Triclinic Labs

Chemical Development and Analysis Expertise

Chemometrics for Semi-Quant and Pure Curve Resolution

Simon Bates, Ph.D.

The Chemometric methods explored up until now have been general methods specifically oriented towards pattern matching and clustering. In addition to the pattern matching capabilities, to cover the range of applications at Triclinic labs, the chemometric approaches must also cover semi-quant analysis and pure curve resolution.

Semi-quant analysis is quantitative phase determination for mixed samples without the use of calibration or validation standards. Very often within the world of quantitative analysis of pharmaceutical systems, the introduction of standard samples introduces a significant amount of error into the quantitative calculations. This error arises as the standard samples are usually not representative of the real materials being analyzed. The real material will have typically been through a production line where it has been subjected to a number of different unit processing steps. The standards on the other hand will usually be made at small scale in a lab often using components from different suppliers. Most often, the standard samples are made by simple mixing of the individual components. In short, pharmaceutical materials often do not follow the linear response law (called Beer's law in absorption spectroscopy). In order for a series of mixed phase samples to follow the linear response theory, the measured signal must be a linear sum of the analytical responses of each of the phase components scaled according to the relative amount of that phase present in the sample. To satisfy the linear response law, three main conditions must be satisfied:

1.) The sample must be homogeneously mixed on the measurement length scale
2.) The analytical probe should have no interaction with the individual phase components except the one being measured. For a diffraction measurement, that means that there must be no significant absorption. And for an absorption spectroscopy measurement there must be no significant diffraction (scatter).
3.) The individual phase components must interact independently with the analytical probe. That is there must be no matrix induced effects (i.e. the sample is more than just the sum of the parts).

For pharmaceutical materials we typically will have matrix effects which for diffraction measurements are caused by local differences in the absorption of the X-rays caused by local inhomogeneity in the electron density of the sample (difference in particle sizes – presence of void space – poor packing). If the analytical method is being calibrated using standard mixtures with the same matrix components as the unknown samples, then the non-linear response can be calibrated out. The non-linear response, however, becomes a significant problem when the standard samples have a different solid state matrix than the unknown samples. For most pharmaceutical materials, the intensive processing tends to move the unknown samples toward an ideal solid state matrix with respect to the linear response requirements. It is the laboratory standard samples that are the main problem. A sample consisting of phases A+B+C may give different analytical responses depending on how the phase components are processed together and where the phase components were sourced. The variations in the analytical response are driven by micro-structure differences and matrix effects which are themselves driven by processing, handing and history (thermodynamic and kinetic) of the individual phase components.

To avoid having to use standard samples to calibrate a semi-quantitative method, the chemometric methods need to be based upon actual measured responses and not some abstract variance component. Methods like PCA and PLS will always need calibration as the computed differences between the measured data sets depends on abstract variables (often tied to the variance in a data ensemble). Pure Curve Resolution methods first proceed by extracting potential pure reference patterns from the mixed data ensemble. With pure curves extracted, the measured data can be quantified using standard full pattern fitting methods. If all the reference patterns have been extracted, the fitting problem is a classic least squares problem, while if only some of the reference patterns have been identified a Predominantly Positive fitting method can be used to extract semi-quant information.

As with all chemometric based analysis, the measured data must first be pre-processed. The obvious pre-processing steps include background removal, smoothing (if required), compression and normalizing. The usual normalization for XRPD data is to scale all patterns to give the same integrated intensity (area). A step beyond this is to normalize all patterns to a unit vector magnitude (i.e. sqrt_sum_of_squares = 1). With unit vector normalization, each powder pattern becomes a unit vector in as many dimensions as there are data points. (With compressed powder patterns, this is usually about 200 data points – which implies that the powder pattern is a unit vector over ~ 200 orthogonal dimensions.) The cross product (correlation) between any two powder patterns is, therefore, the cosine of the angle between the unit vectors in the ~ 200 dimensions (a measure of similarity). Binning is always a useful step not just for saving of computation overhead but after binning, every remaining data point represents about 10 actual measured data points and is a true observable with real information

content - obviously critical for chemometric methods that rely on individual observations. Background removal can be one of the most important pre-processing steps but can be very difficult to perform correctly. The primary reason for removing background is that the background signal scales differently than the actual diffraction signal. What you are trying to achieve is the removal of all background signals that do not arise from the sample (i.e air scatter, main beam bleed through etc..). Background that arises from the sample (i.e amorphous or defect contributions) can be left in the background signal as sample background contributions should scale correctly during normalization.

To investigate the semi-quant capabilities of chemometric methods without having to use standard materials, a series of mixed samples were generated with either 2 or 3 phase components. Nineteen binary samples were generated covering the complete analytical range. Fig. 1 below displays the first 10 binary powder patterns.

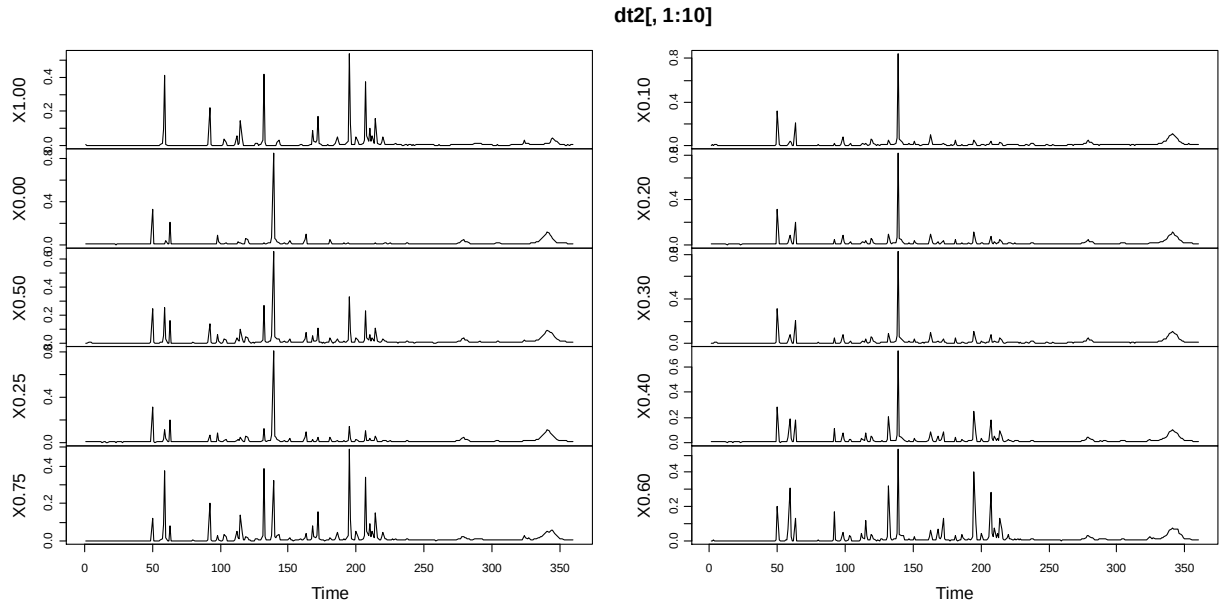

Figure 1: First 10 binary powder patterns. The X?.?? label on the y-axis is the nominal relative weight of phase A. The plotted data has been converted to unit – vector and step size binned from 0.02 to 0.16 degrees 2Theta. Instrumental background has been removed.

As a quick check that the data does indeed represent a binary system, a simple PCA step can be performed to identify the number of components in the variance. The result is shown in Fig. 2. The PCA derivation is also able to isolate the two principle components as factors in real space. That is a powder pattern representation of the principle variance components. This result is shown in Fig. 3. What is immediately apparent from Fig. 3 is that the factors do not really represent real powder patterns. The large negative going peaks are not allowed in reality but

are expected features for PCA. (The data should be mean centered for PCA which will give half the intensity as negative intensity.)

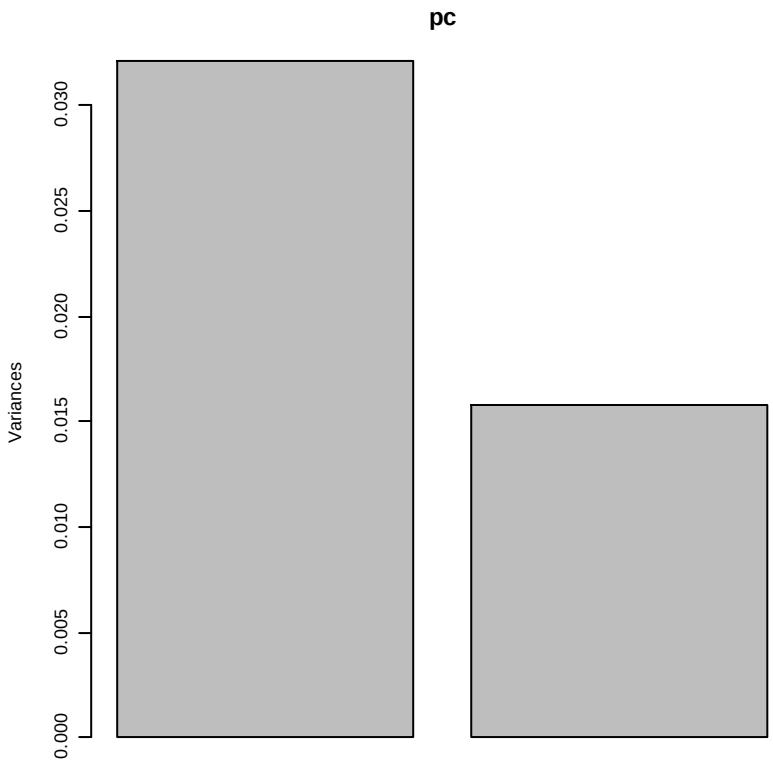

Figure 2: PCA step identifies that the variance in the data can be expressed in terms of 2 components (factors) which is as expected for binary data. pc <- prcomp(dt2, tol=0.1) – where dt2 is raw data matrix in unit vector format – plot(pc).

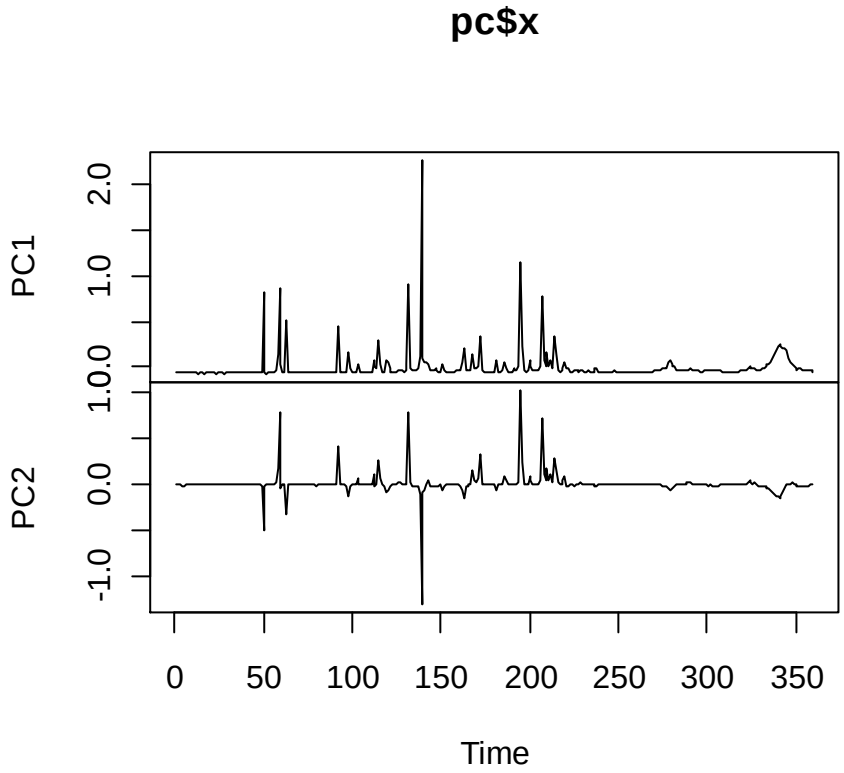

Figure 3: Principle components presented as factors of the measured data (x-axis should read data point #). These two components should ideally represent real powder patterns corresponding to the reference patterns. In this example, the factors of the PCA calculation are not real powder patterns. plot.ts(pc$x). The first principle component -> factor usually represents the 50:50 mixture for binary samples of the 1/3:1/3:1/3 sample for ternaries as these compositions represent the maximum variance. The second principle component -> factor will then be orthogonal to the 50:50 data set and may well have negative peaks. If the PCA could be constrained to select a real reference pattern as the first factor – the second factor will likely not have negative peaks. plot.ts(pc$x).

The table below, Tab. 1, presents the scale factors for PC1 and PC2 required to reconstruct the original data. Note that some of the scale factors are negative for PC2 which requires PC2 to be subtracted from PC1. In this table, the two pure phase reference patterns are entries 1 and 2 in the table. These have concentration X values of 1.0 and 0.0 which represents the relative weight of component A in the mixture. These two entries should represent the extremes of scale factors for PC1 and PC2. All other scale factors should be in-between these first two scaling values. The scaling of PC1 runs from 0.1755 to 0.2186 and the scaling of PC2 runs from

0.3175 to -0.259. These scale factors along with the nominal weight percentages can be used to establish a calibration line for quantitative analysis based upon PCA.

	PC1	PC2
X1.00	0.1755558	0.317507259
X0.00	0.2186390	-0.259011894
X0.50	0.2739518	-0.004547661
X0.25	0.2533361	-0.166778733
X0.75	0.2421723	0.189694557
X0.10	0.2322809	-0.229007006
X0.20	0.2464362	-0.190187181
X0.30	0.2464362	-0.190187181
X0.40	0.2700383	-0.078170246
X0.60	0.2685657	0.075551767
X0.70	0.2530828	0.153851986
X0.80	0.2297902	0.222337073
X0.90	0.2027371	0.276858219
X0.99	0.1781975	0.314012506
X0.98	0.1808710	0.310394641
X0.97	0.1835598	0.306653057
X0.01	0.2199592	-0.256347467
X0.02	0.2212916	-0.253610889
X0.03	0.2226323	-0.250808615

Table 1: Scale factors required for PC1 and PC2 to recreate the individual measured powder patterns (the calculated powder patters match the measured ones very well). pc$rotation – to determine back calculated patterns od <- pc$x %*% t(pc$rotation)

With all the data converted to a unit vector notation, the first obvious step is to take the cross product of each unit vector with respect to every other unit vector. The cross product will be the cosine of the angular separation between the two vectors. The distance between the vector endpoints can be expressed as 2sin(cp/2), where cp = inverse_cosine(cross product). The cross

product is of course the correlation function for multi-dimensional vectors. The correlation matrix for the 19 powder patterns is shown below in Tab. 2:

1.00	0.99	0.98	0.97	0.93	0.89	0.87	0.89	0.83	0.79	0.76	0.71	0.69	0.66	0.58	0.49	0.46	0.44	0.39
0.99	1.00	1.00	0.99	0.96	0.93	0.91	0.93	0.87	0.83	0.80	0.75	0.73	0.70	0.62	0.53	0.50	0.47	0.42
0.98	1.00	1.00	1.00	0.98	0.94	0.93	0.94	0.89	0.86	0.82	0.78	0.75	0.72	0.64	0.55	0.53	0.50	0.45
0.97	0.99	1.00	1.00	0.98	0.96	0.94	0.96	0.90	0.87	0.84	0.79	0.77	0.74	0.66	0.57	0.55	0.52	0.46
0.93	0.96	0.98	0.98	1.00	0.99	0.98	0.99	0.96	0.93	0.90	0.86	0.84	0.82	0.74	0.65	0.62	0.60	0.54
0.89	0.93	0.94	0.96	0.99	1.00	1.00	1.00	0.98	0.97	0.95	0.91	0.90	0.87	0.81	0.72	0.69	0.67	0.61
0.87	0.91	0.93	0.94	0.98	1.00	1.00	1.00	0.99	0.98	0.96	0.93	0.91	0.89	0.83	0.74	0.72	0.69	0.64
0.89	0.93	0.94	0.96	0.99	1.00	1.00	1.00	0.98	0.97	0.95	0.91	0.90	0.87	0.81	0.72	0.69	0.67	0.61
0.83	0.87	0.89	0.90	0.96	0.98	0.99	0.98	1.00	1.00	0.99	0.97	0.95	0.94	0.89	0.81	0.78	0.76	0.70
0.79	0.83	0.86	0.87	0.93	0.97	0.98	0.97	1.00	1.00	1.00	0.98	0.97	0.96	0.92	0.85	0.82	0.80	0.75
0.76	0.80	0.82	0.84	0.90	0.95	0.96	0.95	0.99	1.00	1.00	1.00	0.99	0.98	0.94	0.88	0.86	0.84	0.78
0.71	0.75	0.78	0.79	0.86	0.91	0.93	0.91	0.97	0.98	1.00	1.00	1.00	0.99	0.97	0.92	0.90	0.87	0.82
0.69	0.73	0.75	0.77	0.84	0.90	0.91	0.90	0.95	0.97	0.99	1.00	1.00	1.00	0.98	0.93	0.92	0.89	0.85
0.66	0.70	0.72	0.74	0.82	0.87	0.89	0.87	0.94	0.96	0.98	0.99	1.00	1.00	0.99	0.95	0.93	0.91	0.87
0.58	0.62	0.64	0.66	0.74	0.81	0.83	0.81	0.89	0.92	0.94	0.97	0.98	0.99	1.00	0.98	0.97	0.96	0.91
0.49	0.53	0.55	0.57	0.65	0.72	0.74	0.72	0.81	0.85	0.88	0.92	0.93	0.95	0.98	1.00	1.00	0.99	0.96
0.46	0.50	0.53	0.55	0.62	0.69	0.72	0.69	0.78	0.82	0.86	0.90	0.92	0.93	0.97	1.00	1.00	1.00	0.97
0.44	0.47	0.50	0.52	0.60	0.67	0.69	0.67	0.76	0.80	0.84	0.87	0.89	0.91	0.96	0.99	1.00	1.00	0.98
0.39	0.42	0.45	0.46	0.54	0.61	0.64	0.61	0.70	0.75	0.78	0.82	0.85	0.87					

Table 2: Correlation matrix for the 19 powder patterns representing binary mixtures. Each entry is the cosine of the angle separating the two unit vectors. In this matrix, the powder patterns have been sorted according to the relative weight of Form A present in the mixture. The first entry is 0% A while the last entry is 100% A. The angular cosine between the two pure reference patterns is therefore 0.39 (~67 degrees). This must be the smallest angular cosine and all other powder patterns must have angles below this for a binary system.

Using classic multi-dimension scaling, the distances between the endpoints of all the unit vectors can be converted into physical locations to place each mixed sample. Fig. 4 below

shows a plot of the 2D projection of all the directional cosines expressed as distances for the 19 binary samples.

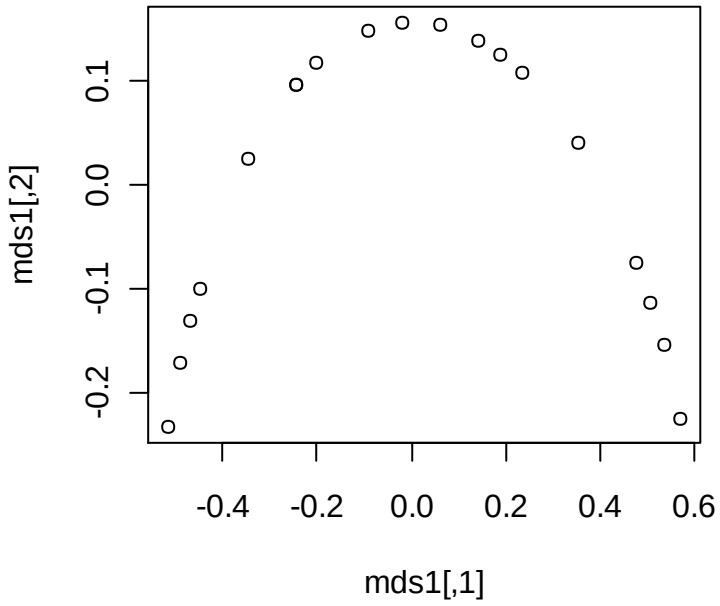

Figure 4: Classic Multi-Dimension scaling applied to the distances between the end points of all the unit vectors determined from the 19 powder patterns of the mixed samples. The plot follows a well-defined curve which could be used as a 'calibration – line' determined from the ensemble of data without the use of standards.

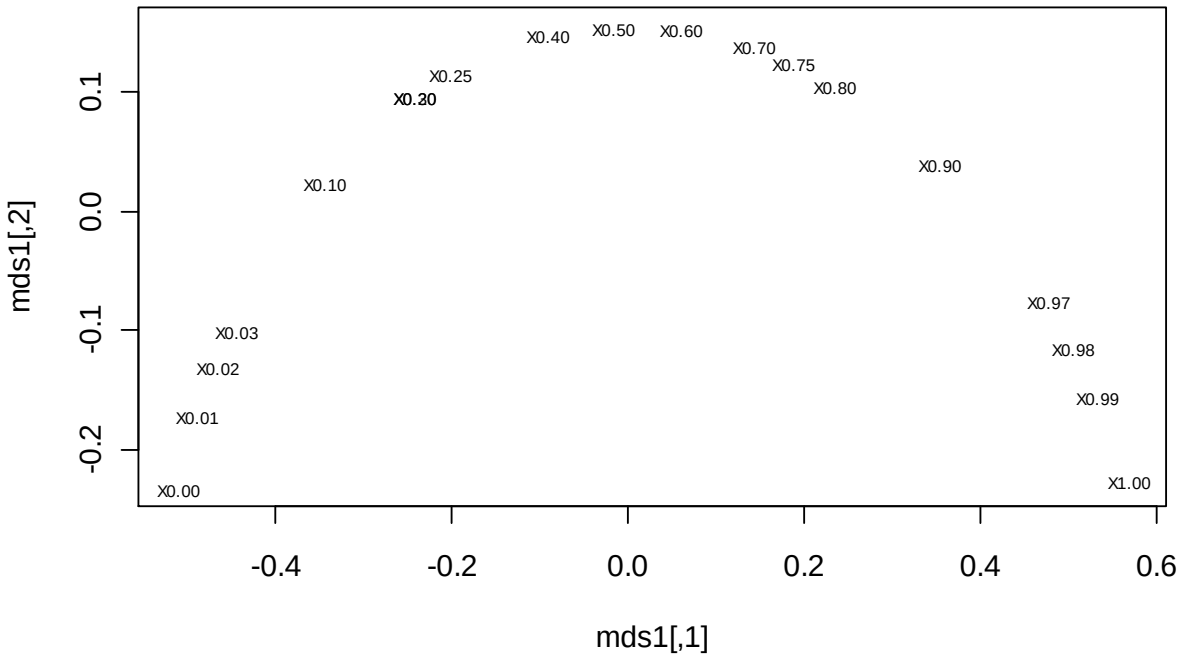

Figure 5: the plot points are replaced by the X label which indicates the nominal relative weight of Form A. The curve shows a continuous evolution from X=0.0 through to X=1.0. The continuous curve functions like a calibration curve. The curve is derived without reference to any standard samples.

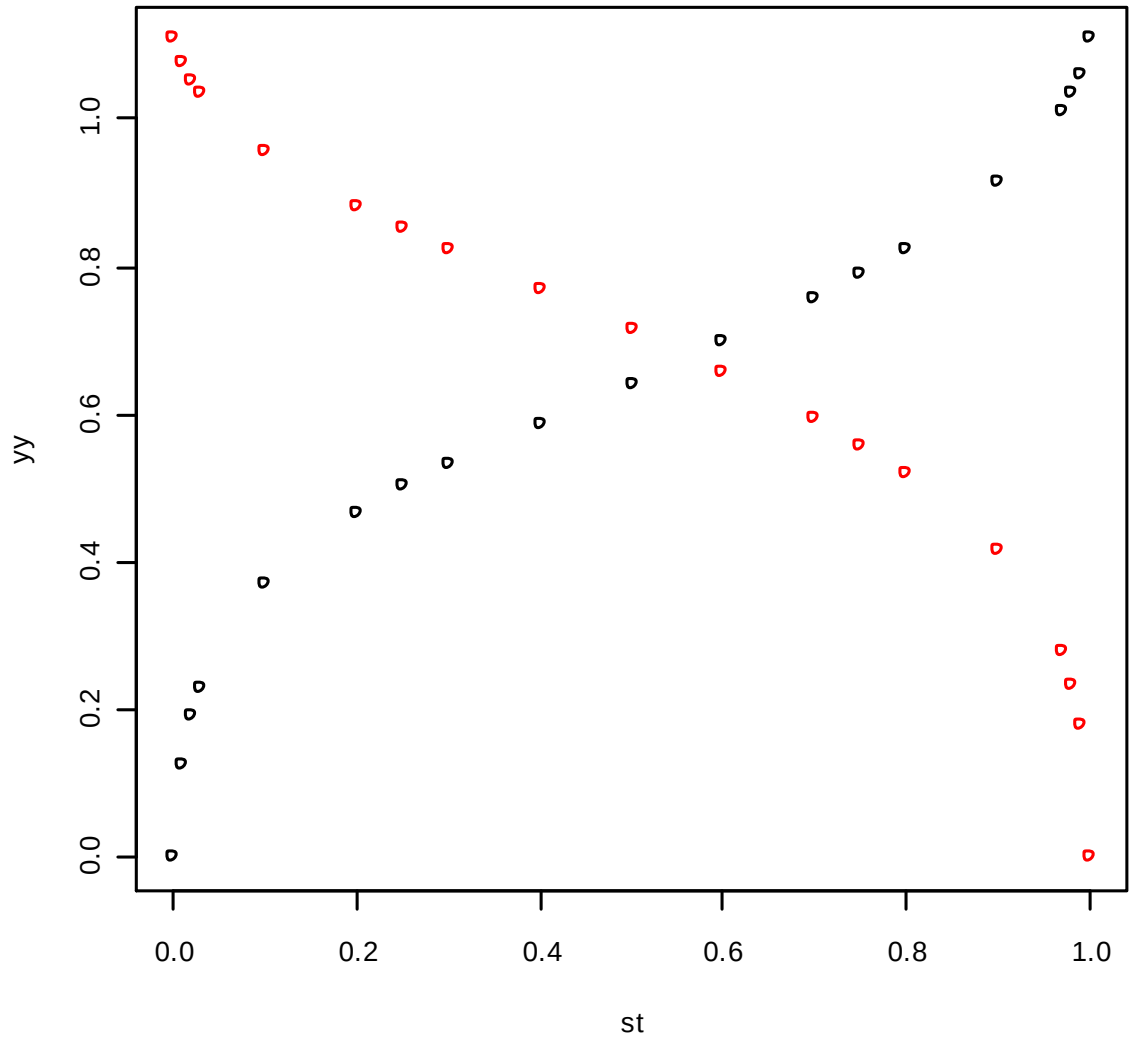

Figure 6: The black points show the directional cosine converted to distance from the Form A reference pattern as a function of relative weight of Form A present. The red points show the same plot but this time from the Form B reference pattern. The linear response region stretches from about 15% through to about 85%. This plot clearly demonstrates the significant non-linearity present at the two concentration extremes close to 0% and 100%.

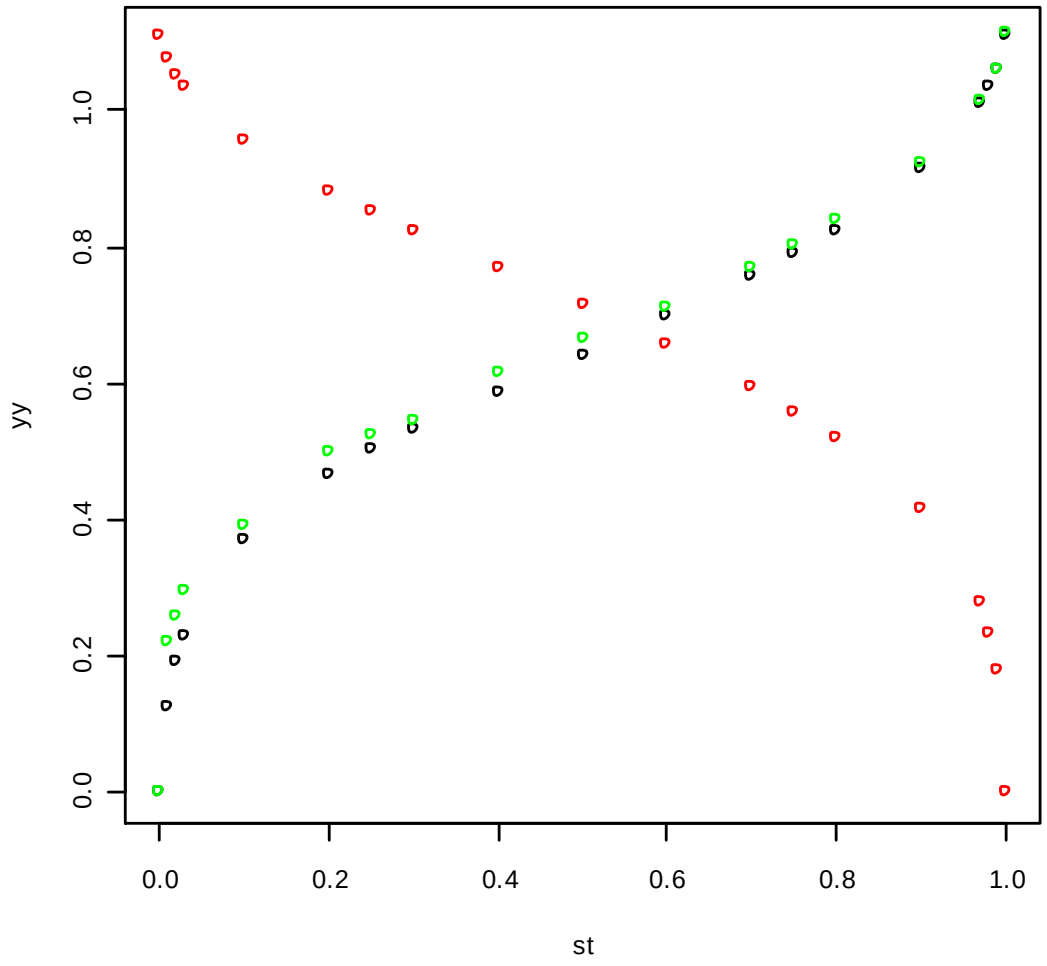

Figure 7: The increase of random noise in the data (green points compared with black points) will lead to a displacement of the individual distances causing a blurring of the calibration line. The blurring increases towards the low concentration end which is to be expected.

For more information or to discuss this White Paper, please contact:

Simon Bates, Ph.D.
Research Fellow
Triclinic Labs, Inc.

1-765-588-5632
sbates@tricliniclabs.com