Editor: Dianne P. O'Leary, oleary@cs.umd.edu

BLIND DECONVOLUTION: A MATTER OF NORM

By Dianne P. O'Leary

WE CONTINUE THE SPECTROSCOPY PROBLEM FROM THE LAST ISSUE, TRYING TO RECONSTRUCT A TRUE SPECTRUM FROM AN OBSERVED ONE. AGAIN, WE'LL USE blind deconvolution, but this time we'll impose some constraints on the error matrix $\boldsymbol{E}$, leading to a more difficult problem to solve but often a more useful reconstruction.

Counts

Recall that we have the counts of Figure 1, measured by a spectrometer (to view the figure, see the solution to last issue's homework on page 63). Suppose we have particles whose energy ranges from e_{lo} to e_{high}, and we define some intermediate energy levels $e_{lo} = e_0 < e_1 < \ldots < e_{n_b-1} < e_{n_b} = e_{high}$. This creates n_b bins, where the count for the jth bin is the number of particles determined to have energies between e_{j-1} and e_j. Our spectrometer records n_b counts, one for each bin.

One way to model this system is to try to determine the correct counts f_j, the correct blurring given the measured counts g_j, $j = 1, \ldots, n_b$.

This gives us a matrix equation $(\boldsymbol{K} + \boldsymbol{E})\mathbf{f} \approx \mathbf{g} + \mathbf{r}$, where $\boldsymbol{E}$ accounts for errors in modeling the spectrometer's blur, and $\mathbf{r}$ accounts for errors in counts. The matrix entry $k_{j\ell}$ is the probability that a particle whose energy is in the interval $[e_{\ell-1}, e_\ell]$ is assigned to bin j ($j, \ell = 1, \ldots, n_b$).

We assume that there is significant error in both $\mathbf{g}$ and $\boldsymbol{K}$, but we note that in our data, each bin's properties are the same, so the rows of $\boldsymbol{K}$ have a pattern: for example, if the $m \times n$ matrix $\boldsymbol{K}$ were 5×5, we would notice that

$$\boldsymbol{K} = \begin{bmatrix} k_5 & k_4 & k_3 & k_2 & k_1 \\ k_6 & k_5 & k_4 & k_3 & k_2 \\ k_7 & k_6 & k_5 & k_4 & k_3 \\ k_8 & k_7 & k_6 & k_5 & k_4 \\ k_9 & k_8 & k_7 & k_6 & k_5 \end{bmatrix},$$

so there would be only $m + n - 1 = 9$ distinct entries in $\boldsymbol{K}$. A matrix of this form is called a *Toeplitz matrix*, and it is determined by the entries in its first row and column. Under this assumption, it might make sense to assume that the error matrix $\boldsymbol{E}$ also has this same structure, and therefore also depends on $m + n - 1$ parameters instead of mn. We will gather these parameters in a vector called $\hat{\mathbf{e}}$.

Using the Euclidean Norm

In the previous homework, we solved the problem

$$\min_{\mathbf{E},\mathbf{r},\mathbf{f}} \frac{1}{2}\|\boldsymbol{E}\|_F^2 + \frac{1}{2}\|\mathbf{r}\|_2^2, \tag{1}$$

where

$$\mathbf{r} = \mathbf{g} - (\boldsymbol{K} + \boldsymbol{E})\mathbf{f}. \tag{2}$$

With our new constraint that $\boldsymbol{E}$ be Toeplitz, our old solution isn't feasible, so we minimize the function, subject to the constraint, over all choices of $\mathbf{f}$ and $\hat{\mathbf{e}}$.

Our goal is to find an effective algorithm to solve this problem, and we'll do it in several steps. The first is to derive a useful alternative expression for the matrix-vector product $\boldsymbol{E}\mathbf{f}$.

PROBLEM 1.

Show that $\boldsymbol{E}\mathbf{f}$ can be written as $\boldsymbol{F}\hat{\mathbf{e}}$, where $\hat{\mathbf{e}}$ is the vector that has entries $\hat{\mathbf{e}}_i$, and $\boldsymbol{F}$ is a matrix whose entries depend on the entries in the vector $\mathbf{f}$. In other words, find a matrix $\boldsymbol{F}$ so that $\boldsymbol{E}\mathbf{f} = \boldsymbol{F}\hat{\mathbf{e}}$.

Let's use Newton's method to solve our minimization problem. Recall that if we minimize some function $s(\mathbf{x})$, then the Newton direction is the solution $\mathbf{p}$ to the linear system

$$\boldsymbol{H}(\mathbf{x})\mathbf{p} = -\nabla s(\mathbf{x}),$$

where $\nabla s(\mathbf{x})$ is the gradient of s with respect to $\mathbf{x}$, and $\boldsymbol{H}(\mathbf{x})$

is the *Hessian* matrix, containing the second derivatives: $h_{ij}(x) = \partial^2 s(\mathbf{x})/\partial x_i \partial x_j$. Let's derive a formula for $\mathbf{p}$.

Problem 2.

Derive the Newton direction for Equation 1. To do this, use the definitions of $\boldsymbol{E}$ (in terms of $\hat{\mathbf{e}}$) and $\mathbf{r}$ (Equation 2), and then differentiate the function in Equation 1 with respect to $\hat{\mathbf{e}}$ and $\mathbf{f}$.

Although the formula from Problem 2 is mathematically correct, it isn't the best computationally; Problem 3 provides a better alternative.

Problem 3.

Show that this Newton direction is approximately the same as the solution to the least squares problem

$$\min_{\Delta\hat{\mathbf{e}},\Delta\mathbf{f}} \left\| \begin{bmatrix} \boldsymbol{F} & \boldsymbol{K}+\boldsymbol{E} \\ \boldsymbol{D} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \Delta\hat{\mathbf{e}} \\ \Delta\mathbf{f} \end{bmatrix} + \begin{bmatrix} -\mathbf{r} \\ \boldsymbol{D}\hat{\mathbf{e}} \end{bmatrix} \right\|_2,$$

where $\boldsymbol{D}$ is a diagonal matrix of size $(m+n-1)\times(m+n-1)$ with entries equal to the square roots of $1, 2, \ldots, n, \ldots, n, n-1, \ldots, 1$. (In particular, the least squares solution is very close to the Newton direction if the model is good, so that $\|\mathbf{r}\|$ is small.)

If we were to use this model on our spectroscopy data, the solution would be quite contaminated by error. Therefore, we make one further modification: we solve the problem

$$\min_{\boldsymbol{E},\mathbf{f}} \frac{1}{2}\|\boldsymbol{E}\|_F^2 + \frac{1}{2}\|\mathbf{r}\|_2^2 + \frac{1}{2}\lambda^2\|\mathbf{f}\|_2^2,$$

where the last term is a *Tikhonov regularization* term (as in "Image Deblurring: I Can See Clearly Now," vol. 5, no. 3, May/June 2003, pp. 82–84), with a fixed parameter λ, added to control the size of $\mathbf{f}$. In this case, the approximate Newton direction is computed from the solution to the least squares problem

$$\min_{\Delta\hat{\mathbf{e}},\Delta\mathbf{f}} \left\| \begin{bmatrix} \boldsymbol{F} & \boldsymbol{K}+\boldsymbol{E} \\ \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \lambda\boldsymbol{I} \end{bmatrix} \begin{bmatrix} \Delta\hat{\mathbf{e}} \\ \Delta\mathbf{f} \end{bmatrix} + \begin{bmatrix} -\mathbf{r} \\ \boldsymbol{D}\hat{\mathbf{e}} \\ \lambda\mathbf{f} \end{bmatrix} \right\|_2,$$

Tools

You can use Matlab's `linprog` to solve the linear programming problems.

Choices of vector norms are discussed in many elementary textbooks, but James Ortega gives a particularly nice discussion.[1]

Armin Pruessner and I discuss the use of regularization plus norm choice.[2] References to earlier work using regularization or norm choice can be found in that article, too.

References

1. J.M. Ortega, *Numerical Analysis: A Second Course*, Academic Press, 1972 (reprinted by SIAM Press).
2. A. Pruessner and D.P. O'Leary, "Blind Deconvolution Using a Regularized Structured Total Least Norm Approach," *SIAM J. Matrix Analysis and Applications*, vol. 24, 2003, pp. 1018–1037.

Now we put these pieces together to solve our problem.

Problem 4.

Write a function `[f,ehat,r,itn] = stls(K,g,lambda,tol)` that uses a variant of Newton's method to solve our Toeplitz-constrained problem in a stable and efficient way. Use the least squares problem from earlier to compute the approximate Newton direction. Start the iteration with $\hat{\mathbf{e}} = \mathbf{0}$ and $\mathbf{f} = \mathbf{0}$. Stop the iteration when the norm of the approximate Newton step is smaller than `tol`, and set `itn` to the number of iterations. Provide documentation for your function. Use it on the data from the Web site (www.computer.org/cise/homework), setting $\lambda = 0.06$ and `tol` $= 10^{-3}$. Plot the solution, and print the residual norm, the solution norm, and the number of iterations.

Using Other Norms

If the errors in our data aren't distributed normally, we have several reasonable alternatives to the choice of the Euclidean norm for the minimization function. For example, instead of minimizing

$$\frac{1}{2}\|\boldsymbol{E}\|_F^2 + \frac{1}{2}\|\mathbf{r}\|_2^2 = \frac{1}{2}\left\| \begin{bmatrix} \mathbf{r} \\ \boldsymbol{D}\hat{\mathbf{e}} \end{bmatrix} \right\|_2^2,$$

we might instead minimize

$$\left\| \begin{bmatrix} \mathbf{r} \\ \boldsymbol{D}\hat{\mathbf{e}} \end{bmatrix} \right\|_p, \tag{3}$$

where if $p = 1$, the norm is defined as the sum of the absolute

values of the components, and if $p = \infty$, the norm is the maximum of the absolute values of the components. Either of these choices has the effect of reducing the effects of outliers in our measurements.

To derive an algorithm to solve this problem for $p = 1$ or ∞, and to match our previous algorithm when $p = 2$, we reason this way. We need to satisfy the constraint

$$\boldsymbol{F}\hat{\mathbf{e}} = \boldsymbol{E}\mathbf{f},$$

even after we replace $\mathbf{f}$ by $\mathbf{f} + \Delta\mathbf{f}$ and $\hat{\mathbf{e}}$ by $\hat{\mathbf{e}} + \Delta\hat{\mathbf{e}}$, so we require

$$(\boldsymbol{F} + \Delta\boldsymbol{F})\,(\hat{\mathbf{e}} + \Delta\hat{\mathbf{e}}) = (\boldsymbol{E} + \Delta\boldsymbol{E})\,(\mathbf{f} + \Delta\mathbf{f}),$$

where $\Delta\boldsymbol{E}$ is formed from $\Delta\hat{\mathbf{e}}$, and $\Delta\boldsymbol{F}$ is formed from $\Delta\mathbf{f}$ so that $\Delta\boldsymbol{F}\hat{\mathbf{e}} = \boldsymbol{E}\Delta\mathbf{f}$ and $\boldsymbol{F}\Delta\hat{\mathbf{e}} = \Delta\boldsymbol{E}\mathbf{f}$. This means that

$$\Delta\boldsymbol{F}\Delta\hat{\mathbf{e}} = \Delta\boldsymbol{E}\Delta\mathbf{f}.$$

Now let's examine the residual after we replace $\mathbf{f}$ by $\mathbf{f} + \Delta\mathbf{f}$ and $\hat{\mathbf{e}}$ by $\hat{\mathbf{e}} + \Delta\hat{\mathbf{e}}$:

$$\begin{aligned}\mathbf{r}_{new} &= \mathbf{g} - (\boldsymbol{K} + \boldsymbol{E} + \Delta\boldsymbol{E})(\mathbf{f} + \Delta\mathbf{f}) \\ &= \mathbf{g} - (\boldsymbol{K} + \boldsymbol{E})\mathbf{f} - \Delta\boldsymbol{E}\mathbf{f} - (\boldsymbol{K} + \boldsymbol{E})\Delta\mathbf{f} - \Delta\boldsymbol{E}\Delta\mathbf{f}.\end{aligned}$$

If both $\Delta\mathbf{f}$ and $\Delta\hat{\mathbf{e}}$ are small, then the last term is negligible, and we can approximate

$$\mathbf{r}_{new} \approx \mathbf{r} - \boldsymbol{F}\Delta\hat{\mathbf{e}} - (\boldsymbol{K} + \boldsymbol{E})\Delta\mathbf{f},$$

so that our minimization function (Equation 3) is approximated by

$$\left\| \begin{bmatrix} \boldsymbol{F} & \boldsymbol{K} + \boldsymbol{E} \\ \boldsymbol{D} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \Delta\hat{\mathbf{e}} \\ \Delta\mathbf{f} \end{bmatrix} + \begin{bmatrix} -\mathbf{r} \\ \boldsymbol{D}\hat{\mathbf{e}} \end{bmatrix} \right\|_p .$$

So, to compute our step, we need to minimize a function of this form; our next task is to develop an algorithm that does this.

PROBLEM 5.

a. Show that when $p = 1$, minimizing

$$\left\| \begin{bmatrix} \boldsymbol{F} & \boldsymbol{K} + \boldsymbol{E} \\ \boldsymbol{D} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \Delta\hat{\mathbf{e}} \\ \Delta\mathbf{f} \end{bmatrix} + \begin{bmatrix} -\mathbf{r} \\ \boldsymbol{D}\hat{\mathbf{e}} \end{bmatrix} \right\|_p$$

over all choices of $\Delta\mathbf{f}$ and $\Delta\hat{\mathbf{e}}$ is equivalent to solving the linear programming problem

$$\min_{\Delta\hat{\mathbf{e}}, \Delta\mathbf{f}, \bar{\sigma}} \quad \bar{\sigma} = \sum_{i=1}^{m} \bar{\sigma}_{1_i} + \sum_{i=1}^{q} \bar{\sigma}_{2_i} + \sum_{i=1}^{n} \bar{\sigma}_{3_i}$$

subject to

$$\begin{array}{ccccc} -\bar{\sigma}_1 & \le & \boldsymbol{F}\Delta\hat{\mathbf{e}} + (\boldsymbol{K} + \boldsymbol{E})\Delta\mathbf{f} - \mathbf{r} & \le & \bar{\sigma}_1 \\ -\bar{\sigma}_2 & \le & \boldsymbol{D}\Delta\hat{\mathbf{e}} + \boldsymbol{D}\hat{\mathbf{e}} & \le & \bar{\sigma}_2 \\ -\bar{\sigma}_3 & \le & \lambda\Delta\mathbf{f} + \lambda\mathbf{f} & \le & \bar{\sigma}_3 \end{array}$$

where $\bar{\sigma}_1 \in \mathbf{R}^{m\times 1}$, $\bar{\sigma}_2 \in \mathbf{R}^{q\times 1}$, and $\bar{\sigma}_3 \in \mathbf{R}^{n\times 1}$.

b. Derive a similar linear program to compute $\hat{\mathbf{e}} + \Delta\hat{\mathbf{e}}$, $\Delta\mathbf{f}$ when $p = \infty$.

Let's see how the choice of norm affects our computed spectrum.

PROBLEM 6.

Write a function `[f,ehat,r,itn] = stln1(K,g, lambda,tol)` that uses a variant of Newton's method to solve the problem when $p = 1$. Use the linear program to compute an approximate Newton direction. Start the iteration with $\hat{\mathbf{e}} = \mathbf{0}$ and $\mathbf{f} = \mathbf{0}$. Stop the iteration when the norm of the approximate Newton step is smaller than `tol`, and set `itn` to the number of iterations. Use it on the data from the Web site (www.computer.org/cise/homework), setting $\lambda = 0.06$ and `tol` $= 10^{-3}$. Plot the solution, and print the residual norm, the solution norm, and the number of iterations.

Repeat for the case $p = \infty$.

Comparing Our Results

Recall that our goal is to reconstruct the spectrum of the particles fed into the spectrometer. Take some time now to compare the results we obtained using various problem formulations.

PROBLEM 7.

Compare the results from Problems 4 and 6 with those of the last issue by answering these two questions: How does the quality of results compare? How does the amount of work compare?

PARTIAL SOLUTION TO
LAST ISSUE'S HOMEWORK ASSIGNMENT

BLIND DECONVOLUTION: ERRORS, ERRORS EVERYWHERE

By Dianne P. O'Leary

CONSIDER FIGURE 1'S DATA, REPRESENTING COUNTS MEASURED BY A SPECTROMETER, MODELED BY $\boldsymbol{K}\mathbf{f} \approx \mathbf{g}$.

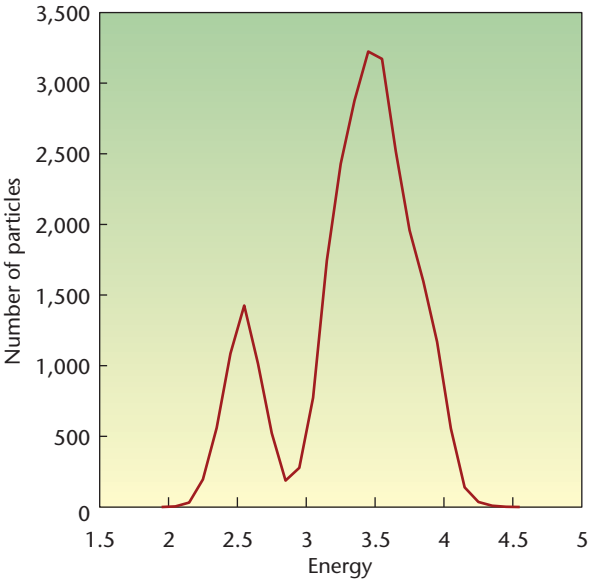

Figure 1. (Simulated) data from a spectrometer. Given that some particles have at most five different energy levels, determine these energies and the relative abundance of the particles.

PROBLEM 1.

Program the least squares algorithm and try it on the data of Figure 1 for various values of $\tilde{n}$, the number of singular values retained. The matrix $\boldsymbol{K}$ is 27×22, and we assume that the true counts for the first two and the last three bins are zero. Note how ill-conditioned the original matrix $\boldsymbol{K}$ is (by recording σ_1/σ_n).

Answer:

See the posted program `problem1_and_3.m` at www.computer.org/cise/homework. The program isn't difficult, but it's important to make sure that you do the singular value decomposition (SVD) only once (at a cost of $O(mn^3)$) and then form each of the trial solutions at a cost of $O(n^2)$. This requires using the associative law of multiplication.

In fact, it's possible to form each solution by updating a previous one (by adding the newly nonneglected term) at a cost of $O(n)$, which would be an even better algorithm, left as an exercise.

PROBLEM 2.

Suppose we have the singular value decomposition (SVD) of $[\boldsymbol{K}, \mathbf{g}] = \tilde{\boldsymbol{U}}\tilde{\Sigma}\tilde{\boldsymbol{V}}^T$. Assume that $\boldsymbol{K}$ has rank n and that $\tilde{v}_{nn} > \tilde{v}_{n+1,n+1} \neq 0$. Show that the solution to

$$\min_{\boldsymbol{E},\mathbf{r}} \left\| [\boldsymbol{E} \quad \mathbf{r}] \right\|_F ,$$

subject to the constraint

$$[\,\boldsymbol{K}+\boldsymbol{E} \quad \mathbf{g}+\mathbf{r}\,] \begin{bmatrix} \mathbf{f} \\ -1 \end{bmatrix} = \mathbf{0}$$

is

$$[\boldsymbol{E} \quad \mathbf{r}] = -\tilde{\sigma}_{n+1}\tilde{\mathbf{u}}_{n+1}\tilde{\mathbf{v}}_{n+1}^T,$$

with

$$\begin{bmatrix} \mathbf{f} \\ -1 \end{bmatrix} = -\frac{1}{\tilde{v}_{n+1,n+1}} \tilde{\mathbf{v}}_{n+1} ,$$

where $\tilde{\mathbf{u}}_{n+1}$ is the $(n+1)$st column of $\tilde{\boldsymbol{U}}$, and $\tilde{\mathbf{v}}_{n+1}$ is the $(n+1)$st column of $\tilde{\boldsymbol{V}}$.

Hint:

a. First show that this solution satisfies the constraint and that the resulting $\| [\boldsymbol{E} \quad \mathbf{r}] \|_F = \tilde{\sigma}_{n+1}$.

b. Show that $\|\tilde{\boldsymbol{U}}^T\boldsymbol{A}\tilde{\boldsymbol{V}}\|_F = \|\boldsymbol{A}\|_F$ for any matrix $\boldsymbol{A}$ of size $m \times (n+1)$.

c. Transform the problem to minimizing $\| [\tilde{\boldsymbol{E}}, \tilde{\mathbf{r}}] \|_F$ subject to $(\tilde{\Sigma} + \tilde{\boldsymbol{E}})\tilde{\mathbf{f}} = \mathbf{0}$ for some vectors $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{r}}$ and matrix $\tilde{\boldsymbol{E}}$, solve the problem in this coordinate system, and show that no solution gives a value of the minimization function smaller than $\tilde{\sigma}_{n+1}$.

Answer:

a. We know that

$$[\boldsymbol{K} \quad \mathbf{g}] = \sum_{i=1}^{n+1} \tilde{\sigma}_i \tilde{\mathbf{u}}_i \tilde{\mathbf{v}}_i^T ,$$

so using the formula for $[\boldsymbol{E} \quad \mathbf{r}]$, we see that

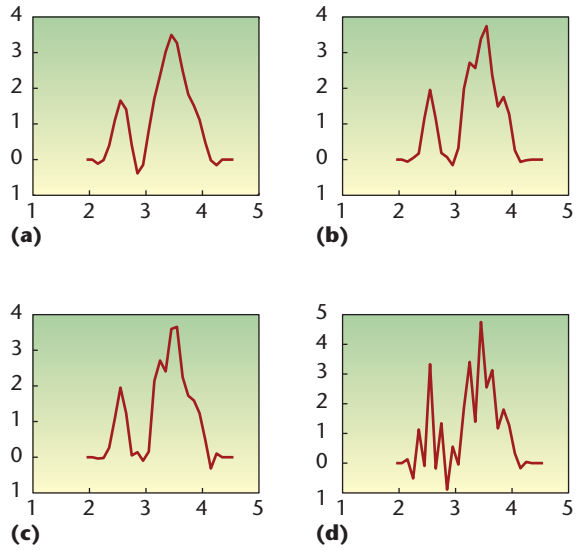

Figure A. Computed least squares solutions (counts versus energies) for various values of the cutoff parameter $\tilde{n}$. The solutions retain (a) 12 singular values, (b) 15 singular values, (c) 17 singular values, and (d) 21 singular values.

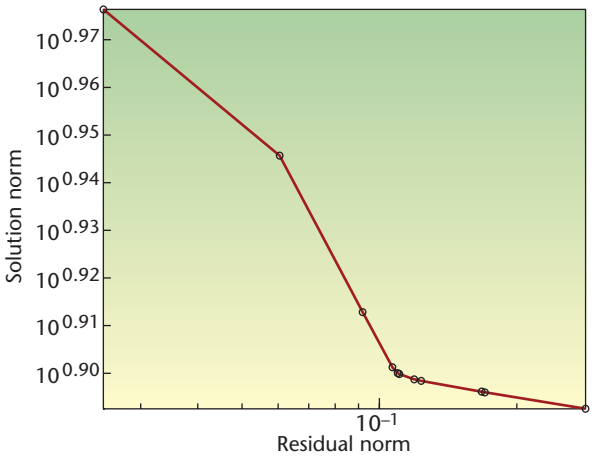

Figure B. The L-curve for least squares solutions.

$$[K \quad g] + [E \quad \mathbf{r}] = \sum_{i=1}^{n+1} \tilde{\sigma}_i \tilde{\mathbf{u}}_i \tilde{\mathbf{v}}_i^T \quad .$$

Now, since $\tilde{\mathbf{v}}_{n+1}$ is orthogonal to $\tilde{\mathbf{v}}_i$ for $k = 1, \ldots, n$, it follows that

$$([K \quad g] + [E \quad \mathbf{r}]) \begin{bmatrix} \mathbf{f} \\ -1 \end{bmatrix} = -\sum_{i=1}^{n} \tilde{\sigma}_i \tilde{\mathbf{u}}_i \tilde{\mathbf{v}}_i^T \frac{1}{\tilde{v}_{n+1,n+1}} \tilde{\mathbf{v}}_{n+1} = \mathbf{0} \, .$$

Note that $\| [\boldsymbol{E}, \boldsymbol{r}] \|_F = \tilde{\sigma}_{n+1}$.

b. This can be proven using the fact that $\| \boldsymbol{A} \|_F^2 = tr(\boldsymbol{A}^T \boldsymbol{A})$, where $tr(\boldsymbol{B})$ is the trace of the matrix $\boldsymbol{B}$, equal to the sum of its diagonal elements (or the sum of its eigenvalues). We can use the fact that $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$.

We can also prove it just from the definition of the Frobenius norm and the fact that $\| \boldsymbol{Ux} \|_2 = \| \boldsymbol{x} \|_2$ for all vectors $\boldsymbol{x}$ and orthogonal matrices $\boldsymbol{U}$. Using this fact, and letting $\boldsymbol{a}_i$ be the ith column of $\boldsymbol{A}$, we see that

$$\| \boldsymbol{UA} \|_F^2 = \sum_{i=1}^{n} \| \boldsymbol{U a}_i \|_2^2 = \sum_{i=1}^{n} \| \boldsymbol{a}_i \|_2^2 = \| \boldsymbol{A} \|_F^2 \, .$$

Similarly, letting $\hat{\boldsymbol{a}}_i^T$ be the ith row of $\boldsymbol{A}$,

$$\| \boldsymbol{AV} \|_F^2 = \sum_{i=1}^{m} \| \hat{\boldsymbol{a}}_i^T \boldsymbol{V} \|_2^2 = \sum_{i=1}^{n} \| \hat{\boldsymbol{a}}_i \|_2^2 = \| \boldsymbol{A} \|_F^2 \, ,$$

and the conclusion follows.

c. From the constraint

$$[\boldsymbol{K} + \boldsymbol{E} \quad \boldsymbol{g} + \boldsymbol{r}] \begin{bmatrix} \boldsymbol{f} \\ -1 \end{bmatrix} = \mathbf{0},$$

we see that

$$\tilde{\boldsymbol{U}}^T [\boldsymbol{K} + \boldsymbol{E} \quad \boldsymbol{g} + \boldsymbol{r}] \tilde{\boldsymbol{V}} \tilde{\boldsymbol{V}}^T \begin{bmatrix} \boldsymbol{f} \\ -1 \end{bmatrix} = \mathbf{0},$$

so

$$(\tilde{\boldsymbol{\Sigma}} + \tilde{\boldsymbol{E}}) \tilde{\boldsymbol{f}} = \mathbf{0}$$

where

$$\tilde{\boldsymbol{E}} = \tilde{\boldsymbol{U}}^T [\boldsymbol{E}, \boldsymbol{r}] \tilde{\boldsymbol{V}} \text{ and } \tilde{\boldsymbol{f}} = \tilde{\boldsymbol{V}}^T \begin{bmatrix} \boldsymbol{f} \\ -1 \end{bmatrix} .$$

From part b, we know that minimizing $\| [\boldsymbol{E}, \boldsymbol{r}] \|_F$ is the same as minimizing $\| \tilde{\boldsymbol{E}} \|_F$.

Therefore, to solve our problem, we want to make the smallest change to $\tilde{\Sigma}$ that makes the matrix $\tilde{\Sigma} + \tilde{E}$ rank deficient, so that the constraint can be satisfied by a nonzero $\tilde{f}$. Changing the $(n + 1, n + 1)$ element of $\tilde{\Sigma}$ from $\tilde{\sigma}_{n+1}$ to 0 certainly makes the constraint feasible (by setting the last component of $\tilde{\boldsymbol{f}}$ nonzero and the other components zero). Any other change gives a bigger $\| \tilde{\boldsymbol{E}} \|_F$. Thus, the smallest value of the minimization function is $\tilde{\sigma}_{n+1}$, and since we verified in part a that our solution has this value, we're finished.

If you don't find this argument convincing, we can be more precise by using a fact found in standard texts. For any matrix $\boldsymbol{B}$ and vector $\boldsymbol{z}$ for which $\boldsymbol{Bz}$ is defined: $\|\boldsymbol{Bz}\|_2 \leq \|\boldsymbol{B}\|_2 \|\boldsymbol{z}\|_2$, where $\|\boldsymbol{B}\|_2$ is defined to be the largest singular value of $\boldsymbol{B}$. Therefore,

- $\|\boldsymbol{B}\|_2 \leq \|\boldsymbol{B}\|_F$, because we can see from part b and the SVD of $\boldsymbol{B}$ that the Frobenius norm of $\boldsymbol{B}$ is just the square root of the sum of the squares of its singular values.
- If $(\tilde{\Sigma} + \tilde{E})\tilde{f} = \mathbf{0}$, then $\tilde{\Sigma}\tilde{f} = -\tilde{E}\tilde{f}$.
- $\tilde{\sigma}_{n+1}^2 \|\tilde{f}\|_2^2 = \sum_{i=1}^{n+1} \tilde{\sigma}_{n+1}^2 \tilde{f}_i^2 \leq \sum_{i=1}^{n+1} \tilde{\sigma}_i^2 \tilde{f}_i^2 = \|\tilde{\Sigma}\tilde{f}\|_2^2$.
- Therefore, $\tilde{\sigma}_{n+1}\|\tilde{f}\|_2 \leq \|\tilde{\Sigma}\tilde{f}\|_2 = \|\tilde{E}\tilde{f}\|_2 \leq \|\tilde{E}\|_2\|\tilde{f}\|_2 \leq \|\tilde{E}\|_F\|\tilde{f}\|_2$, so we conclude that $\|\tilde{E}\|_F \geq \tilde{\sigma}_{n+1}$, and we have found a minimizing solution.

Problem 3.

Write a Matlab function to solve Model 2 using this truncated technique for various values of $\tilde{n}$. The input values should be $\boldsymbol{K}$, $\mathbf{g}$, and a range of $\tilde{n}$ values. Include appropriate documentation, and use your function to solve our problem.

Answer:

See the posted program `problem1_and_3.m` at www.computer.org/cise/homework.

Problem 4.

Write a brief summary of the results you obtained using Model 1 and Model 2 to solve the problem of Figure 1. Give your best estimate of the number of different peaks (energy levels) in the original data $\mathbf{f}_{true}$, the relative heights of the peaks, and the centers of the peaks. Make a convincing argument to justify your estimate and your choice of parameters (δ and $\tilde{n}$) for each method.

Answer:

Model 1: Least Squares

To estimate the variance of the error, note that in the least squares model, the last five components of the right-hand side $\boldsymbol{U}^T\boldsymbol{g}$ can't be zeroed by any choice of $\boldsymbol{f}$, so if we believe the model, we believe that these are entirely due to error. All other components should have at least some data in addition to noise. Therefore, estimate the variance using the last five to get $\delta^2 = 1.2349 \times 10^{-4}$.

The condition number of the matrix, the ratio of largest to smallest singular value, is 61.8455. This is a well-conditioned matrix! Most spectroscopy problems have a very ill-conditioned one (having a condition number of 10^3 or more). This is a clue that an error probably exists in the matrix, moving the small singular values away from zero.

We try various choices of $\tilde{n}$, the number of singular values retained, and show the results in Figure A. The discrepancy principle predicts the residual norm to be $\delta\sqrt{m} = 0.0577$. This is most closely matched by retaining 21 singular values, which gives seven peaks, contradicting the given information that there are at most five peaks. It also produces some rather large magnitude negative peaks, and we know that counts need to be nonnegative. Thus, the least squares model doesn't seem to fit the data well.

An alternate way to pick $\tilde{n}$ is to use the L-curve. This is a plot of the log of the solution norm versus the log of the residual norm. It's called an L-curve because its shape often resembles that of the letter L. What we really want is a small residual and a solution norm that isn't unreasonably big. We could take the value of $\tilde{n}$ at the corner of the L-curve, because if we take a smaller $\tilde{n}$, the residual norm increases quickly, and if we take a larger one, the solution norm increases quickly. This plot, shown in Figure B, advises that we should retain 15 to 17 singular values, and referring to

Tools

You can find more information about the L-curve (originally due to Chuck Lawson and Richard Hanson) in Per Christian Hansen's work;[1] the total least squares (TLS) truncation we used in this problem is discussed in a 1997 article.[2]

The standard reference on TLS is the book by Sabine Van Huffel and Joos Vanderwalle.[3]

An alternate way to "regularize" TLS is given in a 1999 work.

References

1. P.C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM Press, 1998.
2. R. Fierro et al., "Regularization by Truncated Least Squares," *SIAM J. Scientific Computing*, vol. 18, 1997, pp. 1223–1241.
3. S. Van Huffel and J. Vanderwalle, *The Total Least Squares Problem*, SIAM Press, 1991.
4. G.H. Golub, P.C. Hansen, and D.P. O'Leary, "Tikhonov Regularization and Total Least Squares," *SIAM J. Matrix Analysis and Applications*, vol. 21, 1999, pp. 185–194.

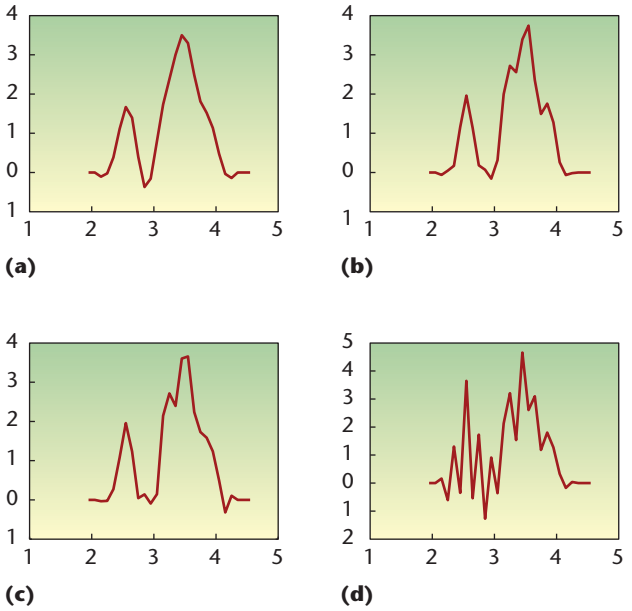

Figure C. Computed total least squares solutions (counts versus energies) for various values of the cutoff parameter $\tilde{n}$. The solutions retain (a) 12 singular values, (b) 15 singular values, (c) 17 singular values, and (d) 21 singular values.

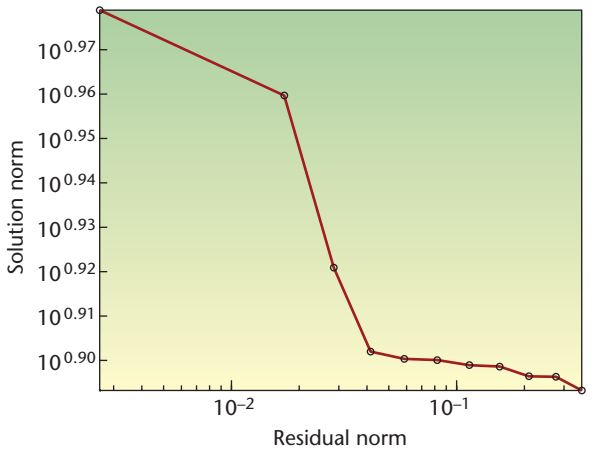

Figure D. The L-curve for total least squares solutions.

Figure A, this yields four peaks, consistent with our given information.

(The theoretical properties of the L-curve, as well as any other method that doesn't require the variance to be given in advance, aren't good, but this method is often more robust to errors in assumptions about the model, such as underestimating the variance or not accounting for errors in the matrix.)

An alternate heuristic is to look for a value of $\tilde{n}$ that makes the residual look most like white noise, but because our error isn't normally distributed, this heuristic doesn't have much meaning for our problem.

An excellent way to approach this problem is to generate your own test data, for which you know the true solution, and use it to gain insight into the choice of $\tilde{n}$.

Model 2: Total Least Squares (TLS)

Figure C shows sample solutions. The discrepancy principle doesn't give much insight for TLS, so we use more heuristic methods, giving us even less confidence in our choice.

For example, from Figure D we see that the L-curve corner occurs when 15 singular values are retained, giving a solution that looks very much like the L-curve least squares solution. Because the number of peaks is reasonable, and because there are only a small number of negative values in the solution (and these have small magnitude), we might accept this solution.

Now we need to extract the energies and estimated counts for the four types of particles. I have normalized them so that the count for the lowest energy peak is one. For the computed estimate to energy levels and counts,

Bin centers	2.55	3.25	3.55	3.85
Relative counts	1.00	1.39	1.91	0.90

A spectroscopist would actually estimate the counts by taking the integral under each of the four peaks, and then estimate the energy by the centroid of the peak, but this is difficult since three of the peaks aren't well separated.

The program used to generate the data is posted at www.computer.org/cise/homework. The variance of the error is 10^{-4}. The true energy levels and counts are

Energy	2.54	3.25	3.53	3.85
Relative counts	1	1.5	2	1

So, despite all the errors, our computed solution estimates the energy levels to two digits and the relative counts to within 10 percent.

Dianne P. O'Leary is a professor of computer science and a faculty member in the Institute for Advanced Computer Studies and the Applied Mathematics Program at the University of Maryland. She received a BS in mathematics from Purdue University and a PhD in computer science from Stanford. She is a member of SIAM, the ACM, and the AWM. Contact her at oleary@cs.umd.edu; www.cs.umd.edu/users/oleary/.