

EE731 Lecture Notes: Matrix Computations for Signal Processing

James P. Reilly©
Department of Electrical and Computer Engineering
McMaster University

November 10, 2006

Lecture 10

In this lecture we investigate how the QR decomposition may be used to solve the LS problem. We find that the QR decomposition yields a very simple solution in the full-rank case, and also leads to an efficient procedure for the LS solution in the rank deficient case. We discuss the necessity for column pivoting during the QR decomposition process when \mathbf{A} is rank deficient.

We then look at a numerically stable technique for solving least-squares in the presence of coloured noise when the noise covariance matrix $\mathbf{\Sigma}$ is known. We have seen previously that an optimal solution is yielded by pre-whitening the data. This step is accomplished by pre-multiplying the data by the inverse Cholesky factor of $\mathbf{\Sigma}$. However, if $\mathbf{\Sigma}$ is ill-conditioned, the inverse is unstable. We examine how to find the optimal solution without explicitly finding an inverse.

13 Solving Least Squares Using the QR Decomposition

13.1 Full Rank LS Using the QR Decomposition

In this section, we look at the structure the QR decomposition reveals in solving the LS problem. We have $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $m > n$, $\text{rank}(\mathbf{A}) = n$, and we wish to solve:

$$\mathbf{x}_{LS} = \arg \min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2.$$

Let the QR decomposition of \mathbf{A} be expressed as

$$\mathbf{Q}^T \mathbf{A} = \mathbf{R} = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix} \begin{matrix} n \\ m-n \end{matrix} \quad (1)$$

where \mathbf{Q} is $m \times m$ orthonormal and \mathbf{R}_1 is upper triangular. Let us partition \mathbf{Q} as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{matrix} n \\ m-n \end{matrix}. \quad (2)$$

From our previous discussion, and from the structure of the QR decomposition $\mathbf{A} = \mathbf{QR}$, we note that \mathbf{Q}_1 is an orthonormal basis for $R(\mathbf{A})$, and \mathbf{Q}_2 is an orthonormal basis for $R(\mathbf{A})_{\perp}$. We now define the quantities \mathbf{c} and \mathbf{d} as

$$\mathbf{Q}^T \mathbf{b} = \begin{bmatrix} \mathbf{Q}_1^T \\ \mathbf{Q}_2^T \end{bmatrix} \mathbf{b} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} \begin{matrix} n \\ m-n \end{matrix}. \quad (3)$$

Then, we may write:

$$\begin{aligned} \min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2 &= \|\mathbf{Q}^T \mathbf{Ax} - \mathbf{Q}^T \mathbf{b}\|_2^2 \\ &= \left\| \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} \right\|_2^2. \end{aligned} \quad (4)$$

It is clear that \mathbf{x} does not affect the “lower half” of the above equation. Eq. (4) may be written

$$\|\mathbf{Ax} - \mathbf{b}\|_2^2 = \|\mathbf{R}_1 \mathbf{x} - \mathbf{c}\|_2^2 + \|\mathbf{d}\|_2^2. \quad (5)$$

Because \mathbf{A} is full rank, \mathbf{R}_1 is invertible, and the above is minimum when

$$\mathbf{x}_{LS} = \mathbf{R}_1^{-1} \mathbf{c}. \quad (6)$$

The LS residual ρ_{LS} is given directly as

$$\rho_{LS} = \|\mathbf{d}\|_2. \quad (7)$$

Thus the LS problem is solved. Note that if a Gram-Schmidt procedure is used to compute the QR decomposition on \mathbf{A} , then there is not enough information to represent the “lower half” in (4). This is because this procedure only gives the partition \mathbf{Q}_1 of \mathbf{Q} , and thus \mathbf{d} and the quantity ρ_{LS} cannot be computed; however, the solution $\mathbf{x}_{LS} = \mathbf{R}_1^{-1}\mathbf{c}$ is still available. In contrast, the Householder or Givens procedure yields a complete $m \times m$ orthonormal matrix $\mathbf{Q} = [\mathbf{Q}_1 \quad \mathbf{Q}_2]$, allowing a complete solution to the LS problem.

The interpretation of (3) is interesting. Recall

$$\begin{array}{l}
 \text{These columns of} \\
 \mathbf{Q} \text{ are in} \\
 R(\mathbf{A}). \\
 \\
 \text{These columns} \\
 \text{are in} \\
 R(\mathbf{A})^\perp
 \end{array}
 \begin{bmatrix}
 \mathbf{Q}_1^T \\
 \text{---} \\
 \mathbf{Q}_2^T
 \end{bmatrix}
 \begin{array}{l}
 n \\
 \\
 m - n
 \end{array}
 \mathbf{b}
 =
 \begin{bmatrix}
 \mathbf{c} \\
 \text{---} \\
 \mathbf{d}
 \end{bmatrix}
 \quad (8)$$

Let us define

$$\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2 \quad (9)$$

where $\mathbf{b}_1 \in R(\mathbf{Q}_1) = R(\mathbf{A})$ and $\mathbf{b}_2 \in R(\mathbf{Q}_2) = R(\mathbf{A})^\perp$. Thus, the elements of \mathbf{c} are the coefficients of \mathbf{b}_1 expressed in the orthonormal basis $\mathbf{Q}_1 \in R(\mathbf{A})$. Likewise, the elements of \mathbf{d} are the coefficients of \mathbf{b}_2 in the basis $\mathbf{Q}_2 \in R(\mathbf{A})^\perp$. From this insight, it follows that the squared norm of the LS residual $\rho_{LS}^2 = \|(\mathbf{Q}_2)^T \mathbf{b}\|_2^2$.

13.2 Rank-Deficient LS Using the QR Decomposition

13.2.1 Computation of the Rank-Deficient QR Decomposition

Before investigating the use of the QR decomposition in the rank-deficient LS problem, we must first examine the structure of the QR decomposition when the matrix \mathbf{A} is rank deficient. If $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m > n$, $\text{rank}(\mathbf{A}) = r < n$, then for the QR decomposition to be of value in solving the LS problem, it is important that the relation $R(\mathbf{A}) = \text{span}[\mathbf{q}_1, \dots, \mathbf{q}_r]$ always holds.

We construct an example to show this is not always true. Suppose the rank 2

matrix \mathbf{A} is defined as follows:

$$\mathbf{A} = \begin{bmatrix} -0.8792 & -0.8792 & -1.1777 \\ -0.4731 & -0.4731 & -0.0769 \\ -0.0567 & -0.0567 & 1.2677 \end{bmatrix}$$

Then the QR decomposition of \mathbf{A} degenerates as follows:

$$\mathbf{A} = \mathbf{QR} = \begin{bmatrix} -0.8792 & 0.1528 & -0.4513 \\ -0.4731 & -0.3926 & 0.7887 \\ -0.0567 & 0.9069 & 0.4174 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}. \quad (10)$$

We see that $R(\mathbf{A}) \neq \text{span}[\mathbf{q}_1, \mathbf{q}_2]$. Further, this QR decomposition is of no value in solving the LS problem, because \mathbf{R} is not full rank. Therefore, from (4), \mathbf{x} does not have a solution in this case. The problem in (10) is that all columns of \mathbf{A} cannot be expressed as a linear combination of any 2 columns of \mathbf{Q} ; i.e.,

$$\text{Range}(\mathbf{A}) \neq \{\text{span}(\mathbf{q}_1, \mathbf{q}_2) \text{ or } \text{span}(\mathbf{q}_1, \mathbf{q}_3) \text{ or } \text{span}(\mathbf{q}_2, \mathbf{q}_3)\}.$$

Therefore for the QR decomposition to be useful for the general LS problem, we must have

$$R(\mathbf{A}) = \text{span}(\mathbf{q}_1 \dots \mathbf{q}_r) \quad (11)$$

where $r = \text{rank}(\mathbf{A})$.

We will show that a column-permutation matrix $\mathbf{\Pi}$ exists such that the QR decomposition on \mathbf{A} may be expressed as

$$\mathbf{Q}^T \mathbf{A} \mathbf{\Pi} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{matrix} r \\ m-r \end{matrix} \quad (12)$$

$r \quad n-r$

where \mathbf{R}_{11} is upper triangular and non-singular and \mathbf{R}_{12} is a rectangular matrix. Then it can be verified that the rank-deficient QR decomposition in the form of (12) indeed satisfies (11). The permutation matrix $\mathbf{\Pi}$ is determined in such a way so that at each stage i , $i = 1, \dots, r$, the diagonal elements r_{ii} of \mathbf{R}_{11} are as large in magnitude as possible, thus avoiding the degenerate form of (10). But what is the procedure to determine $\mathbf{\Pi}$?

To answer this, consider the i^{th} stage of the QR decomposition with column pivoting where the first i columns have been annihilated below the main diagonal by an appropriate QR decomposition procedure, where $i < r$ as shown below:

$$(\mathbf{Q}^{(i)})^T \mathbf{A} \mathbf{\Pi}^{(i)} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{R}_{22} \end{bmatrix} \begin{matrix} i \\ n-i \end{matrix} \quad (13)$$

$i \quad n-i$

where $\mathbf{Q}^{(i)}$ is an $m \times m$ orthonormal matrix at the i th stage of the decomposition, $\mathbf{\Pi}^{(i)}$ is the permutation matrix at the i th stage, and \mathbf{R}_{22} is a rectangular matrix of the dimension indicated.

The $(i + 1)$ th stage of the decomposition proceeds by first, post-multiplying both sides of (13) by a permutation matrix $\mathbf{\Pi}_{i+1}$ (to swap the desired column into the leading position of \mathbf{R}_{22} , as discussed shortly), and then pre-multiplying both sides by an orthonormal matrix $\tilde{\mathbf{Q}}_{i+1}$ ¹ such that the first column of the \mathbf{R}_{22} partition is annihilated below the first element. Since we wish to preserve the first i rows of the decomposition in (13) obtained so far, the orthonormal matrix $\tilde{\mathbf{Q}}_{i+1}$ to execute the $(i + 1)$ th stage is given by

$$\tilde{\mathbf{Q}}_{i+1} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{i+1} \end{bmatrix} \begin{matrix} i \\ n-i \end{matrix} \quad (14)$$

Since \mathbf{Q}_{i+1} is orthonormal, the element $r_{i+1,i+1}$ in the top left position of \mathbf{R}_{22} after the multiplication is complete equals $\left\| \mathbf{r}_1^{(22)} \right\|_2$, where $\mathbf{r}_1^{(22)}$ is the first column of the partition \mathbf{R}_{22} at stage i in (13). It is then clear that to place the elements with the largest possible magnitudes along the diagonal of \mathbf{R} , we must choose the permutation matrix $\mathbf{\Pi}_{i+1}$ at the $(i + 1)$ th stage so that the column of \mathbf{R} in (13) with corresponding maximum $\left\| \mathbf{r}_j^{(22)} \right\|_2$, over $j = i + 1, \dots, n$, is swapped into the first column position of \mathbf{R}_{22} . This procedure ensures that the resulting QR decomposition will have the form of (12) as desired. Effectively, this procedure ensures that no zeros are introduced along the diagonal of \mathbf{R} part way through the process.

It is interesting to note that the elements of \mathbf{R}_{22} are the coefficients of the columns $[\mathbf{a}_{i+1} \dots \mathbf{a}_n]$ in the basis $\mathbf{Q}^{(i)}(i + 1 : n)$, which is an orthonormal basis for the orthogonal complement of $[\mathbf{a}_1 \dots \mathbf{a}_i]$ at the i th stage. Thus, the column of \mathbf{R}_{22} which is annihilated at the $(i + 1)$ th step corresponds to the column of \mathbf{A} which has the largest component in the orthogonal complement subspace of $\text{span}[\mathbf{a}_1 \dots \mathbf{a}_i]$, which corresponds to those columns already annihilated.

To complete the $(i+1)$ th stage of the decomposition, we have $\mathbf{\Pi}^{(i+1)} = \mathbf{\Pi}^{(i)}\mathbf{\Pi}_{i+1}$, and $\mathbf{Q}^{(i+1)} = \tilde{\mathbf{Q}}_{i+1}\mathbf{Q}^{(i)}$. After r stages, the QR decomposition in the form of (12) is complete. Given that the QR decomposition now has the correct structure, we solve:

¹We use a subscript notation to indicate the matrix which performs only the i th stage of the decomposition, and superscript notation to indicate an accumulated matrix at the i th stage; specifically, $\mathbf{\Pi}^{(i)} = \prod_{i=1}^i \mathbf{\Pi}_i$. A similar notation holds for \mathbf{Q} .

13.2.2 The Rank-Deficient LS Problem with QR:

Given $\mathbf{A} \in \mathfrak{R}^{m \times n}$, $m > n$, $\text{rank}(\mathbf{A}) = r < n$, $\mathbf{b} \in \mathfrak{R}^n$, then

$$\|\mathbf{Ax} - \mathbf{b}\|_2^2 = \|(\mathbf{Q}^T \mathbf{A} \mathbf{\Pi}) \mathbf{\Pi}^T \mathbf{x} - \mathbf{Q}^T \mathbf{b}\|_2^2. \quad (15)$$

Note that in the rank deficient case, there is no unique solution for (15). Hence, unless an extra constraint is imposed on \mathbf{x} , the LS solution obtained by a particular algorithm can wander throughout the set of possible solutions, and very large variances can result. As in the pseudo-inverse case, the constraint of minimum norm is a convenient one to apply in this case, in order to reduce the variances to reasonable values. However, unlike the development of the pseudo-inverse solution, we will see that the direct use of the QR decomposition does not lead directly to the minimum norm solution \mathbf{x}_{LS} . However, it is still possible to derive an elegant solution to the LS problem using only the QR decomposition procedure. We now discuss how this is achieved.

Let

$$\mathbf{\Pi}^T \mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \begin{matrix} r \\ n-r \end{matrix}. \quad (16)$$

Substituting (12), (16) and (3) into (15) we have

$$\|\mathbf{Ax} - \mathbf{b}\|_2^2 = \left\| \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} - \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} \right\|_2^2. \quad (17)$$

The minimum residual of norm $\|\mathbf{d}\|_2$ is obtained when

$$\mathbf{R}_{11} \mathbf{y} + \mathbf{R}_{12} \mathbf{z} = \mathbf{c} \quad (18)$$

or when $\mathbf{\Pi}^T \mathbf{x} = [\mathbf{y}, \mathbf{z}]^T$ is defined as

$$\mathbf{\Pi}^T \mathbf{x} = \begin{bmatrix} \mathbf{R}_{11}^{-1}(\mathbf{c} - \mathbf{R}_{12} \mathbf{z}) \\ \mathbf{z} \end{bmatrix}. \quad (19)$$

We see from (19) that the vector \mathbf{z} is arbitrary. We obtain the so-called “basic solution” \mathbf{x}_B by setting $\mathbf{z} = \mathbf{0}$, to give

$$\mathbf{\Pi}^T \mathbf{x}_B = \begin{bmatrix} \mathbf{R}_{11}^{-1} \mathbf{c} \\ \mathbf{0} \end{bmatrix}. \quad (20)$$

It turns out that \mathbf{x}_B is not necessarily the solution \mathbf{x}_{LS} having minimum 2-norm. To see this, we substitute (20) into (19) to obtain:

$$\mathbf{\Pi}^T \mathbf{x} = \left[\mathbf{\Pi}^T \mathbf{x}_B + \begin{pmatrix} -\mathbf{R}_{11}^{-1} \mathbf{R}_{12} \mathbf{z} \\ \mathbf{z} \end{pmatrix} \right] \quad (21)$$

Therefore, we may express \mathbf{x} as

$$\mathbf{x} = \mathbf{x}_B - \mathbf{\Pi} \begin{bmatrix} \mathbf{R}_{11}^{-1} \mathbf{R}_{12} \\ -\mathbf{I} \end{bmatrix} \mathbf{z}. \quad (22)$$

Then, \mathbf{x}_{LS} is that value of \mathbf{x} from above having smallest norm; i.e.,

$$\mathbf{x}_{LS} = \min_{\mathbf{z} \in \mathfrak{R}^{n-r}} \left\| \mathbf{x}_B - \mathbf{\Pi} \begin{bmatrix} \mathbf{R}_{11}^{-1} \mathbf{R}_{12} \\ -\mathbf{I} \end{bmatrix} \mathbf{z} \right\|_2. \quad (23)$$

We see that unless $\mathbf{R}_{12} = \mathbf{0}$, $\mathbf{x}_B \neq \mathbf{x}_{LS}$, and if $\mathbf{R}_{12} \neq \mathbf{0}$, there exists a $\mathbf{z} \neq \mathbf{0}$ such that $\|\mathbf{x}_B\| > \|\mathbf{x}_{LS}\|$. We see that \mathbf{x}_{LS} is not easily determined from (23).

We therefore seek an efficient means of solving (23). We note that the desired solution \mathbf{x}_{LS} to (23) is that solution which minimizes $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ with respect to \mathbf{x} and simultaneously minimizes $\|\mathbf{x}\|_2$. Hence, this solution would possess the same properties as the pseudo-inverse solution, and because of uniqueness, this solution would be identical to the pseudo-inverse solution. To achieve this goal, we solve (23) in an efficient way using:

13.2.3 The Complete Orthogonal Decomposition(COD)

Consider the matrix decomposition resulting from (12):

$$\mathbf{Q}^T \mathbf{A} \mathbf{\Pi} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{matrix} r & n-r \\ r & m-r \end{matrix}$$

The idea is to eliminate \mathbf{R}_{12} ; then finding the \mathbf{x}_{LS} with minimum norm is easy to determine. There exists an orthonormal $\mathbf{Z} \in \mathfrak{R}^{n \times n}$ such that

$$\begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Z} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (24)$$

where \mathbf{T}_{11} is nonsingular and upper triangular of dimension $r \times r$. Therefore,

$$\mathbf{Q}^T \mathbf{A} \mathbf{\Pi} \mathbf{Z} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (25)$$

Eq. (25) is called the *complete orthogonal decomposition* of the matrix \mathbf{A} .

The fact that an orthonormal matrix \mathbf{Z} can exist may be understood by taking the transpose of both sides of (24). Then (24) becomes an ordinary QR decomposition on $\begin{bmatrix} \mathbf{R}_{11}^T \\ \mathbf{R}_{12}^T \end{bmatrix}$, with the exception that the result is \mathbf{T}_{11}^T , which is

lower triangular instead of upper triangular, as expected. However, it is easy to modify the ordinary QR decomposition procedure to yield a lower instead of an upper triangular matrix.

Now solving the LS problem is easy:

$$\begin{aligned} \|\mathbf{Ax} - \mathbf{b}\|_2^2 &= \left\| (\mathbf{Q}^T \mathbf{A} \mathbf{\Pi} \mathbf{Z})(\mathbf{Z}^T \mathbf{\Pi}^T \mathbf{x}) - \mathbf{Q}^T \mathbf{b} \right\|_2^2 \\ &= \left\| \begin{pmatrix} \mathbf{T}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \mathbf{y} \end{pmatrix} - \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} \right\|_2^2 \end{aligned} \quad (26)$$

where

$$\mathbf{Z}^T \mathbf{\Pi}^T \mathbf{x} = \begin{pmatrix} \mathbf{w} \\ \mathbf{y} \end{pmatrix}_{\substack{r \\ n-r}}, \quad (27)$$

and \mathbf{c}, \mathbf{d} are defined in (3) as before. Clearly, \mathbf{y} is arbitrary, and \mathbf{d} is independent of both \mathbf{w} and \mathbf{y} . We can write (26) in the form

$$\|\mathbf{Ax} - \mathbf{b}\|_2^2 = \|\mathbf{T}\mathbf{w} - \mathbf{c}\|_2^2 + \|\mathbf{d}\|_2^2 \quad (28)$$

which is minimum when $\mathbf{w} = \mathbf{T}^{-1}\mathbf{c}$. We also have

$$\mathbf{x}_{LS} = \mathbf{\Pi} \mathbf{Z} \begin{pmatrix} \mathbf{w} \\ \mathbf{y} \end{pmatrix}.$$

which clearly has minimum norm when $\mathbf{y} = \mathbf{0}$.

The \mathbf{x}_{LS} calculated in this way is identical to the pseudo-inverse solution. However, the computational cost with the COD is significantly less. The COD requires only two QR decompositions; the SVD is computed using an iterative procedure involving one QR decomposition per iteration.

13.3 LS in Coloured Noise Using QR

In this case, we have the regression model

$$\mathbf{b} = \mathbf{Ax} + \boldsymbol{\nu} \quad (29)$$

where $\text{cov}(\boldsymbol{\nu}) = \boldsymbol{\Sigma}$ is not diagonal, because the noise $\boldsymbol{\nu}$ is assumed coloured. Recall from the Chapter 7 notes that the normal equation solution

$$\mathbf{x}_{LS} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} \quad (30)$$

does not have minimum variance in this case. However if we pre-whiten the noise by pre-multiplying (29) by \mathbf{G}^{-1} , where $\mathbf{G}\mathbf{G}^T = \boldsymbol{\Sigma}$ (where \mathbf{G}^{-1} could be the inverse Cholesky factor) then

$$\mathbf{G}^{-1}\mathbf{b} = \mathbf{G}^{-1}\mathbf{Ax} + \mathbf{G}^{-1}\boldsymbol{\nu} \quad (31)$$

and the noise is now white. Eq. (31) is a transformation on the original space in (29). Substituting the transformed quantities $\mathbf{G}^{-1}\mathbf{A}$ for \mathbf{A} and $\mathbf{G}^{-1}\mathbf{b}$ for \mathbf{b} in (30), the normal equations become

$$\mathbf{x}_{LS} = (\mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{b} \quad (32)$$

the solution of which does have minimum variance, as demonstrated in Chapter 7. The LS problem in coloured noise is referred to as *generalized least squares*. It is interesting to note that the solution to (32) minimizes $\|\mathbf{Ax} - \mathbf{b}\|$ in the $\boldsymbol{\Sigma}^{-1}$ -metric. Metrics are discussed in more detail in the Appendix of this section.

The problem with the above approach is that if $\boldsymbol{\Sigma}$ is poorly conditioned, then small changes in $\boldsymbol{\Sigma}$ (which can result if $\boldsymbol{\Sigma}$ is not known accurately and must be estimated) can produce relatively large changes in λ_n (the smallest eigenvalue of $\boldsymbol{\Sigma}$), which can result in large changes in $\boldsymbol{\Sigma}^{-1}$. Hence, the normal equations (32) are not numerically stable.

We can rectify this situation by considering the following:

13.3.1 Generalized LS with QR Decompositions

From (31) we have

$$\mathbf{G}^{-1}\mathbf{b} = \mathbf{G}^{-1}\mathbf{Ax} + \mathbf{G}^{-1}\boldsymbol{\nu}$$

where \mathbf{G} is Cholesky factor of $\boldsymbol{\Sigma}$. The LS problem may thus be stated as

$$\min_{\mathbf{x}} \|\mathbf{G}^{-1}(\mathbf{Ax} - \mathbf{b})\|_2^2. \quad (33)$$

As before, we wish to avoid the explicit calculation of \mathbf{G}^{-1} . Along these lines, let us define the argument \mathbf{v} of (33) as

$$\mathbf{v} = \pm \mathbf{G}^{-1}(\mathbf{Ax} - \mathbf{b}). \quad (34)$$

The vector \mathbf{v} is the LS residual for this generalized LS problem. By choosing the negative sign above, we have

$$\mathbf{b} = \mathbf{Ax} + \mathbf{Gv}. \quad (35)$$

Another way to express the LS problem in coloured noise, expressed jointly by (33) and (35) is thus:

$$\min_{\mathbf{b}=\mathbf{Ax}+\mathbf{Gv}} (\mathbf{v}^T \mathbf{v}) \quad (36)$$

which is a constrained minimization problem where \mathbf{x} is the variable. Note that this problem is defined even if \mathbf{A} or \mathbf{G} is rank deficient.

The proposed technique for solving (36) is due to Paige², and is expressed in *Golub and van Loan*, p. 252. It is a very clever idea, since it provides a solution without explicitly computing any inverses.

The first step is to do a QR decomposition on \mathbf{A} :

$$\mathbf{Q}^T \mathbf{A} = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix}. \quad (37)$$

We let

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \quad (38)$$

n $m-n$

Next, we find an orthonormal matrix $\mathbf{Z} \in \mathfrak{R}^{m \times m}$ so that

$$\mathbf{Q}_2^T \mathbf{G} \mathbf{Z} = \begin{bmatrix} \mathbf{0} & \mathbf{S} \end{bmatrix} \quad (39)$$

n $m-n$

where \mathbf{S} is upper triangular. We see that (39) is a transposed QR decomposition on $\mathbf{Q}_2 \mathbf{G}$.

We also partition the orthonormal \mathbf{Z} as

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 \end{bmatrix} \quad (40)$$

n $m-n$

We may now express (35) as

$$\begin{aligned} \mathbf{b} &= \mathbf{A} \mathbf{x} + \mathbf{G} \mathbf{v} \\ \mathbf{Q}^T \mathbf{b} &= \mathbf{Q}^T \mathbf{A} \mathbf{x} + \mathbf{Q}^T \mathbf{G} \mathbf{v} \\ &= \mathbf{Q}^T \mathbf{A} \mathbf{x} + \mathbf{Q}^T \mathbf{G} \mathbf{Z} \mathbf{Z}^T \mathbf{v} \end{aligned} \quad (41)$$

where the 2-norms of terms above are equal. The partitions expressed by (38) to (40) can be incorporated into (41) in the following way:

$$\begin{aligned} \begin{bmatrix} \mathbf{Q}_1^T \mathbf{b} \\ \mathbf{Q}_2^T \mathbf{b} \end{bmatrix} &= \begin{bmatrix} \mathbf{Q}_1^T \mathbf{A} \\ \mathbf{Q}_2^T \mathbf{A} \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{Q}_1^T \mathbf{G} \mathbf{Z}_1 & \mathbf{Q}_1^T \mathbf{G} \mathbf{Z}_2 \\ \mathbf{Q}_2^T \mathbf{G} \mathbf{Z}_1 & \mathbf{Q}_2^T \mathbf{G} \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{Z}_1^T \mathbf{v} \\ \mathbf{Z}_2^T \mathbf{v} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{Q}_1^T \mathbf{G} \mathbf{Z}_1 & \mathbf{Q}_1^T \mathbf{G} \mathbf{Z}_2 \\ \mathbf{0} & \mathbf{S} \end{bmatrix} \begin{bmatrix} \mathbf{Z}_1^T \mathbf{v} \\ \mathbf{Z}_2^T \mathbf{v} \end{bmatrix} \end{aligned} \quad (42)$$

where (39) has been used in the last line.

²C.C. Paige, "Computer solution and perturbation analysis of generalized least squares problems", *Math. Comp.* 33, pp. 171-184, 1979, and
C.C. Paige, "Fast numerically stable computations for generalized linear least squares problems", *SIAM J. Num. Anal.* 16, pp. 165-171, 1979.

In (42), the LS residual \mathbf{v} is isolated in the bottom portion of the equation in a manner equivalent to the way the quantity \mathbf{d} is isolated in the ordinary LS situation. We may therefore solve for the minimum LS residual \mathbf{v} by solving

$$\mathbf{S}\mathbf{u} = \mathbf{Q}_2^T \mathbf{b} \quad (43)$$

for \mathbf{u} , where $\mathbf{u} \triangleq \mathbf{Z}_2^T \mathbf{v}$. Thus, \mathbf{v} can be determined from

$$\mathbf{v} = \mathbf{Z}_2 \mathbf{u}. \quad (44)$$

The \mathbf{x} which corresponds to this minimum residual is found by solving the top half of (42), once \mathbf{v} is determined as above:

$$\begin{aligned} \mathbf{R}_1 \mathbf{x} &= \mathbf{Q}_1^T \mathbf{b} - \left(\mathbf{Q}_1^T \mathbf{G} \mathbf{Z}_1 \mathbf{Z}_1^T + \mathbf{Q}_1^T \mathbf{G} \mathbf{Z}_2 \mathbf{Z}_2^T \right) \mathbf{v} \\ &= \mathbf{Q}_1^T \mathbf{b} - \mathbf{Q}_1^T \mathbf{G} \left(\mathbf{Z}_1 \mathbf{Z}_1^T + \mathbf{Z}_2 \mathbf{Z}_2^T \right) \mathbf{v} \\ &= \mathbf{Q}_1^T \mathbf{b} - \mathbf{Q}_1^T \mathbf{G} \mathbf{v} \end{aligned} \quad (45)$$

where the last term follows because $\mathbf{Z}_1 \mathbf{Z}_1^T + \mathbf{Z}_2 \mathbf{Z}_2^T = \mathbf{I}$. Because all the quantities above are known except \mathbf{x} , the above system is easily solved.

It is clear the choices for \mathbf{x} and \mathbf{v} from (45) and (44) respectively are consistent with the constraint (35). What remains to be shown is that this procedure yields a \mathbf{v} which satisfies (36); i.e., has a minimum norm subject to the constraints.

Clearly the \mathbf{u} satisfying (42) is unique when \mathbf{S} is full rank; (i.e., when \mathbf{G} is full rank). Thus, \mathbf{u} cannot be made smaller in norm without violating (35). From (44) we have

$$\|\mathbf{v}\|_2^2 = \|\mathbf{u}\|_2^2; \quad (46)$$

thus, the \mathbf{v} yielded by this procedure indeed has minimum norm.

This point can also be seen from a different perspective by assigning \mathbf{v} in (42) as

$$\tilde{\mathbf{v}} = \mathbf{v} + \mathbf{v}_1 \quad (47)$$

where \mathbf{v} satisfies (44) (i.e., $\mathbf{v} \in R(\mathbf{Z}_2)$) and $\mathbf{v}_1 \in R(\mathbf{Z}_1)$. It is clear from (42) that an \mathbf{x} can be found for this choice of \mathbf{v} such that (42) is satisfied. In this case, because \mathbf{Z} is orthonormal,

$$\|\tilde{\mathbf{v}}\|_2^2 = \|\mathbf{v}\|_2^2 + \|\mathbf{v}_1\|_2^2 \geq \|\mathbf{v}\|_2^2. \quad (48)$$

Thus the choice of \mathbf{v} given by (44) indeed has a minimum norm amongst all \mathbf{v} which are consistent with the constraint (35).

13.4 Appendix: Discussion of Metrics

In this section we briefly discuss the idea of an algebraic metric; i.e., how distances are measured. The Euclidean metric is the ordinary one. For a given vector \mathbf{x} , we measure its length or “distance” according to the Euclidean metric by evaluating $\mathbf{x}^T \mathbf{x} = \mathbf{x}^T \mathbf{I} \mathbf{x}$.

Now suppose we transform the space as we have done with (31), so that a vector \mathbf{x} in the old space becomes $\mathbf{G}^{-1} \mathbf{x}$ in the new space, where \mathbf{G} is some square full-rank matrix. Then, the length of the transformed vector is $\mathbf{x}^T \mathbf{G}^{-T} \mathbf{G}^{-1} \mathbf{x} = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$, where $\boldsymbol{\Sigma} = \mathbf{G} \mathbf{G}^T$. Thus, the expression for length is now the more general expression $\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$, where we have replaced the \mathbf{I} in the previous case with a more general matrix $\boldsymbol{\Sigma}^{-1}$.

In general, the quadratic form $\mathbf{x}^T \mathbf{A} \mathbf{x}$ is a distance measurement in the metric \mathbf{A} . This expression, denoted $\|\mathbf{x}\|_{\mathbf{A}}$, is referred to as the \mathbf{A} -metric.

The eigendecomposition of $\boldsymbol{\Sigma}^{-1}$ can be expressed as

$$\boldsymbol{\Sigma} = \mathbf{V} \boldsymbol{\Lambda}^{-1} \mathbf{V}^T. \quad (49)$$

By defining the vector \mathbf{y} as $\mathbf{V}^T \mathbf{x}$, $\|\mathbf{x}\|_{\boldsymbol{\Sigma}^{-1}}$ becomes

$$\|\mathbf{x}\|_{\boldsymbol{\Sigma}^{-1}} = \mathbf{y}^T \boldsymbol{\Lambda}^{-1} \mathbf{y} \quad (50)$$

$$= \sum_{i=1}^n \frac{y_i^2}{\lambda_i}. \quad (51)$$

Because \mathbf{y} is the vector \mathbf{x} expressed in the basis \mathbf{V} , we see that distances are now measured in the new metric along the eigenvectors of $\boldsymbol{\Sigma}$, in units of the corresponding eigenvalue. This is in contrast to the ordinary case where distances are measured along the co-ordinate axes in units of one.

INCLUDE FIGURE.

The matrix \mathbf{A} defining the \mathbf{A} -metric must be symmetric and positive semidefinite, otherwise the eigenvalues may be complex or negative, and the notion of distance will not exist in the usual sense.