R&D NOTE

Sub-PCA Modeling and On-line Monitoring Strategy for Batch Processes

Ningyun Lu and Furong Gao
Dept. of Chemical Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Fuli Wang
School of Information Science and Engineering, Northeastern University, Shenyang, Liaoning (110004), P. R. China

Keywords: batch process, online monitoring, PCA, chromatography, statistical analysis.

Multivariate statistical methods such as principal component analysis (PCA) and partial least square (PLS) have been successfully used in modeling multivariable continuous processes (Kaspar and Ray, 1992; Kourti and MacGregor, 1995; Chen and McAvoy, 1998). Several extensions of the conventional PCA/PLS to batch processes have also been reported, among which multiway PCA (MPCA) model is the most widely used (Wold et al., 1987; Nomikos and MacGregor, 1994, 1995; Wise et al., 1999; Smilde, 2001). The MPCA model is ill-suited for multistage batch processes, as it takes the entire batch data as a single object, and it is difficult to reveal the changes of process correlation from stage to stage. Considering that the multiplicity of the operation stage is an inherent nature of many batch processes, each stage has its own underlying characteristics and the process can exhibit significantly different behaviors over different operation stages; it is desirable to develop a stage-based model that can reflect the inherent process stage nature to improve the process understanding and monitoring efficiency. Kosanovich et al. (1994) and Dong and McAvoy (1995) developed two MPCA/nonlinear MPCA models, utilizing the two-stage nature of a jacketed exothermic batch chemical reactor. Their results show that the two-stage models are more powerful than a single model. Their stage models, however, inherit the common weakness of the MPCA model that the unavailable future data in an evolving batch should be estimated for on-line monitoring.

A new stage-based sub-PCA modeling method is proposed in this article for multistage batch processes, based on the recognition of the following: (1) a batch process may be divided into several "operation" stages reflecting its inherent process correlation nature; (2) despite that the process may be time varying, the correlation of its variables will be largely similar within the same "operation" stage. Changes in the correlation may be used to indicate changes in the process "operation" stages. We have placed a quotation mark around "operation" to indicate that the operation referred to in this article may not, and does not have to, have the exact correspondence to the physical operations of the process. Based on the above recognition, a representative model can be built for each stage, using the conventional two-way PCA model. This allows two-way PCA to be "directly" applied to a batch process after a proper stage division; a stage division algorithm is also developed in the article. A three-tank process, as an experimental verification system, is finally introduced to illustrate the effectiveness of the proposed.

Stage-Based Sub-PCA Modeling and On-Line Monitoring

Consider a batch process with J process variables measured over sampling points k ($k = 1, 2, \ldots, K$); a data matrix of dimensions $J \times K$ is generated from each batch run. A set of I number of normal batch runs, hence, result in a three-way process data matrix, $\underline{X}(I \times J \times K)$, which is the most popular data form for batch processes. The horizontal slice $\bar{X}(J \times K)$ is the data matrix from each batch run. The vertical slice $\tilde{X}(I \times J)$, a time-slice matrix that is the basic unit in the proposed modeling method, is used to obtain the process correlation at sampling time k.

In PCA analysis, the loading matrix represents the information of the process correlation. The stage-based sub-PCA mod-

Correspondence concerning this article should be addressed to F. Gao at kefgao@ust.hk.

eling begins with analyzing the loading matrix at each sampling interval. The time-slice matrices $\tilde{X}^k$ will have a similar loading matrix within each "operation" stage. Different "operation" stage results in different loading matrices, reflecting that process correlation changes over different stages. Likewise, changes in the loading matrices, reflecting changes in the underlying process behavior, can be used to determine the "operation" stages. The k-means clustering will be adopted and modified to partition the K number of loading matrices. The clustering results, associated with process time span or indicator variables, can be used to define process "operation" stages.

Modeling procedure

A normal batch implies that the process follows a set of predetermined sequences with acceptable process trajectories. Batch-to-batch variations from the mean trajectories are caused by stochastic factors, that is, the normal batches in the reference data set are deemed to subject to common-cause variation. Nomikos and MacGregor (1995) have shown that it is reasonable to assume that the JK number of process measurements in the unfolded process matrix $X(I \times JK)$ have a multinormal distribution. In this article, we can inherit this conclusion and assume that process measurements over normal batches also follow multinormal distribution at time interval k. With this assumption, PCA can be performed on these time-slice matrices $\tilde{X}^k$ ($k = 1, 2, \ldots, K$), generating a K number of loading matrices, which represent the process correlation at each time interval k

$$\tilde{X}^k = \tilde{T}^k(\tilde{P}^k)^T \quad (k = 1, 2, \ldots, K). \tag{1}$$

Clustering Loading Matrices. The loading matrices $\tilde{P}^k$ are transformed into a weighted form after considering the importance of each column $\mathbf{p}_j^k$

$$\check{P}^k = [\mathbf{p}_1^k \cdot g_1^k, \mathbf{p}_2^k \cdot g_2^k, \ldots, \mathbf{p}_J^k \cdot g_J^k], \tag{2}$$

where $g_j^k = \lambda_j^k / \Sigma_{j=1}^J \lambda_j^k$, and λ_j^k is the eigenvalue of the covariance matrix $(\tilde{X}^k)^T\tilde{X}^k$.

The *Euclidean distance,* the most popular metric, can be used to calculate the dissimilarity between two patterns. A variant k-means algorithm (Jain et al., 1999) is adopted for partitioning the K number of patterns $\check{P}^k$ ($k = 1, 2, \ldots, K$) to determine the optimal number of final clusters by minimizing the local squared error (for patterns within each cluster) and the global squared error (for all the patterns) by specifying a threshold θ of the minimal distance between two clusters' centers, or the maximal radius of a cluster. This algorithm transforms modeling accuracy and complexity into the specification of the threshold. A large threshold results in few clusters, but less accurate modeling. A step is added in the clustering algorithm to eliminate singular clusters that catch few patterns in the iterative clustering procedure to enhance the robustness and reliability of the partition algorithm.

The above improved k-means clustering algorithm can group optimally the K patterns into C number of clusters, representing C kinds of pattern features. Since these patterns are extracted along the sampling time of a batch process, the clustering result can be directly associated with the operation time, which makes the partition of the patterns well interpretable. Normally, each cluster should contain a series of successive samples. The exception may be with a process that has the same underlying characteristics for several disjoint periods of operation time; this would result in a cluster with samples disjoint in time. In most cases, process stages can be determined based on the clustering result associated with operation time.

The number of clusters may be different from the actual operation stages. For example, a process with two or more actual operation stages of similar correlation may be clustered together, resulting in a single representative loading matrix. On the other hand, a long stage of operation having significant changes in the correlation may be divided into several "operation" stages. The proposed method emphasizes the changes of process correlation rather than the physical operation. In this article, C denotes the number of "operation" stages obtained by the clustering algorithm based on the process correlation characteristic. Misclassification may occur at the beginning and end of each stage, because the k-means clustering algorithm is a hard-partition method in dealing with patterns between two neighboring clusters. Such possible misclassification has little influence in the sub PCA model development; however, it may lead to false alarm (type I error) and missing alarm (type II errors) in on-line monitoring due to batch variation. Alternative methods may be used to resolve this problem. One is to relax the monitoring conditions at the beginning and the end of each stage; the other is to associate one or more characteristic process variables with the stage division, rather than using the process time.

Developing Sub PCA Model for each Stage. Define P_c^* ($c = 1, 2, \ldots, C$) as the representative loading matrix for the c^{th} stage as

$$P_c^* = \underset{k}{\text{Min}}(\|\tilde{P}^k - P_{ci}^*\|^2) = \frac{1}{n_{\text{stage\_}c}} \sum_k \tilde{P}^k (c = 1, 2, \ldots, C; \; k = 1, 2, \ldots, n_{\text{stage\_}c}) \tag{3}$$

where $n_{\text{stage\_}c}$ is the number of the process data belonging to stage c. Similarly, define S_c^* ($c = 1, 2, \ldots, C$) as the representative singular-value diagonal matrix for each stage, which will be used for the determination of the control limits latter

$$S_c^* = \frac{1}{n_{\text{stage\_}c}} \sum_{k=1}^{n_{\text{stage\_}c}} \tilde{S}^k = \text{diag}(\lambda_1^{*c}, \lambda_2^{*c}, \ldots, \lambda_J^{*c}) \times (c = 1, 2, \ldots, C) \tag{4}$$

where $\tilde{S}^k = \text{diag}(\lambda_1^k, \lambda_2^k, \ldots, \lambda_J^k)$ is the singular-value diagonal matrix for time-slice matrix in stage c. The number of retained principal components R_c^* for each stage could be determined by the cumulative explained variance rate, defined by $R_c^* = \Sigma_{i=1}^{R_c} \lambda_i^{*c} / \text{trace}(S_c^*) \geq 90\%$.

P_c^* is divided into two parts, $\bar{P}_c^*$ and $\tilde{P}_c^*$, for principal component subspace and residual space, respectively; so does S_c^*. In each stage c ($c = 1, 2, \ldots, C$), the representative loading matrix $\bar{P}_c^*$ is used to construct a sub-PCA model as

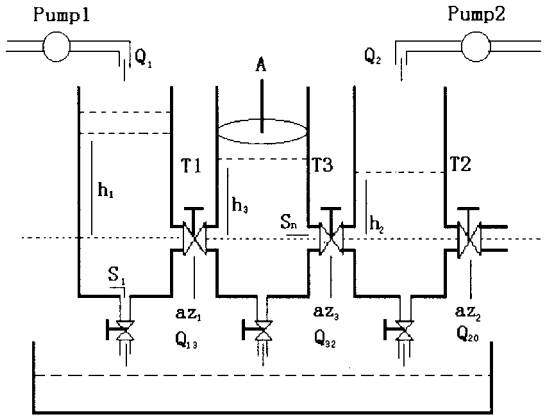

Figure 1. Three-tank process.

$$\tilde{T} = \tilde{X}(\bar{P}_c^*)^T; \quad \hat{\tilde{X}} = \tilde{T}\bar{P}_c^*; \quad \tilde{E} = \tilde{X} - \hat{\tilde{X}} = \tilde{X}(I - \bar{P}_c^*(\bar{P}_c^*)^T). \tag{5}$$

Control limits for on-line monitoring

The control limits for the Hotteling-T^2 and squared prediction error (SPE) charts, which can be estimated from the reference data, should be computed in the modeling procedure.

For principal component scores at time K^1 $\tilde{T}^k$, a generalized T_m^2 statistic is introduced to describe the average variability of process variables over all batches at time interval k. This provides the estimation of the control limits for the principal component scores of future measurements in a new batch

$$T_m^2|^k = I(\bar{\mathbf{t}}^{kT}(\bar{S}_c^*)^{-1}\bar{\mathbf{t}}^k), \bar{\mathbf{t}}^k = \frac{1}{I}\sum_{i=1}^{I}\mathbf{t}_i^k, \tag{6}$$

where $\boldsymbol{t}_i^k$ is the i^{th} row of $\tilde{T}^k$. T_m^2 has the same distribution as the individual T_i^2 (Jackson, 1991)

$$T_i^2 = \mathbf{t}_i^T(\bar{S}_c^*)^{-1}\mathbf{t}_i \sim \frac{R_c(I-1)}{(I-R_c-1)}F_{R_c,(I-R_c-1),\alpha}. \tag{7}$$

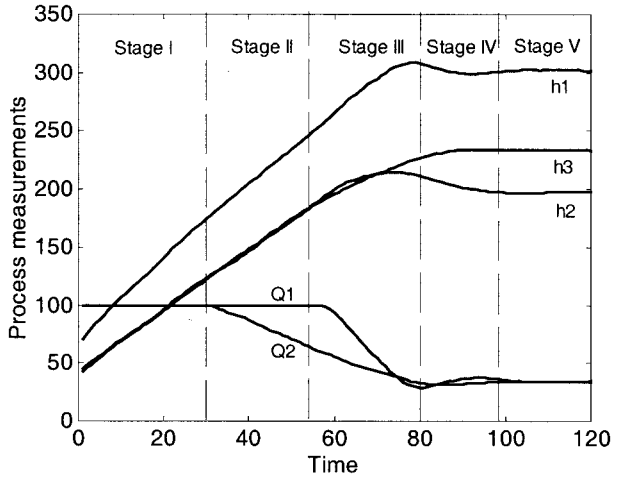

Figure 2. Process variable measurements for three-tank process.

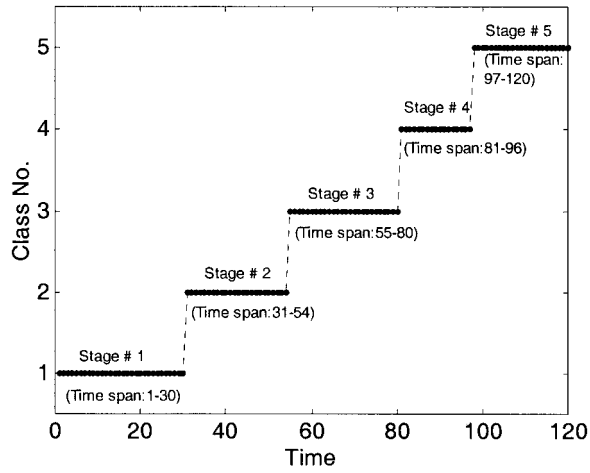

Figure 3. Results of loading matrices clustering algorithm.

Similarly, define Q_m, which has the same distribution and the same degrees of freedom as Q

$$Q_m|^k = I(\bar{\mathbf{x}}^k - \hat{\bar{\mathbf{x}}}^k)^T(\bar{\mathbf{x}}^k - \hat{\bar{\mathbf{x}}}^k), \bar{\mathbf{x}}^k = \frac{1}{I}\sum_{i=1}^{I}\mathbf{x}_i^k, \tag{8}$$

where $\boldsymbol{x}_i^k$ is the i^{th} row of $\tilde{X}^k$. The control limits for Q_m can be calculated by the method outlined by Jackson and Mudholkar (1979) and Jackson (1991), using the representative singular values S_c^* for the stage c.

Monitoring procedure

Using the above procedure, a sub-PCA model can be developed for each stage. For on-line monitoring, the T^2 statistic is calculated using scores obtained by projecting the original process data onto the subspace spanned by the representative loading matrix $\bar{P}_c^*$ for stage c.

Suppose new data $\boldsymbol{x}_{\text{new}}$ belong to stage c

$$\begin{aligned}\mathbf{t} &= \mathbf{x}_{\text{new}}\bar{P}_c^* \\ \mathbf{e} &= \mathbf{x}_{\text{new}} - \mathbf{x}_{\text{new}}\bar{P}_c^*(\bar{P}_c^*)^T\end{aligned} \tag{9}$$

Then, the T^2 statistic is calculated by: $T_{\text{new}}^2 = \boldsymbol{t}^T(\bar{S}_c^*)^{-1}\boldsymbol{t} \sim [R_c(I-1)/I(I-R_c-1)]F_{R_c,I-1,\alpha^*}$, and SPE is calculated by: $SPE_{\text{new}} = \boldsymbol{e}^T\boldsymbol{e}$.

For on-line monitoring, one should first determine which stage new data of the evolving batch belong to before calling the corresponding sub PCA model to obtain the two statistics. As process stages are represented by the process operation time span, one can know which stage the data belong to by checking which time span the current sampling falls in. Process monitoring is conducted by comparing the two statistics with the control limits of the corresponding stage.

Application Illustration

A three-tank system, as shown in Figure 1, is used for the verification of the proposed approach. Closed-loop control is implemented for the levels of tanks 1 and 2. The two levels are

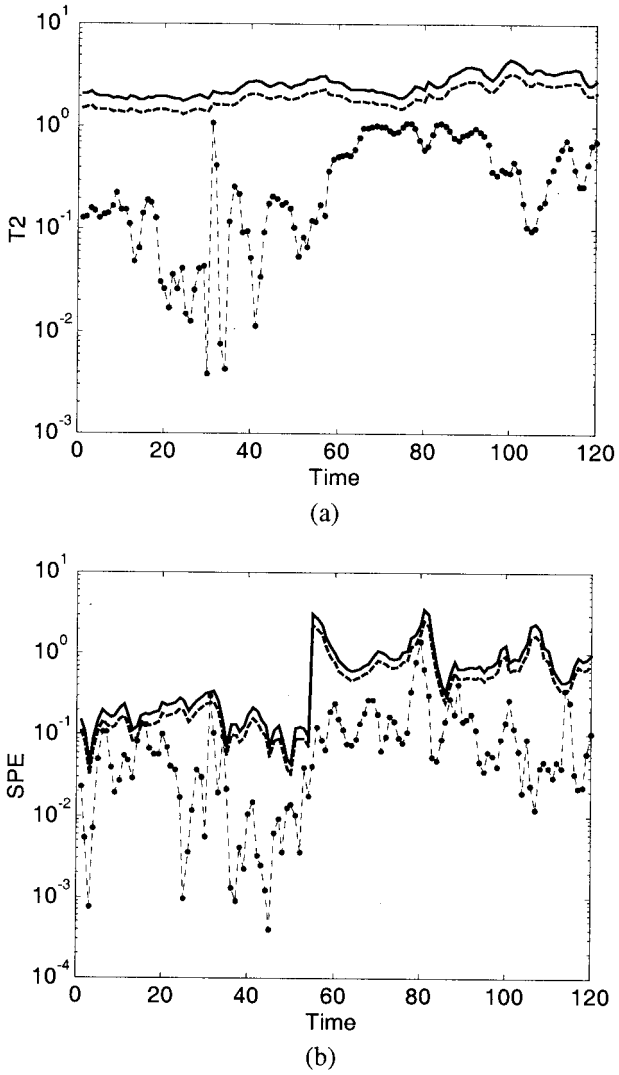

Figure 4. Monitoring charts for a normal batch.

Solid line, 99% control limit; dashed line, 95% control limit; dash dot line, on-line T^2 or SPE. (a) Monitoring chart of T^2 statistic; (b) monitoring chart of SPE statistic.

brought from their initial conditions to the set points of $h_1 = 300$ mm and $h_2 = 200$ mm. The level of h_3 is left to float to reflect the interaction between tanks 1 and 2. The process finishes after the three levels stabilize over a period of time. The raising water levels in the tanks result in a time-varying process dynamics. Five process variables including three levels h_1, h_2, h_3 and two flow rates Q_1 and Q_2 were measured every second. 120 points of historical data were collected in each batch under the normal operation. A typical process curve is shown in Figure 2 for one run.

According to the status of the two manipulated variables Q_1 and Q_2, the process can be classified into the following major stages: stage I, in which both manipulated variables are at saturation; stage II, in which one manipulated variable Q_1 is at saturation; and stages III through V, in which no manipulated variables are at saturation. The last three stages are defined as the decreasing stage (stage III), the tuning stage (stage IV), and the steady stage (stage V) as illustrated in Figure 2.

The data of the 21 normal experiments are used to yield the reference data matrix $\underline{X}$ of dimension of $21 \times 5 \times 120$. The loading matrices calculated from the time-slice matrices are fed to the clustering algorithm, resulting in five groups as shown in Figure 3, which agrees well with the earlier theoretical analysis of Figure 2. This proves that process correlation does remain similar within each stage, changes from stage to stage. Only two or three principal components are needed for each stage to explain over 90% variations, while, for the MPCA analysis of the process, the first three principal components can explain about 50% variations.

The proposed approach is put into on-line monitoring tests. Figure 4 shows the monitoring of a normal batch, where the values of the two statistics, Hotelling T^2 and SPE statistics, are well below the control limits, indicating that the whole batch is free of any process abnormality. For the second case, a fault was introduced to simulate a leakage of tank 1 by opening valve S_1 at the 42nd sampling. According to the above stage division, this fault occurs in the second stage. From the mon-

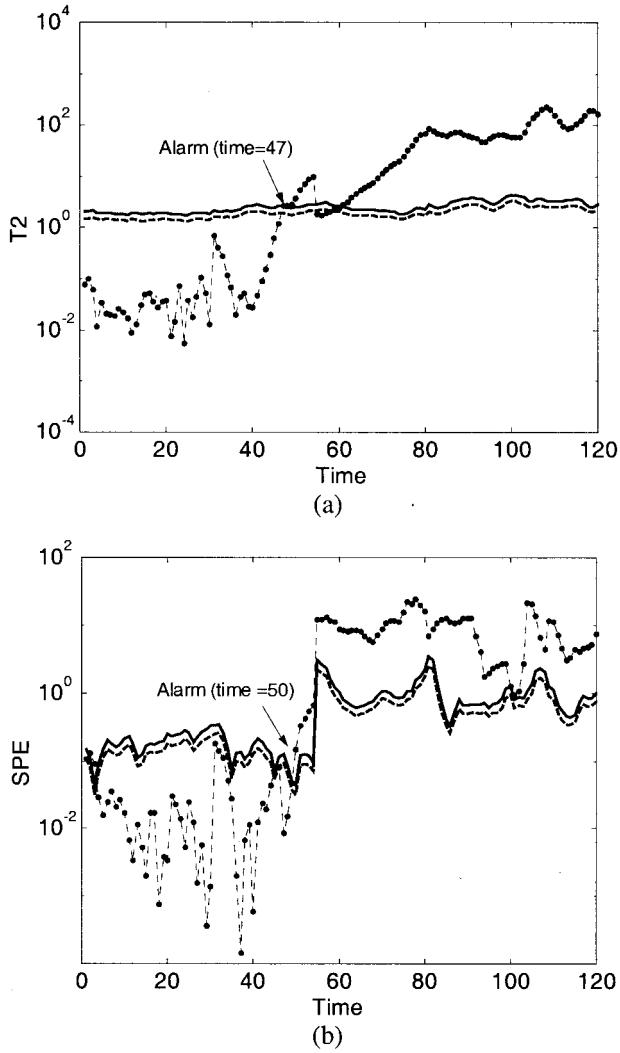

Figure 5. Monitoring charts for an abnormal batch.

Solid line, 99% control limit; dashed line, 95% control limit; dash dot line, on-line T^2 or SPE. (a) Monitoring chart of T^2 statistic; (b) monitoring chart of SPE statistic.

itoring charts shown in Figure 5, the abnormality can be clearly detected in the second stage at 47^{th} sampling, only five samplings after the occurrence.

Conclusions

A new modeling and on-line monitoring scheme for batch process has been developed based on the fact that changes in the process correlation may relate to its "operation" stages. Dividing the process into "operation" stages by analyzing and clustering the loading matrices and constructing sub-PCA model for each stage can apply the conventional two-way PCA "directly" for batch process monitoring, without the need of predicting future data of the evolving batch.

Literature Cited

Chen, G., and T. J. McAvoy, "Predictive On-Line Monitoring of Continuous Processes," *J. Process Control,* **8,** 409 (1998).

Dong, D., and T. J. McAvoy, "Multi-Stage Batch Process Monitoring," *Proc. of American Control Conf.,* 1857 (1995).

Jackson, J. E., *A User's Guide to Principal Components,* Wiley, New York (1991).

Jackson, J. E., and G. S. Mudholkar, "Control Procedures for Residuals Associated with Principal Component Analysis," *TECHNOMETRICS,* **21,** 341 (1979).

Jain, A. K., M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys,* **31,** 264 (1999).

Kaspar, M. H., and W. H. Ray, "Chemometric Methods for Process Monitoring and High-Performance Controller Design," *AIChE J.,* **38,** 1593 (1992).

Kosanovich, K. A., M. J. Piovoso, and K. S. Dahl, "Multi-Way PCA Applied to an Industrial Batch Process," *Proc. of American Control Conf.,* 1294 (1994).

Kourti, T., and J. F. MacGregor, "Process Analysis, Monitoring and Diagnosis, using Multivariate Projection Methods," *Chemometrics and Intelligent Laboratory Systems,* **28,** 3 (1995).

Nomikos, P., and J. F. MacGregor, "Multivariate SPC Charts for Monitoring Batch Process," *TECHNOMETRICS,* **37,** 41 (1995).

Nomikos, P., and J. F. MacGregor, "Monitoring Batch Processes using Multiway Principal Component Analysis," *AIChE J.,* **40,** 1361 (1994).

Smilde, A. K., "Comments on Three-Way Analyses used for Batch Process Data," *J. of Chemometrics,* **15,** 19 (2001).

Wise, B. M., N. B. Gallagher, S. W. Butler, D. D. White, and G. G. Barna, "A Comparison of Principal Component Analysis, Multiway Principal Component Analysis, Trilinear Decomposition and Parallel Factor Analysis for Fault Detection in a Semiconductor Etch Process," *J. of Chemometrics,* **13,** 379 (1999).

Wold, S., K. Esbensen, and P. Geladi, "Principal Component Analysis," *Chemometrics Intelligent Laboratory Systems,* **2,** 37 (1987).

Manuscript received Feb. 24, 2003, and revision received June 12, 2003.