An ADMM for ARMA-State Space Model Estimation via Convex Optimization Using a Nuclear Norm Penalization Approach

Basad Al-Sarray[1], Stéphane Chrétien[2]

[1] University of Baghdad , College of Science, Computer Science, Jadirya, 10071, Baghdad, Iraq
[2] National Physical Laboratory, Teddington, Hampton road, TW11 0LW, UK
basad.a.alsarray@gmail.com , stephane.chretien@npl.co.uk

***Abstract*:** Estimation of State-Space models together with joint model selection, is a difficult computational problem. Recent developments in convex penalization to least squares estimation problems provide an elegant solution to this problem that needs efficient optimization to be put to work in potentially large scale settings. In this paper, we study an Alternating Method of Multipliers for a penalized Subspace-type approach to State Space estimation with a nuclear norm penalty. Our model takes into account possible missing data. More-over, we show how creating artificial missing data at random provides a simple approach to hyper-parameter selection. Numerical experiments are proposed to illustrate the performance of our method.

***Keywords*:** ARMA, Low Rank, Nuclear, Norm, Penalization.

1. Introduction

A real valued random discrete dynamical system $(x_t)_{t\in N}$ admits a State Space representation if there exists a discrete time process $s_{t_{t\in N}}$ such that

$$s_{t+1} = As_t + Ke_t$$
$$x_t = Bs_t + Ke_t$$

Where $(e_t)_{t\in N}$ is the noise, and $A \in \mathcal{R}^{p\times p}$, $B \in \mathcal{R}^{1\times p}$, $K \in \mathcal{R}^{p\times 1}$ are parameter matrices.

The Auto-regressive with moving average (ARMA) processes are sequences of the form $(x_t)_{t\in N}$ that satisfy

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j e_{t-j} + e_t \qquad (1)$$

for all $t \geq \max\{p,q\}$, where $(e_t)_{t\in N}$ is a sequence of independent identically distributed random variables. Time series model are relevant for a wide range of applications in economics, engineering, social science, epidemiology, ecology, signal processing,

It is well known that ARMA processes admit a State Space representation and vice versa [7, 4].

Time series analysis is concerned with two estimation problems.

The first is to select the orders p and q of the model.

The second is to estimate $a = (a_1, a_2, \dots, a_p)$ and $b = (b_1, b_2, \dots, b_q)$.

The model order selection problem is often performed using a penalized log-likelihood approach such as AIC, BIC,....

We refer the reader to the standard text of Shumway and Stoffer [7] for more details on this standard problems.

Turning to the estimation of a and b, it is well known that the log-likelihood is unfortunately not a concave function, and that multiple stationary points exist which can lead to severe bias when using local optimization routines for such as gradient or Newton-type methods for the joint estimation of a and b. In [7,3], an iterative procedure resembling the EM algorithm is proposed, which seems more appropriate for the ARMA model than standard optimization algorithms. However, no convergence grantee towards a global maximizer is provided. A recent advance in the field was the subspace method which turned out to be equivalent to minimizing a convex criterion for the estimation of a State Space model under stability conditions.

Since the recent successes of the LASSO in regression and its multiple generalizations [5], penalization has gained a lot of importance in computational statistics.

In particular, the nuclear norm has played an important role for many problems in engineering, machine learning and statistics such as matrix completion, ...

The goal of the present note is to study the nuclear norm penalization in the subspace method framework for convex minimization based ARMA estimation.

2. The Subspace Method

2.1 Prediction

The problem of predicting x_{t+j} for $j \geq 0$ based on the knowledge of $x_{t'}$, $t' < t$ and s_t can be solved easily following the approach by Bauer [2,8,1].

For given initial values x_0, e_0, the State Space representation gives

$$x_{t+h} = e_{t\_h} + \sum_{j=1}^{h} BA^{j-1}Ke_{t+h-j} + BA^h s_t$$

On the other hand, the State Space representation implies that

$$\begin{aligned} s_t &= As_{t-1} + Ke_{t-1} \\ &= As_{t-1} + K(x_{t-1} - Bs_{t-1}) \\ &= (A-KB)s_{t-1} + Kx_{t-1} \\ &= \cdots \end{aligned}$$

Thus, we obtain

$$s_t = (A - KB)^t s_0 + \sum_{j=0}^{t-1}(A - KB)^j K x_{t-1-j}$$

2.2 Prediction with Hankel matrices

We can rewrite the prediction problem in terms of some

Hankel matrices as explained in [6].
Define

$$\bar{A} = A - KB \quad A_0 = [\bar{A}^t s_0, \bar{A}^{t+1} s_0, \dots, \bar{A}^{T-t+1} s_0]$$

$$K = [\bar{A}^{t-1}K, \dots, \bar{A}^2 K, K], \quad O = \begin{bmatrix} B \\ BA \\ \vdots \\ BA^{t-1} \end{bmatrix}$$

And

$$N = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ BK & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ BA^{t-2}K & BA^{t-3}K & \cdots & \cdots & BK & 1 \end{bmatrix}$$

Define also

$$X_{past} = \begin{bmatrix} x_0 & x_1 & \cdots & x_{T-2t+1} \\ x_1 & x_2 & \cdots & x_{T-2t+2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{t-1} & x_t & \cdots & x_{T-t} \end{bmatrix},$$

$$X_{future} = \begin{bmatrix} x_t & x_{t+1} & \cdots & x_{T-t+1} \\ x_{t+1} & x_{t+2} & \cdots & x_{T-t+2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{2t-1} & x_{2t} & \cdots & x_T \end{bmatrix}$$

Both matrices are Hankel matrices. The first one represents the past values and and second one the future values.
Define also the noise matrix

$$E = \begin{bmatrix} e_t & e_{t+1} & \cdots & e_{T-t+1} \\ e_{t+1} & e_{e+2} & \cdots & e_{T-t+2} \\ \vdots & \vdots & \vdots & \vdots \\ e_{2t-1} & e_{2t} & \cdots & e_T \end{bmatrix}$$

Now, as explained in [6], we have the following relationship

$$X_{future} = O\,K\,X_{past} + O\bar{A}_0 + NE \qquad (2)$$

3. The Estimation Problem

Using equation (2) , it is easy to build a least-squares estimator for the matrix L, [9].
In this section, we describe the nuclear norm-penalized estimator proposed in [6].

3.1 Estimating $O\,K$

The matrix $O\,K$ can be estimated using a least squares approach corresponding to solving

$$\min_{L \in R^{t \times t}} \ \frac{1}{2}\|X_{future} - X_{past}\|_F^2 \qquad (3)$$

This procedure will make sense if the term $O\bar{A}_0$ is small.
This can indeed be justified if t is large and if $\|\bar{A}\|$ is small.
Let us call $\hat{L}$ a solution of equation (3).

3.2 Nuclear Norm penalized least squares for low rank estimation

An interesting property of the matrix $O\,K$ is that its rank is the State's dimension p when A if full rank. Moreover, $O\,K$ has small rank compared to t when t is large compared to p.
Therefore, one is tempted to penalize the least squares problem in equation (3) with a low-rank promoting penalty. One option is to try to solve

$$\min_{L \in R^{t \times t}} \frac{1}{2}\|X_{future} - L\,X_{past}\|_F^2 + \lambda\,\mathrm{rank}(L) \qquad (4)$$

The main drawback of this approach is that the rank function is non continuous and non-convex function.
This renders the optimization problem intractable in practice.
Fortunately, the rank function admits a well-known convex surrogate, which is the nuclear norm, i.e. the sum of the singular values, denoted by $\|.\|_*$.
Thus, a nice convex relaxation of (4) is given by

$$\min_{L \in R^{t \times t}} \frac{1}{2}\|X_{future} - LX_{past}\|_F^2 + \lambda\|L\|_* \qquad (5)$$

As is well known, the penalized least-squares problem (5) can be transformed into the following constrained problem

$$\min_{L \in R^{t \times t}} \|L\|_* \ \text{ subject to } \|X_{future} - LX_{past}\|_F \le \eta$$

for some appropriate choice of η.
The finite sample performance of this estimator was studied in [6].

3.3 The case of missing future data

The problem of handling missing data in the matrix X_{future} is easy to state. Let n_{obs} denote the number of observed entries in X_{future}.
Let $\Omega: R^{t \times T-2t+1} \to R^{n_{obs}}$ denote any operator of the user's choice which extracts the observed entries of X_{future} and stacks them into a real vector.
Then, based on the arguments of the previous section, a reasonable estimator can be proposed as the solution of

$$\min_{L \in R^{t \times t}} \frac{1}{2}\|\Omega(X_{future}) - \Omega(LX_{past})\|_F^2 + \lambda\,\|L\|_* \qquad (6)$$

for some appropriate choice of λ.

4. An ADMM for Computing $\hat{L}$

4.1 The standard case

Notice that equation (5) is equivalent to

$$\min \ \frac{1}{2} \ \|X_f - MX_p)\|_F^2 + \lambda\,\|L\|_* \quad \text{subject to } M = L$$

The Augmented Lagrange function is given by

$$L_\rho(M, L, U) = \frac{1}{2}\|X_f - MX_p\|_F^2 + \|L\|_* + \langle U, M - L\rangle + \frac{1}{2}\rho\|M - L\|_f^2$$

Minimize L_ρ for $M^{(l+1)}$ given $L^{(l)}$ and $U^{(l)}$, by finding the gradient of L_ρ with respect to M

$$\nabla_M L_\rho(M, L^{(l)}, U^{(l)}) = (X_f - MX_p)X_p^T + U^{(l)} + \rho(M - L^{(l)})$$

setting the gradient to 0 gives

$$(X_f - M^{(l+1)}X_p)X_p^T + U^{(l)} + \rho(M^{(l+1)} - L^{(l)}) = 0$$

Therefore,

$$X_f X_p^T - M^{(l+1)}X_p X_p^T + U^{(l)} + \rho M^{(l+1)} - \rho L^{(l)} = 0$$

$$X_f X_p^T + U^{(l)} - \rho L^{(l)} = M^{(l+1)}(X_f X_p^T - \rho\, I)$$

and thus

$$M^{(l+1)} = (X_f X_p^T + U^{(l)} - \rho L^{(l)})(X_p X_p^T - \rho I)^{-1}$$

Now, the next step is performed by computing the approximation of L by solving the following problem of minimization

$$\min_{L \in R^{t \times t}} \frac{1}{2}\rho\|L\|_F^2 - \rho\langle M, L\rangle - \langle U, L\rangle + \lambda\|L\|_*$$

$$= \min \ \frac{1}{2}\rho\|L\|_F^2 - \langle \rho M + U, L\rangle + \lambda\|L\|_*$$

$$= \min \ \frac{1}{2}\rho(\|L\|_F^2 - 2\langle M + \frac{1}{\rho}U, L\rangle) + \lambda\|L\|_*$$

$$= \min \frac{1}{2}\rho\left(\left\|L-(M+\frac{1}{\rho}U)\right\|_F^2\right)+\lambda\|L\|_*$$

Setting $Z = M+\frac{1}{\rho}U$, we obtain the optimization problem

$$= \min \tfrac{1}{2}\|L-Z\|_F^2+\tfrac{1}{\rho}\lambda\|L\|_*$$

Thus, the solution is just defined by the thresholding operator as

$$L^{(l+1)} = \text{Thresh}\left(M^{(l+1)}+\tfrac{1}{\rho}U^{(l)},\tfrac{\lambda}{\rho}\right)$$

The last step consists in updating U, which is simply done by setting

$$U^{(l+1)} = U^{(l)}+\rho(M^{(l+1)}-L^{(l+1)}).$$

4.2 The case of missing data

Notice that equation (6) is equivalent to

$$\min\frac{1}{2}\left\|\Omega X_f-\Omega(MX_p)\right\|_2^2+\lambda\|L\|_*$$

subject to M=L

The Augmented Lagrange function is given by

$$L_\rho(M,L,U) = \frac{1}{2}\left\|\Omega X_f-\Omega(MX_p)\right\|_2^2+\lambda\|L\|_*$$
$$+\langle U,M-L\rangle+\frac{1}{2}\rho\|M-L\|_F^2$$

Minimize L_ρ for $M^{(l+1)}$ given $L^{(l)}$ and $U^{(l)}$, by finding the gradient of L_ρ for M

$$\nabla_M L_\rho(M,L,U) = \Omega^*\left(\Omega(X_f)-\Omega(MX_p)\right)X_p^T+U+\rho(M-L)$$

setting the gradient to 0 gives

$$\Omega^*\left(\Omega(X_f)-\Omega(M^{(l+1)})X_p\right)X_p^T+U^{(l)}+\rho\left(M^{(l+1)}-L^{(L)}\right)=0$$

Therefore, we obtain

$$\Omega^*\circ\Omega(X_f)X_p^T-\Omega^*\circ\Omega(M^{(l+1)}X_p)X_p^T+U^{(l)}+\rho M^{(l+1)}-\rho L^{(L)}=0$$

Which gives

$$\Omega^*\circ\Omega(X_f)\,X_p^T+U^{(l)}-\rho L^{(L)} = \Omega^*\circ\Omega(M^{(l+1)}X_p)X_p^T+\rho M^{(l+1)}$$

This last equation may now be solved using the conjugate gradient method.

Now, the next step is performed exactly as in the previous case by computing the approximation of L by solving the following problem of minimization

$$\min_{L\in R^{txt}}\frac{1}{2}\rho\|L\|_F^2-\rho\langle M^{(l+1)},L\rangle-\langle U^{(l)},L\rangle+\lambda\|L\|_*$$

whose solution is just defined by the thresholding operator as

$$L^{(L+1)} = \text{Thresh}\left(M^{(l+1)}+\tfrac{1}{\rho}U^{(l)},\tfrac{\lambda}{\rho}\right)$$

The last step consists in updating U, which is simply done by setting

$$U^{(l+1)} = U^{(l)}+\rho(M^{(l+1)}-L^{(l+1)})$$

5. Numerical Experiments

In this study we will perform some simulations with the model

$$x_t = 1.4x_{t-1}-.66x_{t-2}+.16x_{t-3}-.023x_{t-4}-.012x_{t-5}+$$
$$e_t+1.7e_{t-1}-4e_{t-2}+2.4e_{t-3}-.86e_{t-4}$$

with e_t, t=1,…,T independent zero mean Gaussian random variables with unit variance.

Figure.1 shows a realization of the signal considered in this section.

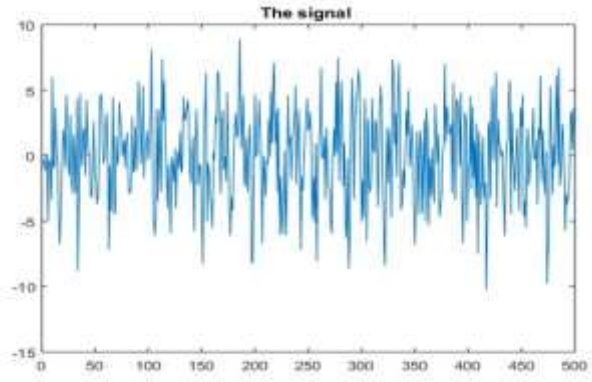

Figure 1. One realization of the signal

Figure.2 illustrates the convergence of the ADMM method.

In all experiments, the stopping criterion was when the relative error in the U variable went below 10^{-4}.

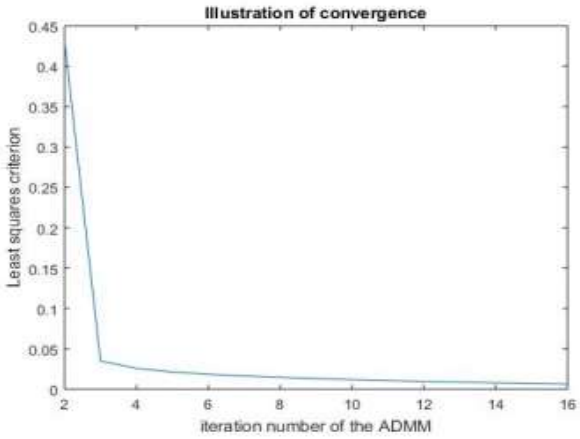

Figure 2. Decrease of the Least squares criterion as a function of the iteration number for 5 missing data and λ =20.

5.1 Choosing the relaxation parameter λ

A very simple way to choose the hyperparameter λ is to create artificially missing data in the set of future observations and tune the value of λ so as to minimize the sum of squares of the errors of the estimator on these observations. Figure.3 shows the error for different values of λ.

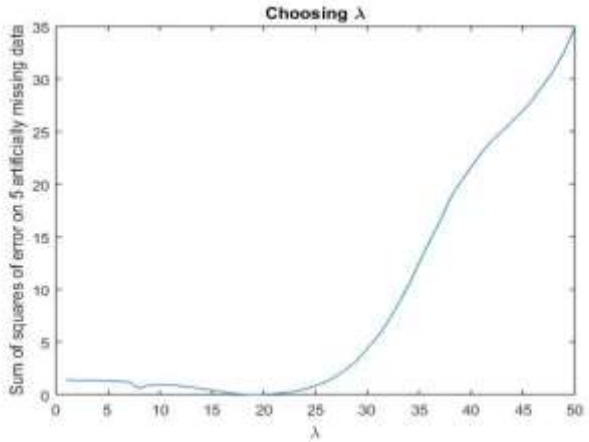

Figure 1. Error on the artificially missing data for selecting the best value for λ. Here, the best value is $\hat{\lambda} = 20$

Conclusion

The goal of the present paper was to present a nuclear norm penalised least-squares estimation procedure for ARMA model selection and estimation where the time series is corrupted by some noise and may have missing data. We proposed an ADMM type algorithm for this problem and studied the performances of the method on simulated data.

References

[1] D. Bauer, Asymptotic properties of subspace estimators, Automatica, Vol 40, No. 3, pp.359–376, 2005.

[2] D. Bauer. Estimating linear dynamical systems using subspace methods, Econometric Theory, Vol 21, No. 01, pp.181–211, 2005.

[3] G.Box, G. Jenkins, G.C Reinsel, and G. M Ljung, Time series analysis: forecasting and control, John Wiley & Sons, 2015.

[4] P.J.Brockwell , R. A. Davis, Time series: theory and methods, Springer Science & Business Media, 2013.

[5] S. Chretien and S. Darses, Sparse recovery with unknown variance: a lasso-type approach, Information Theory, IEEE Transactions on Vol 60 , No. 7, pp. 3970–3988, 2015.

[6] S. Chrétien, T. Wei, and Basad. Al-sarray, Joint estimation and model order selection for one dimensional arma models via convex optimization: a nuclear norm penalization approach, arXiv preprint arXiv:1508.01681 (2015).

[7] Shumway, Robert H., and David S. Stoffer. "Time series regression and exploratory data analysis." Time series analysis and its applications. Springer New York, 2014.

[8] P. Overschee and B. De Moor, N4sid: Subspace algorithms for the identification of combined deterministicstochastic systems, Automatica, Vol 30, No. 1, pp.75–93, 1994.

[9] M.Verhaegen and V. Verdult, Filtering and system Identification: a least squares approach, Cambridge University press, 2007.