Math 642:550 — Summer 2009
MTTh 6:00–8:30 PM   Hill 425
Prof. Bumby

Supplement 7, The Pseudoinverse

1. Introduction

The text uses the **Singular Value Decomposition** to describe the **pseudoinverse**, since it gives an easy proof of existence, but it is not necessary for computation. We describe how to compute values of the pseudoinverse using only rational operations. This makes this operation accessible to hand computation in many problems. However, machine computation is more concerned with controlling error. Here, the use of the SVD may be preferred.

2. Least squares solutions

A special case of the pseudoinverse is the **Least Squares Solution** of an **overdetermined** system. Such systems typically arise when trying to get the **best fit** to data using some special type of function. For example, **linear regression** aims to draw a line $y = ax + b$ through a family of points (x_i, y_i) so that the **set of values** $\{y_i - (ax_i + b)\}$ is small in an appropriate sense. Gauss proposed finding a and b so that

$$\sum_i \bigl(y_i - (ax_i + b)\bigr)^2 \tag{$*$}$$

is minimized. This has the advantage that these values are easily found. The name **least squares** comes from this aim to minimize a sum of squares.

It is common in calculus courses to obtain a **formula** for the desired values of a and b. However, this is a **linear algebra** course, so we shall relate this problem to linear transformations and the geometry of vector spaces.

Thus, we make the desired quantities (a, b) into the components of a vector in $\mathbb{R}^2$, and the values $\langle ax_i + b\rangle$ and $\langle y_i\rangle$ determine vectors in $\mathbb{R}^n$, with n equal to the number of data points. The x_i then determine a linear transformation from $\mathbb{R}^2$ to $\mathbb{R}^n$, and $\langle ax_i + b\rangle$ is an element if the **column space** of the matrix M with rows $[1, x_i]$ that represents the linear transformation.

If $\langle y_i\rangle$ were in this column space, then we could make all $y_i - (ax_i + b)$ equal to zero, but this is unlikely to happen if n is large. Instead, we note the equation $(*)$ is a formula for the square of the distance between $\langle y_i\rangle$ and a point of the column space. The Pythagorean theorem tells us that distance is minimized by the point in the column space on a line through $\langle y_i\rangle$ perpendicular to the column space. To solve a least squares problem, $\langle y_i\rangle$ should be **projected** into the column space of M and that projection expressed in terms of the columns of M. For our linear regression problems, the columns of M will be linearly independent as long as the x_i are not all the same. Then, there will be a unique solution (a, b) expressing the projection into the column space as a linear combination of the columns. The same analysis applies to any matrix M with **linearly independent columns**. We note for future use that the orthogonal projection on the column space of M sends vectors in the **left nullspace** of M to the zero vector.

3. The normal equations

Lemma. *M and $M^\top M$ have the same nullspace.*

Proof. If x belongs to the nullspace of M, $Mx = \mathbf{0}$, which implies that $M^\top M x = \mathbf{0}$, so the nullspace of M is contained in the nullspace of $M^\top M$. If x belongs to the nullspace of $M^\top M$, $M^\top M x = \mathbf{0}$, which implies that $(Mx)^{\mathsf{T}}(Mx) = x^{\mathsf{T}}M^{\mathsf{T}}Mx = 0$, so Mx has length zero. However, the only vector of length zero is the zero vector, so $Mx = \mathbf{0}$, which says that x belongs to the nullspace of M.

It is a **useful fiction** to pretend that we are trying to solve a system of the form $Mx = c$, even if this system is obviously inconsistent, and to multiply this on the left by M^{T} to get $M^{\mathsf{T}}Mx = M^{\mathsf{T}}c$. The choice of M^{T} is a result of the need to remove the component of c lying in the left nullspace of M. If M has linearly independent columns, it has a trivial nullspace, so the lemma tells us that $M^{\mathsf{T}}M$ also has a trivial nullspace. That is, $M^{\mathsf{T}}M$ is nonsingular, and $M^{\mathsf{T}}Mx = M^{\mathsf{T}}c$ has a unique solution. This fiction is only useful is we pick M^{T}, or something closely related to it, as the left multiplier because of the its relation to the projection into the column space of M, so we need to make that choice part of the solution process.

The system $M^{\mathsf{T}}Mx = M^{\mathsf{T}}c$, called the **normal equations** associated with M and c, has a unique solution $x^{(LS)}$. The normal equations say that $M^{\mathsf{T}}(Mx^{(LS)} - c) = 0$, which says that $Mx^{(LS)} - c$ is orthogonal to the column space of M. In other words, $Mx^{(LS)}$ is equal to the desired projection of c into the column space of M.

For the **toy problems** met in homework exercises that can be solved exactly, this is usually the best way to find a least-squares solution.

4. Example

Let

$$M = \begin{bmatrix} 5 & -2 \\ -1 & -6 \\ -3 & -8 \\ 4 & -1 \\ 7 & 2 \end{bmatrix}, \qquad c = \begin{bmatrix} -45 \\ -49 \\ -77 \\ -9 \\ 33 \end{bmatrix}$$

Then,

$$M^{\mathsf{T}}M = \begin{bmatrix} 100 & 30 \\ 30 & 109 \end{bmatrix}, \qquad M^{\mathsf{T}}c = \begin{bmatrix} 250 \\ 1075 \end{bmatrix}$$

so that the normal equations are

$$\begin{bmatrix} 100 & 30 \\ 30 & 109 \end{bmatrix} x = \begin{bmatrix} 250 \\ 1075 \end{bmatrix}$$

whose solution is

$$x = \begin{bmatrix} -1/2 \\ 10 \end{bmatrix}$$

This solution is checked by finding the **residual**

$$c - Mx = \frac{1}{2}\begin{bmatrix} -45 \\ 21 \\ 3 \\ 6 \\ 33 \end{bmatrix}.$$

This isn't particularly small — it doesn't have to be — but it is perpendicular to the column space of M, so it confirms that x is the "least squares solution" of $Mx = c$.

5. Numerical considerations

One thing that should be noticed is that the numbers appearing in the normal equations $M^\top Mx = M^\top c$ are roughly the **squares** of the numbers appearing in M and c. This was not a major concern in our example, since the numbers were rational and could be written exactly with little effort.

However, if the matrices are only known approximately, the behavior of the error depends on the **actual steps** in the computation, **not on the formula** that motivates that computation. The discussion of **condition number** in section 7.2 of the text gives a rough idea of how to predict when there will be a significant magnification of error in a computation. In particular, the application to the "least-squares" problem is discussed on page 356. Note 2 on that page says that, because the **condition number** is the ratio of the largest eigenvalue to the smallest one, squaring the matrix will square this ratio, and this may seriously magnify the errors. A more detailed discussion can be found in section 5.3 of Gene H. Golub & Charles F. Van Loan, *Matrix Computations*, Johns Hopkins, 1989 (ISBN 0-8018-3772-3 [ISBN-13: 9780801837722] (hardcover) and 0-8018-3739-1 [ISBN-13: 9780801837395] (paperback)). After careful analysis of two methods of solving the "least squares" problem, those authors conclude: 'At the very minimum, this discussion should convince you how difficult it can be to choose the "right" algorithm!'. More details will be given in our discussion of chapter 7 of the textbook.

One alternative to the normal equations uses the QR factorization of M. We first note that the role of the factor $M^\top$ in the normal equations is only to assure that Mx and c differ by a vector perpendicular to the column space of M. Hence it could be replaced by the transpose of **any** matrix having the same column space. In particular, if $M = QR$, the normal equations are replaced by

$$Rx = Q^\top QRx = Q^\top Mx = Q^\top c. \qquad (**)$$

These equations are already in triangular form, so they are solved using only the **back substitution** part of Gaussian elimination. Multiplying $(**)$ on the left by $R^\top$ reveals that $R^\top R$ is the LU factorization of $M^\top M$, so the first part of Gaussian elimination on the usual normal equations serves to identify R and $Q^\top c$.

6. The general case

A similar method applies when the columns of M are not linearly independent. Again, the first step is to project c into the column space. This can be done by forming the system $M^\top Mx = M^\top c$, but this system now has a **singular matrix of coefficients**. However, the system is solvable: the column space of $M^\top M$ has been shown to be the same as the column space of $M^\top$ and $M^\top c$ is in this space. Thus, the solutions will fill a **translate of the nullspace** of $M^\top M$. The nullspace of $M^\top M$ is the orthogonal complement of the row space of $M^\top M$, which is the same as the column space of $M^\top M$ since $M^\top M$ is a symmetric matrix. This has been seen to be the same as the column space of $M^\top$ which is orthogonal to the nullspace of M. Thus, $M^\top M$ and M have the same nullspace, and we have assumed that the nullspace of M is not trivial. This means that the normal equations now have infinitely many solutions. To select a **canonical** (or **special**) element of this set, we choose the vector x **of shortest length**. The usual analysis of extrema of distance reveals that this is the vector in this set that is perpendicular to the nullspace of M. Our knowledge of the **four fundamental subspaces** tells us that the space perpendicular to the nullspace is the **row space** of M. To get our special solution, we need to find one solution and project it into the row space of M. Any solution that we find will have the same projection. This analysis shows that the solution of such generalized least-squares problems are unique.

Moreover, the solution of this least-squares problem may be described as the composition of linear transformations. First, project into the column space of M; then, apply the **inverse** of a nonsingular mapping from the r dimensional row space of M to its r dimensional column space; finally, express the resulting vector in terms of the original basis of $\mathbb{R}^n$. An n by m matrix expressing this linear transformation in terms of the given bases of $\mathbb{R}^n$ and $\mathbb{R}^m$ is called the **pseudoinverse** of M. This invention appeared first in

E.H. Moore. Bull. Amer. Math. Soc. **26** (1920), 394–395 (a research announcement with few details), and became more widely known after the appearance of Roger Penrose, Proc. Cambridge Philos. Soc. **51** (1955), 406–413 and **52** (1956), 17–19 (Penrose was then a graduate student at the University of Cambridge). The pseudoinverse often carries the names of these two authors.

7. Computing solutions

The QR factorization of an m by n matrix M of rank r, which can be found either by the **Gram-Schmidt** process or by building an orthogonal matrix Q from **Householder matrices**, produces an **orthogonal** basis for the column space of a given matrix M. The Gram-Schmidt process stops at this point, but the Householder matrices lead to an orthonormal basis for all of $\mathbb{R}^m$ whose first r columns are a basis for the column space, with the remaining $m - r$ vectors being a basis for its **orthogonal complement**, the **left nullspace**. Multiplying any given vector v by $Q^\top$ gives the components of v with respect to this basis, allowing the projection into the column space to be found.

The Gram-Schmidt process is better for hand computation when M has rational entries since it only works with rational numbers, except for square roots needed to normalize the vectors in the orthogonal basis. These normalization factors are useful when Q^{-1} is needed since $Q^{-1} = Q^\top$, but they are mainly a theoretical convenience: they disappear in the expression for the projection. On the other hand, we have seen that Householder matrices sometimes make essential use of irrational quantities. Their effect will also disappear eventually when these matrices are used to find a QR factorization, but the **exact** description of the orthonormal basis used to compute the projection is not guaranteed to be simple.

We have already noted that, if one has the QR factorization of M with an m by r factor Q and an r by n factor R, the vectors x mapping to the projection of c into the column space of M are the solutions of $Rx = Q^\top c$. In order to find our preferred solution, we need to find a solution in the **row space** of M, which is also the row space of R. Since the rows of R are linearly independent, we can use them as a basis for the row space. This suggests writing $x = R^\top z$, so that our equation becomes $RR^\top z = Q^\top c$. If only the r nonzero rows of R are used, $RR^\top$ is nonsingular and this system has a unique solution. That solution is z — the components of x with respect to our chosen basis for the row space on M — so **we need to multiply** by $R^\top$ to obtain x.

This process requires that we solve a system whose coefficient matrix is $RR^\top$, which is no problem when we are solving equations exactly. In particular, this matrix is already factored as a product of triangular matrices. However, as with the **normal equations**, there may be difficulty when solving a large system approximately. If this is a concern, the solution should find any solution of $Rx = Q^\top c$ and project it from $\mathbb{R}^n$ into the **row space** of M, which is also the row space of R. This can be done by finding a QR factorization of $R^\top$. Putting these two factorizations together gives what Golub and Van Loan call a **Complete orthogonal decomposition**. Here, $M = Q_1 T Q_2^\top$, where T is **lower triangular** (since it is the transpose of the R factor in the QR factorization of the transpose of the R factor in the QR factorization of M.

There are many different complete orthogonal decompositions of M, but they all lead to the same pseudoinverse since **the underlying linear transformation is uniquely defined**. The **singular value decomposition**, described in Section 6.3 of the textbook, is an example of a complete orthogonal decomposition. The construction in the textbook is an example of the general construction that we have just described. When dealing with approximate quantities using a computer, a routine for computing an approximate **SVD** is likely to be available, so the calculation of the pseudoinverse, or even the solution of a single general least-squares problem, can be found in terms of the SVD. However, the exact solution of the **toy problems** used as exercises or exam problems in courses can be found by using the Gram-Schmidt process to compute the projections.

8. The Singular Value Decomposition

The SVD is usually not computed by hand, so it probably suffices to know its properties so you can check that it has been computed correctly when you have assigned that task to a computer (or calculator). However, some comments beyond what is in the textbook are in order. First, knowing that you are looking for $A = U\Sigma V^{\mathsf{T}}$ is the key to finding such a factorization since it implies that $A^{\mathsf{T}}A = V\Sigma^{\mathsf{T}}\Sigma V^{\mathsf{T}}$. Since V is supposed to be an orthogonal matrix, $V^{\mathsf{T}} = V^{-1}$, so V is an **eigenvector matrix** of the **symmetric matrix** $A^{\mathsf{T}}A$ and $\Sigma^{\mathsf{T}}\Sigma$ is the corresponding **diagonal matrix of eigenvalues** (there is a similar result for AA^{T}, but **only one** of these equations should be used). It follows from this that

$$(AV)^{\mathsf{T}}AV = V^{\mathsf{T}}A^{\mathsf{T}}AV = \Sigma^{\mathsf{T}}\Sigma.$$

Thus, the nonzero columns of AV are orthogonal vectors whose lengths are given by diagonal entries in Σ. These columns should be scaled to the first r columns of U. The remaining columns are found by extending this set to an orthonormal basis of $\mathbb{R}^m$. Note that the lengths of a columns of AV are precisely the corresponding diagonal entry of Σ. This verifies that $AV = U\Sigma$, which is what is required.

When building an orthogonal matrix from a list of orthogonal vectors, each vector is scaled to give a unit vector in the same direction, There is no reason to prefer one such vector to its negative, so many choices are involved in the construction of V. If we were to try to build U from the eigenvectors of $A^{\mathsf{T}}A$, we would again have choices leading to many orthogonal matrices, but **only one** is correct after V has been chosen.

The SVD also leads to a form of the **spectral theorem** with

$$A = \sum \sigma_i u_i v_i^{\mathsf{T}},$$

where u_i is column i of U and v_i is column i of V. This sum may be restrict to the r terms for which $\sigma_i > 0$. The bases for the nullspaces of A and A^{T} that had to be constructed arbitrarily do not appear in this expression for A. If A has rank 1, then this version of the SVD scales vectors in a column times row expression for A so that each has length 1 and sets σ_1 equal to the product of the scale factors.

End of Supplement