

Some Structural Dynamic Models

Overcoming Latent Score Indeterminacy

Modelli Strutturali Dinamici Con Osservabilità delle Variabili Latenti

Simona Caterina Minotti

Istituto di Statistica,
Università Cattolica di Milano, Largo Gemelli 1, Milano
minotti.tutor@mi.unicatt.it

Riassunto: Nell'ambito dei modelli strutturali dinamici, si discute l'indeterminatezza della soluzione del filtro di Kalman e si propone una procedura alternativa che produce soluzioni uniche, basata sulla Regression Component Decomposition.

Keywords: Structural Dynamic Models, Latent Variables, Kalman Filter, Indeterminacy, Regression Component Decomposition.

1. Introduction

It is well known that the causality principle of the factor model (i.e. to express the observed variables as a function of the latent variables) leads to indeterminacy of latent scores. The same problem arises in the structural equation models (SEM) and also in the dynamic structural models (SDL), because both use as measurement model the factor model (FA). Referring to the dynamic case, it can be proved that the scores of the latent variables obtained by means of the Kalman filter are not unique, i.e. it is always possible to find an infinite number of solutions also very different from each other, even negatively correlated (Haagen, Vittadini; 1994).

To avoid the arbitrariness in interpreting the common factors, some authors propose an estimation procedure which inverts the natural relationship among the variables, expressing the latent variables as a linear combination of the indicator variables. For example, as an alternative to the FA, Schönemann and Steiger (1976) proposed a decomposition of the datamatrix, called Regression Component Decomposition (RCD), while Haagen and Oberhofer (1999) introduced the Replicated Common Factor Model (RCFM). The RCD was successively extended to SEM by Haagen and Vittadini (1991); another alternative is the Partial Least Squares (PLS) by Wold (1982).

Nevertheless no proposal has been made in the dynamic context. In this paper a correction of the Kalman filter by means of RCD is presented.

2. The structural dynamic model

The SDL, introduced by Otter (1992) as a dynamic generalization of SEM, is a linear model with latent variables, where the observations are represented by a single multivariate time series. The SDL consists of a *transition equation*, which describes the temporal relationship among the latent variables, and a *measurement equation*, that

relates the latent variables to the observed variables. Omitting the deterministic input, the SDL reduces to:

$$\mathbf{x}_t = \mathbf{A} \mathbf{x}_{t-1} + \mathbf{w}_t \quad (t=1, \dots, T) \quad (1.1)$$

$$\mathbf{z}_t = \mathbf{C} \mathbf{x}_t + \mathbf{v}_t \quad (t=1, \dots, T) \quad (1.2)$$

where $\mathbf{x}_t = [\mathbf{h}_t, \mathbf{f}_t]$, with $\mathbf{h}_t$ and $\mathbf{f}_t$ vectors respectively of m_1 endogenous and m_2 exogenous latent variables and finite covariance $\mathbf{S}_{\mathbf{x}}$; $\mathbf{w}_t$ and $\mathbf{v}_t = [\mathbf{e}_t, \mathbf{d}_t]$ vectors respectively of $m = m_1 + m_2$ and $p = p_1 + p_2$ errors with expected value equal to zero and time-invariant covariance matrix $\mathbf{S}_{\mathbf{w}}$ and $\mathbf{S}_{\mathbf{v}}$; $\mathbf{A}$ and $\mathbf{C}$ matrices of parameters respectively of dimension m and $(p \times m)$; $\mathbf{z}_t$ vector of p observed variables with covariance $\mathbf{S}_{\mathbf{z}}$.

The latent scores are estimated by the Kalman filter, which is an algorithm that provides linear least squares estimates of the latent vector in two steps; by first we have:

$$\hat{\mathbf{x}}_{t|t-1} = \hat{\mathbf{E}}(\mathbf{x}_t | \mathbf{Z}_{t-1}) = \mathbf{A} \hat{\mathbf{x}}_{t-1} \quad (1.3)$$

$$\mathbf{S}_{\mathbf{x}_{t|t-1}} = \mathbf{E} \left[\left(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t-1} \right) \left(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t-1} \right)' \right] = \mathbf{A} \mathbf{S}_{\mathbf{x}_{t-1}} \mathbf{A}' + \mathbf{S}_{\mathbf{w}} \quad (1.4)$$

where $\mathbf{Z}_{t-1} = (\mathbf{z}_{t-1}, \dots, \mathbf{z}_0)$ and $\hat{\mathbf{E}}(\mathbf{x}_t | \mathbf{Z}_{t-1})$ denotes the linear projection of $\mathbf{x}_t$ on $\mathbf{Z}_{t-1}$.

Next the inference about the current value of $\mathbf{x}_t$ is updated on the basis of the last observation $\mathbf{z}_t$:

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_t = \hat{\mathbf{E}}(\mathbf{x}_t | \mathbf{z}_t, \mathbf{Z}_{t-1}) = \hat{\mathbf{E}}(\mathbf{x}_t | \mathbf{Z}) = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t (\mathbf{z}_t - \hat{\mathbf{z}}_t) \quad (1.5)$$

$$\mathbf{S}_{\mathbf{x}_t} = (\mathbf{I} - \mathbf{K}_t \mathbf{C}) \mathbf{S}_{\mathbf{x}_{t|t-1}} \quad (1.6)$$

where $\mathbf{K}_t = \mathbf{S}_{\mathbf{x}_{t|t-1}} \mathbf{C}' \mathbf{S}_{\mathbf{z}_t}^{-1}$ and $\hat{\mathbf{z}}_t = \mathbf{E}(\mathbf{z}_t) = \mathbf{C} \hat{\mathbf{x}}_{t|t-1}$. (For more details see Otter, 1992).

Most of the published Kalman filter applications consider the $N=1$ case (single time series), because they refer to fields of application where only one (multivariate) observation at a time is available. However, Kalman filter can also apply to large N data (longitudinal data), as in behavioral sciences (Oud, 2001). It will be clear in the following that the alternative proposed is particularly suitable for such a situation.

3. The indeterminacy of the Kalman filter scores

Haagen and Vittadini (1994) discuss the indeterminacy of the Kalman solution by using the result of Guttman (1955). They prove that, if the following equality holds

$$\mathbf{S}_{\mathbf{z}_t} = \mathbf{C} \mathbf{S}_{\mathbf{x}_t} \mathbf{C}' + \mathbf{S}_{\mathbf{v}_t}, \quad (1.7)$$

an infinite number of vectors $\mathbf{x}_t$, which satisfy the (1.2), always exists. These solutions, called “true solutions”, have the following structure:

$$\mathbf{x}_t = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t [\mathbf{z}_t - \hat{\mathbf{z}}_t] + \mathbf{P}_t \mathbf{w}_t \quad (1.8)$$

where $\mathbf{K}_t = \mathbf{S}_{\mathbf{x}_{t-1}} \mathbf{C}' \mathbf{S}_{\mathbf{Z}_t}^{-1}$, $\mathbf{P}_t = (\mathbf{I} - \mathbf{K}_t \mathbf{C}) \mathbf{F}_t$, $\mathbf{F}_t \mathbf{F}_t' = \mathbf{S}_{\mathbf{x}_0}$, $\mathbf{w}_t$ arbitrary vector of m variables and $E[\mathbf{w}_t] = 0$, $\mathbf{S}_{\mathbf{w}_t, \mathbf{Z}_t} = \mathbf{0}$, $\mathbf{S}_{\mathbf{w}_t} = \mathbf{I}$.

It follows that, as in the static context, the latent scores obtained by means of the Kalman filter are not unique, due to the arbitrary vector $\mathbf{w}_t$. Moreover Haagen and Vittadini (1994) prove that the dynamics makes worse the indeterminacy, due to $\hat{\mathbf{z}}_t$. Only the introduction in the (1.3)-(1.6) of alternative models to the FA permits definitely to overcome the indeterminacy.

4. An alternative to the Kalman filter

Suppose to have N observations at each date t , we can express the model in term of the observations as follows:

$$\begin{aligned}\mathbf{X}_t &= \mathbf{X}_{t-1} \mathbf{A}' + \mathbf{W}_t \\ \mathbf{Z}_t &= \mathbf{X}_t \mathbf{C}' + \mathbf{V}_t\end{aligned}$$

with $\mathbf{X}_t = [\mathbf{F}_t, \mathbf{H}_t]$ ($N \times m$), $\mathbf{A}$ ($m \times m$), $\mathbf{W}_t$ ($N \times m$), $\mathbf{Z}_t$ ($N \times p$), $\mathbf{C}$ ($p \times m$), $\mathbf{V}_t$ ($N \times p$) matrices. The RCD provides the following decomposition of the datamatrix $\mathbf{Z}_t$:

$$\mathbf{Z}_t = \tilde{\mathbf{F}}_t \tilde{\mathbf{A}}_{\mathbf{Z}_t \tilde{\mathbf{F}}_t} + (\mathbf{Z}_t - \tilde{\mathbf{F}}_t \tilde{\mathbf{A}}'_{\mathbf{Z}_t \tilde{\mathbf{F}}_t}) \quad (1.9)$$

which leads to a definition of the latent variables $\mathbf{F}_t$ as a linear combination of the observed variables $\mathbf{Z}_t$

$$\tilde{\mathbf{F}}_t = \mathbf{Z}_t \mathbf{L}'_{\mathbf{F}_t} \quad (1.10)$$

where $\mathbf{L}_{\mathbf{F}_t} = (\tilde{\mathbf{A}}'_{\mathbf{Z}_t \tilde{\mathbf{F}}_t} \mathbf{S}_{\mathbf{Z}_t}^{-1} \tilde{\mathbf{A}}_{\mathbf{Z}_t \tilde{\mathbf{F}}_t})^{-1} \tilde{\mathbf{A}}'_{\mathbf{Z}_t \tilde{\mathbf{F}}_t} \mathbf{S}_{\mathbf{Z}_t}^{-1}$ and $\tilde{\mathbf{A}}_{\mathbf{Z}_t \tilde{\mathbf{F}}_t}$ can be calculated as a factor loading matrix by means of a factor extraction method.

Given the unique approximation $\tilde{\mathbf{F}}_t$, $\tilde{\mathbf{H}}_t$ and $\mathbf{L}_{\mathbf{H}_t}$ are simply determined. The matrix $\mathbf{L}_t = [\mathbf{L}_{\mathbf{F}_t}, \mathbf{L}_{\mathbf{H}_t}]'$ provided by the RCD is then introduced in the (1.3)-(1.6).

Under the assumption that $\mathbf{A}$, $E[\mathbf{x}_0]$ and $\mathbf{S}_{\mathbf{x}_0}$ are known, the first step of the Kalman filtering at each time t becomes:

$$\begin{aligned}\bar{\mathbf{X}}_{t|t-1} &= \bar{\mathbf{X}}_{t-1} \mathbf{A}' \\ \mathbf{S}_{\bar{\mathbf{X}}_{t|t-1}} &= \mathbf{A} \mathbf{S}_{\bar{\mathbf{X}}_{t-1}} \mathbf{A}' + \mathbf{S}_{\mathbf{W}}.\end{aligned} \quad (1.11)$$

In the second step we update $\bar{\mathbf{X}}_{t|t-1}$ on the basis of the new observations $\mathbf{Z}_t$, that is.

$$\begin{aligned}\bar{\mathbf{X}}_t &= \bar{\mathbf{X}}_{t|t-1} + (\mathbf{Z}_t - \bar{\mathbf{Z}}_t) \bar{\mathbf{K}}'_t \\ \mathbf{S}_{\bar{\mathbf{X}}_t} &= \left(\mathbf{I} - \bar{\mathbf{K}}_t (\mathbf{L}_t^{-1})' \right) \mathbf{S}_{\bar{\mathbf{X}}_{t|t-1}},\end{aligned} \quad (1.12)$$

where $\bar{\mathbf{Z}}_t = \bar{\mathbf{X}}_{t|t-1} \mathbf{L}_t^{-1}$, $\left(\tilde{\mathbf{X}}_t = \mathbf{Z}_t \mathbf{L}_t', \mathbf{L}_t = [\mathbf{L}_{\mathbf{F}t}, \mathbf{L}_{\mathbf{H}t}]' \right)$ and $\bar{\mathbf{K}}_t = \mathbf{S}_{\bar{\mathbf{X}}_{t|t-1}} \mathbf{L}_t^{-1} \mathbf{S}_{\mathbf{Z}_t}^{-1}$.

The solution $\bar{\mathbf{X}}_t$ is unique under construction because the reason of indeterminacy has been removed. The indeterminacy of the Kalman solution derives from the definition of the forecasting $\hat{\mathbf{Z}}_t$ through the model (1.2); by the alternative solution, the substitution of the matrix $\mathbf{C}$ by $\mathbf{L}_t$ in the definition of $\hat{\mathbf{Z}}_t$, now indicated by $\bar{\mathbf{Z}}_t$, permits to avoid the arbitrariness of the latent scores, because the result of Guttman is no more applicable. Guttman considers in fact models like the (1.2), in which the indeterminacy of the $\mathbf{x}_t$ and $\mathbf{v}_t$ results from the fact that it is impossible to identify the $m+p$ basis vectors given the p observable variables. The RCD, on the contrary, provides a definition of the latent variables as a linear combination of the observable variables, such that indeterminacy vanishes.

Hence, the introduction of the RCD in the recursive procedure (1.3)-(1.6) brings to a unique approximation for the latent scores.

The comparison between the Kalman filtering and the alternative procedure proposed is accompanied by some empirical results referred to real data about Customer Satisfaction, elaborated through some routines written by the Matlab language.

5. Conclusions

The method presented provides unique values for the latent scores in a dynamic context. The next step of the research will be the building of a recursive procedure which calculates simultaneously latent scores and parameters and the discussion of the properties of the approximations provided.

References:

- Guttman L. (1955) The determinacy of factor score matrices with implications for five other basic problems of common factor theory, *British Journal of Math. Stat. Psych.*, 8(II), pp.65-81.
- Haagen K., Oberhofer W. (1999) The replicated common factor-analysis model, *Metron*, LVII(3-4), pp. 35-45.
- Haagen K., Vittadini G. (1994) Sul problema dell'osservabilità nei modelli strutturali dinamici, in *Atti della XXXVII Riunione Scientifica SIS, Sanremo*, vol.1, pp.247-258.
- Haagen K., Vittadini G. (1991) Regression component decomposition in structural analysis, *Communication in Statistics - Theory and Method*, 20(4), pp.1153-1161.
- Otter P. W. (1992) Dynamic models with latent variables from a system theoretic perspective: theory and applications, *Statistica*, LII, 3, pp.347-363.
- Oud J. H. L. (2001) Quasi-longitudinal designs in SEM state space modeling, *Statistica Neerlandica*, 55(2), pp.200-220.
- Schönemann P.H., Steiger J. (1976) Regression Component Analysis, *British Journal of Math. Stat. Psych.*, 29, pp.175-189.
- Wold H. (1982) Soft modelling: the basic design and some extensions, in Joreskog K., Wold H., eds., *Systems under indirect observation: causality, structure, prediction*, North Holland, vol. II, pp. 1-54.