APPENDIX D: MATRIX NOTATION

1. Matrix Notation

Matrix notation is used to simplify the presentation of calculations that are performed in the linear regression. A matrix is a rectangular array of numbers. Boldface capital letters represent matrices, and lower case letters with subscripts represent individual numbers in the matrices. **X**, below, is a 10 by 3 matrix. It has 11 rows and 3 columns. The rows are numbered 0, 1, 2,...10 and columns are numbered 0, 1, and 2. (Other texts may begin numbering with 1.)

$$\mathbf{X} = \begin{matrix} 1 & 1.002 & 1.0040 \\ 1 & 0.902 & 0.8136 \\ 1 & 0.802 & 0.6432 \\ 1 & 0.701 & 0.4914 \\ 1 & 0.601 & 0.3612 \\ 1 & 0.501 & 0.2510 \\ 1 & 0.401 & 0.1608 \\ 1 & 0.301 & 0.0906 \\ 1 & 0.200 & 0.0400 \\ 1 & 0.100 & 0.0100 \\ 1 & 0.000 & 0.0000 \end{matrix}$$

$\mathbf{X}_{i,j}$ denotes the number that is found in the ith row and the jth column. $\mathbf{X}_{0,1} = 1.002$. The first row and column are numbered zero.

A matrix that has only one column is called a column vector, and a matrix that has only one row is called a row vector.

$$\mathbf{y} = \begin{matrix} 0.999 \\ 0.915 \\ 0.828 \\ 0.738 \\ 0.644 \\ 0.549 \\ 0.448 \\ 0.346 \\ 0.237 \\ 0.122 \\ 0.001 \end{matrix}$$

is a column vector.

Subscripts following vector names denote the row or column of the vector. For example, $\mathbf{y}_1$ is the number in the second row of **y**, 0.915. (Remember that we begin counting rows with zero.)

Matrix operations that come into play for calibration include multiplication, transposition, and inversion. The rules for these operations can be found in any introduction to matrices. We will use the following notation for these operations:

X' denotes the transpose of **X** (the ith column of **X** becomes the ith row of **X**')

For the matrices **X** and **Y** above,

	1	1	1	1	. . .	1
X' =	1.002	0.902	0.802	0.701	. . .	0.000
	1.0040	0.8136	0.6432	0.4914	. . .	0.0000

X'Y denotes multiplication of matrices **X'** and **Y**. **X'** must have the same number of columns as **Y** has rows. For the matrix **X** above,

	11	5.511	3.8658			5.827
X'X =	5.511	3.8658	3.0438	and	**X'Y** =	4.0132
	3.8658	3.0438	2.5544			3.1272

det(**X'X**) = 1.0521 (the determinant of **X'X**)

$(X'X)^{-1}$ denotes the inverse of the product of **X'** and **X**

	0.5800	-2.1962	1.7392
$(X'X)^{-1}$ =	-2.196	12.5026	-11.5744
	1.7392	-11.5744	11.5513

2. Calibration by Linear Regression Using Matrix Notation - Example

The linear regression approach is illustrated below for the simple quadratic curve.

The starting point for regression analysis will be a matrix named **X**. This matrix will have 3 columns (one for each coefficient to be determined). The number of rows will be the same as the number of calibration measurements that are performed by the measurement system. The first column is a vector of 1s. The second column contains the certified concentrations of the calibration standards. The third column contains the squares of the values appearing in the second column. When this matrix is multiplied by the vector of coefficients [b_0, b_1, b_2], the result is a vector of responses, so that:

$$response_i = 1 * b_0 + concentration_i * b_1 + concentration_i^2 * b_2$$

or, letting y represent response and x represent concentration,

$y_i = b_0 + b_1 x_i + b_2 x_i^2$

Now, we're interested in estimating the the coefficients b_0, b_1, and b_2, and we're also interested in computing how much error is involved when we use the information to estimate the concentration in an "unknown."

3. Determining the Calibration Equation

The coefficients of the calibration equation or curve are found by matrix multiplication and inversion:

$b = \mathbf{(X'X)^{-1} X'Y} = [\mathbf{b_0, b_1, b_2}]$

Example

1	1.002	1.0040		0.999	0.9967
1	0.902	0.8136		0.915	0.9151
1	0.802	0.6432		0.828	0.8297
1	0.701	0.4914		0.738	0.7394
X =1	0.601	0.3612	**y** =	0.644	**b'x** = 0.6462
1	0.501	0.2510		0.549	0.5491
1	0.401	0.1608		0.448	0.4482
1	0.301	0.0906		0.346	0.3434
1	0.200	0.0400		0.237	0.2336
1	0.100	0.0100		0.122	0.1210
1	0.000	0.0000		0.001	0.0046

$$\mathbf{b} = \begin{matrix} 0.0046 \\ 1.1837 \\ -0.1932 \end{matrix} \qquad \mathbf{b'} = \begin{matrix} 0.0046 & 1.1837 & -0.1932 \end{matrix}$$

The quadratic calibration curve is: response = 0.0046 + 1.1837 C + -0.1932 * C^2

4. Determining the Estimation and Prediction Error

One assumption that underlies the regression approach is that random error is constant across the measurement range. Sometimes it may be necessary to apply a transformation in order to achieve this characteristic, called homogeneity of variance. An estimate of this variance is obtained using matrix operations:

Var = residual sum of squares / degrees of freedom = **Y'Y - b'X'Y** / df

This estimate's "degrees of freedom" (df) is the number of calibration points less the number of coefficients estimated for the calibration equation.

Another important output of the regression analysis is the "variance-covariance" matrix, **V**:

$\mathbf{V} = \text{Var} * (\mathbf{X'X})^{-1}$

The variance of each coefficient is found in the principal diagonal of **V**. For example, the variance of b_0 is $\mathbf{V}_{0,0}$. Covariances are found as off-diagonal elements of **V**.

Hypothesis tests can be performed and confidence intervals can be estimated for each coefficient using the coefficient's estimate, the coefficient's variance (contained in **V**), and the degrees of freedom, df.

Continuing our example

Var = (**Y'Y - b'X'Y**) / df = 5.91E-06

$$\mathbf{V} = \begin{bmatrix} 3.43E\text{-}06 & -1.3E\text{-}05 & 1.03E\text{-}05 \\ -1.3E\text{-}05 & 7.39E\text{-}05 & -6.8E\text{-}05 \\ 1.03E\text{-}05 & -6.8E\text{-}05 & 6.83E\text{-}05 \end{bmatrix}$$

df = 8

(df = degrees of freedom)

95% Confidence Interval for $\mathbf{b}_0 = \mathbf{b}_0 +/- t(0.05,df) * \text{sqrt}(\mathbf{V}_{0,0})$

95% CI for $\mathbf{b}_0$ = 0.000324 to 0.008865

t(0.05, df) = 2.306006

95% Confidence Interval for $\mathbf{b}_1 = \mathbf{b}_1 +/- t(0.05,df) * \text{sqrt}(\mathbf{V}_{1,1})$

95% CI for $\mathbf{b}_1$ = 1.163855 to 1.203512

95% Confidence Interval for $\mathbf{b}_2 = \mathbf{b}_2 +/- t(0.975,df) * \text{sqrt}(\mathbf{V}_{2,2})$

95% CI for $\mathbf{b}_2$ = -0.21224 to -0.17413

Another use of **V** is in computing the uncertainty in a regression predicted concentration of an individual unknown. The analyzer is subjected to the unknown, and a mean response, R, is produced. A solution for C is found. This is the estimated concentration of the unknown. Deriving the confidence intervals for this estimate requires finding two alternative concentrations, one higher and one lower than the estimate, such that the probability of having produced a lesser or greater average response is sufficiently small. For a 95% confidence interval, the lower bound is a concentration whose response would be less than the observed response with 97.5% probability; the upper bound is a concentration whose response would be less than the observed response with 97.5% probability.

Unfortunately, for quadratic curves, this derivation is not so simple.

R measurements of an unknown produce an average response resp:

R = 6
resp = 0.601

The estimated concentration is found by solving the following quadratic equation:

$0.601 = \mathbf{b}_0 + \mathbf{b}_1 C + \mathbf{b}_2 C^2$
$(\mathbf{b}_0 - 0.601) + \mathbf{b}_1 C + \mathbf{b}_2 C^2 = 0$

The potential solutions are found using the quadratic formula:

C = 0.553935 and 5.573267 (only the first of these is reasonable)

Now, if the concentration really had been at this value, the 95% confidence interval for the mean response of six measurements would be symmetric about the observed response:

As above, t = 2.306006

x = 1 0.553935 0.306843 = [1, resp, resp^2]

xb = 0.601 (check)

var(predicted mean response for x) = [var/R + x' V x]

x'**V** = -6.09E-07 6.97E-06 -6.7E-06

x'**V**x = 1.2E-06

var/6 = 9.86E-07

var(predicted mean response for x) = 2.19E-06

95% confidence interval for predicted response = 0.597588 to 0.604412

This is the observed response -/+: 0.003412 and 0.003412

Solving for concentration, the interval is no longer perfectly symmetric:

0.550418 to 0.557456

This is the estimated concentration -/+:	0.003516	and	0.003521
As a percentage of the concentration, this is -/+:	0.006348	and	0.0063569

Fortunately, even with the quadratic calibration curve, with good precision, the confidence intervals will be within a small enough region that the curve is close to linear and the interval will be very nearly symmetric. The uncertainty criterion for multipoint calibration requires the 95% confidence interval's half-width to be less than 1%. The calibrated range of the analyzer extends across all concentrations for which the criterion is satisfied.

Continuing our example

Concentration	Estimated Response	95% conf. interval for response		95% conf. interval for concentration		% error for concentration	
1.002	0.9967	0.9924	1.0010	0.9966	1.0074	-0.53	0.54
0.902	0.9151	0.9121	0.9181	0.8985	0.9055	-0.39	0.39
0.802	0.8297	0.8273	0.8320	0.7993	0.8047	-0.33	0.33
0.701	0.7394	0.7371	0.7417	0.6985	0.7035	-0.36	0.36
0.601	0.6462	0.6437	0.6487	0.5984	0.6036	-0.43	0.43
0.501	0.5491	0.5466	0.5517	0.4984	0.5036	-0.51	0.52
0.401	0.4482	0.4457	0.4507	0.3986	0.4034	-0.60	0.60
0.301	0.3434	0.3411	0.3457	0.2988	0.3032	-0.72	0.72
0.200	0.2336	0.2313	0.2359	0.1979	0.2021	-1.06	1.06
0.100	0.1210	0.1181	0.1240	0.0974	0.1026	-2.58	2.59
0.000	0.0046	0.0003	0.0089	-0.0036	0.0036	---	---
0.210	0.2446	0.2423	0.2469	0.2079	0.2121	-0.9996	1.0004

The calibration curve's uncertainty is acceptable for concentrations above 0.21 ppm.

5. Stability Test

As discussed in Subsection 2.1.6.2, the stability test requires at least three initial measurements of the candidate standard plus at least three additional measurements following a period of 7 days or more. The standard's concentration must be in the calibrated range of the analyzer per Subsection 2.1.7.2.

Concentrations are estimated using the calibration curve, producing at least three estimates for the initial concentration and at least three estimates for the concentration following the holding time. A student's t-test is applied as follows:

Initial Data **Final Data (after holding time)**

C1	C4
C2	C5
C3	C6

s_1 = standard deviation of (C1, C2, C3) x_1 = (C1 + C2 + C3) / 3
s_2 = standard deviation of (C4, C5, C6) x_2 = (C4 + C5 + C6) / 3

alpha = significance level of the test = 0.05
t(1-alpha/2,df) = value of student's t for which the distribution function value is 0.975 and degrees of freedom = number of observations - 2

$s = \text{sqrt}(s_1^2 + s_2^2)$

If |x1 - x2| / s > t(1-alpha/2,df) then the difference is statistically significant and the candidate standard has failed the initial stability test. The test can be repeated after an additional 7 days or more, using the second and third sets of results in the calculations, as above. If another significant difference is found, then the candidate standard is unusable and is disqualified for further use.

Example:

Initial Data	**Final Data (after 7-day holding time)**
0.995 ppm	0.989 ppm
0.996 ppm	0.989 ppm
0.992 ppm	0.982 ppm

s_1 = 0.0020817 ppm x_1 = 0.9943333 ppm

s_2 = 0.0040415 ppm x_2 = 0.9866667 ppm

s = 0.004546 ppm % difference = 0.77% ppm

|x1 - x2| / s = 1.686442

t(1-alpha/2,df) = 2.7764509

The difference is not statistically significant, so the standard can be certified as stable.

6. Recertification

Per Subsection 2.1.6.3, a standard can be recertified if, after the certification period has elapsed, the mean concentration of at least three assay results is within 1.0 percent of the original certified concentration. Additionally, the difference between the estimated mean and the certified concentration must not be statistically significant at the 1% level.

To determine whether the concentration of the standard has changed since the initial certification, new measurements are made using a measurement system that has been calibrated according to Subsection 2.1.7.5. Original certification data are used to provide an initial estimates of mean (x_1) and standard deviation (s_1). New data are used to estimate a second mean (x_2) and standard deviation (s_2). These are used in a t-test that is similar to that used in the stability test. A critical value for t is based on a significance level of 1% (alpha) and degrees of freedom equal to the number of initial and recertification data minus 2. A pooled

estimate of the standard deviation (s) is derived from s_1 and s_2. If the difference between x_1 and x_2, divided by s, is greater (in absolute value) than the critical value for t, then the initial and new concentrations are significantly different and the standard cannot be recertified.

Example:

Initial Data	**Recertification Data**
0.995 ppm	0.989 ppm
0.996 ppm	0.99 ppm
0.992 ppm	0.994 ppm
0.999 ppm	
0.999 ppm	
0.993 ppm	

s_1 = 0.0029439 ppm x_1 = 0.9956667 ppm

s_2 = 0.0026458 ppm x_2 = 0.991 ppm

s = 0.0028619 ppm % difference = 0.47%

|x1 - x2| / s = 1.6306179

t(1-alpha/2,df) = 2.7764509

The % difference is less than the 1% specification, and the difference in means is not found to be statistically significant. The standard may be recertified. The certified concentration of the standard is the grand mean of the combined data set.

Certified Concentration = mean (initial data + recertification data) = 0.994 ppm