The Korean Communications
in Statistics Vol. 6, No. 2, 1999
pp. 367-373

Determination of the Number of Components in Spectroscopy from the Multilinear Model Fitting[1)]

Choongrak Kim[2)], Byung-Chull Chung[3)], and Choon-Hwan Lee[4)]

Abstract

Biological specimens contain several components, and multilinear models are very useful in analyzing these data. After fitting the model, the number of components are determined by the change of mean squared error, however, this method is quite rule of thumb. In this paper, we suggest a measure to decide the number of components based on the relative change of the mean squared error. Simulations are done and applications to real data set are given as illustrations.

1. Introduction

Biological specimens often contain multiple components with overlapping fluorescene spectra which cannot be physically separated without seriously altering important properties of the native specimens. A protein in which the fluorophores are the amino acids tyrosine and tryptophan is a good example. To deal with these kinds of specimen spectroscopy is a very useful tool.

Spectroscopy is the measurement of the absorption of particles by a specimen, or the emission of particles from a specimen, as a function of the energy of particles. The amount of absorption or emission as a function of particle energy is known as a spectrum.

While the primary independent variable of spectroscopy is the energy or wavelength of particles absorbed or emitted, an experiment may involve additional independent variables. With multiple chromophores f and wavelength i and differing circumstances j in which concentrations of the chromophores vary, we have the bilinear equation

$$\mu_{ij} = \sum_{f=1}^{F} \alpha_{if} \beta_{jf} ,$$

where α_{if} is the extinction coefficient of chromophores f at wavelength λ_i and β_{jf} is the

1) This work was supported by a grant from Cray R & D (1992-1993)
2) Department of Statistics, Pusan National University,Pusan, KOREA (609-735)
3) Department of Molecular Biology, Pusan National University, Pusan, KOREA (609-735)
4) Department of Molecular Biology, Pusan National University, Pusan, KOREA (609-735)

concentration of f in circumstances j. Note that μ_{ij} is the mean of the response y_{ij}, and F is the number of components which is unknown. Therefore, to fit the bilinear model, we first assume F is fixed and repeat the fitting process at various F. The usual way to estimate F so far is comparing the residual sum of squares, and this is very important issue in this area.

The amount of light emission measured is separately linear in the number of photons absorbed and in the fraction of photons absorbed that lead to emission at wavelength λ. With multiple chromophores, we then have the trilinear equation

$$\mu_{ijk} = \sum_{f=1}^{F} \alpha_{if}\beta_{jf}\gamma_{kf} ,$$

where γ_{kf} is the concentration of chromophore f in circumstance k, α_{if} is the relative absorption cross-section of chromophore f at wavelength λ_i and β_{jf} is the relative emission at detection wavelength λ_j.

Based on the same idea, the bilinear and trilinear equation can be extended to the quadrilinear model. There are over a hundred publications in the multilinear models, and the best reference seems to be Leugans and Ross (1992). For the computations of the multilinear models, see Lee, et al (1997).

After fitting the multilinear model, the analyst wants to estimate the number of components F in the system. The most frequently quoted statistic to do this is the square root of the residual sum of squares, and the estimation of the number of components is based on the graphical display of RMS_F, $F = 1, 2, \cdots$, and it is clear that this decision is very subjective.

In this paper we suggest a measure to estimate the number of components F. A measure to estimate F is suggested in Section 2, and simulation results are given in Section 3. An illustrative example based on a real data set is given in Section 4.

2. A Suggested Measure

Let $e_{iF} = y_{iF} - \hat{y}_{iF}$ be the residual of the ith observation when the assumed model has F components, $SSE_F = \sum e_{iF}^2$ be the residual sum of squares, df_F be the corresponding degree of freedoms, and define $RMS_F = \sqrt{SSE_F/df_F}$, $F = 1, 2, 3, \cdots$ be the square root mean squares for the multilinear model with F components. As the number of components F increases, RMS_F decreases. As shown in Figure 1, estimation of F based on the absolute decrease of RMS_F is not satisfactory, i.e., it is very hard to estimate F based on RHS_F. Therefore, estimation of F is very subjective. For the case of Figure 1, it is very

hard to determine F. Instead we propose a relative decrease of RMS_F to estimate F, and in fact this idea was already used by Kim and Storer(1996) in regression diagnostics area. To be more specific, let

$$\delta_F = RMS_F - RMS_{F+1}, \quad F = 1, 2, 3, \cdots$$

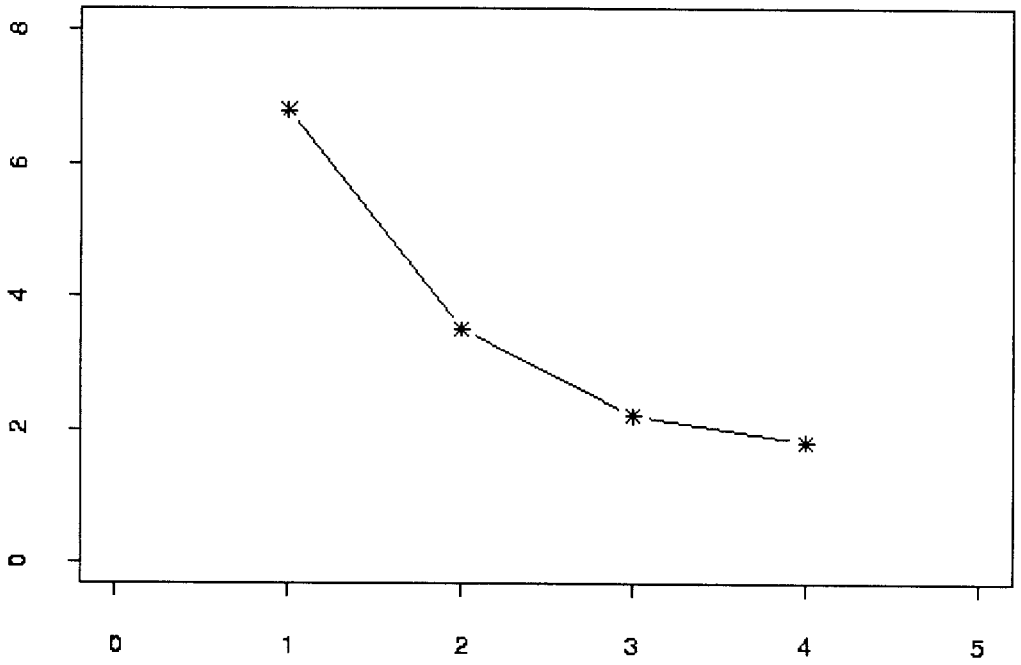
be the absolute decrease, and let

$$RR_F = \delta_{F+1} / \delta_F, \quad F = 1, 2, 3, \cdots$$

be the relative decrease. Finally, as an estimator of F, we suggest

$$\hat{F} = \min_{F \in N} RR_F + 1$$

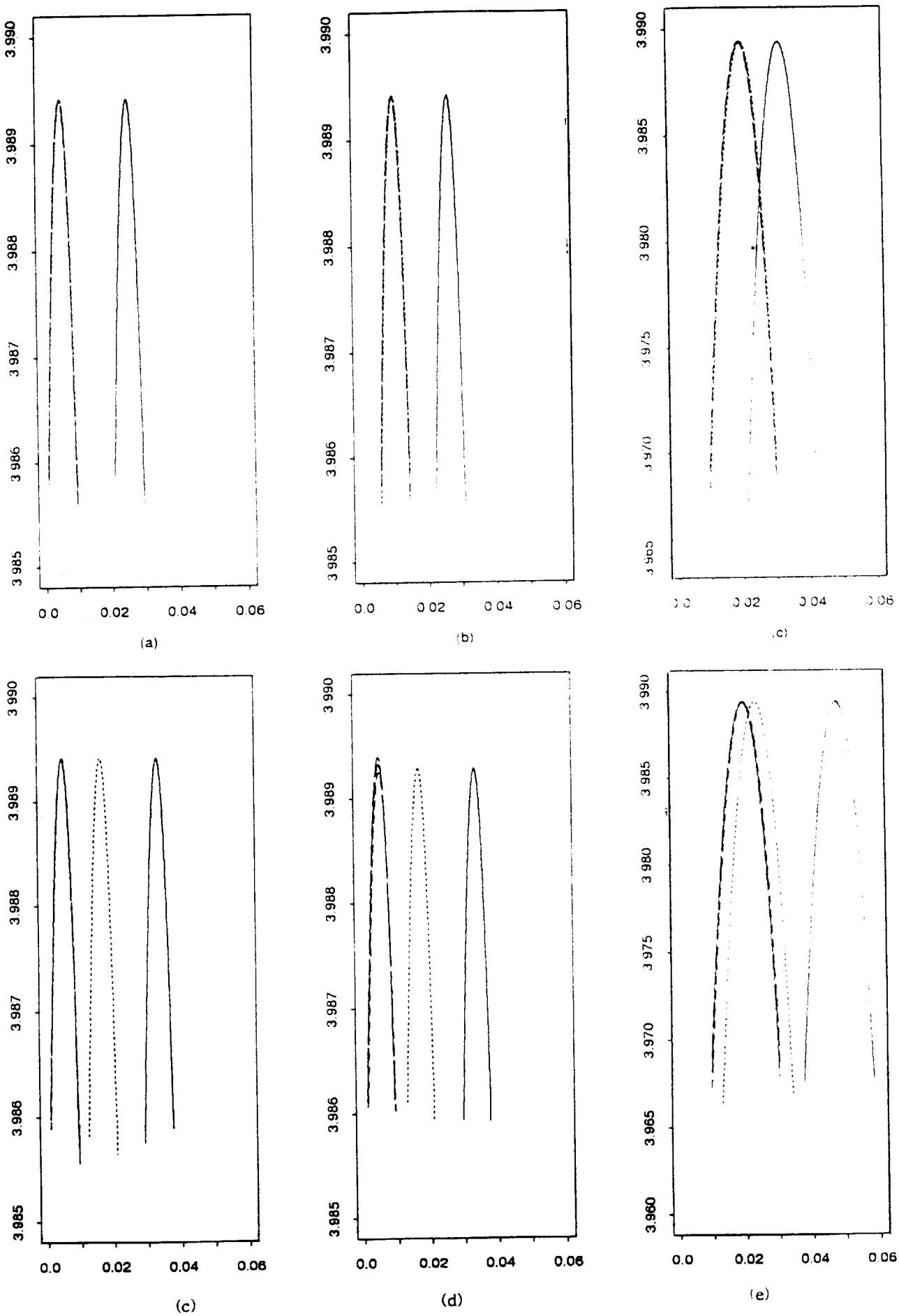
where N denotes set of positive integers. Therefore, $\hat{F}$ gives the point where the minimum relative decrease occurs.



< Figure 1 >

3. Simulation Results

To see the behavior of the suggested measure we perform Monte Carlo studies. When the true number of components is 2, trilinear model is used for F=1,2,3,4 and Gaussian errors are generated with mean 0 and σ = .02, .03, and .04, and 100 replications are done(see Figure 2(a), (b), (c)). Also, we perform trilinear model for F=1,2,3,4,5 when the true number of



< Figure 2 >

Table 1. Simulation results for the number of components F=2 and 3

true F	σ	F	*RMS*	*RR*	$\hat{F}$
2	.005	1	.2493		2
		2	.00500	.00301	
		3	.00494	.83333	
		4	.00489		
	.01	1	.02641		2
		2	.00999	.00609	
		3	.00989	1.10001	
		4	.00978		
	.02	1	.03153		2
		2	.01999	.01820	
		3	.01978	1.00000	
		4	.01957		
3	0.05	1	.03329		3
		2	.01645	.67933	
		3	.00501	.00524	
		4	.00495	.83333	
		5	.00490		
	0.01	1	.03326		3
		2	.01641	.67537	
		3	.00503	.00439	
		4	.00498	1.00000	
		5	.00493		
	0.02	1	.04797		3
		2	.02346	.13953	
		3	.02004	.06140	
		4	.01983	1.04762	
		5	.01961		

Table 2. PARAFAC 3 analysis of fluorescene from a three-dye mixture

F	*RMS*	*RR*	$\hat{F}$
1	.10850		
2	.03074	.37680	
3	.00144	.01900	3
4	.00088	.25080	
5	.00076		

components is 3 with the same Gaussian errors and replications(see Figure 2(d), (e), (f)). Figure 2 shows the kernel density estimation(see Silverman(1985) for details) for 100 RMS_F for F=1,2,3,4 when true F=2, and F=1,2,3,4,5 when true F=3, respectively. In fact, the shape of distributions are very similar except their means. The means or modes of distributions decrease as F increases. In Figure 2(a), (b), (c) RMS_1 is much larger than others, so that it is clear F=2. However, in Figure 2(e), RMS_2 is slightly larger than RMS_F, F=3,4,5. In this case it is not easy to conclude that F=3.

Let us use $\hat{F}$ suggested in Section 2, then as described in Table 1, $\hat{F}$ successfully find the true F. Now, we apply to the real data set where true F is known. The data are obtained by the fluorescene from a three-dye mixture, i.e., F=3(see Lee et al.(1991) for details). We fit the data to trilinear model and the results are summarized in Table 2. Again, $\hat{F}$ successfully find the true F.

4. Example

As an illustrative example, we use the data from the fluorescene from the three chromatographic eluents(see Lee(1992) for details). The number of component F is unknown and we try to estimate it by $\hat{F}$. Results are summarized in Table 3 after fitting to the trilinear model. Here we estimate F by 2.

Table 3. PARAFAC 3 analysis of fluorescene from three chromatographic eluents

F	RMS	RR	$\hat{F}$
1	.02280		
2	.00182	.01620	
3	.00148	1.60882	2
4	.00093	.29616	
5	.00077		

Figure Legends

Figure 1. Root Mean Squared Errors (RMS_F) for F=1, 2, 3, 4 based on artificial data.

$N(0, \sigma^2)$. (a), (b), and (c) correspond to the case of true F=2, and (d), (e), and (f) correspond to the case of true F=3. Also, σ=.005 for (a) and (d), = σ.01 for (b) and (e), and σ=.02 for (c) and (f). ------ (F=1), ····· (F=2), - - - - (F=3), -- -- -- (F=4), --- --- (F=5).

References

[1] Kim, C. and Storer, B.E. (1996) Reference values for Cook's distance, *Communications in Statistics - Simulations and Computations*, 25, 691-708.

[2] Lee, C.H. (1992) Trilinear analysis of fluorescene spectra of chromatographic eluents. *Journal of Science, Pusan National University*, 53, 183-193.

[3] Lee, C.H., Kim, K., and Ross, R.T. (1991) Trilinear analysis for the resolution of overlapping fluorescene spectra. *Korean Biochemistry journal*, 24, 374-379.

[4] Lee, C.H., Ezzeddine, B.M., Kim, C., and Ross, R.T. (1997) An algorithm for weightes bilinear regression. *Communications in Statistics - Simulations and Computations*, 26, 791-804.

[5] Lee, C.H., Kim, K., and Ross, R.T. (1991) Trilinear analysis for the resolution of overlapping flurescence spectra. *Korean Biochemistry journal*, 24, 374-379.

[6] Leugans, S. and Ross, R.T. (1992) Multilinear models : Applications in spectroscopy, *Statistical Science*, 7, 289-319.

[7] Silverman, B.W. (1985) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.