File: /General/MLAB-Text/Papers/svdpaper/svd.tex

Analysis of Absorption Spectra-Titration Data

Gary D. Knott, Ph.D.
Civilized Software, Inc.
12109 Heritage Park Circle
Silver Spring MD 20906
Tel. (301) 962-3711
email: csicivilized.com
URL: http://www.civilized.com

Abstract: A powerful data analysis method based on the singular value decomposition is presented. An example of this method applied to data consisting of an absorption spectra surface is explained in detail. The MLAB mathematical modeling system is a suitable tool for doing a singular value decomposition analysis, and an example of this is presented.

Suppose we have a mixture of substances which we carry through a sequence of changes by varying some state variable. This process is called a *titration* with respect to the state variable, and the successive values of the state variable are called the titration levels. At each titration level, we measure some properties of the mixture; this data can then be modeled by a suitable mathematical description of the chemistry involved. When the measured properties are themselves a spectrum of responses to some sequence of stimuli, such as an absorption spectrum sampled over a range of wavelengths, a particularly elegant method of analysis called the "Singular-Value Decomposition" (SVD) method employed in this context by Richard Shrager and Richard Hendler can be used [1]. The stimulus could even consist of the passage of time; thus kinetics data may be analyzed with this method. We will describe this method below in the context of an example with a titration with respect to hydrogen ion level, pH. (Hendler's actual experiments involved a titration with respect to electron content (i.e., voltage) corresponding to changes in the component substances of the respiratory chain in a mitochondrial membrane.)

For our example, suppose we have data values in a matrix $A[1\colon m, 1\colon n]$ such that A_{ik} is the absorbance at wavelength wl_i of the mixture at titration pH-level pH_k. We will take $m = 70$ and $n = 27$. The 70 wavelengths are equally spaced in the range from 350 nm to 650 nm. The 27 pH levels are equally-spaced in the range from pH 3 to pH 11.

We suppose there are $t-1$ substances in the mixture which appear in greater concentration as the pH level is increased. The value $t-1$ is unknown and is to be estimated from the data. The relative fraction of the jth substance is given by the Henderson-Hasselbalch function:

$$f_j(pH) = 1/(1+10^{(pK_j - pH)}),$$

where pK_j is the pH at which substance j's concentration is 50% of its maximum value; f_j is the *transition function* corresponding to the transition curve of substance j. With a more general choice of f_j, it would be possible to model the disappearance of substance j, as well as its appearance. Let c_j denote the maximum concentration of substance j that occurs at "infinite" pH. We also suppose there is an unchanging "background" complex of materials that we call "substance t". The transition function of substance t is just $f_t(pH) = 1$, and its limiting concentration is the constant c_t.

Substance t and the other $t-1$ substances are each assumed to have a characteristic absorption spectrum; $g_j(wl)$ denotes the absorbance seen for 1 unit of substance j at wavelength wl. Often, but not always, g_j is approximately a Gaussian bell-shaped curve centered at a characteristic central wavelength. Absorption spectra are assumed to be additive; the spectrum for several substances is the sum of the spectra for each individual substance. Now we can compute the overall absorbance b_{ik} at wavelength wl_i and pH-level pH_k as:

$$b_{ik} = \sum_{j=1}^{t} g_j(wl_i) c_j f_j(pH_k), \quad \text{for} \quad 1 \le i \le m \quad \text{and} \quad 1 \le k \le n.$$

Here, $c_j f_j(pH_k)$ is the concentration of substance j at pH-level pH_k, and $g_j(wl_i)c_j f_j(pH_k)$ is the absorbance at wavelength wl_i due to this concentration of substance j. When the pK_j values appearing in the f_j functions are correctly specified, b_{ik} should match the data value A_{ik}.

We can express this model in matrix form as follows. Let F be an $n \times t$ matrix with $F_{kj} = f_j(pH_k)$. Note that F col $t = 1$ identically. Let D be an $m \times t$ matrix with $D_{ij} = g_j(wl_i)c_j$. Note that D col t is the baseline spectrum at the m wavelengths being monitored. Let B be the $m \times n$ matrix with $B_{ik} = b_{ik}$. Then $B = DF^T$, and $B \approx A$. (We assume B would equal A exactly, except for measurement error.)

Given the data A, we wish to estimate t, and $pK_1, \ldots, pK_{t-1}$, and D_{ij} for $1 \le i \le m$ and $1 \le j \le t$, and F_{kj} for $1 \le k \le n$ and $1 \le j \le t$. This may be done using the matrix singular-value decomposition. Note that if we know $pK_1, \ldots, pK_t$, we then know the elements of the matrix F, and conversely.

Any $m \times n$ real matrix A can be written as $A = USV^T$, where U is an $m \times m$ orthogonal matrix, S is an $m \times n$ diagonal matrix with decreasing non-negative diagonal values called the singular values of A, and V is an $n \times n$ orthogonal matrix. The number of positive singular values exhibited in S is the rank of A.

In our example, we thus write $A = USV^T$. Now we discard the zero or almost zero singular values from S, leaving t positive singular values in the modified $t \times t$ diagonal matrix S. We have thus determined the value t; recall that the number of non-background substances is $t - 1$. We also discard the last $m - t$ columns of U and the last $n - t$ columns of V that correspond to the discarded singular values to obtain $U = U$ col $1 : t$ and $V = V$ col $1 : t$. We now have $A \approx USV^T$. "Almost zero" can be defined as less than about 1% to 2% of the maximum singular value; however, this can be a dangerous assumption. Shrager suggests we check the noise apparent in the corresponding columns of V; only singular values corresponding to significantly noisy columns of V should be discarded. (By noisy, we mean the graph of V col j with respect to $1 : n$ is jagged.)

Now we may factor V as $V = FH$ where H is a $t \times t$ matrix, so $A = USH^T F^T$, and hence we may obtain $D = USH^T$, since $A = DF^T$. In order to factor V into explicitly-known matrices F and H, we use curve-fitting. Let:

$$q_j(x) = H_{tj} + \sum_{r=1}^{t-1} H_{rj} f_r(x; pH_r)$$

Note that $[q_j(pH_1), \ldots, q_j(pH_k)]^T = V$ col j. Thus we may estimate $H_{1j}, H_{2j}, \ldots, H_{(t-1)j}, H_{tj}$, and $pK_1, \ldots, pK_{t-1}$ by fitting the model q_j to the $n \times 2$ data matrix $X_j := (3 : 11!27)\&'(V \text{ col } j)$, for $j = 1, \ldots, t$. (Recall that in our example we are using 27 equally-spaced pH-values between 3 and 11.) This is best done by simultaneously fitting the functions q_j to the data points X_j using the constant weight values $(S_{jj})^2$ for $1 \leq j \leq t$.

Now having estimated t, and $pK_1, \ldots, pK_{t-1}$, and the t^2 elements of H, we may directly obtain the matrix F and then D.

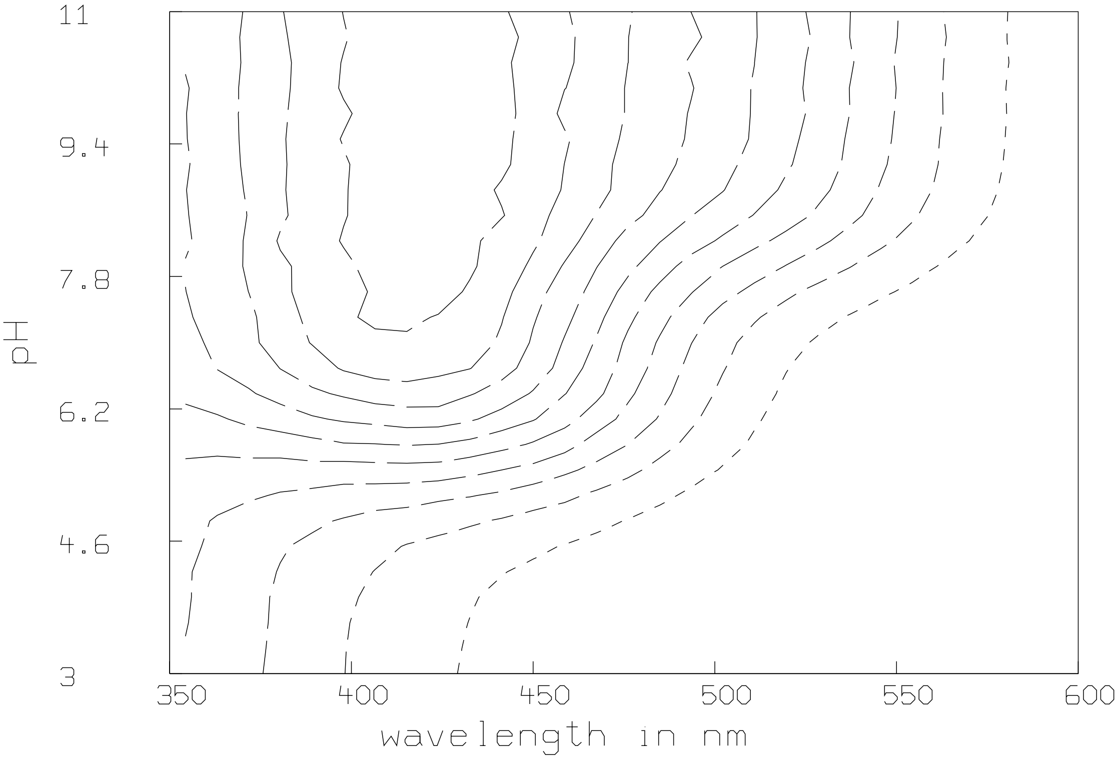
The columns of D are the spectra of the t substances taken at their maximum concentrations; thus $(Dcolj)^T = c_j[g_j(wl_1), \ldots, g_j(wl_m)]$. If we know these spectral curves independently, we can deduce the concentration scale factors $c_1, \ldots, c_t$. Looking at these curves also serves as a check on the entire modeling process.

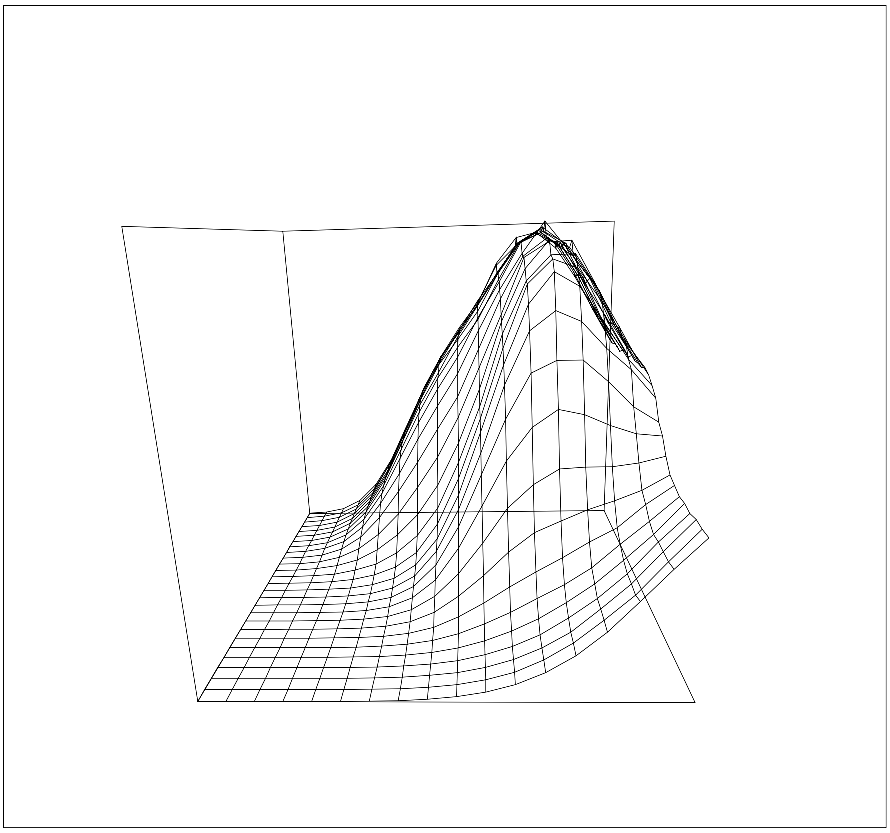
MLAB is an advanced mathematical and statistical modeling system, originally developed at the NIH, that is particularly suited for handling modeling and estimation problems in physiology and analytical chemistry among other domains (see `www.civilized.com`. MLAB contains all the facilities needed to carry out the SVD analysis described above. Below is a particular MLAB dialog showing the SVD method applied for our example.

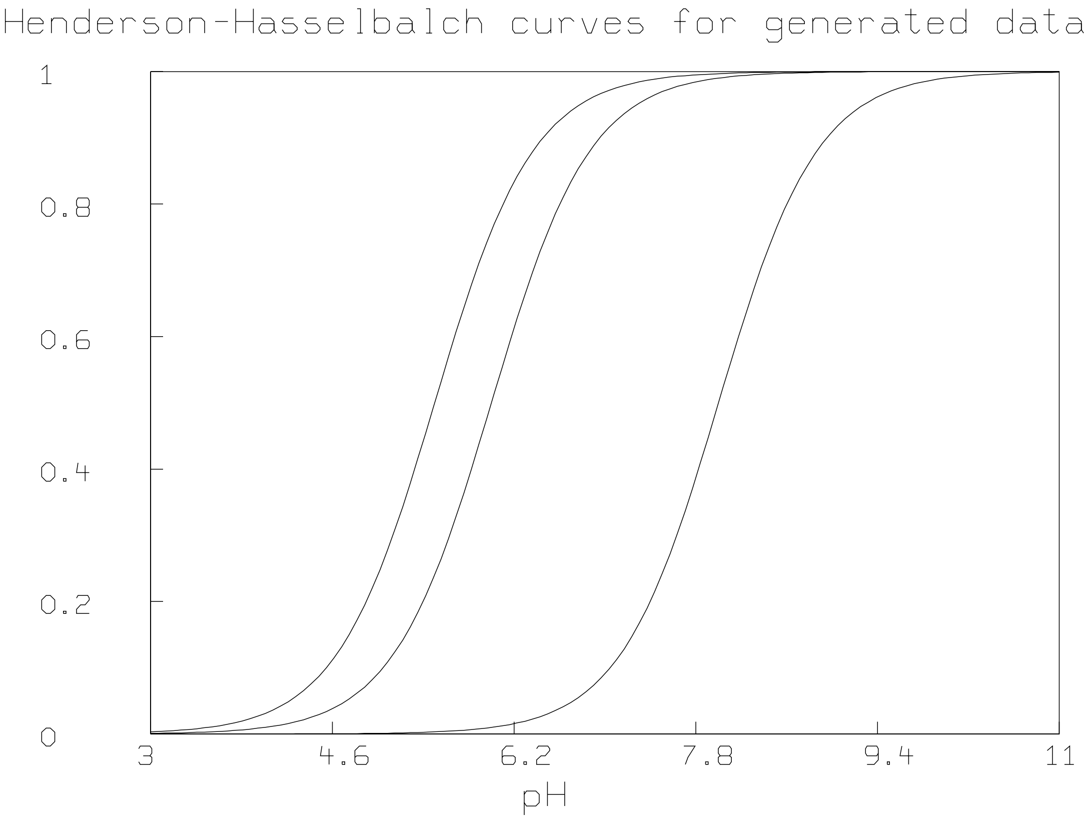
First we read in our data matrix A from a text-file `dataf`, where A is a 70×27 matrix whose rows correspond to the wavelengths $350 : 600!70$ (nm), and whose columns correspond to the pH values $3 : 11!27$. A_{ij} is the absorbance observed at wavelength-level i and pH-level j.

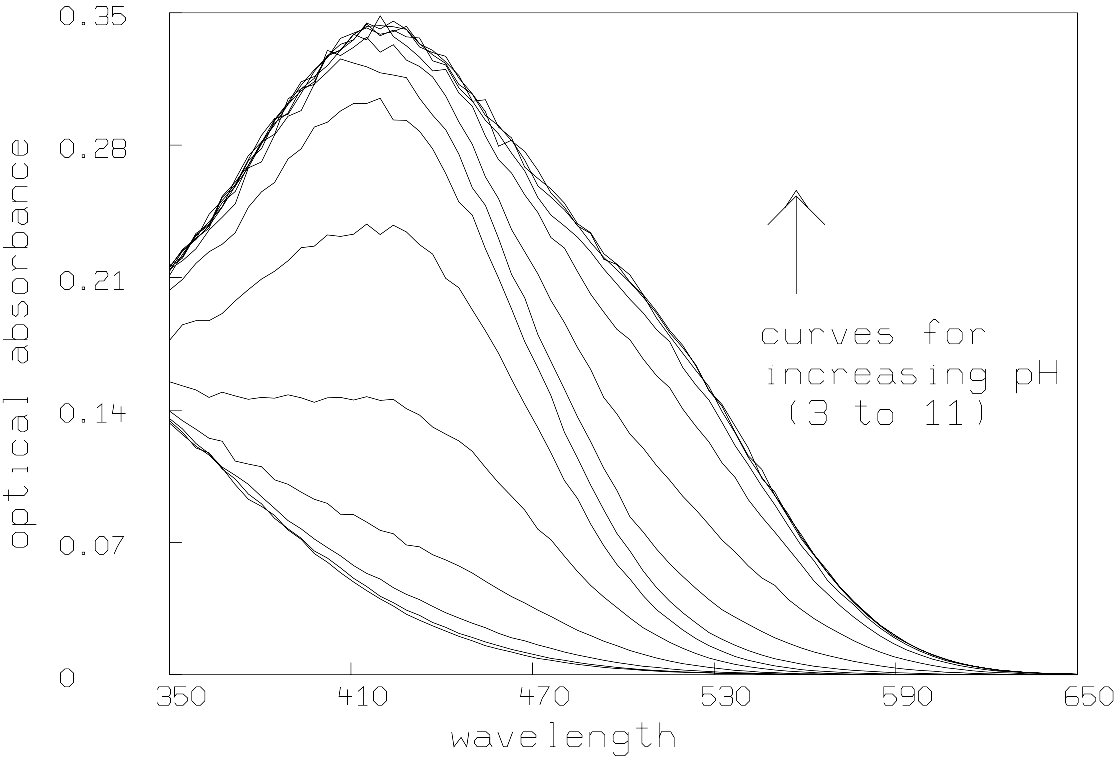
```
* a = read(dataf,70,27)
```

Some graphs which show the nature of the data A are given below. All of these graphs were drawn with MLAB.





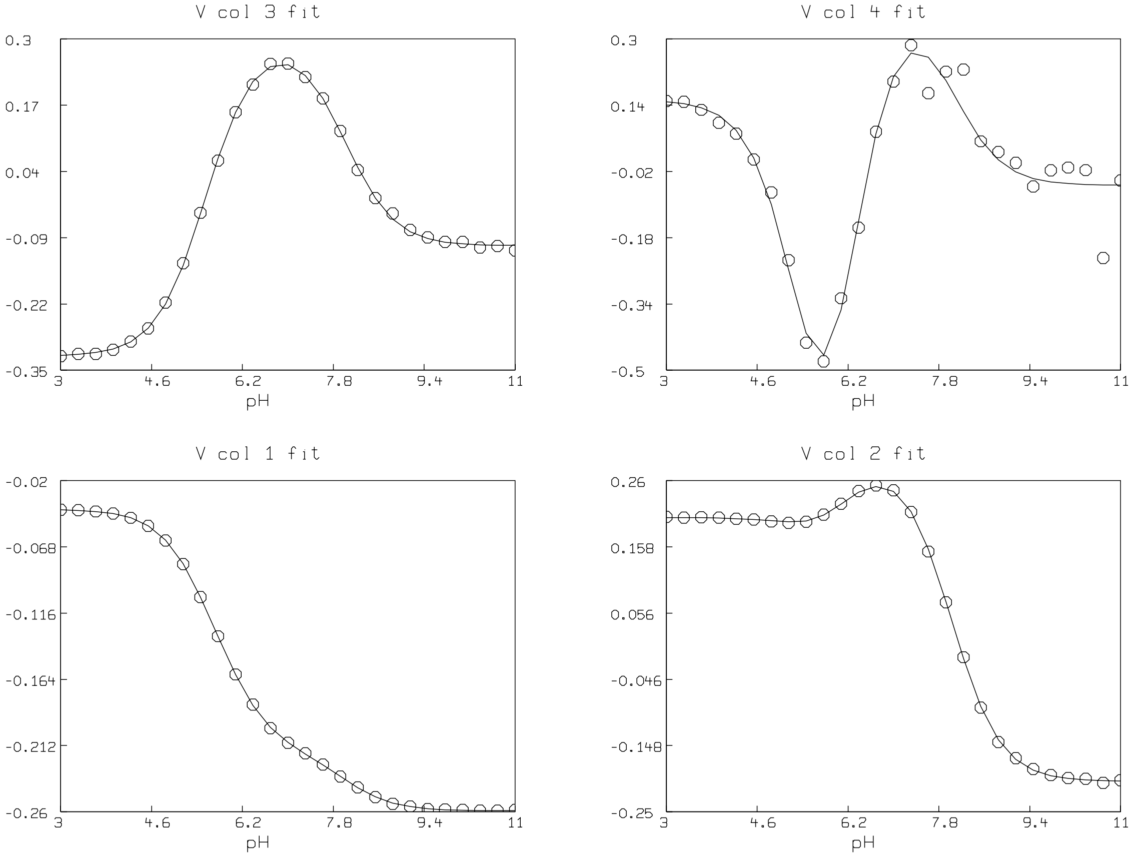

Henderson-Hasselbalch curves for generated data
1
0.8
0.6
0.4
0.2
0
3
4.6
6.2
7.8
9.4
11
pH

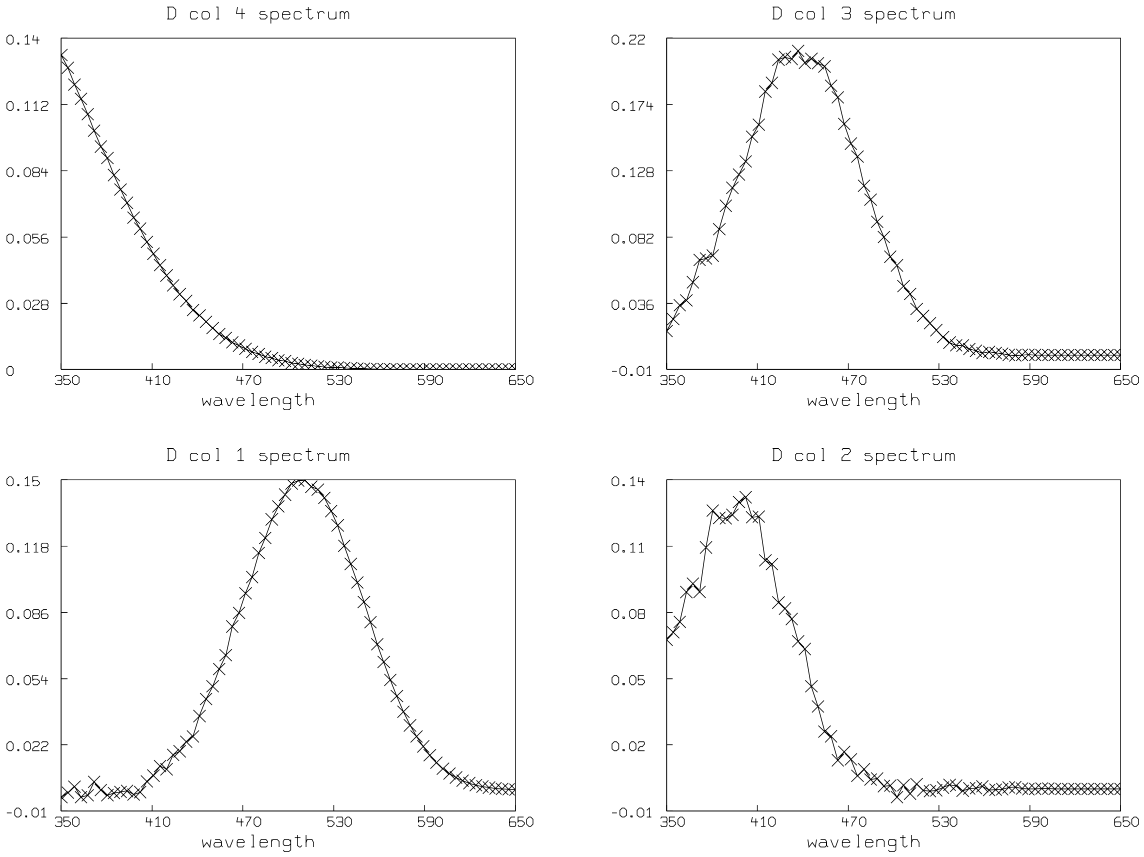


Now we compute the singular value decomposition of A and decide the value of t.

```
* m=nrows(a); n=ncols(a)
*
* u=svd(a)
* s=(u row 1)'
* v=u row (m+2):(n+m+1)
* u = u row 2:(m+1)
* t=nrows(compress(((s col 1)>.025)&'s));
* type s row 1:(t+3); "type the singular values 3 beyond the t-th"

    : a  7 by 1 matrix

  1: 6.90549919
  2:  .964371539
  3:  .458168815
  4:  .0474121723
  5:  .023760869
  6:  .0221695634
  7:  .020740888

* v=v col 1:t; u= u col 1:t
*
* fct f1(ph)=1/(1+10^(pk1-ph))
* fct f2(ph)=1/(1+10^(pk2-ph))
* fct f3(ph)=1/(1+10^(pk3-ph))
* fct b1(ph)=h[t,1]+h[1,1]*f1(ph)+h[2,1]*f2(ph)+h[3,1]*f3(ph)
* fct b2(ph)=h[t,2]+h[1,2]*f1(ph)+h[2,2]*f2(ph)+h[3,2]*f3(ph)
* fct b3(ph)=h[t,3]+h[1,3]*f1(ph)+h[2,3]*f2(ph)+h[3,3]*f3(ph)
* fct b4(ph)=h[t,4]+h[1,4]*f1(ph)+h[2,4]*f2(ph)+h[3,4]*f3(ph)
*
* h=.3*shape(4,4,ran on 0^^16); "generate random guesses for h"
* pk1= 7; pk2 = 7.2; pk3= 6; "choose guesses for pk1,pk2,pk3"
*
* wl=350:650!70;
* phv=3:11!27
*
* maxiter=30; symdsw=0
* fit(h,pk1,pk2,pk3),b1 to phv&'(v col 1) with wt s[1]^^n,\
> b2 to phv&'(v col 2) with wt s[2]^^n,\
> b3 to phv&'(v col 3) with wt s[3]^^n,\
> b4 to phv&'(v col 4) with wt s[4]^^n
final parameter values
      value              error                 dependency    parameter
 -0.04313096632      0.002075088787         0.8681882157   H[1]
  -0.5091112274      0.006380218701         0.9001594031   H[2]
  -0.4141844706      0.008460681648          0.880494897   H[3]
  -0.5124442081       0.02935544387         0.9040697608   H[4]
 -0.06083337034       0.01144532059         0.9974810125   H[5]
   0.2027353714       0.03290432573         0.9978176388   H[6]
  0.00747866897       0.07009215024         0.9989876919   H[7]
    2.848373276        0.4970101962         0.9998054387   H[8]
  -0.1143730484       0.01060867204         0.9973495801   H[9]
 -0.09979224416       0.03273355056          0.998006569   H[10]
   0.6236741131       0.06485462555         0.9989311313   H[11]
   -2.542220886        0.5055039097         0.9998299828   H[12]
 -0.04089842681       0.0009856221674          0.8092088134   H[13]
   0.2031173917       0.002508541316             0.78909474   H[14]
  -0.3226308862       0.003843686277         0.8109168102   H[15]
    0.152892387        0.01220313549         0.8187234648   H[16]
    8.006947711        0.01676401014          0.720544985   PK1
    6.058377382        0.05911112374         0.9828623017   PK2
    5.553356498        0.04694757727         0.9790372858   PK3
20 iterations
CONVERGED
best weighted sum of squares = 3.075587e-03
weighted root mean square error = 5.878534e-03
weighted deviation fraction = 3.081091e-03
R squared = 9.844744e-01
```

We have estimated t $(= 4)$ and $pK1$, $pK2$, and $pK3$ ($pK1 = 8.007$, $pK2 = 6.058$, $pK3 = 5.553$). We have also computed the matrix D. Now we present some pictures drawn in MLAB showing the results after this





Although the fit we show here is very good, in general the fit we obtain can be very sensitive to the initial guesses for the parameters. Shrager and Hendler have developed a collection of heuristic techniques to acquire reasonable initial guesses. Also, the fact that the H_{ij} parameters appear linearly means that a two-stage fitting approach consisting of separately fitting just the linear and then the nonlinear parameters could be used [2].

Information about MLAB is available at *www.civilized.com*

[1] Hendler R, Shrager R.; "Analysis of the Spectra and Redox Properties of the Pure Cytochromes aa3", Biophysical J. Vol. 49, pp. 717–729, March 1986.

[2] Golub and Peryra, "Separable Least-Squares", SIAM J. of Numerical Analysis, Vol. 10, No. 2, pp. 413:432, April 1973.