Memo about outliers elimination

L. Petrov
pet@leo.gsfc.nasa.gov

October 1, 1997

Abstract:

Discussion about the optimal strategy for outliers elimination in LSQ for the case of full normal matrix and B_3D normal matrix. Algorithm implemented in SOLVE is described.

When we deal with real observations we used to disclose that the set of the observations can be partitioned onto two subsets: normal observations and abnormal observations. These two datasets have different statistics. Abnormal observations — outliers usually stem from some misfunction of appliance (although sometimes they may stem from the setbacks of the theoretical model). The presence of strong outliers may distort both the adjustments and the estimates of their uncertainties. It is essential to be able to detect outliers and to remove them from the solution. Various statistical criteria are used to separate normal and abnormal observations. The approaches how to do it will be discussed in chapter three. We'll discuss how to update the solution for removing outliers without re-calculation solution anew in the first two chapters. The way how automatic outliers elimination procedure is implemented in SOLVE is discussed in the last chapter.

1 Correction of solution for the case of full normal matrix

Assume that we have estimates of the parameters and their covariance matrix and we would like to update our solution for elimination of the k-th observation. Revised normal matrix without the k-th equation can be written in the form

$$N_r = N - a_k w_k^2 {a^\top}_k \tag{1}$$

where N — is initial normal matrix, a_k — expelled equation of conditions, w_k — its weight.

We can transform (1) using lemma about inversion of expanded matrix[1]

$$\left(N - a_k\, w_k^2\, {a^\top}_k\right)^{-1} = N^{-1} - N^{-1} a_k \left({a^\top}_k N^{-1} a_k - w^{-2}\right)^{-1} {a^\top}_k N^{-1} \tag{2}$$

Having noticed that $\left({a^\top}_k N^{-1} a_k - w^{-2}\right)$ is scalar we can rewrite (2) in the form:

$$\left(N - a_k\, w_k^2\, {a^\top}_k\right)^{-1} = N^{-1} + \frac{N^{-1} a_k\, w_k^2\, {a^\top}_k N^{-1}}{1 - w_k^2 {a^\top}_k N^{-1} a_k} \tag{3}$$

[1] See Appendix

We can find correction to the vector of adjustments obtained using all equations for removing the k-th equation . Corrected vector of adjustments $\hat{x}_r$ is

$$\hat{x}_r = N_r^{-1}(z - w_k^2\, y_k\, a_k) \tag{4}$$

where z — normal vector calculated for all equations. Having substituted expression (3) for N_r^{-1} and made some algebra we obtain

$$\begin{aligned}
x_r \; &= \; N^{-1}z \; - \; w_k^2\, y_k\, N^{-1}a_k \; + \; \frac{N^{-1}a_k\, w_k^2\, a^\top{}_k N^{-1}z}{1 - w_k^2 a^\top{}_k N^{-1}a_k} \; - \; w_k^2\, y_k \frac{N^{-1}a_k\, w_k^2\, a^\top{}_k N^{-1}a_k}{1 - w_k^2\, a^\top{}_k N^{-1}a_k} \; = \\
N^{-1}z \; &- \; \frac{w_k^2\, y_k N^{-1}a_k(1 - w_k^2 a^\top{}_k N^{-1}a_k) - N^{-1}a_k\, w_k^2 a^\top{}_k N^{-1}z + w_k^2\, y_k N^{-1}a_k\, w_k^2\, a^\top{}_k N^{-1}a_k}{1 - w_k^2 a^\top{}_k N^{-1}a_k} \; = \\
N^{-1}z \; &- \\
&\frac{w_k^2\, y_k N^{-1}a_k - w_k^2\, y_k N^{-1}a_k\, w_k^2\, a^\top{}_k N^{-1}a_k - N^{-1}a_k\, w_k^2 a^\top{}_k N^{-1}z + w_k^2\, y_k N^{-1}a_k\, w_k^2\, a^\top{}_k N^{-1}a_k}{1 - w_k^2 a^\top{}_k N^{-1}a_k} \; = \\
&= \; N^{-1}z \; - \; \frac{w_k^2\, y_k\, N^{-1}a_k - N^{-1}a_k\, w_k^2\, a^\top{}_k N^{-1}z}{1 - w_k^2\, a^\top{}_k N^{-1}a_k} \\
&= \; N^{-1}z \; - \; N^{-1}a_k\, w_k^2\, \frac{y_k - a^\top{}_k N^{-1}z}{1 - w_k^2\, a^\top{}_k N^{-1}a_k} \\
&= \; \hat{x} \; + \; N^{-1}a_k\, w_k\, \frac{a^\top{}_k\, w_k\, \hat{x} - w_k\, y_k}{1 - w_k^2\, a^\top{}_k N^{-1}a_k}
\end{aligned} \tag{5}$$

The asymptotic number of pairs of operations: multiplication and addition needed for updating adjustments and their covariance matrix is $Op^u(F) \approx \frac{3}{2}m^2$ where m is the number of parameters. The number of operations needed for obtaining adjustments in full solution is $Op^s(F) \approx \frac{1}{6}m^3$ (and $\frac{1}{3}m^3$ for obtaining covariance matrix). Thus, updating solution using expressions (3), (5) is faster by $m/3$ times than entire recalculation it anew.

2 Correction of solution for the case of bordered block tridiagonal normal matrix (B_3D)

When we solve normal equations using B_3D method we can try to exploit the sparseness of rejected equation of condition. Let's apply expressions (3), (5) to the blocks of submatrices of bordered block tridiagonal normal matrix. Given the d-th observation in the i-th block is rejected, the following algorithm can be proposed:

I. Calculation of $N^{-1}\, w_d\, a_d$ — a set of vectors q:

1) $\quad q_o = \mathrm{Cov}(\hat{x}_o,\, \hat{x}_o^\top)\, a_{id}^o\, w_{id} + \, \mathrm{Cov}^\top(\hat{x}_i,\, \hat{x}_o^\top)\, a_{id}^{l1}\, w_{id} + \, \mathrm{Cov}^\top(\hat{x}_{i+1},\, \hat{x}_o^\top)\, a_{id}^{l2}\, w_{id}$

2) $\forall\, j = 1, 2, \ldots n \quad q_j^l = \mathrm{Cov}(\hat{x}_j,\, \hat{x}_o^\top)\, a_{id}^o\, w_{id} + \, \mathrm{Cov}(\hat{x}_j,\, \hat{x}_i^\top)\, a_{id}^{l1}\, w_{id} + \, \mathrm{Cov}^\top(\hat{x}_{j+1},\, \hat{x}_i^\top)\, a_{id}^{l2}\, w_{id}$

II. Calculation of gain:

$$\begin{aligned}
3)\quad c_q \; &= \; a_{id}^{o\,\top} w_{id}\, q_o \; + \; a_{id}^{l1\,\top} w_{id}\, q_i^l \; + \; a_{id}^{l2\,\top} w_{id}\, q_{i+1}^l \\
4)\quad c_e \; &= \; a_{id}^{o\,\top} w_{id}\, \hat{x}_o \; + \; a_{id}^{l1\,\top} w_{id}\, \hat{x}_i^l \; + \; a_{id}^{l2\,\top} w_{id}\, \hat{x}_{i+1}^l \\
5)\quad \mathrm{g} \; &= \; \frac{1}{1 - c_q}
\end{aligned}$$

III. Calculation of corrected global adjustments and global-global block of the covariance matrix:

$$\begin{array}{lrcl} 6) & \hat{x}_o & = & \hat{x}_o + \mathrm{g}\,(c_e - y_d\,w_d)\,q_o \\ 7) & \mathrm{Cov}(\hat{x}_o,\,\hat{x}_o^\top) & = & \mathrm{Cov}(\hat{x}_o,\,\hat{x}_o^\top) + \mathrm{g}\,q_o q^\top{}_o \end{array}$$

IV. Calculation of corrected estimates of the local parameters, local-global and local-local blocks of the covariance matrix:

$$\begin{array}{llrcl} 8) & \forall\, j = 1, 2, \ldots n & \hat{x}_j & = & \hat{x}_j + \mathrm{g}\,(c_e - y_d\,w_d)\,q_j \\ 9) & \forall\, j = 1, 2, \ldots n & \mathrm{Cov}(\hat{x}_j,\,\hat{x}_o^\top) & = & \mathrm{Cov}(\hat{x}_j,\,\hat{x}_o^\top) + \mathrm{g}\,q_j q^\top{}_o \\ 10) & \forall\, j = 1, 2, \ldots n & & & \\ & \forall\, k = j, j+1, \ldots n & \mathrm{Cov}(\hat{x}_k,\,\hat{x}_j^\top) & = & \mathrm{Cov}(\hat{x}_k,\,\hat{x}_j^\top) + \mathrm{g}\,q_k q^\top{}_j \end{array}$$

It is worth to notice that off-diagonal blocks[2] of the covariance matrix are used in the proposed algorithm. The expressions for these blocks were omitted in [1]. Arbitrary off-diagonal block of the covariance matrix can be obtained by the following way:

$$\begin{aligned} \mathrm{Cov}(\hat{x}_{i+k}, \hat{x}_i^\top) &= \mathrm{Cov}(C_{i+k}^{-1} z_{i+k} - C_{i+k}^{-1} B_{i+k} \hat{x}_o - C_{i+k}^{-1} D_{i+k} \hat{x}_{i+k-1},\, \hat{x}_i^\top) = \\ &\mathrm{Cov}(C_{i+k}^{-1} z_{i+k},\, \hat{x}_i^\top) - \mathrm{Cov}(C_{i+k}^{-1} B_{i+k} \hat{x}_o,\, \hat{x}_i^\top) - \mathrm{Cov}(C_{i+k}^{-1} D_{i+k} \hat{x}_{i+k-1},\, \hat{x}_i^\top) = \\ &-\Big(C_{i+k}^{-1} B_{i+k} \mathrm{Cov}^\top(\hat{x}_i,\, \hat{x}_o^\top) + C_{i+k}^{-1} D_{i+k}\, \mathrm{Cov}(\hat{x}_{i+k-1},\, \hat{x}_i^\top)\Big) \end{aligned} \tag{6}$$

Thus, calculation of blocks of the covariance matrix goes down of the down-diagonal block row by row and then column by column. Each column of the blocks is calculated by the following recurrent algorithm:

$$\forall\, k = 2, 4, \ldots n \quad \mathrm{Cov}(\hat{x}_{i+k},\, \hat{x}_i^\top) - \Big(C_{i+k}^{-1} B_{i+k} \mathrm{Cov}^\top(\hat{x}_i,\, \hat{x}_o^\top) + C_{i+k}^{-1} D_{i+k}\, \mathrm{Cov}(\hat{x}_{i+k-1},\, \hat{x}_i^\top)\Big) \tag{7}$$

Let's calculate asymptotic computational expenses of the proposed algorithms under assumption that $n \gg 1$.

Calculation of the set of vectors q takes $Op^q(B_3D) \approx nl(g + 2l)$ and updating covariance matrix takes $Op^v(B_3D) \approx nl(g + \frac{n}{2}l)$ pairs of operations: multiplication and addition. Thus, computational complexity of updating the solution in B_3D mode is about $Op^u(B_3D) \approx \frac{nl}{2}((4+n)l + g)$ pairs of operations. But we should also take into account computational expenses for calculation of all off-diagonal blocks of the covariance matrix. They are not usually needed for ordinary solution and these expenses should be considered as pure overheads of proposed scheme of outliers rejection. $Op^o(B_3D) \approx \frac{n^2l^2}{2}(l+g)$. Table 1 contains a summary of computational expenses for the different steps of B_3D algorithm:

Table 1

$$\begin{aligned} \mathrm{Op}^s(B_3D) &\approx \frac{nl(2l+g)^2}{2} + \frac{g^3}{2} \\ \mathrm{Op}^o(B_3D) &\approx \frac{nl}{2}\, n\, l\, \Big(l + g\Big) \\ \mathrm{Op}^u(B_3D) &\approx \frac{nl}{2}\Big((n+4)\, l + 4g\Big) \\ \mathrm{Op}^c(B_3D) &\approx \frac{nl}{2}\Big((2l+g)^2 - l^2 + lg\Big) \end{aligned} \tag{8}$$

[2] I mean blocks away from the main diagonal, main down subdiagonal and main up subdiagonal.

where l — the number of local parameters in one group, n — the number of groups of local parameters, g — the number of global parameters.

Which strategy of outliers rejection for the case of B_3D has advantages: to recalculate solution anew or to update it? Computational time for the first strategy can be expressed as

$$T_a = s + k * s + c_c \tag{9}$$

where s — time for obtaining adjustments (without calculation of the covariance matrix), k — the number of outliers and c_c — time for calculation of the covariance matrix (only blocks which correspond to non-zero blocks of the normal matrix). Computational time for the second strategy is expressed as

$$T_u = s + c_o + k * u \tag{10}$$

where c_o — computational expenses for calculation of the off-diagonal blocks of the covariance matrix, u — computational time for one update of the adjustments and covariance matrices.

Let's try to express all constituents of the expressions (9), (10) via s for the typical values of n, l, g which we have in the problem of parameters estimation in VLBI data analysis.

First of all, notice that

$$\frac{c_c}{s} \approx \frac{(2l+g)^2 - l^2 + lg}{(2l+g)^2} \tag{11}$$

If $l = g$, then this ratio is about 1, if $l \gg g$ then it is $\frac{3}{4}$, and $\frac{c_c}{s} \approx 1$ when $l \ll g$.

$$\frac{c_o}{s} \approx n \frac{l(l+g)}{(2l+g)^2} \tag{12}$$

If $l = g$ then $\frac{c_o}{s} \approx 0.22n$, if $l \gg g$ then $\frac{c_o}{s} \approx 0.25n$, if $5l = g$ then $\frac{c_o}{s} \approx 0.12n$ and computational expenses for calculation of the off-diagonal submatrices become negligible when $l \ll g$.

$$\frac{u}{s} \approx n \frac{l}{(2l+g)^2} \tag{13}$$

If $l = g$ then $\frac{u}{s} \approx 0.11\frac{n}{l}$, if $l \gg g$ then $\frac{u}{s} \approx 0.25\frac{n}{l}$, if $g = 5l$ then $\frac{u}{s} \approx 0.10\frac{n}{g}$.

We have n=25, l=20, g=20 for typical session. Then $\frac{u}{s} \approx 0.1$

Now we can easily evaluate T_a and T_u for typical session (when n=25, l=20, g=20): $T_a = (k+2)s$, $T_u = 7s + 0.1k$

We obtained rather unexpected result: there is some critical value of the number of eliminated outliers. If the number of outliers which we are going to eliminate less than this critical value it is more profitable to recalculate the solution anew. If the number of outliers exceeds this critical number then updating solution takes less time. This magic number is **6** for the typical session. It increases linearly with increasing the number of blocks of local parameters and decreases with increasing the ratio $\frac{g}{l}$.

3 How to find outliers?

There is no definite answer. We assume that normal observations belongs to one statistics but abnormal observations belongs to another statistics. Analyzing residuals we can try to separate them. One strategy is to declare a priori statistic. We can state that our residuals belong to normal distribution with certain dispersion. And then having stated statistical confidence level we can put aside the observations which are out of considered statistical ensemble. In practice in that case we set up the threshold and all observations with residuals higher than this threshold we consider as outliers. The setback of this way is that we usually don't know statistical properties of postfit residuals.

The second approach is to use some kind of a posteriori statistics. We can calculate *the estimate* of the dispersion of distribution and to express the threshold in sigmas. Under assumption of normal distribution the probability of the event that the observation with residual deviating higher than 2σ belongs to the statistical ensemble is 5% and deviating higher than 3σ is 0.3%. The last criterion seems to be sufficient for practical needs. However we should keep in mind that normal distribution implies statistical independence of the residuals. It is not quite true due to the contribution of unmodeled effects and the actual distribution is nearer to χ^2 distribution which is known to converge to the normal distribution when the number of degrees of freedom goes to infinity but it has larger "tails" than normal distribution. For this reason it appears sometimes that 3σ criterion is too stiff and should be increased to 3.5, or 4σ. My preference for the case of analysis of VLBI observations is 3.5σ. As a rule of thumb we should take special care to the statistical criteria when the pattern of out residuals shows strong systematic behavior.

If our statistical ensemble is not homogeneous it is worth to partition the set of residuals onto some subsets. If we didn't apply baseline or station dependent reweighting we should calculate dispersion of residuals for each baseline separately. From the other hand the presence of outliers is able to distort reweighting constants and to force reweighting to add too much noise to absorb the influence of outliers. Reweighting and outliers rejection are in some extent concurrent processes.

It is essential that the outliers should be removed in strict order: first the most strong outlier. Otherwise one strong outlier may force to rejection of good observations (in some pathological cases all observations may appear to be rejected!) since the presence of outliers distorts adjustments and therefore the residuals itself.

4 Generalizations

Outliers rejection was mentioned in the first two chapters. It is worth noting that we are able to change slightly the problem. Let we have solution of LSQ problems for K equations of conditions. Let's find the correction to the solution for adding new K+1 -th equation.

To do it notice that the only thing which we should do is to change the sign in (1).

The only consequence of such a change is that the sign in denominator in (3)–(5) will differ: $-(1 + w_k a^\top{}_k N^{-1} a_k)$ instead of $1 - w_k a^\top{}_k N^{-1} a_k$ as well as in denominator in expression for gain g used in B_3D scheme: $-(1 + c_q)$ instead of $1 - c_q$

The problem considered can be easily generalized to the following problem: let we have the LSQ solution for K equations of conditions. And let the k-th equation had the weight w_k^o. How to find the correction to the solution when we change the weight from w_k^o to w_k^n? The answer this question is reduced to the cases considered above.

5 Implementation elimination/restoration algorithm in SOLVE

5.1 Outliers elimination

After obtaining the parameters estimates, full covariance matrix and calculation of postfit residuals square root from dispersion of postfit residuals is found:

$$D = \sqrt{\frac{\sum\limits_{i=1}^{i=N} (p_i w_i)^2}{N - 1}}$$

where p_i is postfit residual for the i-th observation, w_i its weight and summing is done either over all used observations for all baselines or over all used observations for the certain b-th baseline.

The quantity $p_i^n = \dfrac{p_i\, w_i}{D}$ we call here normalized residual. Two kinds of normalized residuals can be used depending on what kind of dispersion is used for normalization: baseline-dependent dispersion calculated for all observations of the baseline where the i-th observation has been done or dispersion calculated for all used observations of the database.

Two criteria are used for outlier detection: threshold criterion and nσ cutoff criterion. Observations with postfit residuals larger in modulo than the specified threshold and observations with normalized residuals exceeding in modulo the specified cutoff limit are marked as outliers. One of these criteria or both can be used. The outlier with maximal postfit residual is found among the set of outliers[3].

Then estimates of the parameters, full covariance matrix are updated for elimination of the most considerable outlier, all postfit residuals are recalculated, statistics are calculated anew and the observation yielded this outlier is suppresed for further participation in estimation. The process is iterated until no one outlier will be detected.

5.2 Restoration previously suppressed observations

All observations can be partitioned onto three categories: 1) observations used in solution; 2) observations which were not used in solution but are in principle restorable; 3) observation which were not used in solution and which are marked as not restorable in principle (no fringe detected, the lack of ionosphere correction and etc). Inverted threshold and cutoff nσ criterion is applied to the observations from the second group: observations with postfit residuals less in modulo than the specified threshold and observations with normalized residuals not exceeding in modulo

[3]If both criteria are used then the observation with maximal normalized residual will be considered as the most considerable outlier.

the specified cutoff limit are marked as candidates in restoration. One of these criteria or both can be used. The candidate in restoration with minimal postfit residual is found among the set of candidates in restoration.

Then the estimates of the parameters, full covariance matrix are updated for restoration of the most favorable considered candidate in restoration, all postfit residuals are recalculated and statistics are calculated anew as well as for the case of outliers elimination procedure. Restored observation will be used in further solutions. The process is iterated until no one candidate in restoration will be detected.

5.3 Solution update

Since SOLVE uses such a scheme of parameterization that the values of the estimated parameters vary in very large range (up to 20 orders!) it is critical to use scaling by proper way. The normal equations which are solved by SOLVE can be written in such a form:

$$(A^\top w^2 A) \cdot S^2 = A^\top w \cdot S \cdot y \tag{14}$$

where A — is the matrix of equation of conditions, w — vector of weights, y — vector of right parts of equation of conditions (o-c of observable under consideration), S — diagonal matrix of scales. Matrix S is chosen by such a manner that the matrix $(A^\top w^2 A) \cdot S^2$ has a unit main diagonal. As a result we determine in "raw solution" not the vector of parameter estimates $\hat{x}$ and their covariance matrix $\mathrm{Cov}(\hat{x},\, \hat{x}^\top)$, but scaled results: $S^{-1}\hat{x}$ and $S^{-2}\,\mathrm{Cov}(\hat{x},\, \hat{x}^\top)$.

Taking into account scaling scheme the following algorithm for update of the solution for elimination or restoration the d-th observation in the i-th block is proposed:

I. Calculation of scaled weighted equation of conditions:

$$1) \qquad \tilde{a}_i = w_i\, a_i * s$$

where s — the vector of scales located at the main diagonal of the matrix S, and sign $*$ denotes element-by-element multiplication of the elements of two vectors.

II. Calculation of $(N^{-1}\, S^{-2})\, S\, w_d\, a_d$ — a set of vectors q:

$$2) \qquad q_o = S^{-2}\,\mathrm{Cov}(\hat{x}_o,\, \hat{x}_o^\top)\, \tilde{a}^o_{id} + \, S^{-2}\mathrm{Cov}^\top(\hat{x}_i,\, \hat{x}_o^\top)\, \tilde{a}^{l1}_{id} + \, S^{-2}\mathrm{Cov}^\top(\hat{x}_{i+1},\, \hat{x}_o^\top)\, \tilde{a}^{l2}_{id}$$

$$3)\ \forall\, j = 1, 2, \ldots n \qquad q^l_j = S^{-2}\,\mathrm{Cov}(\hat{x}_j,\, \hat{x}_o^\top)\, \tilde{a}^o_{id} + \, S^{-2}\,\mathrm{Cov}(\hat{x}_j,\, \hat{x}_i^\top)\, \tilde{a}^{l1}_{id} + \, S^{-2}\mathrm{Cov}^\top(\hat{x}_{j+1},\, \hat{x}_i^\top)\, \tilde{a}^{l2}_{id}$$

III. Calculation of gain:

$$\begin{array}{rrcl}
4) & c_q & = & \tilde{a}^{o}_{id}{}^\top q_o \,+\, \tilde{a}^{l1}_{id}{}^\top q^l_i \,+\, \tilde{a}^{l2}_{id}{}^\top q^l_{i+1} \\
5) & c_e & = & \tilde{a}^{o}_{id}{}^\top \hat{x}_o \,+\, \tilde{a}^{l1}_{id}{}^\top \hat{x}^l_i \,+\, \tilde{a}^{l2}_{id}{}^\top \hat{x}^l_{i+1} \\
6a) & \text{Elimination: g} & = & \dfrac{1}{1 - c_q} \\
6b) & \text{Restoration: g} & = & -\dfrac{1}{1 + c_q}
\end{array}$$

IV. Calculation of corrected global adjustments and global-global block of the covariance matrix:

$$\begin{array}{rrcl}
7) & S^{-1}\hat{x}_o & = & S^{-1}\hat{x}_o \,+\, \mathrm{g}\,(c_e - y_d\, w_d)\, q_o \\
8) & S^{-2}\,\mathrm{Cov}(\hat{x}_o,\, \hat{x}_o^\top) & = & S^{-2}\,\mathrm{Cov}(\hat{x}_o,\, \hat{x}_o^\top) \,+\, \mathrm{g}\, q_o {q^\top}_o
\end{array}$$

V. Calculation of corrected estimates of the local parameters, local-global and local-local blocks of the covariance matrix:

$$\begin{array}{llrcl}
9) & \forall\, j = 1, 2, \ldots n & S^{-1}\hat{x}_j & = & S^{-1}\hat{x}_j \;+\; \mathrm{g}\,(c_e - y_d\, w_d)\, q_j \\
10) & \forall\, j = 1, 2, \ldots n & S^{-2}\,\mathrm{Cov}(\hat{x}_j,\, \hat{x}_o^\top) & = & S^{-2}\,\mathrm{Cov}(\hat{x}_j,\, \hat{x}_o^\top) \;+\; \mathrm{g}\, q_j {q^\top}_o \\
11) & \forall\, j = 1, 2, \ldots n & & & \\
 & \forall\, k = j, j+1, \ldots n & S^{-2}\,\mathrm{Cov}(\hat{x}_k,\, \hat{x}_j^\top) & = & S^{-2}\,\mathrm{Cov}(\hat{x}_k,\, \hat{x}_j^\top) \;+\; \mathrm{g}\, q_k {q^\top}_j
\end{array}$$

6 Appendix

Lemma *about inversion of expanded matrix:*

$$\Big(A - B^\top C B\Big)^{-1} = A^{-1} - A^{-1} B^\top \Big(B A^{-1} B^\top - C^{-1}\Big)^{-1} B A^{-1} \qquad (15)$$

when matrices A^{-1}, C^{-1}, $\Big(A - B^\top C B\Big)^{-1}$ and $\Big(B A^{-1} B^\top - C^{-1}\Big)^{-1}$ exist.

This lemma can be proven by direct calculation. Let's multiply both parts of the (15) by $(A - B^\top C B)$ and do some algebra:

$$\begin{array}{l}
\Big(A^{-1} - A^{-1}B^\top\Big(BA^{-1}B^\top - C^{-1}\Big)^{-1}BA^{-1}\Big) \cdot \Big(A - B^\top CB\Big) = \\
I - A^{-1}B^\top\Big(BA^{-1}B^\top - C^{-1}\Big)^{-1}B - A^{-1}B^\top CB + A^{-1}B^\top\Big(BA^{-1}B^\top - C^{-1}\Big)^{-1}BA^{-1}B^\top CB = \\
I - A^{-1}B^\top CB + \\
\qquad \Big[A^{-1}B^\top\Big(BA^{-1}B^\top - C^{-1}\Big)^{-1} \cdot BA^{-1}B^\top \cdot CB - A^{-1}B^\top\Big(BA^{-1}B^\top - C^{-1}\Big)^{-1} \cdot C^{-1} \cdot CB\Big] = \\
I - A^{-1}B^\top CB + \Big[A^{-1}B^\top\Big(BA^{-1}B^\top - C^{-1}\Big)^{-1} \cdot \Big(BA^{-1}B^\top - C^{-1}\Big) \cdot CB\Big] = \\
I - A^{-1}B^\top CB + A^{-1}B^\top CB = \\
I
\end{array}$$

Lemma is proven.

References

[1] Petrov L. Multigroup LSQ method and its generalization. *(Submitted to Computational Statistics and Data Analysis)*, 1997.