EFFICIENT FILTERING AND SAMPLING FOR A CLASS OF TIME-VARYING LINEAR SYSTEMS

James Murphy, Simon Godsill

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge, United Kingdom CB2 1PZ
Emails:{jm362, sjg30}@cam.ac.uk

ABSTRACT

This paper presents an $O(n^4)$ time method for filtering and sampling of a time-varying $n \times n$ system matrix A_t in a restricted class of time-varying linear systems of the form $X_t = A_t X_{t-1} + C_t + \varepsilon_t$, via a matrix-variate normal formulation. This allows larger systems within this class to be inferred via Gibbs sampling in reasonable time than is possible with methods that rely on vectorization of the system matrix, followed by standard Kalman filtering, which run in $O(n^6)$ time. It is shown how to apply the method to vector autoregression problems with time-varying system matrices (TVP-VAR problems). Noisy observations of the underlying system state are also accommodated in a straightforward way.

Index Terms— Matrix-variate normal, TVP-VAR, time-varying, vector autoregression, linear systems

1. INTRODUCTION

This paper presents efficient algorithms for filtering and conditionally sampling a time-varying system matrix in a linear system of the form given in equations (1) to (3). This is a restricted, but still potentially useful class of such models, for which filtering and sampling is possible in $O(n^4)$ time, as opposed to the $O(n^6)$ time required by the (more general) algorithm used in e.g. [1] and [2], which is based on a standard Kalman filter [3] via vectorization of the system matrix (see section 1.2 and e.g. [2]). Such conditional sampling is necessary as part of a Gibbs sampler to jointly infer hidden states, system matrix elements and parameters in time-varying systems, and thus faster sampling, albeit in a restricted case, is of interest when estimation of such systems is required. Time-varying linear systems have also been used for robust Kalman filtering [4]. The filtering portion of the method given here is also useful for Rao-Blackwellization (see e.g. [5] for an overview) in mixed linear/nonlinear systems with conditionally linear portions of the form in equations (1) to (3). In this case, efficient filtering can be performed with a combination of particle filtering (for nonlinear portions of the state) and the filtering method given here (for conditionally linear, matrix-variate normal portions).

Vector autoregression (VAR) models with time-varying system matrices (sometimes known as time-varying parameter vector autoregression, or TVP-VAR) have been applied in econometrics, particularly examining macroeconomic factors such as monetary policy [2], the business cycle [6] and inflation dynamics [1]. The form of these is slightly different to the system in equations (1) to (3), but can readily be put into a suitable form, as shown in section 4.

1.1. Problem Formulation

The time-varying parameter linear system considered here for an n dimensional X_t (i.e. $X_t \in \mathbb{R}^n$) is given by

$$X_t = A_t X_{t-1} + C_t + \varepsilon_t \qquad (t \geq 2) \tag{1}$$
$$A_t = A_{t-1} + \eta_t \qquad (t \geq 3) \tag{2}$$
$$y_t = H_t X_t + \nu_t \qquad (t \geq 1) \tag{3}$$

with $\varepsilon_t \sim \mathcal{N}(0, \gamma_t Q)$ with γ_t scalar, $\eta_t \sim \mathcal{MN}(0, \lambda_t Q, V_t)$ with λ_t scalar, and $\nu_t \sim \mathcal{N}(0, R_t)$. The $A_t \in \mathbb{R}^{n \times n}$ matrix is a time varying $n \times n$ system matrix, perturbed at each step with matrix-variate normal distributed noise η_t. The probability density function of the matrix-variate normal distribution for an $n \times p$ variate matrix A is given by

$$\mathcal{MN}(A; M, U, V) = (2\pi)^{-\frac{np}{2}} |V|^{-\frac{n}{2}} |U|^{-\frac{p}{2}} \times \exp\left\{ -\frac{1}{2} \mathrm{tr}\left[V^{-1}(A-M)' U^{-1}(A-M) \right] \right\}$$

with M an $n \times p$ matrix of element means, U an $n \times n$ matrix of among-row covariance and V a $p \times p$ matrix of among-column covariance. It is related to the standard vector-valued normal distribution by

$$\mathcal{MN}(A; M, U, V) = \mathcal{N}(\mathrm{vec}(A); \mathrm{vec}(M), V \otimes U), \tag{4}$$

and is used for the estimation of linear systems and their parameters in [7].

In equation (2), the among-row covariance matrix of η_t is restricted to be a scaled version of the state transition noise

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) grant number EP/K020153/1

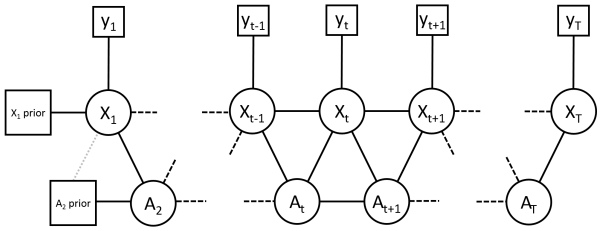

Fig. 1. Dependence structure between A, X and y variables

covariance Q. This is a limitation of this method and defines the class of problems for which the method is applicable.

The problem addressed in this paper is the joint inference of $X_{1:T}$ and $A_{2:T}$ for all t via Gibbs sampling. As part of this, a conditional filter is developed for $p(A_{2:t} \mid X_{1:t}, \theta_{1:t})$, where $\theta_{1:t}$ is the collection of parameters $C_{1:t}, Q, V_{2:t}$, $\lambda_{2:t}$ and $\gamma_{1:t}$.

1.2. Gibbs Sampler for A_t and X_t

In order to sample both $X_{1:T}$ and $A_{2:T}$ from the posterior distribution $p(X_{1:T}, A_{1:T} \mid y_{1:T}, \theta_{1:T})$, a Gibbs sampler [8, 9] can be used, sampling $X_{1:T}$ and $A_{2:T}$ in turn from the conditional distributions $p(X_{1:T} \mid A_{2:T}, y_{1:T}, \theta_{1:T})$ and $p(A_{2:T} \mid X_{1:T}, \theta_{1:T})$ (since, as shown in figure 1, $A_{2:T}$ are conditionally independent of $y_{1:T}$ given $X_{1:T}$).

Due to the linear nature of the system, *both* of these conditional distributions can be sampled via forward Kalman filtering, followed by a backward sampling pass of the form described in e.g. [10]. For the $X_{1:T}$ variables, the form is standard and sampling takes $O(n^3)$ time. In the case of the $A_{2:T}$ variables, a linear system of standard form with respect to the elements of A_t is given by

$$\begin{aligned} \text{vec}(A_{t+1}) &= \text{vec}(A_t) + \bar{\eta}_t \\ X_t &= H_t \text{vec}(A_t) + \varepsilon_t \end{aligned}$$

with $H_t = X_t \otimes I_n = [X_{1,t} I_n,\ X_{2,t} I_n, ..., X_{n,t} I_n]$ (where I_n is the $n \times n$ identity matrix and $X_{i,t}$ is the i^{th} element of the X_t vector), $\bar{\eta}_t \sim \mathcal{N}(0, U_t)$ and $\varepsilon_t \sim \mathcal{N}(0, S_t)$, as given in [2]. This is a linear system with the n^2 elements of A as its state variables, and thus takes $O(n^6)$ time to filter and sample (assuming naive algorithms for matrix operations, so that e.g. matrix multiplication has $O(n^3)$ complexity for $n \times n$ matrices; faster algorithms exist e.g. [11], although often with prohibitive non-dominating costs or constant factors). In the case when $U_t = V_t \otimes \lambda_t Q$ and $S_t = \gamma_t Q$, this system is equivalent to the one given in equations (1) and (2) and in that special case it is possible to filter and sample this system in $O(n^4)$ time, as is shown below.

The parameters composing $\theta_{1:T}$ can also be sampled via Gibbs sampling (see e.g. [2]), although this is not examined in detail here.

2. FILTER FOR TIME-VARYING SYSTEM MATRIX

The matrix-variate normal filter for the system matrices $A_{2:T}$ in equation (1) (where T is the total number of steps; A_1 does not exist) can most easily be derived in two parts, a *predict* and *correct* step, similar to those for a standard Kalman filter.

Assuming that the filtering distribution of A_{t-1} is given by $p(A_{t-1} \mid X_{1:t-1}) = \mathcal{MN}(A_{t-1}; M_{t-1}, Q, V_{t-1})$, then the *predictive* distribution of A_t is given by

$$\begin{aligned} p(A_t \mid X_{1:t-1}) &= \int p(A_t \mid A_{t-1}) p(A_{t-1} \mid X_{1:t-1}) dA_{t-1} \\ &= \int Z_* \mathcal{MN}(A_{t-1}; M_*, Q, V_*)\, dA_{t-1} \\ &= Z_* \end{aligned}$$

using identity 7.1 (equation (7)) and where V_*, M_* and the normalizing constant Z_* are given there. The expression for Z_* can be rearranged so that it is a matrix-variate normal with respect to A_t (proportionality being sufficient, since $p(A_t \mid X_{1:t-1})$ is a probability distribution and thus integrates to 1). With some algebra, this gives

$$p(A_t \mid X_{1:t-1}) = \mathcal{MN}\left(A_t; M_{t|t-1}, Q, V_{t|t-1}\right)$$

with

$$\begin{aligned} V_{t|t-1} &= \lambda_t V_t + V_{t-1|t-1} \\ M_{t|t-1} &= M_{t-1} V_{t-1|t-1}^{-1} (V_t^{-1} + \lambda V_{t-1|t-1}^{-1})^{-1} V_t^{-1} V_{t|t-1}. \end{aligned}$$

The *correction* step uses this predictive distribution to obtain $p(A_t \mid X_{1:t})$, the filtering distribution of A_t, as

$$\begin{aligned} p(A_t \mid X_{1:t}) &\propto p(X_t \mid A_t, X_{t-1}) p(A_t \mid X_{1:t-1}) \\ &= \mathcal{N}(X_t; A_t X_{t-1} + C_t, \gamma_t Q)\, \mathcal{MN}\left(A_t; M_{t|t-1}, Q, V_{t|t-1}\right) \\ &= \mathcal{MN}\left(A_t; M_{t|t}, Q, V_{t|t}\right) \end{aligned}$$

with

$$V_{t|t} = (V_{t|t-1}^{-1} + \gamma_t^{-1} X_{t-1} X_{t-1}')^{-1} \quad (5)$$

$$M_{t|t} = (M_{t|t-1} V_{t|t-1}^{-1} + \gamma_t^{-1} (X_t - C_t) X_{t-1}') V_{t|t}, \quad (6)$$

using identity 7.2 (equation (8)).

2.1. Prior

In order to initialize the filter for $A_{2:T}$, a prior for the distribution of A_2 is required. To ensure conjugacy, this must take the form of a matrix-variate normal distribution, so that $p(A_2 \mid X_1) = \mathcal{MN}\left(A_2; M_{2|1}, \alpha Q, V_{2|1}\right) = \mathcal{MN}\left(A_2; M_{2|1}, Q, \alpha V_{2|1}\right)$, where $M_{2|1}$, $V_{2|1}$ and α are the parameters of the prior (which can depend on X_1). The correction step defined above can then be used to obtain the distribution $p(A_2 \mid X_{1:2})$ and the prediction/correction steps defined above used to find the filtering distributions of $A_{3:T}$.

3. BACKWARD SAMPLING

In order to perform Gibbs sampling, samples must be drawn from $p(A_{2:T} \mid X_{1:T}, \theta_{1:T})$. This requires drawing samples of $A_{2:T}$ from its *smoothing* distribution, conditioned on all $X_{1:T}$. This can be achieved by a running the forward filter as given in section 2, followed by a backward sampling step similar to the standard backward sampling for linear systems given in [10]. The smoothing distribution $p(A_{2:T} \mid X_{1:T})$ can be decomposed as

$$p(A_{2:T} \mid X_{1:T}) = p(A_T \mid X_{1:T}) \prod_{t=2}^{T-1} p(A_t \mid A_{t+1}, X_{1:t}),$$

due to the dependence structure between the A and X variables shown in figure 1. The distribution $p(A_T \mid X_{1:T})$ is the final filtering distribution found using the filter in section 2. This can be sampled using the procedure in section 3.1.

Given a sample of A_{t+1}, A_t can be sampled from the conditional distribution $p(A_t \mid A_{t+1}, X_{1:t})$, allowing all $A_{2:T}$ to be sampled in a backward pass from A_{T-1} to A_2. The distribution to be sampled at each step is given by

$$\begin{aligned} p(A_t \mid A_{t+1}, X_{1:t}) &\propto p(A_{t+1} \mid A_t) p(A_t \mid X_{1:t}) \\ &= \mathcal{MN}(A_{t+1}; A_t, \lambda_t Q, V_t)\, \mathcal{MN}(A_t; M_t, Q, V_t) \\ &\propto \mathcal{MN}\left(A_t; M_{t|T}, Q, V_{t|T}\right) \end{aligned}$$

with, by identity 7.1,

$$\begin{aligned} V_{t|T} &= \left({\lambda_t}^{-1} V_t^{-1} + V_{t|t}^{-1}\right)^{-1} \\ M_{t|T} &= \left(M_{t|t} V_t^{-1} + {\lambda_t}^{-1} A_{t+1}^* V_t^{-1}\right) V_{t|T} \end{aligned}$$

where A_{t+1}^* is the sampled value of A_{t+1}.

3.1. Sampling Matrix-Variate Normal Distributions

Matrix-variate normal distributions can be sampled by making use of the identity in equation (4), which reduces their sampling to the standard problem of sampling from the vector-valued normal distribution. This can be achieved by calculating the Cholesky decomposition of the covariance matrix and pre-multiplying a vector of independent standard Gaussian variables (i.e. distributed $\mathcal{N}(0,1)$) by this.

In order to draw samples of an $n \times p$ matrix-variate normal distribution in $O(n^2p^2)$ time, however, the Cholesky factorization of the $np \times np$ covariance matrix $V \otimes Q$ should be calculated by first calculating the Cholesky decomposition and taking the Kronecker product of these, using the fact that

$$\text{chol}(V \otimes Q) = \text{chol}(V) \otimes \text{chol}(Q).$$

This is readily seen by letting $L_Q = \text{chol}(Q)$ and similarly L_V, so that $V \otimes Q = (L_V L_V') \otimes (L_Q L_Q') = (L_V \otimes L_Q)(L_V \otimes L_Q)'$. Calculation of the individual Cholesky decompositions takes $O(n^3)$ and $O(p^3)$ time for Q and V, respectively, and calculation of the Kronecker product takes $O(n^2p^2)$ time, which can dominates as p and n increase.

4. TIME VARYING PARAMETER VECTOR AUTOREGRESSION (TVP-VAR)

An important application for the estimation of time-varying system matrices is in the inference of time-varying parameter vector autoregression systems [2], of the form

$$X_t = C_t + B_{1,t} X_{t-1} + B_{2,t} X_{t-2} + ... + B_{k,t} X_{t-k} + \varepsilon_t,$$

with k being the maximum lag in the autoregression and, as before, $\varepsilon_t \sim \mathcal{N}(0, Q)$. The time-varying $B_{1:k,t}$ matrices can be estimated by writing the system as

$$X_t = C_t + B_t \bar{X}_{t-1} + \varepsilon_t,$$

where $\bar{X}_{t-1} = [X_{t-1}', ..., X_{t-k}]'$ and $B_t = [B_{1,t}, ..., B_{k,t}]$. If the dynamics of B are assumed to be of the form

$$B_t = B_{t-1} + \eta_t$$

with $\eta_t \sim \mathcal{MN}(0, Q, V)$, then, given a filtering distribution of B_{t-1} of the form $p(B_{t-1} \mid X_{1:t-1}) \sim \mathcal{MN}\left(M_{t-1}, Q, V_{t-1|t-1}\right)$, the predictive distribution of B_t can be found using the prediction step given in section 2, which applies directly as shown. The correction step is also identical aside from the replacement of X_{t-1} with $\bar{X}_{t-1}$ in equations (5) and (6).

To model the elements of B at each lag as evolving independently, i.e.

$$B_{i,t} = B_{i,t-1} + \eta_{i,t}$$

with $\eta_{i,t}$ independent of $\eta_{j,t}$ for $j \neq i$, the V_t matrix should be chosen to be block diagonal with k blocks on the main diagonal $V_{1,t}, ..., V_{k,t}$, each of which is an $n \times n$ covariance matrix. In this case $\eta_{i,t} \sim \mathcal{MN}(0, Q, V_i) = \mathcal{MN}\left(0, \lambda_{i,t} Q, \lambda_{i,t}^{-1} V_i\right)$.

5. RESULTS

To demonstrate the operation of the matrix-variate normal filter and backward sampler, its output can be compared to that of the standard Kalman filter approach as given in section 1.2. Figure 2 shows a comparison of the ± 2 standard deviation range of 2000 samples drawn using the two methods. There are slight discrepancies between these due to the randomly drawn samples, but the results are consistent with the fact that the two methods sample the same distribution. This comparison used $Q = 0.1 I_n$, $\gamma_t = 1$, $V_t = 0.01 I_n$ and $\lambda_t = 10$ for all t as both filter parameters and for data generation (during data generation, the A_t matrix was only allowed to take stable values by ensuring the magnitude of all eigenvalues was less than 1). It can be seen that the true value of A_t almost always falls within the 2 standard deviation range.

In order to compare the running times of the two methods with increasing system dimension n, both methods were timed running on (stable) systems of increasing dimension.

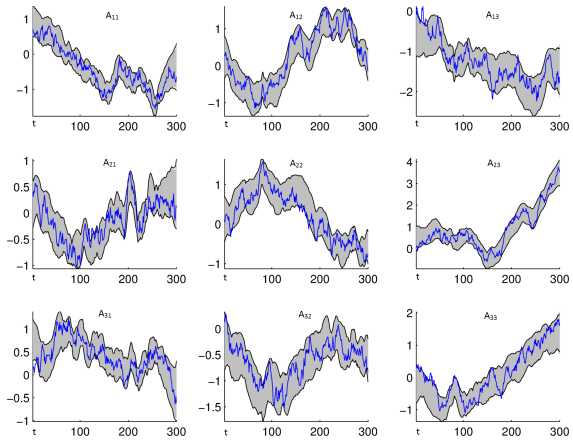

Fig. 2. Comparison of ± 2 standard deviation range of 2000 samples of $A_{2:300}$ drawn using the Kalman filter method (grey shading) and matrix-variate normal method (black lines) for each element of a 3×3 A matrix. Blue lines show true value of element of A_t at each time

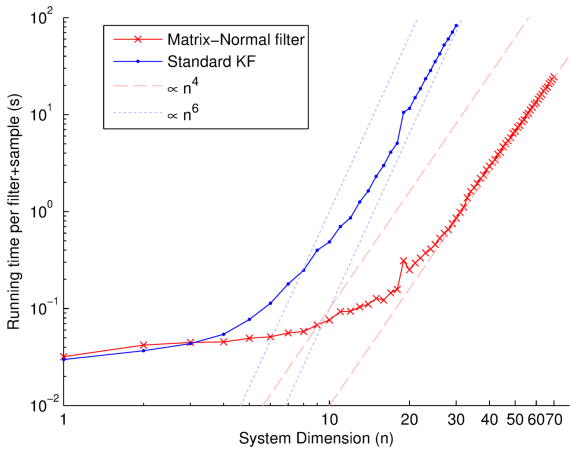

Fig. 3. Running times against system dimension for a single forward filter and backward sampling step using a standard Kalman filter approach and the matrix-variate normal approach proposed here. Faint dashed lines are proportional to n^4 (long dashes) and n^6 (short dashes)

Timing results are shown in figure 3, averaging over 10 samples at each dimension. It can be seen that the matrix-variate normal method runs faster than the Kalman filter method for all $n > 3$ and, as expected, its runtime appears to scale with n^4 for $n > 30$. Due to its slow running time the Kalman filter method was only tested up to $n = 25$ at this dimension it exhibits slightly lower than n^6 scaling of its runtime, probably due to optimizations possible below this scale such as exploiting vector processing capabilities on the CPU (an Intel i7-3770); a similar effect is observed in the matrix-variate normal case at $n < 30$.

Matlab code for the results in this section, including matrix-variate normal filter and backward sampling algorithms can be found at **www-sigproc.eng.cam.ac.uk/Main/JM362**.

6. CONCLUSION

This paper has presented an $O(n^4)$ time method for filtering and sampling for a time-varying system matrix in a restricted class of time-varying linear systems. This allows larger systems within this class to be inferred in reasonable time than is possible with a standard $O(n^6)$ Kalman filter-based method.

The derivation of the filter is similar to that of the standard Kalman filter (in a Bayesian formulation) and thus the methods given here can be extended further in several ways not shown. It is straightforward to deal with systems in which the dynamics of A take the form $A_{t+1} = F_t A_t + \eta_t$. A simple expression for the state likelihood $p(X_{1:T} \mid \theta)$ can be obtained in a way analogous to the prediction error decomposition of the Kalman filter (e.g. [12]). It is also possible to derive a smoother analogous to the RTS smoother [13], to give expressions for the smoothing *distributions* $p(A_t \mid X_{1:T}, \theta)$.

7. APPENDIX: MATRIX-VARIATE NORMAL IDENTITIES

7.1. Product of Two Matrix-Variate Normal PDFs

For X an $n \times p$ matrix,

$$\mathcal{MN}(X; M_1, Q, V_1)\,\mathcal{MN}(X; M_2, \lambda Q, V_2) = Z_* \mathcal{MN}(X; M_*, Q, V_*) \quad (7)$$

with

$$\begin{aligned} V_* &= \left(V_1^{-1} + \lambda^{-1} V_2^{-1}\right)^{-1} \\ M_* &= \left(M_1 V_1^{-1} + \lambda^{-1} M_2 V_2^{-1}\right) V_* \\ Z_* &= (2\pi)^{np/2} \left(\frac{|V_1||V_2|}{|V_*|}\right)^{n/2} |Q|^{p/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(T)\right\} \\ T &= V_1^{-1} M_1' Q^{-1} M_1 + \lambda^{-1} V_2^{-1} M_2' Q^{-1} M_2 \\ &\quad - V_*^{-1} M_*' Q^{-1} M_* \end{aligned}$$

7.2. Product of Normal and Matrix-Variate Normal PDFs

For X an $n \times 1$ vector, Y a $p \times 1$ vector and A and $n \times p$ matrix

$$\mathcal{N}(X; AY, \gamma Q)\,\mathcal{MN}(A; M, Q, V) \propto \mathcal{MN}(A; M_*, Q, V_*) \quad (8)$$

with

$$\begin{aligned} V_* &= \left(V^{-1} + \gamma^{-1} YY'\right)^{-1} \\ M_* &= \left(MV^{-1} + \gamma^{-1} XY'\right) V_* \end{aligned}$$

This relies on the fact that $xAy = \mathrm{tr}(Ayx')$ for appropriately sized vectors x and y and matrix A.

8. REFERENCES

[1] T. Cogley and T.J. Sargent, "Evolving post-World War II US inflation dynamics," in *NBER Macroeconomics Annual 2001, Volume 16*, pp. 331–388. MIT Press, 2002.

[2] G.E. Primiceri, "Time varying structural vector autoregressions and monetary policy," *The Review of Economic Studies*, vol. 72, no. 3, pp. 821–852, 2005.

[3] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Fluids Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[4] F. Wang and V. Balakrishnan, "Robust Kalman filters for linear time-varying systems with stochastic parametric uncertainties," *Signal Processing, IEEE Transactions on*, vol. 50, no. 4, pp. 803–813, 2002.

[5] O. Cappé, S.J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, 2007.

[6] J.H. Stock and M.W. Watson, "Has the business cycle changed and why?," in *NBER Macroeconomics Annual 2002, Volume 17*, pp. 159–230. MIT press, 2003.

[7] A. Wills, T.B. Schön, F. Lindsten, and B. Ninness, "Estimation of linear systems using a Gibbs sampler," in *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*, 2012.

[8] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 6, pp. 721–741, 1984.

[9] A.E. Gelfand and A.F.M. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 398–409, 1990.

[10] S. Frühwirth-Schnatter, "Data augmentation and dynamic linear models," *Journal of Time Series Analysis*, vol. 15, no. 2, pp. 183–202, 1994.

[11] V.V. Williams, "Multiplying matrices faster than Coppersmith-Winograd," in *Proceedings of the 44th annual ACM Symposium on Theory of Computing*. ACM, 2012, pp. 887–898.

[12] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration: A Statistical Model-Based Approach*, Berlin: Springer, ISBN 3 540 76222 1, Sept. 1998.

[13] H.E. Rauch, C.T. Striebel, and F. Tung, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, no. 8, pp. 1445–1450, 1965.