Iranian Journal of Science & Technology, Transaction A, Vol. 30, No. A2
Printed in The Islamic Republic of Iran, 2006

SIMILARITY MEASURE FOR TWO DENSITIES*

A. R. SOLEIMANI[1**] AND J. BEHBOODIAN[2]

[1]Department of Statistics, College of Sciences, Shiraz University, 71454 Shiraz, I. R of Iran
[2]Department of Mathematics, Shiraz Islamic Azad University, Shiraz, I. R of Iran
Emails: a.soleimani@sttu.ac.ir, Behboodian@stat.susc.ac.ir

Abstract – Scott and Szewczyk in Technometrics, 2001, have introduced a similarity measure for two densities f_1 and f_2, by

$$sim(f_1,f_2)=\frac{<f_1,f_2>}{\sqrt{<f_1,f_1><f_2,f_2>}}$$

where

$$<f_1,f_2>=\int_{-\infty}^{+\infty} f_1(x,\theta_1)f_2(x,\theta_2)dx.$$

$sim(f_1,f_2)$ has some appropriate properties that can be suitable measures for the similarity of f_1 and f_2. However, due to some restrictions on the value of parameters and the kind of densities, discrete or continuous, it cannot be used in general.

The purpose of this article is to give some other measures, based on modified Scott's measure, and Kullback information, which may be better than $sim(f_1,f_2)$ in some cases. The properties of these new measures are studied and some examples are provided.

Keywords – Mixed model, similarity measure, kullback information, poisson distribution, normal distribution

1. INTRODUCTION

Scott and Szewezyk [1] have introduced a similarity measure for two densities, which is defined and denoted by

$$sim(f_1,f_2)=\frac{<f_1,f_2>}{\sqrt{<f_1,f_1><f_2,f_2>}}\ , \tag{1}$$

where $<f_1,f_2>=\int_{-\infty}^{+\infty} f_1(x,\theta_1)f_2(x,\theta_2)dx$, if f_1 and f_2 are continuous densities, and $<f_1,f_2>=\sum_x f_1(x,\theta_1)f_2(x,\theta_2)$, if f_1 and f_2 are discrete densities.

Their motivation for giving this measure was to reduce the number of components in a finite mixture that you find, for example, in McLachlan and Peel [2]. This similarity measure by itself can be used for different aspects of statistical inference.

It is easy to show that $sim(f_1,f_2)$ has the following appropriate properties:

a) (Symmetry) $sim(f_1,f_2)=sim(f_2,f_1)$

b) (By Cauchy-Schwartz) $0\le sim(f_1,f_2)\le 1$

c) $sim(f_1,f_2)$=1 if and only if $f_1=f_2$

*Received by the editor August 27, 2005 and in final revised form November 8, 2006
**Corresponding author

When $sim(f_1, f_2)$ is close to one, we can assume that $f_1 = f_2$.

It is shown in [1] that when f_1 and f_2 are normal densities, $sim(f_1, f_2)$ has a closed–form expression. However, we found that when f_1 and f_2 are $Bin(n, \theta_1)$ and $Bin(n, \theta_2)$, or $Poiss(\theta_1)$ and $Poiss(\theta_2)$, $sim(f_1, f_2)$ cannot be computed easily. On the other hand, when f_1 and f_2 are $Beta(\theta_1, 1)$ and $Beta(\theta_2, 1)$, we have

$$sim(f_1, f_2) = \frac{\sqrt{(2\theta_1 - 1)(2\theta_2 - 1)}}{\theta_1 + \theta_2 - 1}, \quad \theta_1 \geq 0.5, \theta_2 \geq 0.5 \tag{2}$$

Thus, we should have some restriction on the parameters. The above examples show that $sim(f_1, f_2)$, due to the nature of the densities, cannot be used in general. In this article, we introduce two new measures, which have more general scopes and have the same properties as measure sim(.,.). In Section 2 a modified and new version of $sim(f_1, f_2)$ is introduced. Section 3 is devoted to the Kullback similarity measure based on the known Kullback information. In Section 4 we study these similarity measures in general, for an exponential family. Finally these similarity measures are compared with each other by some numerical examples.

2. A MODIFIED VERSION OF $sim(f_1, f_2)$

We define and denote a modified and new version of $sim(f_1, f_2)$, for densities f_1 and f_2, by

$$simm(f_1, f_2) = [< f_1, \frac{f_1}{f_2} >< f_2, \frac{f_2}{f_1} >]^{-1} \tag{3}$$

$simm(f_1, f_2)$ has the same properties as $sim(f_1, f_2)$, i.e.,

a) $simm(f_1, f_2) = simm(f_2, f_1)$

b) $0 < simm(f_1, f_2) \leq 1$

c) $simm(f_1, f_2) = 1$ if $f_1 = f_2$

The properties (a) and (c) are obvious and (b) is concluded from

$$< f_1, \frac{f_1}{f_2} > = \int_{-\infty}^{+\infty} \frac{[f_1(x)]^2}{f_2(x)} dx = E_2[\frac{f_1(X)}{f_2(X)}]^2 \geq \{E_2[\frac{f_1(X)}{f_2(X)}]\}^2 = 1,$$

where $E_i[.]$ means expectation of a random variable with respect to f_i.

By some numerical examples in Section 5, it is conjectured that $simm(f_1, f_2) \leq sim(f_1, f_2)$, but it is not easy to investigate this conjecture. However, the following examples show that the computation of $simm(f_1, f_2)$ is quicker than the computation of $sim(f_1, f_2)$ for many densities.

Example 1. If f_1 and f_2 are $Poiss(\theta_1)$ and $Poiss(\theta_2)$, then

$$simm(f_1, f_2) = \exp[-\frac{(\theta_1 + \theta_2)(\theta_1 - \theta_2)^2}{\theta_1 \theta_2}],$$

while $sim(f_1, f_2)$ is too complicated.

Example 2. If f_1 and f_2 are $Bin(n, \theta_1)$ and $Bin(n, \theta_2)$, then

$$simm(f_1, f_2) = [1 + \frac{(\theta_2 - \theta_1)^2}{\theta_1(1 - \theta_1)} + \frac{(\theta_2 - \theta_1)^2}{\theta_2(1 - \theta_2)} + \frac{(\theta_2 - \theta_1)^4}{\theta_1 \theta_2 (1 - \theta_1)(1 - \theta_2)}]^{-n}.$$

For $\theta_1 = \theta_2$ we have $simm(f_1, f_2) = 1$, i.e., $f_1 = f_2$.

Example 3. If f_1 and f_2 are two components of the following mixture

$$f(x) = \sum_{i=1}^{m} p_i \frac{1}{\beta_i} e^{-\frac{(x-\alpha_i)}{\beta_i}} I(x \geq \alpha_i),$$

then

$$simm(f_1, f_2) = \frac{(2\beta_2 - \beta_1)(2\beta_1 - \beta_2)}{\beta_1\beta_2} \exp(-\frac{(\beta_1 - \beta_2)(\alpha_1 - \alpha_2)}{\beta_1\beta_2}).$$

3. KULLBACK SIMILARITY MEASURE

This measure, which is based on Kullback information, is defined and denoted by

$$simk(f_1, f_2) = [1 + D(f_1 \,//\, f_2) + D(f_2 \,//\, f_1)]^{-1} , \tag{4}$$

where

$$D(f_1 \,//\, f_2) = \int_{-\infty}^{+\infty} f_1(x) \ln[\frac{f_1(x)}{f_2(x)}] dx = E_1\{\ln[\frac{f_1(X)}{f_2(X)}]\},$$

is non-negative and zero if $f_1 = f_2$ (see Zeevi and Meir [3]and Schervish [4]). $simk(f_1, f_2)$ has the same properties as $sim(f_1, f_2)$ and $simm(f_1, f_2)$, i.e.,

a) $simk(f_1, f_2) = simk(f_2, f_1)$

b) $0 < simk(f_1, f_2) \leq 1$

c) $simk(f_1, f_2) = 1$ if $f_1 = f_2$

The proof of (b) is obtained from the fact that $\ln z \geq 1 - \frac{1}{z}$ for $z > 0$ and as a result

$$D(f_1 \,//\, f_2) = E_1\{\ln[\frac{f_1(X)}{f_2(X)}]\} \geq E_1[1 - \frac{f_2(X)}{f_1(X)}] = 0$$

Theorem: we have

$$simm(f_1, f_2) \leq simk(f_1, f_2). \tag{5}$$

Proof:

$$D(f_1 \,//\, f_2) = E_1\{\ln[\frac{f_1(X)}{f_2(X)}]\} \leq \ln\{E_1[\frac{f_1(X)}{f_2(X)}]\} \qquad \text{(By Jensen's inequality)}$$

and similarly we have

$$D(f_2 \,//\, f_1) \leq \ln\{E_2[\frac{f_2(X)}{f_1(X)}]\}.$$

Therefore, using $\ln z \geq 1 - \frac{1}{z}$, we obtain

$$1 + D(f_1 \,//\, f_2) + D(f_2 \,//\, f_1) \leq 1 + \ln\{E_1[\frac{f_1(X)}{f_2(X)}] E_2[\frac{f_2(X)}{f_1(X)}]\}$$

$$[simm(f_1, f_2)]^{-1} \leq 1 + \ln[simk(f_1, f_2)]^{-1}$$

$$simm(f_1, f_2) \leq simk(f_1, f_2).$$

Inequality (5) may say that $simm(f_1, f_2)$ better shows the similarity of f_1 and f_2 rather than $simk(f_1, f_2)$.

Example 4. If f_1 and f_2 are $Poiss(\theta_1)$ and $Poiss(\theta_2)$, then

$$simk(f_1, f_2) = [1 + (\theta_1 - \theta_2)\ln(\frac{\theta_1}{\theta_2})]^{-1}.$$

Example 5. If f_1 and f_2 are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, we have

$$simk(f_1, f_2) = [1 + \frac{(\sigma_1^2 - \sigma_2^2)^2 + (\sigma_1^2 + \sigma_2^2)(\mu_1 - \mu_2)^2}{2\sigma_1^2\sigma_2^2}]^{-1}.$$

Example 6. If f_1 and f_2 are $Bin(n, \theta_1)$ and $Bin(n, \theta_2)$, then

$$simk(f_1, f_2) = [1 + n(\theta_2 - \theta_1)\ln\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}]^{-1}.$$

We observe that in the above examples, as the parameters disperse, the similarity measures go to zero.

4. SIMILARITY MEASURES FOR AN EXPONENTIAL FAMILY

Definition: A family of distributions on the real line with probability mass function or density $f(x/\theta)$, $\theta \in \Theta$ (θ may be a vector) is said to be an exponential family of distributions, if $f(x/\theta)$ is of the following form:

$$f(x/\theta) = c(\theta)h(x)\exp[\sum_{i=1}^{k} \pi_i(\theta)t_i(x)], \tag{6}$$

where

$$c(\theta) = \{\sum_x h(x)\exp[\sum_{i=1}^{k} \pi_i(\theta)t_i(x)]\}^{-1},$$

in the discrete case and

$$c(\theta) = \{\int_x h(x)\exp[\sum_{i=1}^{k} \pi_i(\theta)t_i(x)]dx\}^{-1},$$

in the absolutely continuous case (see Ferguson [5] and Lehmann [6]).

The following table gives the values of similarity measures for two densities, f_1 and f_2, from an exponential family.

Table 1. Similarity measures for an exponential family

Measure	**Value**
$sim(f_1, f_2)$	$\frac{C1(\theta_1, \theta_2)}{\sqrt{C1(\theta_1, \theta_1)C1(\theta_2, \theta_2)}}$
$simm(f_1, f_2)$	$\frac{C2(\theta_1, \theta_2)C2(\theta_2, \theta_1)}{C(\theta_1)C(\theta_2)}$
$simk(f_1, f_2)$	$\{1 + \sum_{i=1}^{k} [\pi_i(\theta_1) - \pi_i(\theta_2)][E_{f_1}(t_i(x)) - E_{f_2}(t_i(x))]\}^{-1}$

The notations in Table 1 are defined as follows:

$$C1(\theta_1,\theta_2) = \int_x h^2(x)\exp(\sum_{i=1}^{k}[\pi_i(\theta_1)+\pi_i(\theta_2)]t_i(x))dx$$

$$C2(\theta_1,\theta_2) = \{\int_x h\ (x)\exp(\sum_{i=1}^{k}[2\pi_i(\theta_1)+\pi_i(\theta_2)]t_i(x))dx\}^{-1}$$

$$C3(\theta_1,\theta_2) = \{\int_x h\ (x)\exp(\frac{1}{2}\sum_{i=1}^{k}[\pi_i(\theta_1)+\pi_i(\theta_2)]t_i(x))dx\}^{-1}.$$

$$E_{f_j}[t_i(X)] = -\frac{\partial}{\partial\pi_i(\theta_j)}\ln[C(\theta_j)]; \quad j=1,2.$$

5. SOME NUMERICAL EXAMPLES

If f_1 and f_2 are $Poiss(\theta_1)$ and $Poiss(\theta_2)$, we obtain

$$sim(f_1,f_2) = \frac{\sum_{x=0}^{+\infty}(\theta_1\theta_2)^x\frac{\exp[-(\theta_1+\theta_2)]}{x!x!}}{\sqrt{\sum_{x=0}^{+\infty}\theta_1^{2x}\frac{\exp(-2\theta_1)}{x!x!}\sum_{x=0}^{+\infty}\theta_2^{2x}\frac{\exp(-2\theta_2)}{x!x!}}}.$$

$$simm(f_1,f_2) = \exp[-\frac{(\theta_1+\theta_2)(\theta_1-\theta_2)^2}{\theta_1\theta_2}],$$

$$simk(f_1,f_2)=[1+(\theta_1-\theta_2)\ln(\frac{\theta_1}{\theta_2})]^{-1}.$$

If f_1 and f_2 are $N(\theta_1,1)$ and $N(\theta_2,1)$, we have

$$sim(f_1,f_2) = \exp(-\frac{(\theta_1-\theta_2)^2}{4}).$$

$$simm(f_1,f_2)=\exp[-2(\theta_1-\theta_2)^2].$$

$$simk(f_1,f_2) = \frac{1}{1+(\theta_1-\theta_2)^2}.$$

For different values of θ_1 and θ_2 we obtain the following tables by using Maple. These tables give the different values of measures for numerically comparing them and determining the approximate equality of f_1 and f_2. For example, for θ_1=0.50 and θ_2=0.45 in Table 2, we have $simk(f_1,f_2)$=0.99476. This shows that f_1 and f_2 are close to each other.

Acknowledgments- The authors would like to thank the editor and referees for their helpful suggestions that helped to improve the presentation of the paper.

Table 2. Similarity measures for two Poisson Densities
$\theta_1 = 0.50$

Measures \ θ_2	0.40	0.45	0.50	0.55	0.60
$sim(f_1, f_2)$	0.99585	0.99898	1.00000	0.99902	0.99616
$simm(f_1, f_2)$	0.95599	0.98950	1.00000	0.99050	0.96400
$simk(f_1, f_2)$	0.97817	0.99476	1.00000	0.99526	0.98209

Table 3. Similarity measures for two Poisson Densities
$\theta_1 = 1.00$

Measures \ θ_2	0.80	0.85	0.90	0.95	1.00
$sim(f_1, f_2)$	0.98881	0.99384	0.99732	0.99934	1.00000
$simm(f_1, f_2)$	0.91393	0.95221	0.97911	0.99488	1.00000
$simk(f_1, f_2)$	0.95728	0.97620	0.98957	0.99744	1.00000

Table 4. Similarity measures for two Normal Densities
$\theta_1 = 0.50$

Measures \ θ_2	0.50	0.80	1.10	1.40	2.00
$sim(f_1, f_2)$	1.0000	0.9778	0.9139	0.8167	0.5698
$simm(f_1, f_2)$	1.0000	0.8353	0.4868	0.1979	0.0111
$simk(f_1, f_2)$	1.0000	0.9174	0.7353	0.5525	0.3077

Table 5. Similarity measures for two Normal Densities
$\theta_1 = 1.00$

Measures \ θ_2	0.80	0.85	0.90	0.95	1.00
$sim(f_1, f_2)$	0.9900	0.9944	0.99861	0.9975	1.00000
$simm(f_1, f_2)$	0.9231	0.9560	0.97778	0.9802	1.00000
$simk(f_1, f_2)$	0.9615	0.9780	0.98901	0.9901	1.00000

REFERENCES

1. Scott, D. W. & Szewezyk, W. F. (2001). From Kernel to Mixture. *Technometrics*, *43*(3), 323.
2. McLachlan, G. & Peel, D. (2000). Finite Mixture Models. *John Wiley*.
3. Zeevi, A. J. & Meir, R. (1991). Density estimation through convex combinations of densities; approximation and estimation bounds. *Journal of Statistics and Probability Letters, 10*(1).

4. Schervish, M. J. (1995). Theory of Statistics. *Springer Verlag*, 115-116.
5. Ferguson, T. S. (1996). A Course in Large Sample Theory. *Chapman and Hall*, 59.
6. Lehmann, E. L. & Casella, G. (1998). Theory of Point Estimation. Second Edition, *Springer*, 259, 47.