# A NOTE ON THE EFFICIENT SIMULATION IN STATE SPACE MODELS

Miroslav Plašil*[1]

Czech National Bank, University of Economics, Prague

November 24, 2016

* E-mail: miroslav.plasil@cnb.cz

[1] I would like to thank Michal Andrle for numerous discussions and helpful comments.

# 1 Introduction

State space models have become one of the most popular tools in modern Bayesian econometrics. The state-space representation can encompass a wide class of diverse models, including structural time series models, time-varying parameter models, dynamic factor models and many others. Bayesian estimation of state space models builds on the Markov chain Monte Carlo (MCMC) methods and feature many well-established algorithms for the simulation of latent-states (see e.g. Carter and Kohn, 1994; de Jong and Shepard, 1995; Durbin and Koopman, 2002). These algorithms, however, rely on multiple loops through time, which may complicate their effective code implementation and, perhaps even more importantly, limit their conceptual transparency. Motivated by these considerations, Chan and Jeliazkov (2009) build on the alternative derivation of the joint density of the states and discuss their effective estimation.

This short note provides a parallel derivation of the sampler proposed by Chan and Jeliazkov (2009). It draws on the fact that Kalman recursions are just a computationally effective solution of the least squares problem. The equivalence between least-squares estimation theory and the probabilistically based mean-square estimation was well understood in the old days, however it is somewhat neglected by a wider economic audience nowadays. That is to our detriment since (as Gauss puts it himself) the least-squares method gives rise to several elegant analytical investigations (cf. Sorenson, 1970).

The purpose of the note is purely methodological. The alternative derivation outlined below is provided with a hope that it can be found more accessible and easier to follow by some readers. On the route to its ultimate goal, the note may also offer some additional insights on the relation to other existing methods while ticking off the possibilities for further practical enhancements.

# 2 Chan and Jeliazkov: alternative derivation

## 2.1 A primer on the least-square estimation

For the sake of completeness and to set up the notation it is useful to recapitulate some basic (and very well-known) facts from the least-squares theory. Suppose that the data and parameters are related according to the model

$$y = Xb + e$$

where $e$ are measurement errors, $y$ and $X$ are formed by the observed data and $b$ is a set of parameters to be estimated. The idea is to make the errors 'small' in some sense, i.e. to make them close to zero as possible:

$$e = y - Xb \approx 0$$

The method of least squares meets this objective by minimizing the sum of the squared residuals, which is equivalent to minimizing the square of the norm:

$$\min f = \|Xb - y\|^2 = \|e\|^2 = e_1^2 + \cdots + e_T^2 \tag{1}$$

The minimization problem (1) is known to have a closed-form solution:

$$\hat{b} = (X'X)^{-1}X'y \tag{2}$$

Moreover, if the measurement errors are assumed to be Gaussian, the covariance matrix of the estimator is equal to

$$Var(\hat{b}) = \sigma^2(X'X)^{-1} \tag{3}$$

where $\sigma^2$ the variance of the error term.

In some applications, the researchers may wish to work with several objectives, $f_1, f_2, \ldots, f_k$, all of which should be minimized. In this case, a standard solution for finding the values of the unknown parameter vector, $b$, is to use a weighted sum objective:

$$f = \lambda_1 f_1 + \cdots + \lambda_k f_k = \lambda_1 \|X_1 b - y_1\|^2 + \cdots + \lambda_k \|X_k b - y_k\|^2 \tag{4}$$

where $\lambda_1, \ldots, \lambda_k$ are positive constants representing the weights attached to individual objectives. The higher is the value of $\lambda_i$, the stronger is our desire for $f_i$ to be small. Since scaling all the weights by any positive number does not change the minimum of (4), it is possible to set $\lambda_1 = 1$ with no loss of generality.

Ridge regression is a prototypical example of the bi-objective least-squares problem with $X_2 = I$ and $y_2 = \mathbf{0}$. It seeks to minimize the norm $\|b\|^2$ alongside to the traditional minimization of the sum of the squared residuals. A well-known Hodrick-Prescott filter can also be interpreted as an application of a ridge regression with $X_1 = I$, $X_2$ equal to a second-differencing matrix and $y_2 = \mathbf{0}$.[2]

[2] This interpretation of the Hodrick-Prescott filter immediately allows for the construction of the confidence intervals, see Giles (2013).

The weighted-sum objective (4) can be minimized using the data augmentation approach also known as stacking. In particular, one can solve (4) by stacking the objectives one below another and expressing $f$ as a norm of a single vector:

$$f = \left\| \begin{bmatrix} \sqrt{\lambda_1}(X_1 b - y_1) \\ \vdots \\ \sqrt{\lambda_k}(X_k b - y_k) \end{bmatrix} \right\|^2.$$

In other words, forming stacked matrices

$$\tilde{X} = \begin{bmatrix} \sqrt{\lambda_1} X_1 \\ \vdots \\ \sqrt{\lambda_k} X_k \end{bmatrix}, \quad \tilde{y} = \begin{bmatrix} \sqrt{\lambda_1} y_1 \\ \vdots \\ \sqrt{\lambda_k} y_k \end{bmatrix} \tag{5}$$

one can reduce (4) to a standard least-squares problem where $b$ can be found as

$$\tilde{b} = \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\tilde{y} \tag{6}$$

## 2.2 The sampler

Throughout, I consider a simple state space model where, for $t = 1, \ldots, T$, $n \times 1$ vector of observations $y_t$ is assumed to depend on the $q \times 1$ vector of latent states, $\eta_t$:[3]

$$y_t = G_t \eta_t + \varepsilon_t \tag{7}$$

$$\eta_t = F_t \eta_{t-1} + \ \nu_t \tag{8}$$

and (8) is initialized with $\eta_1 \sim N(x_1, D)$, and

$$\begin{pmatrix} \varepsilon_t \\ \nu_t \end{pmatrix} \sim N\left(0, \begin{pmatrix} \Omega_{11} & 0 \\ 0 & \Omega_{22} \end{pmatrix}\right).$$

Using standard matrix notation: $y = (y_1', \ldots, y_T')', \eta = (\eta_1', \ldots, \eta_T')'$,

[3] I use the same notation as the original paper by Chan and Jeliazkov (2009) to make potential cross-check easier. Note however, that I consider a slightly less general model than the authors above. This is mainly done for the expositional ease, but the derivation provided below can be easily extended to a more general model (at the cost of slightly heavier notation).

$$G = \begin{bmatrix} G_1 & & \\ & \ddots & \\ & & G_T \end{bmatrix}; H = \begin{bmatrix} I_q & & & & \\ -F_2 & I_q & & & \\ & -F_3 & I_q & & \\ & & \ddots & \ddots & \\ & & & -F_T & I_q \end{bmatrix}.$$

$$\begin{array}{l} \varepsilon \sim N(0, I_T \otimes \ \Omega_{11}) \\ \nu \sim N(0, S), \end{array}$$

where

$$S = \begin{bmatrix} D & & & & \\ & \Omega_{22} & & & \\ & & \Omega_{22} & & \\ & & & \ddots & \\ & & & & \Omega_{22} \end{bmatrix},$$

the state-space model (7)-(8) can be written in the compact form as:

$$\begin{array}{l} y = G\eta + \varepsilon \\ H\eta = \nu \end{array}$$

From the least-square perspective, any estimate of the state sequence $\eta$ should make discrepancies in the measurement as well as the state equation as small as possible (i.e. $y_t - G_t\eta_t \approx 0$ and $\eta_t - F_t\eta_{t-1} \approx 0$, for $t = 1, \ldots, T$). This can be achieved by minimizing the bi-objective[4] least square problem:

$$\lambda_1 \|G\eta - y\|^2 + \lambda_2 \|H\eta - \mathbf{0}\|^2. \tag{9}$$

Pursuing the data-augmentation approach outlined in (5), we obtain:

$$\tilde{X} = \begin{bmatrix} \sqrt{\lambda_1} G \\ \sqrt{\lambda_2} H \end{bmatrix}, \tilde{y} = \begin{bmatrix} \sqrt{\lambda_1} y \\ \sqrt{\lambda_2} \mathbf{0} \end{bmatrix}.$$

[4] In fact, the weighted objective function comprises of 3 objectives as one should also minimize the term: ${x_1}' D^{-1} x_1$ that corresponds to initial-conditions constraint. To follow Chan and Jeliazkov (2009) as close as possible, I only work with two objectives, since the objective for initial conditions has already been stacked in their matrices $H$ and $S$. For simplicity, I assume that $x_1$ is set to zero vector. An extension to a more general case of specific initial conditions is straightforward and is provided in Appendix for convenience.

Making the use of (6), its standard OLS solution can be expressed as:

$$\left( [\sqrt{\lambda_1}G \quad \sqrt{\lambda_2}H] \begin{bmatrix} \sqrt{\lambda_1}G \\ \sqrt{\lambda_2}H \end{bmatrix} \right)^{-1} [\sqrt{\lambda_1}G \quad \sqrt{\lambda_2}H] \begin{bmatrix} \sqrt{\lambda_1}y \\ \sqrt{\lambda_2}\mathbf{0} \end{bmatrix} = \quad (10)$$

$$= (\lambda_1 G'G + \lambda_2 H'H)^{-1}\lambda_1 G'y$$

If the errors are not tied by any stochastic restrictions and distributional assumptions (that may not hold in practice), the equation (10) just defines a flexible least squares (FLS) estimator of Kalaba and Tesfatsion (1989) who set $\lambda_1 = 1$. The authors do not suggest any specific value for $\lambda_2$ and rather use it as a tuning parameter.[5] On the other hand, if the errors are presumed to be Gaussian, then the weights $\lambda_1$, $\lambda_2$ can be set optimally to minimize the mean square error. In particular, the weights should equal to the inverse of the error variance in the measurement and the state equation, respectively: i.e. $\lambda_1 = (I \otimes \Omega_{11}^{-1})$ and $\lambda_2 = S^{-1}$.

Under this setting, the mean of $\eta$ emerges as

$$\hat{\eta} = (G'(I \otimes \Omega_{11}^{-1})G + H'S^{-1}H)^{-1}G'(I \otimes \Omega_{11}^{-1})y.$$

and the variance is given by

$$P^{-1} = \ (G'(I \otimes \Omega_{11}^{-1})G + K)^{-1}, \quad (11)$$

where $K = H'S^{-1}H$. To derive the variance, we applied its standard OLS expression $\sigma^2\left(\tilde{X}'\tilde{X}\right)^{-1}$. Since all 'observations' in $\tilde{X}$ are scaled by the inverse variance of the error terms, the symbol $\sigma^2$ actually represents a unit variance and the expression reduces to $\left(\tilde{X}'\tilde{X}\right)^{-1}$. One can easily check that this is exactly the same result as that obtained by Chan and Jeliazkov (2009).

Using (11) it is further possible to simplify the expression for the mean as

$$\hat{\eta} = P^{-1}(G'(I \otimes \Omega_{11}^{-1})y), \quad (12)$$

[5] They rather argue that the whole collection of paths for $\eta$ can be of interest. It is also interesting to note, that (despite their awareness of the possibility) Kalaba and Tesfatsion (1989) do not take the advantage of the special form of the matrix to be inverted and propose sequential procedure to solve the problem. Montana et al. (2009) have shown that FLS recursions are largely similar to traditional Kalman filter/smoother.

which is closely related, but perhaps more intuitive expression for the mean than its original counterpart.

Also note that it is always possible to rescale the weights in such a way that $\lambda_1 = I$ (see above). This leads to alternative weights for $\lambda_2 = (I \otimes \Omega_{11})S^{-1}$, which simply corresponds to the signal-to-noise ratio of the state space model. It is a well-known fact (often alluded to within the context of the Hodrick-Prescott filter, for example) that variability of $\eta$ does not depend on the absolute magnitude of variances, but is only driven by their relative magnitude (i.e. by signal-to-noise ratio).

In order to generate samples from the distribution $\eta \sim N(\hat{\eta}, P^{-1})$ efficiently, it is necessary 'invert' the large matrix $P$ in a computationally inexpensive way.[6] As noted by Chan and Jeliazkov (2009) this goal can be handled by the application of the sparsity-aware algorithms due to the specific form of the matrix $P$. It can be shown that $P$ is a block band tridiagonal matrix that contains a lot of zeros (see Kalaba and Tesfatsion, 1989 and Aravkin et al., 2013 for its general form). Since under mild assumptions the matrix is positive definite (Lütkepohl and Herwartz, 1996) its fast inverse can be obtained via a Cholesky decomposition and recursive inversion of a triangular matrix. The algorithms designed for speedy sparse-matrix computations are readily available on many modern software platforms, including Matlab or R, which makes the code implementation of the sampler very fast[7] and conceptually elegant.

# 3 Summary and possible extensions

This short note provided complementary look on the latent-state sampler proposed by Chan and Jeliazkov (2009). The formulation along the lines of the least-squares paradigm can be easier to follow for some readers while providing closer insights on some interesting links to other existing methods. Indeed, the smoothed estimate of the latent states can be seen as little more than the application of the popular ridge regression (or Tikhonov regularization).[8]

---

[6] Since we essentially solve a least square problem, one can potentially use any general sparse least-squares solvers.

[7] In other domains (Gaussian Markov Random Fields), this approach is known as a Cholesky factor algorithm. McCausland et al. (2011) investigate its performance vis-à-vis Kalman filter-based sampling and find substantial efficiency gains over the methods relying on the Kalman filter.

[8] Secondary objectives $f_2, \ldots, f_k$ can be interpreted as regularizing terms. This implies that multi-objective least squares formulation can be also seen as a *regularized estimation* problem.

Another advantage of the least-square formulation is that it instantly provides a fertile and fully general ground for many relevant extensions. The researcher can freely add another objectives (penalty terms) into the cost function (6) to express her desire for the additional constraints on the sequence of states, $\eta$. Such restrictions may help to accommodate relevant prior knowledge about the model behavior, for example. Lütkepohl and Herwartz (1999) use additional penalties to account for seasonal patterns in time-varying coefficients, Andrle (2014) uses additional objective to obtain uncorrelated structural shocks in the DSGE model (presented in a state-space form) and Andrle and de Wind (2017) use the stacked least-squares formulation to elicit system priors within the time-varying VAR context. In general, any prior of the sort $\|A\eta - \eta^{des}\|^2$ where $A$ is some convenient matrix for the problem at hand and $\eta^{des}$ stands for some desired sequence of $\eta$ can be easily implemented. In some respect, such an approach is close to optimal control theory. It should be also stressed that the objectives do not need to take quadratic form and the method can be made much more general. However, costly numerical optimization would be usually necessary if the specific conditions are not met.

# Literature


Andrle, M. (2014): Estimating Structural Shocks with DSGE models, *manuscript*.

Andrle, M. and de Wind, J. (2017): Time-varying VARs and System Priors, *manuscript*, preliminary.

Aravkin, A. Y., Bell, B. B., Burke J. V. and Pillonetto, G. (2013): Kalman smoothing and block triagonal systems: new connections and numerical stability results, arXiv.org.

Boyd, S. and Vandenberghe L. (2016): Vectors, Matrices, and Least Squares (working title), *textbook draft*.

Carter, C.K. and Kohn, R. (1994): On Gibbs sampling for state space models, *Biometrika*, Vol. 81, pp. 541-553.

Chan, J. C.C. and Jeliazkov, I. (2009): Efficient simulation and integrated likelihood estimation in state space models, *Mathematical Modelling and Numerical Optimisation*, Vol. 1, Nos. 1/2.

de Jong, P. and Shepard, N. (1995): The simulation smoother for time series models, *Biometrika*, Vol. 82, pp. 339-350.

Durbin, J. and Koopman, S. J. (2002): A simple efficient simulation smoother for state space time series analysis, *Biometrika*, Vol. 89, pp. 603-615.

Giles, D. E. (2013): Constructing conficence bands for the Hodrick-Prescott filter, *Applied Economic Letters*, Vol. 20 (5), pp. 480-484.

Kalaba, R. and L. Tesfatsion (1989): Time-varying linear regression via flexible least squares, *Computers and Mathematics with Applications*, Vol. 17, pp. 1215-1245.

Lütkepohl, H. and Herwartz, H. (1996): Specification of varying coefficient time series models via generalized flexible least squares, *Journal of Econometrics*, Vol. 70, pp. 261-290.

McCausland, W. J., Miller, S. and Pelletier, D. (2011): Simulation smoothing for state-space models: A computational efficiency analysis, *Computational Statistics and Data Analysis*, Vol. 55, pp. 199-212.

Montana, G., Triantafyllopoulos, K. and Tsagaris, T. (2009): Flexible least squares for temporal data mining and statistical arbitrage, *Expert Systems with Applications*, Vol. 36, pp. 2819-2830.

Sorenson, H.W. (1970): Least-squares estimation: from Gauss to Kalman, *IEEE Spectrum*, Vol. 7 (7), pp. 63-68.

# Appendix

Let us consider a case when the mean of the initial state, $x_0$, is not set to zero. Define $y_2 = (x_0, 0, ..., 0)$ and analogously to (9) solve the bi-objective least squares problem:

$$\lambda_1 \|G\eta - y\|^2 + \lambda_2 \|H\eta - y_2\|^2.$$

Its solution, again, is analogous to Equation (10):

$$\left( [\sqrt{\lambda_1}G \quad \sqrt{\lambda_2}H] \begin{bmatrix} \sqrt{\lambda_1}G \\ \sqrt{\lambda_2}H \end{bmatrix} \right)^{-1} [\sqrt{\lambda_1}G \quad \sqrt{\lambda_2}H] \begin{bmatrix} \sqrt{\lambda_1}y \\ \sqrt{\lambda_2}y_2 \end{bmatrix}$$

and in the case of Gaussian errors it can be expressed as

$$\hat{\eta} = (G'(I \otimes \Omega_{11}^{-1})G + H'S^{-1}H)^{-1}(G'(I \otimes \Omega_{11}^{-1})y + H'S^{-1}y_2) =$$

$$= P^{-1}(G'(I \otimes \Omega_{11}^{-1})y + H'S^{-1}y_2).$$

The variance of the states remains unchanged.