Research Article

Fixed-Effects Modeling of Cohen's Weighted Kappa for Bivariate Multinomial Data: A Perspective of Generalized Inverse

Jingyun Yang[1] and Vernon M. Chinchilli[2]

[1] *Department of Biostatistics and Epidemiology, College of Public Health, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA*

[2] *Department of Public Health Sciences, Penn State College of Medicine, 600 Centerview Drive, Suite 2200, Hershey, PA 17033-0855, USA*

Correspondence should be addressed to Vernon M. Chinchilli, vchinchi@hes.hmc.psu.edu

Cohen's kappa and weighted kappa statistics are the conventional methods used frequently in measuring agreement for categorical responses. In this paper, through the perspective of a generalized inverse, we propose an alternative general framework of the fixed-effects modeling of Cohen's weighted kappa, proposed by Yang and Chinchilli (2011). Properties of the proposed method are provided. Small sample performance is investigated through bootstrap simulation studies, which demonstrate good performance of the proposed method. When there are only two categories, the proposed method reduces to Cohen's kappa.

1. Introduction

Measurement of agreement is used widely in diverse areas of scientific research to assess the reproducibility of a new assay, instrument, or method, the acceptability of a new or generic process and methodology, as well as in method comparison. Examples include the agreement when two or more methods or raters simultaneously assess a response [1, 2] or when one rater makes the same assessment at two times [3], the agreement of a newly developed method with a gold standard method [4], and the agreement of observed values with predicted values [5].

Traditionally, kappa [6] is used for measurement of agreement for categorical responses and weighted kappa [7] for ordinal responses. The concordance correlation coefficient [8, 9] is often used when the responses are continuous [10, 11]. In this paper, we focus on the measurement of agreement for categorical responses, that is, when the responses are either nominal or ordinal.

Table 1: The bivariate probability table.

$X \setminus Y$	0	1	2	$\cdots$	i	$\cdots$	j	$\cdots$	k	$P_X(x)$
0	p_{00}	p_{01}	p_{02}	$\cdots$	p_{0i}	$\cdots$	p_{0j}	$\cdots$	p_{0k}	$p_{0\cdot}$
1	p_{10}	p_{11}	p_{12}	$\cdots$	p_{1i}	$\cdots$	p_{1j}	$\cdots$	p_{1k}	$p_{1\cdot}$
2	p_{20}	p_{21}	p_{22}	$\cdots$	p_{2i}	$\cdots$	p_{2j}	$\cdots$	p_{2k}	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	p_{i0}	p_{i1}	p_{i2}	$\cdots$	p_{ii}	$\cdots$	p_{ij}	$\cdots$	p_{ik}	$p_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	p_{j0}	p_{j1}	p_{j2}	$\cdots$	p_{ji}	$\cdots$	p_{jj}	$\cdots$	p_{jk}	$p_{j\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	p_{k0}	p_{k1}	p_{k2}	$\cdots$	p_{ki}	$\cdots$	p_{kj}	$\cdots$	p_{kk}	$p_{k\cdot}$
$P_Y(y)$	$p_{\cdot 0}$	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot i}$	$\cdots$	$p_{\cdot j}$	$\cdots$	$p_{\cdot k}$	1

Note: there are $k + 1$ categories in the response. $P_X(x)$ and $P_Y(y)$ denote the marginal probabilities.

Suppose there is a bivariate response (X, Y), where each of X and Y yields a categorical response. For convenience, the categories are denoted as $0, 1, 2, \ldots, k$. Suppose the bivariate distribution is defined as in Table 1.

Cohen [6] proposed a coefficient of agreement, called the kappa coefficient, for nominal scales of responses, which is defined by

$$\kappa = \frac{p_o - p_c}{1 - p_c}, \tag{1.1}$$

where p_o is the observed proportion of agreement and p_c is the proportion of agreement expected by chance.

To deal with ordinal responses, Cohen [7] proposed a weighted version of the kappa statistic:

$$\kappa_w = 1 - \frac{q'_o}{q'_c} = 1 - \frac{\sum v_{ij} p_{oij}}{\sum v_{ij} p_{cij}}, \tag{1.2}$$

with

$$q'_o = \frac{\sum v_{ij} p_{oij}}{v_{\max}},$$
$$q'_c = \frac{\sum v_{ij} p_{cij}}{v_{\max}}, \tag{1.3}$$

where v_{ij} is the nonnegative weight assigned to the disagreement for the cell ij, $v_{\max}$ is the weight assigned to the maximum disagreement, p_{oij} is the proportion of joint judgement in the ij cell, and p_{cij} is the proportion expected by chance in the same cell.

Despite the fact that they may fail to work well under certain situations [12, 13], Cohen's kappa and weighted kappa coefficients have been widely used in various areas as

a measure of agreement for categorical responses, partly due to their ease in calculation; see, for example, Justice et al. [14], Kerosuo and Ørstavik [15], Landers et al. [16], and Suebnukarn et al. [17]. In the meanwhile, researchers have been exploring alternative methods for measuring agreement for categorical data; see Agresti [18], Barlow [19], Broemeling [20], Carrasco and Jover [21], Dou et al. [22], Fanshawe et al. [23], Graham and Jackson [24], King and Chinchilli [25], Kraemer [26], Landis and Koch [27], Laurent [28], Lin et al. [29], and Svensson [30], to name only a few. It is worth noting that, for ordinal or binary data, the concordance correlation coefficient (CCC) [8] based on the squared function of distance reduces to Cohen's weighted kappa when the Fleiss-Cohen weight function is used and the CCC based on the absolute value function of distance to the power $\delta = 1$ is equivalent to Cohen's weighted kappa if Cicchetti-Allison weight function is used [25]. In addition, variance of the CCC is identical to the one given by Fleiss et al. [31] if the Fleiss-Cohen weight function and the GEE approach are used [29].

Yang and Chinchilli developed a fixed-effects modeling of kappa and weighted kappa [32, 33]. Their simulation studies and illustrative examples show good performance of the proposed methods. In this paper, we propose another version of fixed-effects modeling of Cohen's kappa and weighted kappa through a perspective of a generalized inverse. We want to stress that this paper is not to demonstrate the advantage of the proposed method in constructing an agreement coefficient. Rather, it provides an alternative form of a general framework and direction under which new and better matrix functions could be explored.

In Section 2 we provide the derivation and properties of the proposed method. In Section 3, we compare the small sample performance of the proposed method with that of three other agreement coefficients through bootstrap simulation studies. Examples are given in Section 4 followed by a conclusion.

2. The Generalized Inverse Version of Fixed-Effects Modeling of Kappa and Weighted Kappa

2.1. Derivation

Using similar notation as in Yang and Chinchilli [32], we let X and Y be categorical variables, with the categories being designated as $0, 1, 2, \ldots, k$, as displayed in Table 1. Let $p_{ij} = \Pr(X = i, Y = j)$, $i, j = 0, 1, 2, \ldots, k$, denote the bivariate probability, and let $p_{i \cdot} = \Pr(X = i)$ and $p_{\cdot j} = \Pr(Y = j)$ denote the marginal probabilities. Let $I(\cdot)$ denote the indicator function, and define the $(k + 1) \times 1$ vectors as

$$Z_X = \begin{bmatrix} I(X = 0) \\ I(X = 1) \\ I(X = 2) \\ \vdots \\ I(X = k) \end{bmatrix},$$

$$Z_Y = \begin{bmatrix} I(Y = 0) \\ I(Y = 1) \\ I(Y = 2) \\ \vdots \\ I(Y = k) \end{bmatrix}.$$

(2.1)

Define the $(k + 1) \times (k + 1)$ matrices P_D and P_I as

$$P_D = E\left[(Z_X - Z_Y)(Z_X - Z_Y)^T\right],$$

$$P_I = E_{\text{Ind}}\left[(Z_X - Z_Y)(Z_X - Z_Y)^T\right],$$

(2.2)

where $E_{\text{Ind}}[\cdot]$ denotes expectation under the assumption that X and Y are independent. It easily can be verified that P_D and P_I are nonnegative definite matrices, denoted by $P_D \geq 0$ and $P_I \geq 0$.

Yang and Chinchilli [32] show that

$$P_I - P_D$$

$$= \begin{bmatrix} 2(p_{00} - p_{0\cdot}p_{\cdot0}) & (p_{01} - p_{0\cdot}p_{\cdot1}) + (p_{10} - p_{1\cdot}p_{\cdot0}) & \cdots & (p_{0k} - p_{0\cdot}p_{\cdot k}) + (p_{k0} - p_{k\cdot}p_{\cdot0}) \\ (p_{01} - p_{0\cdot}p_{\cdot1}) + (p_{10} - p_{1\cdot}p_{\cdot0}) & 2(p_{11} - p_{1\cdot}p_{\cdot1}) & \cdots & (p_{1k} - p_{1\cdot}p_{\cdot k}) + (p_{k1} - p_{k\cdot}p_{\cdot1}) \\ \vdots & \vdots & \ddots & \vdots \\ (p_{0k} - p_{0\cdot}p_{\cdot k}) + (p_{k0} - p_{k\cdot}p_{\cdot0}) & (p_{1k} - p_{1\cdot}p_{\cdot k}) + (p_{k1} - p_{k\cdot}p_{\cdot1}) & \cdots & 2(p_{kk} - p_{k\cdot}p_{\cdot k}) \end{bmatrix}.$$

(2.3)

P_I and P_D are $(k + 1) \times (k + 1)$ matrices of rank k because $P_I \cdot 1 = 0$ and $P_D \cdot 1 = 0$, where 1 is a $(k + 1) \times 1$ vector of unit values. Thus, $P_I \cdot 1 = 0 \cdot 1$, indicating that 0 is an eigenvalue for P_I with corresponding (standardized) eigenvector $e_0 = 1/\sqrt{k + 1}$. The same is true for P_D as well. Let $E_I = [e_0 e_1 \cdots e_k]$ denote the $(k+1) \times (k+1)$ matrix of eigenvectors for P_I. By definition, E_I is orthogonal, so $E_I(E_I)^T = (E_I)^T E_I = I$, which yields that $1^T e_1 = 0, \ldots, 1^T e_k = 0$. Because P_I is not of full rank, its inverse does not exist, but it does have a Moore-Penrose generalized inverse [34, 35], denoted by $(P_I)^+$. If $\Lambda_I = \text{Diag}(0, \lambda_1, \ldots, \lambda_k)$ denotes the $(k + 1) \times (k + 1)$ diagonal matrix of eigenvalues for P_I, then the Moore-Penrose generalized inverse of Λ_I is $(\Lambda_I)^+ = \text{Diag}(0, (\lambda_1)^{-1}, \ldots, (\lambda_k)^{-1})$ and the Moore-Penrose generalized inverse of P_I is $(P_I)^+ = E_I(\Lambda_I)^+(E_I)^T$.

It easily can be verified that $-P_I \leq P_I - P_D \leq P_I$, which implies that

$$-((P_I)^+)^{1/2}P_I((P_I)^+)^{1/2} \leq ((P_I)^+)^{1/2}P_I((P_I)^+)^{1/2} - ((P_I)^+)^{1/2}P_D((P_I)^+)^{1/2}$$

$$\leq ((P_I)^+)^{1/2}P_I((P_I)^+)^{1/2},$$

(2.4)

where $((P_I)^+)^{1/2} = E_I((\Lambda_I)^+)^{1/2}(E_I)^T$.

Since $((P_I)^+)^{1/2} P_I ((P_I)^+)^{1/2} = I - 1 \cdot 1^T/(k+1)$, (2.4) can be rewritten as

$$-\left\{I - 1 \cdot 1^T/(k+1)\right\} \leq \left\{I - 1 \cdot 1^T/(k+1)\right\} - ((P_I)^+)^{1/2} P_D ((P_I)^+)^{1/2} \leq \left\{I - 1 \cdot 1^T/(k+1)\right\}. \tag{2.5}$$

Suppose $g(A)$ is a function of the symmetric matrix A and $g(\cdot)$ satisfies the following two definitions [32].

Definition 2.1. Suppose $g(\cdot)$ is a $\mathbb{R}^{(k+1)\times(k+1)} \mapsto \mathbb{R}$ function. $g(\cdot)$ is said to be nondecreasing if for all nonnegative definite matrices $A_{(k+1)\times(k+1)}$, $B_{(k+1)\times(k+1)}$, where $A \leq B$, one has $g(A) \leq g(B)$.

Definition 2.2. Suppose $g(\cdot)$ is a $\mathbb{R}^{(k+1)\times(k+1)} \mapsto \mathbb{R}$ function. $g(\cdot)$ is said to be a scale-equivariant function [36] if $g(CA) = cg(A)$ for all matrices and constant c.

Then a class of agreement coefficients is given by

$$\kappa_g^* = 1 - \frac{g\left\{(P_I^+)^{1/2} P_D (P_I^+)^{1/2}\right\}}{g\left\{(I - (1/(k+1))1 \cdot 1^T)\right\}}. \tag{2.6}$$

If $g(A) = \text{tr}(A)$, then

$$\kappa_{\text{tr}}^* = 1 - \frac{1}{k} \text{tr}(P_D (P_I)^+). \tag{2.7}$$

If we set $g(A) = \text{tr}(W * A)$, then

$$\kappa_{\text{tr}}^*(W) = 1 - \frac{\text{tr}(W * P_D * (P_I)^+)}{\text{tr}(W) - (1/(k+1))\sum_{i=0}^{k}\sum_{j=0}^{k} W_{ij}}, \tag{2.8}$$

where W is a nonnegative definite $(k+1) \times (k+1)$ symmetric matrix of agreement weights with $w_{ij} = 1$ for $i = j = 0, 1, 2, \ldots, k$ and $0 \leq w_{ij} < 1$ for $i \neq j = 0, 1, 2, \ldots, k$. Two frequently used weighting schemes are the Cicchetti-Allison weights [37]

$$w_{ij} = 1 - \frac{|i-j|}{k}, \quad \text{for } i, j = 0, 1, 2, \ldots, k, \tag{2.9}$$

and the Fleiss-Cohen weights [38]

$$w_{ij} = 1 - \frac{(i-j)^2}{k^2}, \quad \text{for } i, j = 0, 1, 2, \ldots, k. \tag{2.10}$$

We can also use $g(A) = \lambda_{A_1}$, the largest eigenvalue of A as another function for $g(\cdot)$, which leads to a new agreement coefficient $\kappa_{\text{le}}^*(W)$.

Therefore, including the coefficients that we propose in this paper, there are four novel coefficients to assess agreement for categorical data: (1) $\kappa_{\text{tr}}(W)$, (2) $\kappa_{\text{le}}(W)$, (3) $\kappa_{\text{tr}}^*(W)$, and

(4) $\kappa_{\text{le}}^*(W)$. In practice, all these indices can be estimated using their sample counterparts. In this paper, a "^" is used to denote the estimate of an index.

2.2. Properties

In general, if $g(\cdot)$ satisfies the properties in the two aforementioned definitions, then $\kappa_g^*(W)$ has the following properties:

(1) $-1 \leq \kappa_g^*(W) \leq 1$,

(2) $\kappa_g^*(W) = 1$ if $X = Y$ with probability one, that is, $p_{ij} = 0$ for all $i \neq j$,

(3) $\kappa_g^*(W) = -1$ if and only if $p_{ii} = 0$ for each $i = 0, 1, 2, \ldots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \neq j$, excluding the set of degenerate cases,

(4) $\kappa_g^*(W) = 0$ if X and Y are independent.

Proof. Without loss of generality, we set $W = I$. The proofs for other weight matrices W are slightly more complex.

(1) From the previous proof, we know that

$$0 \leq (P_I^+)^{1/2} P_D (P_I^+)^{1/2} \leq 2 \left(I - \frac{1}{k+1} 1 \cdot 1^T \right). \tag{2.11}$$

Since $g(\cdot)$ is a nondecreasing and scale-equivariant function, we have

$$0 \leq g\left\{ (P_I^+)^{1/2} P_D (P_I^+)^{1/2} \right\} \leq 2g\left\{ \left(I - \frac{1}{k+1} 1 \cdot 1^T \right) \right\}. \tag{2.12}$$

It is then straightforward to see that $-1 \leq \kappa_g^* \leq 1$.

(2) If $X = Y$ with probability one, that is, $p_{ij} = 0$ for all $i \neq j$, then $P_D = 0$ [32]. It is then straightforward to see that $\kappa_g^* = 0$.

(3) Excluding the set of degenerate cases, if $p_{ii} = 0$ for each $i = 0, 1, 2, \ldots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \neq j$, then $P_I - P_D = -P_I \Rightarrow P_I = (1/2)P_D$ [32]. Hence, $(P_I)^+ = ((1/2)P_D)^+ = 2(P_D)^+$ and $((P_I)^+)^{1/2} = \sqrt{2}((P_D)^+)^{1/2}$. Therefore, $(P_I^+)^{1/2} P_D (P_I^+)^{1/2} = 2(P_D^+)^{1/2} P_D (P_D^+)^{1/2}$ and $(P_I^+)^{1/2} P_I (P_I^+)^{1/2} = (P_D^+)^{1/2} P_D (P_D^+)^{1/2}$. Hence, $\kappa_g^* = -1$.

(4) If X and Y are independent, then $P_I = P_D$ [32]. It is straightforward to see that $\kappa_g^* = 0$.   □

Lemma 2.3. *When $k = 1$, that is, when there are only two categories, the four coefficients are equivalent and they all reduce to Cohen's kappa.*

For proof of the lemma, please see the appendix.

3. Small Sample Performance

3.1. Simulation Design

In this section, we compare the performance of Cohen's weighted kappa ($\kappa_{\mathrm{tr}}(W)$), $\kappa_{\mathrm{le}}(W)$, $\kappa_{\mathrm{tr}}^*(W)$, and $\kappa_{\mathrm{le}}^*(W)$. We used the bootstrap to estimate the sample variances of the last three coefficients because the asymptotic distributions of their estimates are too complex.

We used Matlab (version 7.10.0.499, The MathWorks Inc., Natick, Mass) to generate samples following multinominal distributions. Twelve different distributions were employed, representing different levels of agreement (from poor to excellent) and different numbers of categories in the responses (from 3 to 5). For details about the distributions and the design of the simulations, see Yang and Chinchilli [32]. We calculated the true coefficient values for κ, $\kappa_{\mathrm{le}}(W)$, $\kappa_{\mathrm{tr}}^*(W)$, and $\kappa_{\mathrm{le}}^*(W)$ as well as their estimate, bias, mean square error, the empirical variance, and the bootstrap variance. The weights we used are Cicchetti-Allison weights.

We did 5000 simulations for each sample size in each distribution and 1000 bootstrap replicates for each simulation. The observed variances of the proposed method and the mean bootstrap variances are compared. We follow Yang and Chinchilli [32] in the interpretation of the results. That is, agreement measures greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance and values below 0.40 or so may be taken to represent poor agreement.

3.2. Simulation Results and Discussion

Table 2 gives the simulation results for the four coefficients with the weighting scheme being Cicchetti-Allison weights (the simulation results are split into two tables because it is unwieldy to fit in one page).

In almost all the cases, the four agreement coefficients give the same classification in terms of the magnitude of the degrees of agreement. An interesting case is case 6 where the degree of agreement is around the boundary of poor and medium agreement. Cohen's weighted kappa and $\kappa_{\mathrm{tr}}^*(W)$ indicate close to medium agreement while $\kappa_{\mathrm{le}}(W)$ and $\kappa_{\mathrm{le}}^*(W)$ indicate poor agreement, according to our rule of interpretation of the results. Biases of all the four coefficients are small, and, as sample sizes increase, they become almost negligible. The bootstrap turns out to be a good alternative as an estimation of the sample variances, which in many cases gives the estimates of the variances equivalent to the observed values.

It also should be noted that the true values of coefficients within Table 2(a) can vary, especially for cases 5–9 ($k = 3$) and cases 10–12 ($k = 4$). This suggests that, as table size $(k + 1) \times (k + 1)$ increases, the coefficients could lead to different interpretations.

4. Examples

In this section, we use two empirical examples to illustrate the application and compare the result of the four agreement coefficients. The first example is from Brenner et al. [39] who examined the agreement of cause-of-death data for breast cancer patients between Ontario Cancer Registry (OCR) and classification results from Mount Sinai Hospital (MSH). OCR, established in 1964, collects information about all new cases of cancer in the province of Ontario while MSH has a systematic and rigorous monitoring and follow-up of confirmed

Table 2: Simulation results.

(a)

Case	n	k	κ_1	$\widehat{\kappa}_1$	b_1	κ_2	$\widehat{\kappa}_2$	b_2	κ_3	$\widehat{\kappa}_3$	b_3	κ_4	$\widehat{\kappa}_4$	b_4
1	80	2	0.495	0.490	-0.005	0.467	0.458	-0.009	0.493	0.487	-0.007	0.488	0.482	-0.007
1	160	2	0.495	0.495	0.000	0.467	0.466	-0.002	0.493	0.492	-0.001	0.488	0.488	-0.001
2	80	2	-0.128	-0.127	0.001	-0.030	-0.036	-0.006	-0.122	-0.121	0.002	-0.097	-0.096	0.001
2	160	2	-0.128	-0.126	0.001	-0.030	-0.031	-0.002	-0.120	-0.120	0.002	-0.097	-0.095	0.002
3	80	2	0.851	0.850	-0.001	0.841	0.834	-0.007	0.848	0.845	-0.002	0.845	0.843	-0.003
3	160	2	0.851	0.851	0.000	0.841	0.838	-0.003	0.848	0.847	-0.001	0.845	0.844	-0.001
4	80	2	0.538	0.535	-0.004	0.556	0.554	-0.002	0.583	0.580	-0.004	0.606	0.602	-0.004
4	160	2	0.538	0.537	-0.002	0.556	0.556	0.000	0.583	0.581	-0.002	0.606	0.604	-0.002
5	80	3	0.056	0.053	-0.003	-0.016	-0.019	-0.003	0.141	0.138	-0.003	0.171	0.167	-0.004
5	160	3	0.056	0.053	-0.003	-0.016	-0.020	-0.004	0.141	0.138	-0.003	0.171	0.168	-0.003
6	80	3	0.398	0.395	-0.003	0.305	0.299	-0.006	0.392	0.387	-0.005	0.378	0.371	-0.007
6	160	3	0.398	0.397	-0.001	0.305	0.303	-0.002	0.392	0.390	-0.002	0.378	0.374	-0.003
7	80	3	-0.174	-0.172	0.002	-0.254	-0.253	0.001	-0.166	-0.165	0.001	-0.190	-0.188	0.002
7	160	3	-0.174	-0.174	0.000	-0.254	-0.254	0.000	-0.166	-0.166	0.000	-0.190	-0.189	0.000
8	80	3	0.148	0.150	0.002	-0.058	-0.053	0.005	0.018	0.021	0.004	-0.013	-0.009	0.004
8	160	3	0.148	0.148	0.000	-0.058	-0.056	0.002	0.018	0.019	0.001	-0.013	-0.011	0.002
9	80	3	-0.319	-0.313	0.007	-0.411	-0.404	0.007	-0.296	-0.290	0.006	-0.322	-0.316	0.006
9	160	3	-0.319	-0.316	0.003	-0.411	-0.408	0.003	-0.296	-0.293	0.003	-0.322	-0.319	0.003
10	80	4	0.564	0.561	-0.003	0.543	0.532	-0.010	0.515	0.505	-0.009	0.522	0.511	-0.011
10	160	4	0.564	0.562	-0.001	0.543	0.538	-0.005	0.515	0.510	-0.005	0.522	0.516	-0.005
11	80	4	-0.075	-0.075	0.000	-0.087	-0.089	-0.002	-0.075	-0.076	-0.001	-0.087	-0.087	0.000
11	160	4	-0.075	-0.076	-0.001	-0.087	-0.090	-0.003	-0.075	-0.076	-0.001	-0.087	-0.088	-0.001
12	80	4	0.791	0.790	0.000	0.769	0.767	-0.002	0.703	0.699	-0.004	0.701	0.684	-0.017
12	160	4	0.791	0.792	0.001	0.769	0.770	0.000	0.703	0.702	-0.001	0.701	0.694	-0.007

Note: n is the sample size, $k+1$ is equal to the number of categories in the response. κ_1 is the true coefficient value for Cohen's weighted kappa, κ_2 stands for $\kappa_{le}(W)$, κ_3 stands for $\kappa_{le}^*(W)$, and κ_4 stands for $\kappa_{tr}^*(W)$. $\widehat{\kappa}_1$, $\widehat{\kappa}_2$, $\widehat{\kappa}_3$, and $\widehat{\kappa}_4$ are their estimates, respectively. b_1, b_2, b_3, and b_4 are the biases, respectively.

(b)

Case	n	k	κ_{b_2}	κ_{b_3}	κ_{b_4}	ms_1	ms_2	ms_3	ms_4	v_{2e}	v_{2b}	v_{3e}	v_{3b}	v_{4e}	v_{4b}
1	80	2	0.450	0.482	0.477	0.007	0.011	0.008	0.009	0.011	0.011	0.008	0.008	0.009	0.009
1	160	2	0.462	0.490	0.485	0.004	0.006	0.004	0.004	0.006	0.006	0.004	0.004	0.004	0.004
2	80	2	−0.041	−0.119	−0.094	0.004	0.007	0.004	0.005	0.007	0.007	0.004	0.004	0.005	0.004
2	160	2	−0.034	−0.119	−0.094	0.002	0.004	0.002	0.002	0.004	0.003	0.002	0.002	0.002	0.002
3	80	2	0.827	0.843	0.839	0.003	0.005	0.004	0.004	0.005	0.005	0.004	0.004	0.004	0.004
3	160	2	0.835	0.845	0.843	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
4	80	2	0.547	0.576	0.598	0.005	0.007	0.004	0.004	0.007	0.007	0.004	0.004	0.004	0.004
4	160	2	0.554	0.579	0.602	0.002	0.004	0.002	0.002	0.004	0.004	0.002	0.002	0.002	0.002
5	80	3	−0.022	0.135	0.163	0.008	0.016	0.007	0.008	0.016	0.015	0.007	0.007	0.008	0.008
5	160	3	−0.022	0.137	0.166	0.004	0.008	0.004	0.004	0.008	0.008	0.004	0.003	0.004	0.004
6	80	3	0.295	0.382	0.365	0.008	0.015	0.008	0.010	0.015	0.014	0.008	0.007	0.010	0.009
6	160	3	0.301	0.387	0.371	0.004	0.007	0.004	0.005	0.007	0.007	0.004	0.004	0.005	0.005
7	80	3	−0.252	−0.163	−0.186	0.004	0.008	0.003	0.004	0.008	0.008	0.003	0.003	0.004	0.004
7	160	3	−0.254	−0.166	−0.188	0.002	0.004	0.002	0.002	0.004	0.004	0.002	0.002	0.002	0.002
8	80	3	−0.049	0.025	−0.006	0.006	0.010	0.003	0.004	0.010	0.010	0.003	0.003	0.004	0.004
8	160	3	−0.054	0.021	−0.009	0.003	0.005	0.001	0.002	0.005	0.005	0.001	0.001	0.002	0.002
9	80	3	−0.399	−0.284	−0.311	0.003	0.008	0.004	0.006	0.008	0.008	0.004	0.005	0.006	0.007
9	160	3	−0.406	−0.291	−0.317	0.002	0.004	0.002	0.003	0.004	0.004	0.002	0.002	0.003	0.003
10	80	4	0.522	0.496	0.500	0.006	0.009	0.008	0.008	0.009	0.009	0.008	0.008	0.008	0.008
10	160	4	0.533	0.505	0.511	0.003	0.004	0.004	0.004	0.004	0.005	0.004	0.004	0.004	0.004
11	80	4	−0.091	−0.075	−0.087	0.003	0.007	0.003	0.004	0.007	0.007	0.003	0.003	0.004	0.004
11	160	4	−0.091	−0.076	−0.088	0.001	0.004	0.001	0.002	0.004	0.004	0.001	0.001	0.002	0.002
12	80	4	0.763	0.696	0.671	0.004	0.007	0.005	0.007	0.007	0.007	0.005	0.005	0.007	0.008
12	160	4	0.768	0.700	0.686	0.002	0.004	0.002	0.003	0.004	0.003	0.002	0.002	0.003	0.003

Note: n is the sample size, $k + 1$ is equal to the number of categories in the response. κ_{b_2}, κ_{b_3}, and κ_{b_4} are the mean bootstrap estimates for $\kappa_{le}^*(W)$, $\kappa_{tr}^*(W)$, and $\kappa_{le}^*(W)$, respectively. ms_1, ms_2, ms_3, and ms_4 are the MSE for the estimates of the four coefficients, respectively. v_{2e} and v_{2b} are the empirical variance and the bootstrap variance for $\kappa_{le}^*(W)$, respectively. v_{4e} and v_{4b} are the empirical variance and the bootstrap variance for $\kappa_{tr}^*(W)$, respectively. v_{3e} and v_{3b} are the empirical variance and the bootstrap variance for $\kappa_{le}^*(W)$, respectively.

Table 3: Data for example 1.

Ontario Cancer Registry	MSH study			
	No death information	Competing cause of death	Breast cancer	Total
No death information	1331	6	6	1343
Competing cause of death	19	129	7	155
Breast cancer	5	21	124	150
Total	1355	156	137	1648

Note: OCR and MSH studies classify cause of death for 1648 breast cancer patients as having no cancer-related information, competing cause of death, and breast cancer death.

Table 4: Data for example 2.

OGTT at revascularization	OGTT at one month later			
	Normal	Glucose intolerance	DM	Total
Normal	17	2	3	22
Glucose intolerance	22	10	4	36
DM	10	11	9	30
Total	49	23	16	88

Note: OGTT classifies 88 patients admitted with an acute coronary syndrome with no previous DM as having DM, being glucose intolerance or normal at the day after revascularization and one month later.

Table 5: Results for the two examples.

	κ_W	95% CI	κ_{le}	95% CI	$\kappa_{\text{tr}}^*(W)$	95% CI	$\kappa_{\text{le}}^*(W)$	95% CI
Example 1	0.900	(0.843, 0.957)	0.924	(0.899, 0.944)	0.872	(0.840, 0.902)	0.881	(0.848, 0.910)
Example 2	0.203	(0.079, 0.327)	0.257	(0.078, 0.417)	0.202	(0.052, 0.352)	0.218	(0.058, 0.374)

Note: The 95% CI was for κ_W calculated using the large sample variance derived by Fleiss et al. [31]. The CIs for the other three coefficients were calculated using bootstrap method with 5000 replicates.

node-negative breast cancer patients. Due to its relative complete and accurate data, MSH also provides specialist-determined cause of death. 1648 patients entered the analysis through linking the OCR data to the MSH study patients via the OCR standard procedure. The purpose of the study is to examine the degree of agreement between OCR, which is often used in cancer studies, and MSH, which was taken as the reference standard. A missed cancer-related death is considered of greater importance. The data are summarized in Table 3.

The second example is from Jiménez-Navarro et al. [40] who studied the concordance of oral glucose tolerance test (OGTT), proposed by the World Health Organization, in the diagnosis of diabetes mellitus (DM) which may possibly lead to future epidemic of coronary disease. OGTT classifies a diagnosis result as DM, glucose intolerance, or normal. Eighty-eight patients admitted with an acute coronary syndrome who had no previous diagnosis of DM underwent percutaneous coronary revascularization and received OGTT the day after revascularization and one month after that. Researchers are interested in the reproducibility of OGTT performed at these two different time points. The data are summarized in Table 4.

We calculated all the four agreement coefficients as well as their corresponding 95% confidence intervals (CI). The 95% CI for Cohen's weighted kappa was based on the large sample variance derived by Fleiss et al. [31]. 5000 bootstrap replicates were used in the calculation of the CI for the other three coefficients. Table 5 gives the results for the above two examples. In all the calculations, the Cicchetti-Allison weights were used (see (2.9)).

As can be seen from Table 5, there seems to be no substantial difference among all the four coefficients using the aforementioned interpretation rule. For example 1, all coefficients indicate that there is very strong agreement between OCR and MSH studies in classifying the cause of death for the 1648 cancer patients. The second example, however, indicates that OGTT, when taken at two different time points with a span of one month, has very poor reproducibility in the diagnosis of DM.

5. Conclusion

We developed a fixed-effects modeling of Cohen's kappa and weighted kappa for bivariate multinomial data through the perspective of generalized inverse, which provides an alternative form of framework to the one proposed by Yang and Chinchilli [33]. The proposed method also allows the application of different matrix functions, as long as they satisfy certain conditions. When there are only two categories in the response, the coefficient reduces to Cohen's kappa. The proposed method is new and promising in that it allows the use of other matrix functions, such as the largest eigenvalue of a matrix, that might also yield reasonable, and possibly better, result. Exploration of such matrix functions is our future research topic.

Appendix

Proof of Lemma 2.3

Proof. (1) When $k = 1$, the Cicchetti-Allison weights become an identity matrix. Therefore,

$$
\begin{aligned}
\kappa_{\mathrm{tr}}(W) &= 1 - \frac{\mathrm{tr}(WP_D)}{\mathrm{tr}(WP_I)} \\
&= 1 - \frac{\mathrm{tr}(P_D)}{\mathrm{tr}(P_I)} \\
&= \kappa.
\end{aligned}
\tag{A.1}
$$

(2) Since the weighting matrix is a 2×2 matrix, the two eigenvalues of P_D can be denoted by 0 and λ_1. Therefore, the largest eigenvalue is equal to $\lambda_1 = \mathrm{tr}(P_D)$. Similarly, it can be shown that the largest eigenvalue of P_I is also equal to $\mathrm{tr}(P_I)$. Thus, $\kappa_{\mathrm{tr}}(W) = \kappa_{\mathrm{le}}(W)$.

(3) From the previous proof, we know that $(P_I)^+ = E_I(\Lambda_I)^+ E_I^T$. When $k = 1$, (Λ_I) is

$$
(\Lambda_I) = \begin{bmatrix} 0 & 0 \\ 0 & \lambda_{I1} \end{bmatrix}
\tag{A.2}
$$

and $(\Lambda_I)^+$ is

$$
(\Lambda_I)^+ = \begin{bmatrix} 0 & 0 \\ 0 & \dfrac{1}{(\lambda_{I1})} \end{bmatrix}.
\tag{A.3}
$$

Because $k = 1$,

$$
\begin{aligned}
\kappa_{\mathrm{tr}}^*(W) &= 1 - \mathrm{tr}\big(P_D(P_I)^+\big) \\
&= 1 - \mathrm{tr}\left(P_D \frac{1}{\lambda_{I1}} e_{I1} e'_{I1}\right) \\
&= 1 - \frac{\mathrm{tr}\big(P_D e_{I1} e'_{I1}\big)}{\lambda_{I1}} \\
&= 1 - \frac{\mathrm{tr}\big(P_D e_{I1} e'_{I1}\big)}{\mathrm{tr}(P_I)},
\end{aligned}
\tag{A.4}
$$

where e_{I1} is the eigenvector of P_I associated with the largest eigenvalue λ_{I1} of P_I.

But for the case $k = 1$, because $e_{I1} = e_{D1} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$, e_{I1} also is the eigenvector of P_D associated with the largest eigenvalue λ_{D1} of P_D. Thus,

$$
\begin{aligned}
\kappa_{\mathrm{tr}}^*(W) &= 1 - \frac{\mathrm{tr}\big(P_D e_{D1} e'_{D1}\big)}{\mathrm{tr}(P_I)} \\
&= 1 - \frac{\mathrm{tr}\big(e'_{D1} P_D e_{D1}\big)}{\mathrm{tr}(P_I)} \\
&= 1 - \frac{\lambda_{D1}}{\mathrm{tr}(P_I)} \\
&= 1 - \frac{\mathrm{tr}(P_D)}{\mathrm{tr}(P_I)} \\
&= \kappa.
\end{aligned}
\tag{A.5}
$$

(4) From the derivation of $\kappa_{\mathrm{le}}^*(W)$, where $W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, we know that it can be written as

$$
\kappa_{\mathrm{le}}^*(W) = 1 - \frac{\text{largest eigenvalue of } \left[\big((P_I)^+\big)^{1/2} P_D \big((P_I)^+\big)^{1/2}\right]}{\text{largest eigenvalue of } \left[I - (1/2) 1 \cdot 1^T\right]}.
\tag{A.6}
$$

But the largest eigenvalue for the case $k = 1$ is equal to the trace, so $\kappa_{\mathrm{le}}^*(W) = \kappa_{\mathrm{tr}}^*(W) = \kappa.$ □

References

[1] S. M. Gregoire, U. J. Chaudhary, M. M. Brown et al., "The Microbleed Anatomical Rating Scale (MARS): reliability of a tool to map brain microbleeds," *Neurology*, vol. 73, no. 21, pp. 1759–1766, 2009.

[2] A. Riaz, F. H. Miller, L. M. Kulik et al., "Imaging response in the primary index lesion and clinical outcomes following transarterial locoregional therapy for hepatocellular carcinoma," *Journal of the American Medical Association*, vol. 303, no. 11, pp. 1062–1069, 2010.

[3] J. Johnson and J. A. Kline, "Intraobserver and interobserver agreement of the interpretation of pediatric chest radiographs," *Emergency Radiology*, vol. 17, no. 4, pp. 285–290, 2010.

[4] J. F. Hamel, D. Foucaud, and S. Fanello, "Comparison of the automated oscillometric method with the gold standard doppler ultrasound method to access the ankle-brachial pressure index," *Angiology*, vol. 61, no. 5, pp. 487–491, 2010.

[5] K. Y. Bilimoria, M. S. Talamonti, J. S. Tomlinson et al., "Prognostic score predicting survival after resection of pancreatic neuroendocrine tumors: analysis of 3851 patients," *Annals of surgery*, vol. 247, no. 3, pp. 490–500, 2008.

[6] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[7] J. Cohen, "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.

[8] L. I. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[9] L. I. K. Lin, "Assay validation using the concordance correlation coefficient," *Biometrics*, vol. 48, no. 2, pp. 599–604, 1992.

[10] P. B. Barbosa, M. M. Franco, F. O. Souza, F. I. Antônio, T. Montezuma, and C. H. J. Ferreira, "Comparison between measurements obtained with three different perineometers," *Clinics*, vol. 64, no. 6, pp. 527–533, 2009.

[11] X. C. Wang, P. Y. Gao, J. Xue, G. R. Liu, and L. Ma, "Identification of infarct core and penumbra in acute stroke using CT perfusion source images," *American Journal of Neuroradiology*, vol. 31, no. 1, pp. 34–39, 2010.

[12] D. V. Cicchetti and A. R. Feinstein, "High agreement but low kappa: II. Resolving the paradoxes," *Journal of Clinical Epidemiology*, vol. 43, no. 6, pp. 551–558, 1990.

[13] A. R. Feinstein and D. V. Cicchetti, "High agreement but low kappa: I. the problems of two paradoxes," *Journal of Clinical Epidemiology*, vol. 43, no. 6, pp. 543–549, 1990.

[14] A. C. Justice, J. A. Berlin, S. W. Fletcher, R. H. Fletcher, and S. N. Goodman, "Do readers and peer reviewers agree on manuscript quality?" *Journal of the American Medical Association*, vol. 272, no. 2, pp. 117–119, 1994.

[15] E. Kerosuo and D. Ørstavik, "Application of computerised image analysis to monitoring endodontic therapy: Reproducibility and comparison with visual assessment," *Dentomaxillofacial Radiology*, vol. 26, no. 2, pp. 79–84, 1997.

[16] S. Landers, W. Bekheet, and L. C. Falls, "Cohen's weighted kappa statistic in quality control-quality assurance procedures: application to network-level contract pavement surface condition surveys in British Columbia, Canada," *Transportation Research Record*, no. 1860, pp. 103–108, 2003.

[17] S. Suebnukarn, S. Ngamboonsirisingh, and A. Rattanabanlang, "A systematic evaluation of the quality of meta-analyses in endodontics," *Journal of Endodontics*, vol. 36, no. 4, pp. 602–608, 2010.

[18] A. Agresti, "A model for agreement between ratings on an ordinal scale," *Biometrics*, vol. 44, no. 2, pp. 539–548, 1988.

[19] W. Barlow, "Measurement of interrater agreement with adjustment for covariates," *Biometrics*, vol. 52, no. 2, pp. 695–702, 1996.

[20] L. D. Broemeling, *Bayesian methods for measures of agreement*, Chapman & Hall/CRC, Boca Raton, Fla, USA, 2009.

[21] J. L. Carrasco and L. Jover, "Concordance correlation coefficient applied to discrete data," *Statistics in Medicine*, vol. 24, no. 24, pp. 4021–4034, 2005.

[22] W. Dou, Y. Ren, Q. Wu et al., "Fuzzy kappa for the agreement measure of fuzzy classifications," *Neurocomputing*, vol. 70, no. 4–6, pp. 726–734, 2007.

[23] T. R. Fanshawe, A. G. Lynch, I. O. Ellis, A. R. Green, and R. Hanka, "Assessing agreement between multiple raters with missing rating information, applied to breast cancer tumour grading," *PLoS ONE*, vol. 3, no. 8, Article ID e2925, 2008.

[24] P. Graham and R. Jackson, "The analysis of ordinal agreement data: Beyond weighted kappa," *Journal of Clinical Epidemiology*, vol. 46, no. 9, pp. 1055–1062, 1993.

[25] T. S. King and V. M. Chinchilli, "A generalized concordance correlation coefficient for continuous and categorical data," *Statistics in Medicine*, vol. 20, no. 14, pp. 2131–2147, 2001.

[26] H. C. Kraemer, "Measurement of reliability for categorical data in medical research," *Statistical methods in medical research*, vol. 1, no. 2, pp. 183–199, 1992.

[27] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[28] R. T. S. Laurent, "Evaluating agreement with a gold standard in method comparison studies," *Biometrics*, vol. 54, no. 2, pp. 537–545, 1998.

[29] L. Lin, A. S. Hedayat, and W. Wu, "A unified approach for assessing agreement for continuous and categorical data," *Journal of Biopharmaceutical Statistics*, vol. 17, no. 4, pp. 629–652, 2007.

[30] E. Svensson, "A coefficient of agreement adjusted for bias in paired ordered categorical data," *Biometrical Journal*, vol. 39, no. 6, pp. 643–657, 1997.

[31] J. L. Fleiss, J. Cohen, and B. S. Everitt, "Large sample standard errors of kappa and weighted kappa," *Psychological Bulletin*, vol. 72, no. 5, pp. 323–327, 1969.

[32] J. Yang and V. M. Chinchilli, "Fixed-effects modeling of cohen's kappa for bivariate multinomial data," *Communications in Statistics—Theory and Methods*, vol. 38, no. 20, pp. 3634–3653, 2009.

[33] J. Yang and V. M. Chinchilli, "Fixed-effects modeling of Cohen's weighted kappa for bivariate multinomial data," *Computational Statistics and Data Analysis*, vol. 55, no. 2, pp. 1061–1070, 2011.

[34] E. H. Moore, "On the reciprocal of the general algebraic matrix," *Bulletin of the American Mathematical Society*, vol. 26, pp. 394–395, 1920.

[35] R. A. Penrose, "A generalized inverse for matrices," *Proceedings of the Cambridge Philosophical Society*, vol. 51, pp. 406–413, 1955.

[36] E. L. Lehmann, *Theory of Point Estimation*, Springer, New York, NY, USA, 1998.

[37] D. V. Cicchetti and T. Allison, "A new procedure for assessing reliability of scoring EEG sleep recordings," *American Journal of EEG Technology*, vol. 11, pp. 101–109, 1971.

[38] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and Psychological Measurement*, vol. 33, pp. 613–619, 1973.

[39] D. R. Brenner, M. C. Tammemägi, S. B. Bull, D. Pinnaduwaje, and I. L. Andrulis, "Using cancer registry data: agreement in cause-of-death data between the Ontario Cancer Registry and a longitudinal study of breast cancer patients," *Chronic Diseases in Canada*, vol. 30, no. 1, pp. 16–19, 2009.

[40] M. F. Jiménez-Navarro, J. M. Garcia-Pinilla, L. Garrido-Sanchez et al., "Poor reproducibility of the oral glucose tolerance test in the diagnosis of diabetes during percutaneous coronary intervention," *International Journal of Cardiology*, vol. 142, no. 3, pp. 245–249, 2010.

Advances in
Operations Research

Advances in
Decision Sciences

Mathematical Problems
in Engineering

Algebra

Journal of
Probability and Statistics

The Scientific
World Journal

International Journal of
Differential Equations

International Journal of
Combinatorics

Advances in
Mathematical Physics

Hindawi

Submit your manuscripts at
http://www.hindawi.com

Journal of
Complex Analysis

Journal of
Mathematics

International Journal of
Stochastic Analysis

Abstract and
Applied Analysis

Discrete Dynamics in
Nature and Society

International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces

Journal of
Applied Mathematics

Journal of
Optimization