Application of Variable Selection for Prediction of Target Concentration

Seonwoo Kim,† Yoen-Joo Kim, Jong-Won Kim,‡ and Gilwon Yoon*

†*Clinical Research Center, Samsung Biomedical Research Institute, Seoul 135-230, Korea*
Biomedical Engineering Research Center, Samsung Advanced Institute of Technology, Suwon 440-600, Korea
‡*Sungkyunkwan University, College of Medicine, Department of Clinical Pathology,*
Samsung Medical Center, Seoul 135-230, Korea
Received Octoeber 25, 1998

Many types of chemical data tend to be characterized by many measured variables on each of a few observations. In this situation, target concentration can be predicted using multivariate statistical modeling. However, it is necessary to use a few variables considering size and cost of instrumentation, for an example, for development of a portable biomedical instrument. This study presents, with a spectral data set of total hemoglobin in whole blood, the possibility that modeling using only a few variables can improve predictability compared to modeling using all of the variables. Predictability from the model using three wavelengths selected from all possible regression method was improved, compared to the model using whole spectra (whole spectra: SEP = 0.4 g/dL, 3-wavelengths: SEP=0.3 g/dL). It appears that the proper selection of variables can be more effective than using whole spectra for determining the hemoglobin concentration in whole blood.

Introduction

Regression has a goal to model the predictive relationships between a dependent variable (e.g. analyte concentration) and the measured independent variables (e.g. spectral data) from a set of training observations on which all of the variables have been measured. The most popular regression method is Partial Least Squares Regression (PLSR),[1-5] to a somewhat lesser extent, Principal Component Regression (PCR),[1,6] ridge regression,[7-9] and ordinary least squares regression. These methods have their own strengths and weaknesses.[2,10]

Many types of chemical data from organic and analytical chemistry, food research, and environmental studies tend to be characterized by many measured variables with a few observations. Often the number of such variables greatly exceeds the observation counts. In this situation, statistical multivariate quantitative methods, such as PLSR and PCR are necessary to analyze full spectra or a specific range of spectra.

In practice, it becomes to critically consider to select only an appropriate subset when miniaturization or cost of instrument is an important issue. For example, for the development of a portable biomedical instrument, it is better to use a few diodes as a source without monocromator to decrease the size and cost of instrumentation. In this case, it is essentially required to select a few variables. Moreover, the model using a few variables should predict target concentration as good as the model using all of the variables. The predictability of model depends on the types of calibration method, data preprocessing as well as the set of variables. Optimal data preprocessing removes noise and redundant information on data, which can improve predictability. Various methods of removing scattering, adjusting baseline variations, and enhancing the spectral features belong to data preprocessing method.[11-19]

The predictability of model using a few variables has been evaluated for the determination of hemoglobin concentration in blood in the study. There are many different ways for selection of variables,[20-26] however all possible regression method which evaluates all possible subsets produces the mathematically optimal result.

Therefore, in this study, using all possible regression method three variables were selected and PLSR was used as a method of using all variables. The predictability of the model using the selected variables was compared to model using all of the variables. The results present the modeling using a few variables is as good as modeling using all of the variables in the data.

Experimental Section

Whole blood specimens that contained EDTA for anticoagulant were obtained from 95 individual out-patients. The total hemoglobin concentrations were determined by a hemiglobinocyanide method (HiCN) with a SE8000 (Sysmex, Kobe, Japan). With this method, blood is diluted in a solution of potassium ferricyanide and potassium cyanide. The absorbance of the solution is measured at 540 nm and then compared to that of a standard solution. The resulting distribution of total hemoglobin concentrations ranged from 6.6 to 17.2 g/dL. The prepared blood samples were rolled and rotated with Hematology Mixer (Fisher Scientific) for consistently homogeneous suspension. The absorbance spectra, $-\log(I/I_0)$, of the 500-900 nm region were collected with a Cary5E spectrophotometer (Varian, Melbourne, Australia), which equipped with a R928 PMT detector and a visible lamp. A whole blood sample (26 μL) was placed into SUPRASIL cuvette (Hellma, Rijswijk, Netherlands) with a pathlength of 0.1 mm that had detachable windows. The spectral bandwidth was 2 nm with spectral collection at 1 nm interval. Two scans were averaged for each sample and

air was used as a reference. Each measurement was accomplished by double beam mode. Acquisition of a single scan took 27 s. Samples were scanned without temperature control.

Data Analysis

We used PLSR with two different ways. At first, PLSR model was developed using the whole range of spectra and predicted the hemoglobin concentrations. At second, PLSR model was developed using only the selected wavelengths. For this purpose, all possible regression method was used to select variables. In either case, no data preprocessing method was used.

PLSR. PLSR was introduced by Wold[4] and has been greatly promoted in the statistical analysis of chemical data as an alternative to ordinary least squares regression in ill-conditioned problems.[10] PLSR is a factor-based multivariate calibration method whose properties and merits were presented previously.[2,27-28] During calibration it is important to determine the appropriate number of PLSR factors to retain in the calibration model. Using too few factors can miss important spectral information while using too many factors draws too much noise into a calibration model. A common method to determine an optimal number of factors is cross validation. The prediction set, which are not included in the calibration set but are similar to the range of analyte concentrations and sample variability, is used to predict the quantity of analyte using an obtained PLSR model. The prediction result is compared with the reference analyte concentration. The Standard Error of Prediction (SEP) is mainly used to judge the predictability of a model.

All possible regression method. This method fits all possible regression models to the spectral data. If n variables are available in the data and p variables among them are selected, the nCp models are possible. Prior to the advent of high-speed computer, such calculation was out of the question when involves more than a few variables. The availability of rapid computation has inspired efforts in this direction and various efficient algorithms for evaluating all possible regressions[29-32] are now available. However, it still requires huge computing time, especially when the number of variables to select increases.

Application of PLSR to the spectroscopic diagnosis of total hemoglobin in whole blood. Seventy out of total 95 spectra from 95 individual patients were randomly chosen as the calibration set and the remained 25 spectra were used as the prediction set. Figure 1 shows the average spectrum in calibration set and all spectra in prediction set. PLSR was performed with Pirouette 2.03 (InfoMetrix, Woodinville, WA, U. S. A.) software. PLSR calibration models were derived by using both the reference total hemoglobin concentration and the corresponding spectral data.

In the case of using whole range of spectra, the optimal number of factor was determined to eight by the one-in-one-out cross-validation and F-test. For the use of a few variables, the 3-factor PLSR model was used to each subset of

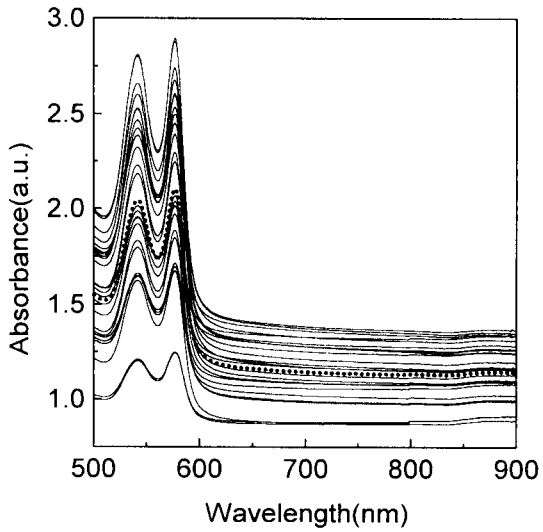

Figure 1. Average spectrum in calibration set (dotted line) and all spectra in prediction set. Whole blood samples from 95 patients were prepared, where the total hemoglobin concentrations were varied between 6.6 and 17.2 g/dL.

three wavelengths. By each model, the total hemoglobin concentrations were predicted and the SEP was estimated to evaluate the predictability. To search the optimal subset of wavelengths, the 10,666,600, $_{401}C_3$, computations were performed by all possible regression method. The 560, 826, 856 nm were determined as the best optimal three wavelengths with the minimum SEP among all possible calibration models. The SEP and SEC (Standard Error of Calibration) from

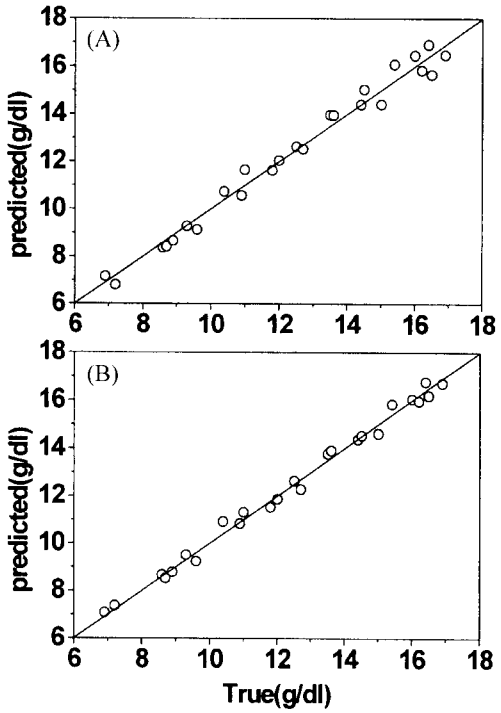

Figure 2. (A) Scatter plot showing correlation between reference data and NIR prediction data using whole range of spectra. The solid line corresponds to the ideal line. (B) Scatter plot showing correlation between reference data and NIR prediction data using 3-wavelengths (560, 826, 856 nm).

this result were 0.3 and 0.4 g/dL, respectively. Addtionally, the SEP and SEC from the 8-factor PLSR model using all of the wavelengths were 0.4 and 0.3 g/dL, respectively. This results show that the model using a few variables possibly gives the smaller SEP compared to the model developed with much more variables. Additionally, during all possible regression procedure, a large number of models using three wavelengths produced the value of SEP less than 0.4 g/dL. Scatter plots of the reference vs predicted concentrations are presented in Figure 2A (using all wavelengths) and Figure 2B (using 3 wavelengths).

Discussion

This study presents the possibility to develop a calibration model with the good predictability only using a few variables compared to that using all of the variables. This fact may be partly due to the redundancy of data in case of modeling with all of the variables, although some noise was removed using PLSR. Absorbances between wavelengths in the spectral data are usually very much correlated each other. Therefore, the predictability can be maintained or improved if a subset of variables is well selected.

We had applied mean centering, variance scaling, and auto scaling to whole range of spectra to enhance spectral features. The SEPs using mean centering with 4-factors and auto scaling with 3 factors were 0.4 and 0.4 g/dL, respectively. However the SEP using variance scaling with 4 factors was 0.5 g/dL. The predictability was maintained using mean centering or auto scaling, but it is not as much good as the predictability using three wavelengths.

It is still remained to a problem how many variables should be selected. There have already been several methods to give the solution to such a problem.[20,33-37] However, in practice, it depends on environment of experiment or instrument to develop. For the development of non-invasive home monitoring device for blood components, for example, the system without monochromator should be made. This helps to decrease the size and cost of instrumentation. One of the possible system is utilizing LED as a light source which emits only two or three wavelengths. In this case, the number of selected wavelengths should be limited to two or three.

References

1. Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley & Sons: New York, U. S. A., 1989.
2. Helland, I. S. *Communications in Statistics-Simulation and Computation* **1988**, *17*, 581.
3. Wold, H. In *Perspectives in Probability and Statistics*; Gani, J., Ed.; Academic Press: London, 1975.
4. Lorber, A.; Wangen, L. E.; Kowalski, B. R. *Journal of Chemometrics* **1987**, *1*, 19.
5. Stone, M.; Brooks, R. J. *Journal of Royal Statistical Society, Series B* **1990**, *52*, 237.
6. Massy, W. F. *Journal of American Statistical Association* **1965**, *60*, 234.
7. Hoerl, A. E.; Kennard, R. W. *Technometrics* **1970**, *8*, 27.
8. Hoerl, A. E.; Kennard, R. W. *Technometrics* **1970**, *12*, 55.
9. Hoerl, A. E.; Kennard, R. W. *Technometrics* **1970**, *12*, 69.
10. Frank, I. E.; Friedman, J. H. *Technometrics* **1993**, *35*, 109.
11. Geladi, P.; MacDougall, D.; Martens, H. *Applied Spectroscopy* **1985**, *39*, 491.
12. Barnes, B. J.; Dhanoa, M. S.; Lister, S. J. *Applied Spectroscopy* **1989**, *43*, 772.
13. *PLSplus/IQ User's Guide*; Galactic Industries Corporation: 1966.
14. Fuller, M. P.; Ritter, G. L.; Draper, C. S. *Applied Spectroscopy* **1988**, *42*, 217.
15. Norris, K. H.; Williams, P. C. *Cereal Chemistry* **1984**, *62*, 158.
16. Norris, K. H. In *Proceedings of the 1982 IUFST Symposium*; Applied Science Publishers: Oslo, 1983.
17. Savitsky, A.; Golay, M. J. E. *Analytical Chemistry* **1964**, *36*, 1627.
18. Steiner, J.; Termonia, Y.; Deltour, J. *Analytical Chemistry* **1972**, *44*, 1906.
19. Madden, H. M. *Analytical Chemistry* **1978**, *50*, 1383.
20. Gorman, J. W.; Toman, R. J. *Technometrics* **1966**, *8*, 27.
21. Draper, N. R.; Smith, H. *Applied Regression Analysis*; Wiley: New York, U. S. A., 1966.
22. Hocking, R. R.; Leslie, R. N. *Technometrics* **1967**, *9*, 531.
23. Beale, E. M. L.; Kendall, M. G.; Mann, D. W. *Biometrika* **1967**, *54*, 357.
24. LaMotte, L. R.; Hocking, R. R. *Technometrics* **1970**, *12*, 83.
25. Barr, A. J.; Goodnight, J. H. *Strategic Analysis System*; SAS Institute: 1971.
26. Furnival, G. M.; Wilson, R. W., Jr. *Technometrics* **1974**, *16*, 499.
27. Ham, F. M.; Kostanic, I. N.; Cohen, G. M.; Patel, K.; Gooch, B. R. *IEEE* **1997**, *44*, 475.
28. Beebe, K. R.; Kowalski, B. R. *Anal. Chem.* **1987**, *59*, 1007A.
29. Newton, R. G.; Spurrell, D. J. *Applied Statististics* **1967**, *16*, 51.
30. Newton, R. G.; Spurrell, D. J. *Applied Statististics* **1967**, *16*, 165.
31. Furnival, G. M. *Technometrics* **1971**, *13*, 403.
32. Morgan, J. A.; Tatar, J. F. *Technometrics* **1972**, *14*, 317.
33. Tukey, J. W. *Journal of Royal Statistical Society* **1967**, *29*, 47.
34. Rothman, D. *Technometrics* **1968**, *10*, 432.
35. Crocker, D. C. *The American Statistician* **1972**, *26*, 31.
36. Mallows, C. L. *Technometrics* **1973**, *15*, 661.
37. Stone, M. *Journal of Royal Statistical Society* **1974**, *36*, 111.