SAN096-1973C
CONF-960815--1

Insights into Multivariate Calibration Using Errors-in-Variables Modeling

Edward V. Thomas
Sandia National Laboratories
Albuquerque, NM 87185-0829
USA

Keywords: Causal model, chemometrics, maximum likelihood estimation, principal components regression

ABSTRACT

A q-vector of responses, y, is related to a p-vector of explanatory variables, x, through a causal linear model. In analytical chemistry, y and x might represent the spectrum and associated set of constituent concentrations of a multicomponent sample which are related through Beer's Law. The model parameters are estimated during a calibration process in which both x and y are available for a number of observations (samples/specimens) which are collectively referred to as the calibration set. For new observations, the fitted calibration model is then used as the basis for predicting the unknown values of the new x's (concentrations) from the associated new y's (spectra) in the prediction set. This prediction procedure can be viewed as parameter estimation in an errors-in-variables (EIV) framework. In addition to providing a basis for simultaneous inference about the new x's, consideration of the EIV framework yields a number of insights relating to the design and execution of calibration studies. A particularly interesting result is that predictions of the new x's for individual samples can be improved by using seemingly unrelated information contained in the y's from the other members of the prediction set. Furthermore, motivated by this EIV analysis, this result can be extended beyond the causal modeling context to a broader range of applications of multivariate calibration which involve the use of principal components regression.

This work was supported by the United States Department of Energy under Contract DE-AC04-94AL85000.

1. *INTRODUCTION*

Recently, calibration involving a multidimensional response has received significant attention. Two significant references covering multivariate calibration are: Martens and Naes (1989) from a practical perspective, and Brown (1993) from a theoretical viewpoint. Multivariate calibration is used in a number of applications, often in analytical chemistry. Many of these applications involve specialized instrumentation (e.g., spectrometers) that result in response variables of a very high dimension (e.g., spectra). The discussion presented here is particularly relevant for this situation. Martens and Naes (1989) and references therein provide a number of relevant examples.

Specifically in the case of spectroscopic applications, measurements such as absorbance and reflectance, are taken at one or more spectral wavelengths. These measurements are obtained from a number of specimens in which the amount of the analyte of interest has been determined by some independent assay (e.g., wet chemistry). Together, the spectral measurements and results from the independent assays are used to construct a model that relates the amount of the analyte of interest to the spectral measurements. This step is referred to as calibration. Then in the prediction step, the model is used to predict analyte concentrations of future samples solely on the basis of the spectral measurements. Quantitative analysis based on spectral data has advantages over some of the more traditional methods of analysis because it is often much quicker, less labor intensive (hence cheaper), and can be nondestructive/noninvasive.

The basis for the many calibration models using absorbance spectroscopy is Beer's Law. That is, in the limiting case of dilute component concentrations in a nonabsorbing medium, the absorbance of a sample at wavelength λ, $y(\lambda)$, depends upon the concentration of the multiple (p) chemical species in the sample through the equation

$$y(\lambda) = a_\lambda + x_1 k_1(\lambda) + x_2 k_2(\lambda) + \ldots + x_p k_p(\lambda) + \varepsilon_\lambda, \qquad (1)$$

where a_λ is a spectral intercept (or baseline), x_i is the concentration of the i^{th} chemical species, k_i is the product of the optical pathlength with the absorptivity of the i^{th} chemical species, and ε_λ is the measurement error of the absorbance at wavelength λ.

Because of the cost of developing a calibration model, a single fitted calibration model is often used repeatedly on a number of new observations to predict some characteristic of interest. For example, in analytical chemistry a calibration model is often constructed to predict some characteristic (e.g., chemical concentration) in a batch or stream comprising a number of new specimens (prediction set). A natural ordering in this stream can be developed from the sequence in which the specimens are measured by the appropriate analytical instrument (e.g., spectrometer). Prediction follows the acquisition of the instrumental measurements. Traditionally, predictions are developed sequentially for each new specimen. It is also traditional that each prediction is developed through consideration of only the model and the instrumental measurements associated with that single new specimen. Thus, instrumental measurements from other new specimens, a source of information regarding the pattern of variation among the various instrumental measurements, are ignored.

The model presented in eqn. 1 is an example of an idealized causal model in which all deterministic factors are explicitly accounted for. The successful use of this or any other causal model depends on a complete knowledge of the factors that affect the response, y. This requirement greatly restricts the applicability of such models in complex chemical problems where, often, only an incomplete knowledge of the causal factors is available. However, when a causal model is applicable, an analytical assessment of the uncertainty of predictions based on the model is possible. Furthermore, as will be demonstrated in this paper, insights developed from the causal situation may be extended to more commonly used soft-modeling, empirically-based calibration methods such as principal components regression (PCR).

In the soft-modeling approach, the characteristic of interest (e.g., analyte concentration) is modeled directly as a linear combination of the instrumental measurement(s). For example in spectroscopic applications,

$$x = b_0 + b_1 y(\lambda_1) + b_2 y(\lambda_2) + \ldots + b_q y(\lambda_q) + \varepsilon, \tag{2}$$

where, x denotes the concentration of the analyte of interest , $y(\lambda_q)$ is the measured intensity of the q^{th} wavelength, and the b_i are parameters. In contrast to methods which are based on an explicit causal model, these methods do not require a complete knowledge of the chemical system involved and are thus applicable to a great variety of problems in science and industry. For example, the various soft-modeling approaches are used in combination with spectral data to estimate the amount of protein in wheat (e.g., see Fearn 1983), composition of thin films used in semiconductor manufacturing (e.g., see Haaland 1988), and in various medical applications such as the noninvasive measurement of blood components (e.g., see Robinson et al. 1992).

The remainder of this paper consists of the following. Section 2 explains how, in the causal modeling context, the prediction procedure can be viewed as parameter estimation in an EIV setting. Given this EIV analogy (and by assuming normally distributed measurement errors), standard results from the EIV literature are used in Section 3 to obtain the joint maximum likelihood estimates (MLE's) of the new x's. It is observed that this joint estimation procedure, which uses the responses (e.g., spectral measurements) from all available members of the prediction set, can outperform an analogous individual estimation procedure which uses only the response from the current member of the prediction set. Motivated by this result, in Section 4 it is demonstrated how the concept of improving the prediction of a single new specimen by using the instrumental measurements from the other new specimens can be extended to a soft-modeling approach using PCR. A short conclusion follows.

2. *EIV CONNECTION IN CAUSAL MODELING CONTEXT*

In the case of a multivariate linear causal model, the calibration step consists of estimating the $(p+1)$ by q parameter matrix B_0 in the model $[1; X_0] B_0 = Y_0$, where Y_0 is the n by q matrix of the true, but unobservable responses and X_0 denotes the n by p matrix of causal factors (see Thomas 1991). The observables, $Y = Y_0 + \Delta Y$ and $X = X_0$ are used to estimate B_0 (note that it is assumed $\Delta X = 0$). Here, q is the number of response variables (e.g., wavelengths), n is the number of calibration samples, and p is the number of explanatory variables (e.g., chemical components). Note that this model represents controlled calibration, as X is assumed to be fixed (see Brown 1982). For notational simplicity it is also assumed that the columns of X have mean zero. Further, it will be assumed that the matrix of random errors, ΔY, is multivariate normal with independent elements having mean zero and variance, σ^2. Note that the assumption of independent elements can be relaxed somewhat (see Thomas 1991).

By using least-squares regression, an estimate of B_0 is

$$B = [b_1; b_2; ...; b_q] = ([1; X_0]^T [1; X_0])^{-1} [1; X_0]^T Y. \quad (3)$$

It follows that the columns of $\Delta B = B - B_0$ are independent and normally distributed with mean zero and variance $([1; X_0]^T [1; X_0])^{-1} \sigma^2$.

In the prediction step, measurements from $r \geq 1$ new samples are obtained. Note that this generalizes the setup in Thomas (1991) from $r = 1$ to a general r (see Thomas 1994). These measurements will be denoted by the r by q matrix $W = W_0 + \Delta W = [w_1; w_2; ...; w_r]^T$. The measurement errors, ΔW, are assumed to be independent of Y, but have the same distribution as ΔY. The unobservable entity W_0 is assumed to be related to the

underlying *r* by *p* matrix of causal factors $V_0 = \left[v_{0_1}; v_{0_2}; \ldots; v_{0_r}\right]^T$ through the model $[1; V_0]\, B_0 = W_0$. Note that V_0 and W_0 are analogous to X_0 and Y_0 in the calibration step. Here, however, the objective is to predict V_0 given B and the measurements W_0. Here, no distribution of V_0 will be assumed. Note that there are two sources of error when predicting V_0, ΔB and ΔW.

Thus, it is straightforward to cast the prediction of the *r* new observations represented by V_0 into the framework of a linear functional model (e.g., see Fuller 1987) where $B_0^T\,[1; V_0]^T = W_0^T$. Let $[a_0; C_0] = B_0^T$ where a_0 is *q* by 1, and $Z_0 = W_0^T - a_0\, 1_r$. Then, $C_0 V_0^T = Z_0$. Denoting the observables of Z_0 and C_0 by $Z = W^T - a\, 1_r$, and $C = [c_1; c_2; \ldots; c_q]^T$ respectively, define

$$\mathcal{E} = [\varepsilon_1; \varepsilon_2; \ldots; \varepsilon_q]^T = [Z - Z_0; C - C_0] = [\Delta Z; \Delta C].$$

From earlier assumptions, we have that the $\varepsilon_t = [\mathcal{E}_{t1}, \mathcal{E}_{t2}, \ldots, \mathcal{E}_{t,r+p}]$ for $t = 1, 2, \ldots, q$ are *i.i.d. multivariate normal* with mean zero and covariance $\Sigma = \sigma^2\, \Omega$, where

$$\Omega = \left[\begin{array}{c|c} \Omega_{11} & 0 \\ \hline 0 & \left[X_0^T X_0\right]^{-1} \end{array}\right] \text{ and } \Omega_{11} = \frac{1}{n}\left(n \cdot I_r + 1 1^T\right).$$

3. *MAXIMUM LIKELIHOOD ESTIMATION*

Given the knowledge of the distribution of the ε_t, the maximum likelihood estimates (MLEs) of V_0 and C_0 are obtained by maximizing the log likelihood,

$$-\frac{q}{2} \cdot \log\left|2\pi\,\sigma^2\,\Omega\right| - \left(2\sigma^2\right)^{-1} \sum_{t=1}^{q} \left((z_t; c_t) - \left(z_{0_t}; c_{0_t}\right)\right)\, \Omega^{-1}\, \left((z_t; c_t) - \left(z_{0_t}; c_{0_t}\right)\right)^T,$$

with respect to V_0 and C_0, where z_t, c_t, z_{0_t}, and c_{0_t} are the t^{th} rows of Z, C, Z_0, and C_0 respectively (see e.g., Fuller 1987, Chapter 4). Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{p+r}$ be the eigenvalues of $\Omega^{-1/2} M \Omega^{-1/2}$ and let $G = (G_1, G_2)$ be the matrix of corresponding orthonormal eigenvectors such that $\Omega^{-1/2} M \Omega^{-1/2} G_2 = G_2 R$, where

$M = q^{-1} \sum_{t=1}^{q} S_t S_t^T$, $S_t = (Z_{t1}, Z_{t2}, \ldots, Z_{tr}, c_t)^T$, $R = \mathrm{diag}(\lambda_{p+1}, \lambda_{p+2}, \ldots, \lambda_{p+r})$, and $\Omega^{-1/2}$ is the matrix square root of Ω^{-1}. Further, let $D = \Omega^{-1/2} G_2$. Then the MLE of V_0 is $\tilde{V}_0 = -D_{rr}^{-1} D_{pr}^T$, where D_{rr} consists of the first r rows of D, while D_{pr} consists of the last p rows of D.

In the context of the limit $q \to \infty$, $\tilde{V}_0$ is a consistent estimator of V_0 and has a limiting normal distribution (see e.g., Fuller 1987, Chapter 4). That is,

$F^{-1/2} \mathrm{vec}\left(\tilde{V}_0^T - V_0^T\right) \xrightarrow{Dist.} Normal\,(0, I)$, as $q \to \infty$, where

$F = \frac{\sigma^2}{q} \cdot \Psi \otimes \left(L^{-1} + \sigma^2 \cdot L^{-1} \Theta \cdot L^{-1}\right)$, $L = q^{-1} \sum_{t=1}^{q} c_{0_t}^T c_{0_t}$,

$\Psi = (I_r; -V_0)\, \Omega\, (I_r; -V_0)^T$, and $\Theta = \left\{\left(V_0^T; I_p\right) \Omega^{-1} \left(V_0^T; I_p\right)^T\right\}^{-1}$. The limiting distribution of $\tilde{V}_0$ can be used as a basis for statistical inference for V_0 (see Thomas 1994). Furthermore, the form of the limiting distribution can provide a number of useful insights relating to the design and execution of calibration studies. For example, one can evaluate the effect on the asymptotic covariance matrix, F, by varying the set of response variables that are used in the model. Clearly, the sensitivity and selectivity of response variables affect L and therefore F. Other factors that affect F are: 1. the measurement error variance (σ^2), 2. the design matrix defined by the calibration set (X_0) through its influence on Ω, and 3. the underlying levels of causal factors in the prediction set (V_0).

The joint estimation of the r rows of V_0 by $\tilde{V}_0$ involves using the responses from all members of the prediction set. That is, each row of $\tilde{V}_0$ is influenced by all rows of W. This differs from traditional calibration practice where the prediction of the j^{th} specimen is developed through consideration of only the model (e.g., B) and the instrumental measurements associated with that single new specimen (e.g., w_j).

Thus, it is interesting to compare $\tilde{V}_0$ with the sequence of estimators that result from individual maximum likelihood estimation of each row of V_0. In this case, which is consistent with traditional calibration practice, the individual MLEs of C_0 and the j^{th} row of V_0 are obtained by maximizing the log likelihood

$$-\frac{q}{2}\cdot\log\left|2\pi\sigma^2\Omega_a\right| - \left(2\sigma^2\right)^{-1}\sum_{t=1}^{q}\left(\left(Z_{tj},c_t\right)-\left(Z_{0_{tj}},c_{0_t}\right)\right)\Omega_a^{-1}\left(\left(Z_{tj},c_t\right)-\left(Z_{0_{tj}},c_{0_t}\right)\right)^T,$$

where $\Omega_\alpha=\begin{bmatrix}\frac{n+1}{n} & 0\\ 0 & \left[X_0^T X_0\right]^{-1}\end{bmatrix}$. Thomas (1994) shows through analysis of the respective asymptotic covariances of the individual and joint MLEs and through simulation that the variability of the individual MLE of a given row of V_0 can be much greater than that of the corresponding row of $\tilde{V}_0$ when: 1. the noise-to-signal ratio, $\sigma^2 L^{-1}$, is poor (large), 2. n is small, implying that $\left[X_0^T X_0\right]^{-1}$ is relatively large, 3. the specificity of the response variables is poor (i.e., L is poorly conditioned), and 4. r is large and the rows of V_0 are well dispersed.

The basis for the superiority of the joint MLE over the individual MLE is the common relationship among the response variables (i.e. model parameters) which exists whether or not the associated values of the explanatory variables are known. To illustrate, consider the following synthetic example (which meets the above conditions) in a

spectroscopic context where $\sigma = .1$, $\left([1; X_0]^T [1; X_0]\right)^{-1} = \text{diag}\,(.2,\ 1,\ 1)$, $p = 2$, $q = 50$, $r = 20$, $v_{0_1} = (0,0)$, and the 38 elements of last r-1 rows of V_0 were independently sampled from the uniform distribution on [-.9, .9]. Figures 1 and 2 display the MLEs of C_0 (Δ's for column 1 of C_0 and +'s for column 2) obtained by independent and joint estimation, respectively. Superimposed for comparison are the solid and dashed Gaussian-shaped lines representing the two columns of C_0. Clearly joint estimation provides superior estimates of the elements of C_0. Because the estimation of V_0 is connected to the implicit estimation of C_0 through the estimation procedure, the superiority of $\tilde{V}_0$ over the individual estimator of V_0 follows.

4. EXTENSION TO PRINCIPAL COMPONENTS REGRESSION

From the previous section, in the case of an underlying causal linear model, it is observed that there can be some benefit associated with using the responses from all members of the prediction set to predict an individual sample. As mentioned earlier, however, the applicability of such causal models may be significantly limited in practice. Thus, it is interesting to consider whether a similar benefit can be realized when using a soft-modeling approach.

Motivated by the results presented in the previous section, Thomas (1995) demonstrates that in the case of PCR a similar benefit can be realized. Here, we will assume that both the calibration and prediction set specimens are drawn at random from some population of specimens. Briefly, the idealized PCR model is $x = b_0 + b_1 \cdot T_1 + b_2 \cdot T_2 + \ldots + b_h \cdot T_h$, where the b_i are parameters, and T_i is the latent variable obtained by projecting the q-dimensional idealized noise-free response variable on the i^{th} population eigenvector which represents the i^{th} largest source of response variation. Only the h largest sources of variation are used to define the model. In practice, the model prediction is given by

$\hat{x} = \hat{b}_0 + \hat{b}_1 \cdot \hat{T}_1 + \hat{b}_2 \cdot \hat{T}_2 + \ldots + \hat{b}_h \cdot \hat{T}_h$, where $\hat{b}_i$ denotes an estimate of the model parameters, and $\hat{T}_i$ denotes the projection of the measured q-dimensional response variable onto the i^{th} sample eigenvector. Given that the model form is appropriate, the prediction error (i.e., $\hat{x} - x$) is due to: 1. errors in the estimated model parameters, ($\hat{b}_i - b_i$), and 2. errors in the latent variables ($\hat{T}_i - T_i$) which are due to the combined effects of measurement errors of the response variable and differences between the sample and population eigenvectors.

Traditionally, the basis set of sample eigenvectors is based exclusively on the responses associated with the calibration set. Alternately, if the sample eigenvectors are based on the combined calibration and prediction sets, the resulting sample eigenvectors will tend to be closer to the population eigenvectors than those based only on the calibration set. The reduction, in magnitude, of the resulting prediction errors can be substantial if the differences between the sample and population eigenvectors is a significant source of error and $\frac{r}{n}$ is large, where n is the number of calibration samples and r is the number of prediction samples (see Thomas 1995). Again, the basis for improvement is the common relationship among the response variables that exists throughout the the calibration *and* prediction sets and which defines the basis set that is used for modeling.

5. CONCLUSIONS

This paper has demonstrated the usefulness of an EIV model as a device for providing insights into the multivariate calibration problem. By using the EIV model in the context of an underlying causal model, analysis of the asymptotic distribution of prediction errors and subsequent simulation established the value of a nontraditional concept for prediction. This concept uses the responses from all available specimens of the prediction set to improve the prediction of individual specimens. Motivated by this result, this concept was extended to PCR, which is a soft-modeling approach and therefore generally more applicable than a causal modeling approach. In practice, this new concept

could be useful in the analytical chemistry laboratory, where measurements of the prediction set specimens are often obtained close in time so as to be available to assist in the prediction of an individual specimen.

REFERENCES

Brown, P. J., *Measurement, Regression, and Calibration*, Oxford University Press, New York, 1993.

Fearn, T., "A Misuse of Ridge Regression in the Calibration of a Near Infrared Reflectance Instrument," *Applied Statistics*, 32, 1983, pp. 73-79.

Fuller, W. A., *Measurement Error Models*, John Wiley, New York, 1987.

Haaland, D. M., "Quantitative Infrared Analysis of Borophosposilicate Films using Multivariate Statistical Methods," *Analytical Chemistry*, 60, 1988, pp. 1208-1217.

Martens, H., and Naes, T., *Multivariate Calibration*, John Wiley, Chichester, 1989.

Robinson M. R., Eaton, R. P., Haaland, D. M., Koepp, G. W., Thomas, E. V., Stallard, B. R., and Robinson, P. L., "Noninvasive Glucose Monitoring in Diabetic Patients: A Preliminary Evaluation," *Clinical Chemistry*, 38, 1992, pp. 1618-1622.

Thomas, E. V., "Errors-in-Variables Estimation in Multivariate Calibration," *Technometrics*, 33, 1991, pp. 405-413.

Thomas, E. V., "Simultaneous Inference in Multivariate Calibration," *Sandia National Laboratories Technical Report SAND94-2642*, 1994.

Thomas, E. V., "Incorporating Auxiliary Predictor Variation in Principal Component Regression Models," *Journal of Chemometrics*, 9, 1995, pp. 471-481.

FIGURE 1 - INDIVIDUAL MLE OF PARAMETERS

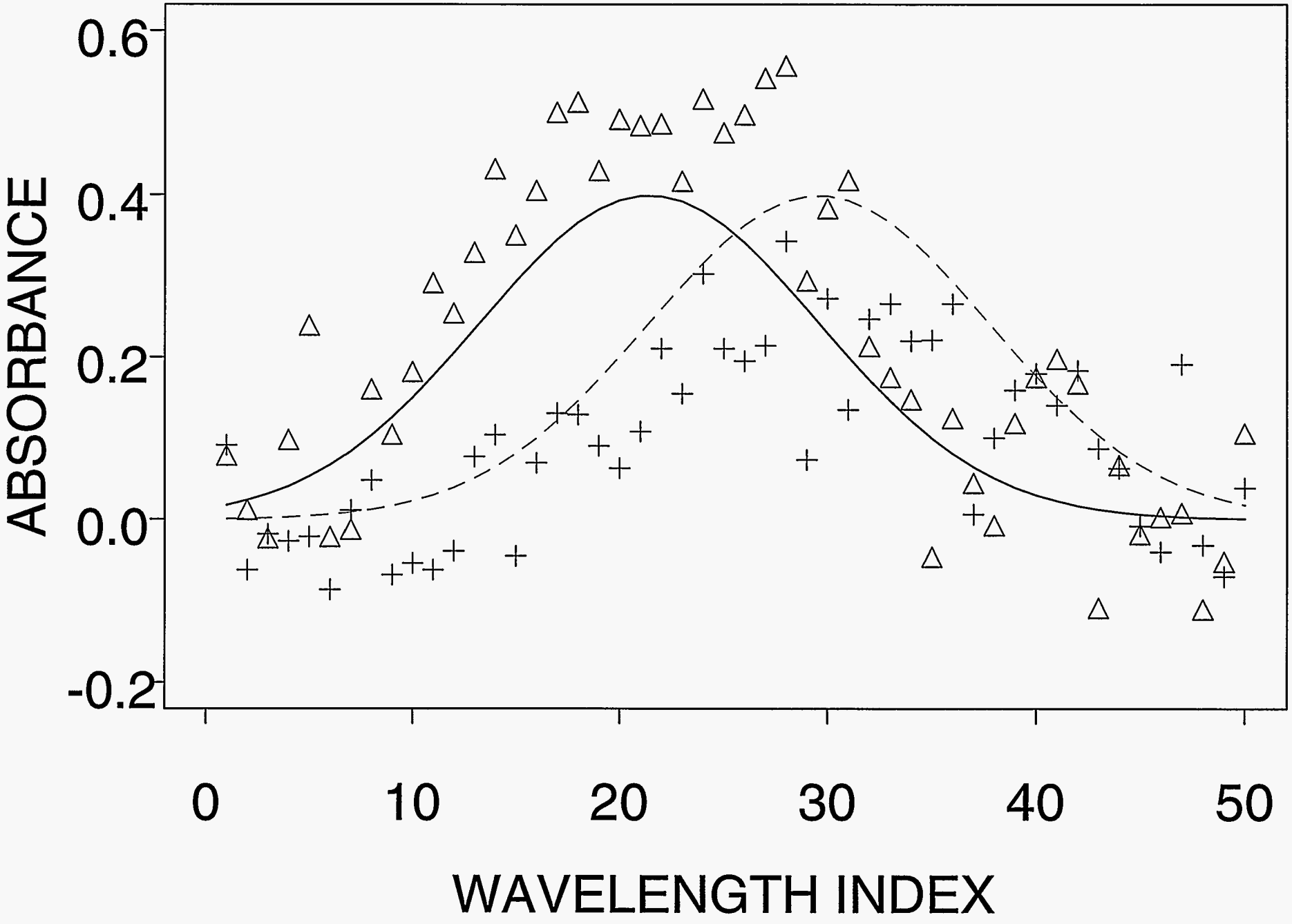

FIGURE 2 - JOINT MLE OF PARAMETERS