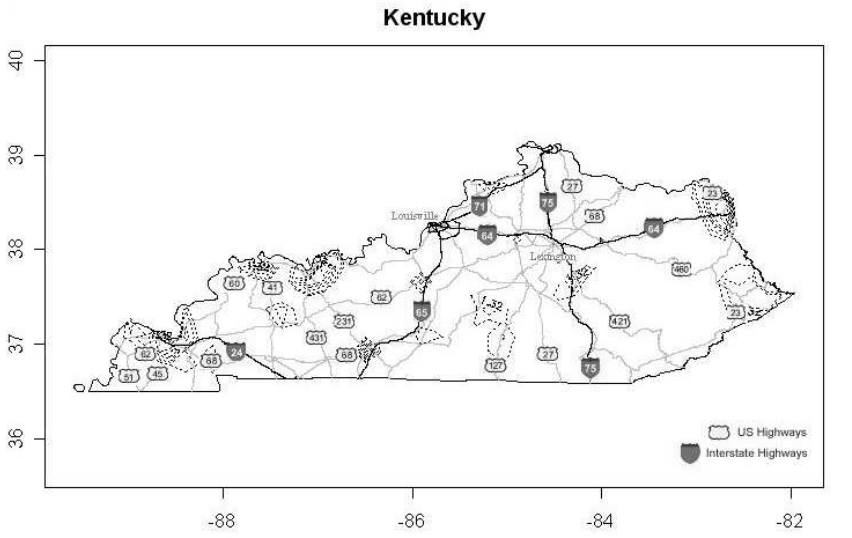
(k_2, k_1). Thus, from the last equation, we have $\phi(t_\alpha, 1) = 1 - \eta_{1-\alpha}$. Solving for t_α, we reject H_0 and claim excessive risk at X when $\hat{\gamma}(X) \geq t_\alpha = \dfrac{k_1}{k_2} \dfrac{\eta_{1-\alpha}}{1-\eta_{1-\alpha}}$.

5. Application. The kNN relative risk estimator has been implemented as an exploratory tool in the Kentucky Occupational Safety and Health Surveillance (KOSHS) program to create a continuous relative risk representation based on existing discrete data for the purpose of hypothesis generation. KOSHS is part of a program, sponsored by the US National Institute for Occupational Safety and Health (NIOSH), to conduct surveillance of 19 basic state-wide indicators for occupational injury and health and to build capacity for state-based occupational surveillance. Occupational injury surveillance is defined as the routine, ongoing collection, analysis and dissemination of data for the purpose of developing injury prevention programs in the workplace. More information can be found at `http://www.kiprc.uky.edu/projects/KOSHS` or `http://www.cdc.gov/niosh/topics/surveillance/`.

One emphasis area of the KOSHS program is on older drivers of large commercial trucks because one in six of the US long-haul truck drivers is 55 years of age or older (US Census, 2000). We want to identify geographical areas where the older large truck drivers are at higher risk for creating collisions in order to improve our injury prevention education programs. For the first step of the exploratory study, we obtained data from the Kentucky State Police Collision Report Analysis for Safer Highways (CRASH) data set. This electronic file contains information for all motor vehicle collisions in Kentucky: drivers, passengers, and roadway conditions, human or environmental factors contributing to the collision, and geographical coordinates of the collisions. Cases were identified as male drivers 50 years of age or older in at-fault large truck collisions and controls were male drivers 50 years of age or older in not-at-fault large truck collisions, using the CRASH electronic database from 2002 to 2006. Unit type classifications included: trucks and trailers, truck-single unit, truck-tractor and semi-trailer, and truck-other combination. All vehicles were designated as commercial vehicles in the CRASH file.

We used the kNN method ($n_1 = 5528$, $n_2 = 1584$, $k_1 = 100$, $k_2 = 50$) to estimate the density of the cases and the density of the controls and to construct the relative risk for being at-fault versus being not-at-fault for the older commercial drivers on the Kentucky roadways. The lower boundary t_α of the critical region of the right-tailed test at a level $\alpha = 0.05$ had a value of 1.32. Therefore, when the value of the estimated relative risk at a particular location was higher that 1.32, the geographical location was considered to present significantly higher relative risk for the older large truck drivers to be at-fault. The areas of high risk are marked with dashed lines on the contour map (Fig. 1).

Our exploratory analysis showed that the older commercial drivers are at higher risk for creating collisions mainly in two geographical areas – in the western and eastern ends of the state, off the interstates. A follow-up study, investigating the characteristics of collisions involving at-fault and not-at-fault large truck older male drivers and the effect of the passengers on working drivers, found that curvy and graded/hillcrest roads, and roads with one or two lanes increased the odds that the driver would be at fault in a large truck collision, and that passengers were protective in the prevention of at-fault large truck collisions among older drivers (a detailed manuscript is in progress).



Geographical areas of elevated relative risk for at-fault vs. not-at-fault commersial motor vehicle collisions

Fig. 1

REFERENCES

[1] K. Fukunaga, L. D. Hostetler. Optimization of k-nearest neighbor density estimates. *IEEE Transactions on Information Theory*, **19**, No 3 (1973), 320–326.

[2] D. Pavlov, S. Slavova, R. Kryscio. Estimating Relative Risk on the Line Using Nearest Neighbor Statistics. Methodology and Computing in Applied Probability, 2007, DOI 10.1007/s11009-007-9039-1.

[3] D. Pavlov. Identifying Spatial Disease Clusters in Nonhomogeneous populations. PhD Dissertation, 2001, University of Kentucky.

Svetla Slavova
Department of Statistics
University of Kentucky
817 Patterson Office Tower
Lexington, KY 40506, USA
e-mail: ssslav2@email.uky.edu

Richard J. Kryscio
Department of Statistics and Biostatistics
University of Kentucky
817 Patterson Office Tower
Lexington, KY 40506, USA
e-mail: kryscio@email.uky.edu

Terry L. Bunn
Kentucky Injury Prevention and Research Center
University of Kentucky
333 Waller Ave
Lexington, KY 40504, USA
e-mail: tlbunn2@email.uky.edu

СЪСТОЯТЕЛНОСТ НА ОЦЕНКАТА ПО МЕТОДА НА k-ТИЯ НАЙ-БЛИЗЪК СЪСЕД ЗА ОТНОСИТЕЛНИЯ РИСК И НЕЙНОТО ПРИЛОЖЕНИЕ В СИСТЕМАТА ЗА ПРОСЛЕДЯВАНЕ НА НАРАНЯВАНИЯТА

Светла Славова, Ричард Крисио, Тери Бън

Функцията на относителния риск може да се дефинира като отношение на две функции на вероятностна плътност. За оценка на плътността се използва непараметричният подход на k-тия най-близък съсед (kNN). Доказана е асимптотична състоятелност на kNN оценката за относителния риск във фиксирана точка. Обсъдено е приложението на kNN оценката за относителен риск в системата за проследяване на нараняванията.