

Wavelength selection by loading vector analysis in determining total protein in human serum using near-infrared spectroscopy and Partial Least Squares Regression

YOEN-JOO KIM*, GILWON YOON

Samsung Advanced Institute of Technology, P. O. Box 111, Suwon 440-600, Korea

In multivariate analysis, absorbance spectrum is measured over a band of wavelengths. One does not often pay attention to the size of this wavelength band. However, it is desirable that spectrum is measured at only necessary wavelengths as long as the acceptable accuracy of prediction can be met. In this paper, the method of selecting an optimal band of wavelengths based on the loading vector analysis was proposed and applied for determining total protein in human serum using near-infrared transmission spectroscopy and PLSR. Loading vectors in the full spectrum PLSR were used as reference in selecting wavelengths, but only the first loading vector was used since it explains the spectrum best. Absorbance spectra of sera from 97 outpatients were measured at 1530~1850 nm with an interval of 2 nm. Total protein concentrations of sera were ranged from 5.1 to 7.7 g/dl. Spectra were measured by Cary 5E spectrophotometer (Varian, Australia). Serum in the 5 mm-pathlength cuvette was put in the sample beam and air in the reference beam. Full spectrum PLSR was applied to determine total protein from sera. Next, the wavelength region of 1672~1754 nm was selected based on the first loading vector analysis. Standard Error of Cross Validation (SECV) of full spectrum (1530~1850 nm) PLSR and selected wavelength PLSR (1672~1754 nm) was respectively 0.28 and 0.27 g/dl. The prediction accuracy between the two bands was equal. Wavelength selection based on loading vector in PLSR seemed to be simple and robust in comparison to other methods based on correlation plot, regression vector and genetic algorithm. As a reference of wavelength selection for PLSR, the loading vector has the advantage over the correlation plot since the former is based on multivariate model whereas the latter, on univariate model. Wavelength selection by the first loading vector analysis requires shorter computation time than that by genetic algorithm and needs not smoothing.