Atkins, P. J., & Cummins, M. (2020).
A dynamic Cholesky data imputation method for correlation structure consistency.
Applied Economics Letters, 29(4), 311-315.
https://doi.org/10.1080/13504851.2020.1866153

A dynamic Cholesky data imputation method for correlation structure consistency

Philip J. Atkins[a] and Mark Cummins[b]

[a]Credit Portfolio Management, BP Oil International Limited, 20 Canada Square, E14 5NJ, London, UK; [b]DCU Business School, Dublin City University, Ireland

ABSTRACT

In the context of data that is missing completely at random, we propose a new data imputation method that exploits Cholesky decomposition. The data imputation method falls within the multiple imputation framework and is designed to ensure consistency with the correlation structure of the available data. The advantage is an accessible and computationally efficient approach to managing missing data that avoids the model risk associated with applying complex model based data imputation methods. The non-recursive nature of our data imputation method further avoids the convergence issues associated with recursive approaches.

KEYWORDS

missing data imputation; dynamic Cholesky; correlation structure consistency

JEL CLASSIFICATION: C1; C8

1. Introduction

Little (1992), Pigott (2001) and Allison (2001) provide useful surveys of missing data methods. Rubin (1987, 1996) advocates multiple imputation on the basis of accessibility and computational efficiency. Multiple imputation has received various treatments in the literature. Wang and Robins (1998) derive the asymptotic variance structures of alternative methods in a large sample setting. Robins and Wang (2000) derive an estimator of the asymptotic variance of both single and multiple imputation estimators. Degree of freedom adjustments in the small sample setting are proposed by Barnard and Rubin (1999) and Reiter (2007).

Aerts et al. (2002) propose nonparametric and semiparametric methods based on local resampling procedures.

Our method - "Dynamic Cholesky" - falls within the multiple imputation framework. We contribute through a new tractable data imputation method that exploits Cholesky decomposition and retains the correlation structure of the available data. The approach avoids introducing model risk from applying complex model based data imputation methods. The non-recursive design avoids the convergence issues associated with recursive methods, such as expectation-maximization (Dempster et al., 1977).

We deal with a scenario where data is missing completely at random (Rubin, 1976). This setting has received considerable attention in the academic literature. Zhang (2013), for example, proposes a model averaging approach. This contrasts to the situation where the occurrence of the missing data is not completely random. In such a missing not at random case the researcher would need to investigate the source of sample selection bias in order to avoid applying an inappropriate data imputation method.

We therefore confine ourselves to the missing completely at random case. Our practice-relevant context is that of energy markets where incomplete datasets are common. We further assume a risk management application that prioritizes the correlation structure over point estimates. However, there are several pertinent settings where missing data is a problem and where our data imputation method could be useful. These include cross-country emerging equity markets (e.g. Abd Majid, M.S. and Kassim, S.H., 2009), emerging interest rate markets (e.g. Nagy, 2020), and credit derivatives markets (e.g. Hüttner et al., 2020).

The paper is organized as follows. Section 2 considers the problem of imputing continguous missing data, presenting some important concepts used in the Dynamic Cholesky method proposed in Section 3. Section 4 concludes.

2. Cholesky method for imputing contiguous missing data

Before presenting our Dynamic Cholesky method, we consider the problem of filling contiguous gaps in data. The set up is artificial in the sense that we impose the contiguous gaps on some complete dataset. We highlight some important concepts for the Dynamic Cholesky method.

Assume an $N \times n$ matrix H_n of historical price returns, with each row a cross-section of price returns and each column a different price returns series over time. We calculate the covariance matrix C_n. Now suppose that we artificially zero out the returns data for price returns series beyond some column position $m < n$, i.e. columns $(m+1)$-n are assumed to be missing. So we assume we only have an

$N \times m$ matrix H_m of historical price returns, but we know the covariance matrix C_n. We create an extended quasi-history $H_n^{'}$ from H_m.

Define C_m to be the top-left $m \times m$ block of C_n, and $\widetilde{C}_m \equiv chol(C_m)$ its upper triangular Cholesky matrix. Consider now the following matrix:

$$K_m = H_m \widetilde{C}_m^{-1}$$

K_m is a version of H_m in which the returns have been standardized, by calculating the associated covariance matrix C_{K_m}:

$$\begin{aligned} C_{K_m} &= \frac{K_m^\top K_m}{N} = \frac{(H_m \widetilde{C}_m^{-1})^\top (H_m \widetilde{C}_m^{-1})}{N} = (\widetilde{C}_m^{-1})^\top \frac{(H_m^\top H_m)}{N} \widetilde{C}_m^{-1} \\ &= (\widetilde{C}_m^{-1})^\top C_m \widetilde{C}_m^{-1} = (\widetilde{C}_m^{-1})^\top (\widetilde{C}_m^\top \widetilde{C}_m) \widetilde{C}_m^{-1} = \left[(\widetilde{C}_m^\top)^{-1} \widetilde{C}_m^\top \right] \left[\widetilde{C}_m \widetilde{C}_m^{-1} \right] = I \end{aligned} \tag{1}$$

We extend the dimension of matrix K_m to n columns by post-appending an $N \times (n-m)$ matrix Φ_{n-m} of uncorrelated standard normal random numbers:

$$K_n = [K_m \ \Phi_{n-m}]$$

We can now compute the analogue of (1):

$$\begin{aligned} C_{K_n} &= \frac{K_n^\top K_n}{N} = \frac{1}{N} \begin{bmatrix} K_m^\top \\ \Phi_{n-m}^\top \end{bmatrix} [K_m \Phi_{n-m}] = \frac{1}{N} \begin{bmatrix} K_m^\top K_m & K_m^\top \Phi_{n-m} \\ \Phi_{n-m}^\top K_m & \Phi_{n-m}^\top \Phi_{n-m} \end{bmatrix} \\ &\approx \begin{bmatrix} I_m & 0_{m,n-m} \\ 0_{n-m,m} & I_{n-m} \end{bmatrix} = I_n \end{aligned}$$

The extended quasi-history then follows:

$$H_n^{'} = K_n \widetilde{C}_n$$

If we compute the corresponding covariance matrix:

$$C_{H_n^{'}} = \frac{H_n^{'\top} H_n^{'}}{N} = \frac{(K_n \widetilde{C}_n)^\top (K_n \widetilde{C}_n)}{N} = \widetilde{C}_n^\top \frac{(K_n^\top K_n)}{N} \widetilde{C}_n \approx \widetilde{C}_n^\top \widetilde{C}_n = C_n$$

This shows that we can impute an extended quasi-history $H_n^{'}$ from the subset

H_m that is perfectly consistent with the covariance matrix C_n.

3. Dynamic Cholesky method for imputing random missing data

We move to the real-world problem of imputing missing data positioned randomly in a dataset. We develop our Dynamic Cholesky method. Working sequentially, row-by-row, through a given data structure, we manoeuvre the complete data to the left, into a contiguous block, and the missing data to the right, into an adjacent contiguous block. The Cholesky method of Section 2 can then be applied.

Consider an $N \times n$ price returns matrix H with randomly positioned data gaps and fill the data gaps with zeros. We create an $N \times n$ matrix, D, to identify the complete data: 1 (0) indicates a position in H with complete (missing) data. We call D the 'data indicator matrix'.

Take a specific row h of H and corresponding row d of D. Manoeuvring the complete (missing) data to the left (right) can be achieved by post-multiplying h by a rearranged version of the $n \times n$ identity matrix I_n, which we call the 'rearranger matrix' R.

To illustrate: assume a log returns structure $h = (r_1, 0, r_3, 0, r_5)$, where zeros initially fill the missing data. The data indicator vector is $d = (1, 0, 1, 0, 1)$. Data clearly occurs in positions 1, 3 and 5, so take I_5 and move the corresponding columns to the left to form R:

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Post-multiplication by R achieves the required rearrangement of h and d.

We first use R to create a rearranged version of the correlation matrix C, which may be derived from the price returns matrix H with missing data using an available data approach, or from its nearest correlation matrix counterpart that imposes the positive semi definite property. We denote this rearranged correlation matrix by C_R, defined as:

$$C_R = R^\top C R$$

We also create a rearranged version of the indicator vector d:

$$d_R = dR$$

This allows us to define the partial correlation matrix required to decorrelate the data:

$$C_P = diag(d_R)\, C_R\, diag(d_R) + (I - diag(d_R))$$

By appropriate use of the rearranger matrix R, we can define, what we term, the 'generator matrix':

$$G = R\, chol(C_P)^{-1}\, chol(C_R)\, R^{\top}$$

The final step of the rearrangement process is to take a copy of h in which the gaps have been filled with *uncorrelated* normal random variates, scaled with the relevant volatility (e.g. from the covariance matrix derived), and then post-multiply this by G, to return the version of h with imputed data that leads to the desired correlation consistency with C.

Looping through each row h of H, imputes the missing data. A single iteration of the procedure allows for single imputation, while multiple imputation is possible with multiple iterations.

3.1. Convergence: Oil Futures Data

A pertinent question in respect of the Dynamic Cholesky imputation method is: how does the data imputation process perform with varying degrees of missing data? To test convergence, we use data drawn from Bloomberg for the Brent crude oil price series CO1, CO6, CO12, CO18 and CO24. These contracts allow market participants to lock in a price immediately for future delivery of crude oil in 1, 6, 12, 18 and 24 months respectively. The data spans the period from 8 October 2014 to 7 October 2019.

We run the data imputation process 10,000 times and compute the standard deviations of the resultant correlations (which constitute the standard errors). Tables 1 to 3 show the results for different percentages of data gaps *imposed* on the observed data; namely, 10%, 30% and 50% respectively. The standard deviations can be seen to be negligible, although error magnitude does increase as the percentage of data gaps increases.

3.2. Convergence: Oil Stock Data

The Dynamic Cholesky imputation method is appropriate to many different applications. To illustrate this, we repeat our analysis using stock returns data for five Oil Majors drawn from Yahoo Finance, namely BP, Shell, Chevron, Total, and Exxon (respective tickers: BP, RDSB, CVX, TOT, and XOM). The data spans the period from 14 November 2017 to 11 November 2020.

As before, we run the data imputation process 10,000 times and compute the standard deviations of the resultant correlations (which constitute the standard errors). Tables 4 to 6 show the results for different percentages of data gaps imposed on the observed data; namely, 10%, 30% and 50% respectively. Again, the standard deviations are negligible but their magnitude increases as the percentage of data gaps increases.

4. Conclusion

We develop a novel data imputation method that draws on the simplicity and elegance of Cholesky decomposition, when the objective is to be consistent with the correlation structure of the available data. The method is suitable in a data missing completely at random setting. A useful direction for future research is a comprehensive comparison of the Dynamic Cholesky method against alternative methods, which would be ideally framed around the replication of prior studies. Potential also lies in the application and assessment of the data imputation method for specific risk management problems, such as risk factor modelling.

Table 1. Futures - standard deviation of percentage correlation (10% gaps)

	$CO1$	$CO6$	$CO12$	$CO18$	$CO24$
$CO1$	0.000	0.031	0.040	0.048	0.061
$CO6$	0.031	0.000	0.004	0.009	0.018
$CO12$	0.040	0.004	0.000	0.003	0.009
$CO18$	0.048	0.009	0.003	0.000	0.004
$CO24$	0.061	0.018	0.009	0.004	0.000

Table 2. Futures - standard deviation of percentage correlation (30% gaps)

	$CO1$	$CO6$	$CO12$	$CO18$	$CO24$
$CO1$	0.000	0.033	0.073	0.106	0.169
$CO6$	0.033	0.000	0.010	0.035	0.082
$CO12$	0.073	0.010	0.000	0.011	0.042
$CO18$	0.106	0.035	0.011	0.000	0.015
$CO24$	0.169	0.082	0.042	0.015	0.000

Table 3. Futures - standard deviation of percentage correlation (50% gaps)

	$CO1$	$CO6$	$CO12$	$CO18$	$CO24$
$CO1$	0.000	0.081	0.158	0.212	0.267
$CO6$	0.081	0.000	0.029	0.073	0.115
$CO12$	0.158	0.029	0.000	0.018	0.051
$CO18$	0.212	0.073	0.018	0.000	0.018
$CO24$	0.267	0.115	0.051	0.018	0.000

Table 4. Stocks - standard deviation of percentage correlation (10% gaps)

	BP	$RDSB$	CVX	TOT	XOM
BP	0.000	0.154	0.325	0.222	0.343
$RDSB$	0.154	0.000	0.317	0.369	0.367
CVX	0.325	0.317	0.000	0.552	0.413
TOT	0.222	0.369	0.552	0.000	0.510
XOM	0.343	0.367	0.413	0.510	0.000

Table 5. Stocks - standard deviation of percentage correlation (30% gaps)

	BP	$RDSB$	CVX	TOT	XOM
BP	0.000	0.301	0.462	0.366	0.469
$RDSB$	0.301	0.000	0.634	0.599	0.672
CVX	0.462	0.634	0.000	0.722	0.518
TOT	0.366	0.599	0.722	0.000	0.730
XOM	0.469	0.672	0.518	0.730	0.000

Table 6. Stocks - standard deviation of percentage correlation (50% gaps)

	BP	$RDSB$	CVX	TOT	XOM
BP	0.000	0.324	0.636	0.566	0.521
$RDSB$	0.324	0.000	0.571	0.679	0.586
CVX	0.636	0.571	0.000	0.714	0.567
TOT	0.566	0.679	0.714	0.000	0.669
XOM	0.521	0.586	0.567	0.669	0.000

References

[1] Abd Majid, M.S. and Kassim, S.H. (2009). Impact of the 2007 US financial crisis on the emerging equity markets. *International Journal of Emerging Markets*, **4(4)**, 341–357.

[2] Aerts, M., Claeskens, G., Hens, N. and Molenberghs, G. (2002). Local multiple imputation. *Biometrika* **89(2)**, 375–388.

[3] Allison, P.D. (2001). Missing data. Sage Publications.

[4] Balakrishnan, S., Wainwright, M.J. and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics* **45(1)**, 77–120.

[5] Barnard, J. and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86(4)**, 948–955.

[6] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39(1)**, 1–22.

[7] Higham, N.J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis* **22(3)**, 329–343.

[8] Hüttner, A., Scherer, M. and Gräler, B. (2020). Geostatistical modeling of dependent credit spreads: Estimation of large covariance matrices and imputation of missing data. *Journal of Banking Finance* **118**, 1–13.

[9] Jennrich, R.I. (1962). Missing data correlation computations. *Mathematics of Computation* **16(80)**, 496–497.

[10] Little, R.J. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association* **87(420)**, 1227–1237.

[11] Nagy, K. (2020). Term structure estimation with missing data: Application for emerging markets. *The Quarterly Review of Economics and Finance* **75**, 347–360.

[12] Pigott, T.D. (2001). A review of methods for missing data. *Educational Research and Evaluation* **7(4)**, 353–383.

[13] Reiter, J.P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika* **94(2)**, 502–508.

[14] Robins, J.M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87(1)**, 113–124.

[15] Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63(3)**, 581–592.

[16] Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons.

[17] Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91(434)**, 473–489.

[18] Wang, N. and Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85(4)**, 935–948.

[19] Zhang, X. (2013). Model averaging with covariates that are missing completely at random. *Economics Letters* **121(3)**, 360–363.