Algorithm for Sequential Estimation of the Covariance Matrix and some Applications

Edgard M. Maboudou[1] and Douglas M. Hawkins[2]

[1] Department of Statistics, University of Central Florida,
4000 Central Florida Blvd,
Orlando, FL 32816, USA
emaboudo@mail.ucf.edu

[2] School of Statistics, University of Minnesota,
224 Church Street S.E. ,
Minneapolis, MN 55455, USA
dhawkins@umn.edu

Abstract. Multivariate control charts are valuable tools in industrial quality control. They are designed to detect changes in any direction. Changes can occur in either the location or the variability of the correlated multivariate quality characteristics calling for methodologies for detecting changes in the covariance matrix. We recall Shewart's famous dictum that proper quality control involves both the mean and the variance of the process. This paper develops a reasonably fast algorithm to compute the new value of the covariance matrix each time a new row is added to the data set. We will show later how the computation of the generalized variance and the trace can be easily accelerated from this algorithm.

Keywords. Matrix factorization, Nonsingularity, Positive definite matrix.

1 Introduction

Several methodologies for monitoring the covariance matrix are available. The generalized variance of Montgomery and Wadsworth (1972) involves charting the determinant of $\mathbf{S}$. Another methodology consists of charting the trace of the covariance matrix $\mathbf{S}$, Reynolds and Cho (2006). Another alternative is the generalized likelihood ratio (GLR) statistic, which was proposed as a chart quantity by Alt (1984). In each case, we have to compute the covariance matrix $\mathbf{S}$ of the data.

When dealing with individual observations rather than rational groups, one has to recompute the covariance matrix each time a new observation is added to the data. This task can be a burdensome process essentially when we have a huge data set.

Let $\mathbf{S}_i$ be the conventional sample covariance matrix of the data up to observation i, $i = p+1, \ldots, n$ where n is the number of observations in the data. Some applications using the sample covariance matrix are:

1. Generalized variance of $\mathbf{S}_i$
2. Trace of $\mathbf{S}_i$
3. Generalized likelihood ratio (omitting unneeded constants), $GLR = tr(\mathbf{S}_i) - log|\mathbf{S}_i| - p$
4. Criterion for multiple change-point of multivariate data, $C = \sum_{i=1}^{n} r_i \log|\mathbf{S}_i|$

We estimate the covariance matrix in the first section. In the second section, we discuss some applications and the algorithm is presented in the third section.

2 Estimation of the covariance matrix

Assume that the p component vector $\mathbf{X}_i, \quad i = 1, 2, ..., n$ is a sequence of independent distributed p-dimensional random vectors with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. An unbiased estimator of $\boldsymbol{\Sigma}$ is the sample covariance matrix.

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' \tag{1}$$

where $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$.

The sample covariance matrix gives a picture of the correlation between each variable in the sample. Now,

$$\mathbf{S} = \frac{1}{n-1}\mathbf{C}\mathbf{C}' \tag{2}$$

where $\mathbf{C}$ is the unique $p \times p$ lower triangular matrix satisfying (2). Since $n \geq p+1$, $\mathbf{S}$ is clearly a symmetric positive definite matrix with probability 1 (w.p.1), so that $\mathbf{C}$ in (2) is unique with probability 1– sea Eaton (1983). Then $\mathbf{C}$ is the square root matrix of $\mathbf{S}$ also known as the Cholesky factor of $\mathbf{S}$

3 Applications

3.1 Total sample variance

The total sample variance is defined as the trace of the sample covariance matrix $\mathbf{S}$. It describes the variability of the data without taking into account the covariances. Let s_{ii} be the i^{th} diagonal element of the sample covariance matrix, the trace of $\mathbf{S}$ is $tr(\mathbf{S}) = \sum_{i=1}^{p} s_{ii}$. Also, the trace of $\mathbf{S}$ can be constructed from the Cholesky factor by using the following result

$$\begin{aligned} tr(\mathbf{S}) &= \frac{1}{n-1}tr(\mathbf{C}\mathbf{C}') \\ &= \frac{1}{n-1}\sum_{i=1}^{p}\sum_{k=1}^{i} c_{ik}^2 \end{aligned}$$

where the c_{ij} are the elements of the matrix $\mathbf{C}$.

3.2 Generalized sample variance

The generalized sample variance provided a single-number summary of the sample covariance matrix. It is defined as the determinant of the sample covariance matrix $|\mathbf{S}|$. Another way to compute the generalized sample variance is to use the Choloseky decomposition of $|\mathbf{S}|$. Let c_{ii} be the i^{th} diagonal element of the Cholesky factor $\mathbf{C}$, then

$$|\mathbf{S}| = \frac{1}{(n-1)^p}(\prod_{i=1}^{p} c_{ii})^2 \tag{3}$$

Proof.

$$\begin{aligned} |\mathbf{S}| &= \frac{1}{(n-1)^p}|\mathbf{C}\mathbf{C}'| \\ &= \frac{1}{(n-1)^p}|\mathbf{C}|^2 \\ &= \frac{1}{(n-1)^p}(\prod_{i=1}^{p} c_{ii})^2 \end{aligned}$$

So, the generalized sample variance can be obtained as the product of the square of the diagonal element of the Cholesky factor $\mathbf{C}$ divided by $(n-1)^p$.

3.3 Logarithm of the generalized sample variance

When dealing with likelihood function or generalized likelihood ratio, the logarithm of the generalized sample variance is the quantity of interest.

$$log(|\mathbf{S}|) = -plog(n-1) + 2\sum_{i=1}^{p} log(c_{ii}) \tag{4}$$

The logarithm of the generalized variance is twice the sum of the logarithm of the diagonal elements of the Cholesky matrix $\mathbf{C}$ minus the term $plog(n-1)$

4 Computation

It is as important to monitor the process mean vector as the process variability. Several methods were proposed to monitor the process variability. One of them is the generalized variance. This yields to compute a sequential generalized variance. The procedure is as follows:

1. For each new observation, compute the new value of the covariance matrix $\mathbf{S}_n$,
2. compute the determinant of $\mathbf{S}_n$,

The reader is referred to Montgomery and Wadsworth (1972) for more details.

Recomputing the covariance matrix each time a new observation is added can be burdensome.

A reasonably fast numerically stable approach is given by use of Cholesky factorizations. Append a 1 to each data vector $\mathbf{X}_i$, writing

$\mathbf{Z}_i = (1, \mathbf{X}_i')'$

Define

$\mathbf{B}_{h,m} = \sum_{j=h+1}^{m} \mathbf{Z}_j \mathbf{Z}_j'$

and form the lower triangular Cholesky factorization

$\mathbf{B}_{h,m} = \mathbf{C}_{h,m} \mathbf{C}_{h,m}'$

where the Cholesky factor matrix $\mathbf{C}_{h,m}$ is of dimension $p+1$. Standard results then show that (writing $c_{j,j}$ for the j^{th} diagonal element of $\mathbf{C}_{h,m}$)

$|\mathbf{A}_{h,m}| = \prod_{j=2}^{p+1} c_{j,j}^2$

The attraction of the Cholesky factorization is that adding a new data vector $\mathbf{X}_{m+1}$, corresponding $\mathbf{C}_{h,m+1}$ can be computed from $\mathbf{C}_{h,m}$ and $\mathbf{Z}_{m+1}$ with a fast, stable update (see for example Chambers, 1971).

This leads to the outline algorithm: For each $h = 1$ to $n-p$, initialize $\mathbf{C}_{h,h-1} = 0$. Then for $m = h$ to n, calculate $\mathbf{C}_{h,m}$ from $\mathbf{C}_{h,m-1}$ and $\mathbf{Z}_m$ using the update. Once enough observations are included to achieve nonsingularity, use the diagonal elements of $\mathbf{C}_{h,m}$ to compute $|A_{h,m}|$ and from this $\log|A_{h,m}|$ for example.

The following algorithm can be used to compute the trace and the generalized likelihood ratio

5 Example

We give an illustration of the algorithm on a small data. The data is presented in the table below.

X_1	X_2	X_3
12	10	7
8	12	8
11	9	9
9	13	10
13	7	11

This data has three variables X_1, X_2, and X_3, and five observations. A direct computation of the covariance matrix, the unnormalized covariance matrix $\sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$, the determinant, the log determinant, and the Cholesky factor gives:

```
>#  Covariance matrix S of the data
> s<-cov(data);s
       x1    x2    x3
x1   4.30 -4.40  0.75
x2  -4.40  5.70 -1.25
```

```
x3  0.75 -1.25  2.50

> # Determinant of S
> d<-det(s);d
[1] 11.2

> Log determinant of S
> ldet<-log(d);ldet
[1] 2.415914

> Unnormalized covariance matrix S
> ss<-(n-1)*cov(data);ss
      x1    x2 x3
x1  17.2 -17.6  3
x2 -17.6  22.8 -5
x3   3.0  -5.0 10

> Cholesky factor of the numerator
> ch<-chol(ss);ch
         x1        x2         x3
x1 4.147288 -4.243737  0.7233642
x2 0.000000  2.188766 -0.8818815
x3 0.000000  0.000000  2.9494117
```

Next, we will apply the algorithm to the data to compute the determinant and log determinant of S without computing neither the covariance matrix nor the numerator of the covariance matrix.

```
> q<-p+1
> rr<-matrix(rep(0,q*q),q,q)
> lndet<-0
> for(m in 1:n){
+       xx<-data[m,];x<-cbind(1,t(xx))
+       rr<-s(x,rr,1,n)
+        }

> # Cholesky factor using the updating algorithm
> print(rr)
         [,1]      [,2]      [,3]       [,4]
[1,] 2.236068 23.702321 22.807893 20.1246118
[2,] 0.000000  4.147288 -4.243737  0.7233642
[3,] 0.000000  0.000000  2.188766 -0.8818815
[4,] 0.000000  0.000000  0.000000  2.9494117

> # Remove the first value of the diagonal of the Cholesky
> dchl<-diag(rr);dch<-dchl[-1]

> # Computing the determinant of S
> dets<-prod(dch^2)/(n-1)^p;dets
[1] 11.2

> # Computing the log determinant of S
> lndet<-2*sum(log(dch))-p*log(n-1);lndet
```

```
[1] 2.415914
```

The function "s(x,rr,...)" used in the code above is responsible for the update algorithm and is not provided in this paper. A Fortran code of the updating algorithm of the Cholesky is available in Hawkins and Maboudou-Tchao (2007). Removing the first row and first column of the updated Cholesky "rr" yields the Cholesky factor denoted "ch" obtained using the direct approach. Also, the determinant of $\mathbf{S}$, "dets" is computed using equation (3) and the log determinant of $\mathbf{S}$, "lndet" is obtained using equation (4). The results match the ones obtained using the direct computation.

6 Conclusion

We show in this paper that when we have to do a sequential estimation of the covariance matrix for a huge data set, it is interesting to use the Cholesky factorization. As a new case is added, we don't need to recompute the new Cholesky factor. We just need to update the Cholesky using a numerically stable udpate algorithm. Also, we explained how to derive some quantities of interest like the trace and the determinant of the covariance matrix. This method can be also extended to the generalized likelihood ratio.

References

Beasly, J. D. and Smith, S. G. (1985). The Percentage Points of the Normal Distribution. In: *Applied Statistics Algorithms* (P. Griffiths and I.D. Hill, Eds). Ellis Horwood, London, 238–242.

Alt, F.A. (1984). Multivariate Quality Control. In:*The Encyclopedia of Statistical Sciences* (Ed N. L. Johnson, S. Kotz and C.R. Read, New York, Wiley).**6**, 110–122.

Anderson, T. W.(2003) *An Introduction to Multivariate Statistical Analysis 3rd Ed*. John Wiley and Sons New York.

Chambers, J. M., (1971). Regression Updating. *Journal of the American Statistical Association*, **66**, 744–748.

Eaton, M. L.(1983) *Multivariate Statistics: A vector approach*. John Wiley and Sons New York.

Hawkins, D. M. and Olwell, D. H. (1997) *Cumulative Sum Charts and Charting for quality Improvement*. New York, Springer-Verlag.

Hawkins, D. M., Maboudou-Tchao, E. M., (2007) Self-starting multivariate exponentially weighted moving average control charting for location. *Technometrics*, **49**, 199–209.

Hawkins, D. M. and Maboudou-Tchao, E. M. Fitting a multiple change-point model to a multivariate Gaussian model. manuscript submitted for publication.

Montgomery, D. C. (2005), *Introduction to Statistical Quality Control*. 5th edition, Wiley

Montgomery, D.C., Wadsworth, H.M., (1972). Some Techniques for Multivariate Quality Control Applications. Transactions of the ASQC, Washington, D.C.

Reynolds, M. R., Cho, G.-Y. (2006). Multivariate Control Charts for Monitoring the Mean Vector and Covariance Matrix. *Journal of Quality Technology.* **38**, 230–253.