Available online at www.sciencedirect.com

ScienceDirect

Statistics & Probability Letters 78 (2008) 257–264

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Missing data in time series: A note on the equivalence of the dummy variable and the skipping approaches

Tommaso Proietti

Dipartimento S.E.F. e ME.Q, Via Columbia 2, 00133 Rome, Italy

Received 13 March 2006; received in revised form 22 February 2007; accepted 23 May 2007
Available online 16 June 2007

Abstract

This note shows the equivalence of the dummy variable approach and the skipping approach for the treatment of missing observations in state space models. The equivalence holds when the coefficient of the dummy variable is considered as a diffuse rather than a fixed effect. The equivalence concerns both likelihood inference and smoothed inferences.

Keywords: Kalman filter; Smoothing; Influence; Cross-validation

1. Introduction

A well-known result is that estimating a missing observation by skipping the Kalman filter (KF) updating step is equivalent to introducing a dummy variable (additive outlier) in the measurement equation, filling the missing value arbitrarily. This result (in different frameworks) appears in a number of papers: Sargan and Drettakis (1974), Bruce and Martin (1989), Ljung (1993). A detailed discussion can be found in Fuller (1996, Section 8.7). However, if the additive outlier is treated as a fixed effect, with zero covariance matrix, the likelihood is defined differently and a correction has to be computed in the second case, see Gómez et al. (1999). The correction factor is related to the determinantal term of the likelihood and depends in a simple fashion from quantities computed under the model for the complete observations, requiring a single run of the KF and smoothing filter.

To our knowledge, a proof the equivalence of the skipping approach and the dummy approach for the definition of the likelihood and for smoothing is not available. This note aims at bridging the gap, providing a simple proof that when the additive outlier is treated as diffuse, with arbitrarily large covariance matrix, the correction to the likelihood takes place automatically. This is convenient, as no extra programming effort is necessary once a programme handling diffuse initial conditions and regression effects has been implemented.

The equivalence is also carried forward to smoothed inferences, concerning the estimation of the states and the disturbances. The derivation of analytical expressions for the influence of an observation on these quantities, made in De Jong (1996), is greatly simplified in the dummy variable setup as they depend in a simple fashion on the output of the KF and smoother run on intervention variables.

E-mail address: tommaso.proietti@uniroma2.it

0167-7152/$ - see front matter
doi:10.1016/j.spl.2007.05.031

The plan of the paper is the following: Section 2 introduces the dummy variable approach for stationary state space models with no regression effects, under fixed and diffuse conditions, and derives the prediction error decomposition form of the likelihood under the latter. In Section 3 we present the alternative strategy of handling missing observations, known as the skipping approach, and prove that the likelihood for this model is equivalent to the dummy variable one. In Section 4 the equivalence is extended to smoothed estimates of the states and the disturbances, and measures of influence of an observations are given, which depends in a simple way on the output of the KF and smoothing filter run on the intervention variable.

2. The Dummy variable approach

Let $\boldsymbol{y}_t$ denote a vector stationary time series with N elements; the state space model is

$$\boldsymbol{y}_t = \boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{G}_t\boldsymbol{\varepsilon}_t, \quad t = 1, 2, \ldots, T, \tag{1}$$

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{T}_t\boldsymbol{\alpha}_t + \boldsymbol{H}_t\boldsymbol{\varepsilon}_t, \quad t = 1, 2, \ldots, T, \tag{2}$$

with $\boldsymbol{\alpha}_1 \sim \mathrm{N}(\boldsymbol{a}_1, \sigma^2\boldsymbol{P}_1)$, where $\boldsymbol{a}_1$ and $\sigma^2\boldsymbol{P}_1$ denote the unconditional mean and covariance matrix of $\boldsymbol{\alpha}_t$, and $\boldsymbol{\varepsilon}_t \sim \mathrm{NID}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The system matrices, $\boldsymbol{Z}_t, \boldsymbol{G}_t, \boldsymbol{T}_t, \boldsymbol{H}_t$, are functionally related to a vector of hyperparameters, $\boldsymbol{\theta}$.

The Kalman filter (KF) is a well-known recursive algorithm for computing the minimum mean square estimator of $\boldsymbol{\alpha}_t$ and its mean square error (MSE) matrix conditional on $\boldsymbol{Y}_{t-1} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{t-1}\}$. Defining

$$\boldsymbol{a}_t = \mathrm{E}(\boldsymbol{\alpha}_t|\boldsymbol{Y}_{t-1}), \quad \mathrm{MSE}(\boldsymbol{a}_t) = \sigma^2\boldsymbol{P}_t = \mathrm{E}[(\boldsymbol{\alpha}_t - \boldsymbol{a}_t)(\boldsymbol{\alpha}_t - \boldsymbol{a}_t)'|\boldsymbol{Y}_{t-1}],$$

the filter consists of the following recursions:

$$\boldsymbol{v}_t = \boldsymbol{y}_t - \boldsymbol{Z}_t\boldsymbol{a}_t, \quad \boldsymbol{F}_t = \boldsymbol{Z}_t\boldsymbol{P}_t\boldsymbol{Z}_t' + \boldsymbol{G}_t\boldsymbol{G}_t',$$

$$q_t = q_{t-1} + \boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t, \quad \boldsymbol{K}_t = (\boldsymbol{T}_t\boldsymbol{P}_t\boldsymbol{Z}_t' + \boldsymbol{H}_t\boldsymbol{G}_t')\boldsymbol{F}_t^{-1},$$

$$\boldsymbol{a}_{t+1} = \boldsymbol{T}_t\boldsymbol{a}_t + \boldsymbol{K}_t\boldsymbol{v}_t, \quad \boldsymbol{P}_{t+1} = \boldsymbol{T}_t\boldsymbol{P}_t\boldsymbol{L}_t' + \boldsymbol{H}_t\boldsymbol{J}_t' \tag{3}$$

with $\boldsymbol{L}_t = \boldsymbol{T}_t - \boldsymbol{K}_t\boldsymbol{Z}_t$ and $\boldsymbol{J}_t = \boldsymbol{H}_t - \boldsymbol{K}_t\boldsymbol{G}_t$; $\boldsymbol{v}_t = \boldsymbol{y}_t - \mathrm{E}(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1})$ are the filter innovations, with MSE matrix $\sigma^2\boldsymbol{F}_t$. The filter is started off with $\boldsymbol{a}_1 = \boldsymbol{0}$, $\boldsymbol{P}_1 = \boldsymbol{H}_0\boldsymbol{H}_0'$ and $q_0 = 0$. The log-likelihood for the model is, apart from a constant term,

$$\mathscr{L}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T; \boldsymbol{\theta}) = -\frac{1}{2}\left[NT\ln\sigma^2 + \sum_{t=1}^{T}\ln|\boldsymbol{F}_t| + \sigma^{-2}q_T\right], \tag{4}$$

where $q_T = \sum_{t=1}^{T}\boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t$.

Suppose that an intervention is included at $t = i$ so that the measurement equation becomes

$$\boldsymbol{y}_t = \boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{I}_t(i)\boldsymbol{\delta} + \boldsymbol{G}_t\boldsymbol{\varepsilon}_t, \tag{5}$$

where $\boldsymbol{I}_t(i)$ is an indicator variable taking value 1 for $t = i$ and 0 elsewhere. For its statistical treatment, the KF (3) at $t = i$ is augmented by the following recursions:

$$\boldsymbol{V}_t^+ = I_t(i)\boldsymbol{I} - \boldsymbol{Z}_t\boldsymbol{A}_t^+,$$

$$\boldsymbol{A}_{t+1}^+ = \boldsymbol{T}_t\boldsymbol{A}_t^+ + \boldsymbol{K}_t\boldsymbol{V}_t^+ = \boldsymbol{K}_i I_t(i) + \boldsymbol{L}_t\boldsymbol{A}_t^+,$$

$$\boldsymbol{S}_t^+ = \boldsymbol{S}_{t-1}^+ + \boldsymbol{V}_t^{'+}\boldsymbol{F}_t^{-1}\boldsymbol{V}_t^+,$$

$$\boldsymbol{s}_t^+ = \boldsymbol{s}_{t-1}^+ + \boldsymbol{V}_t^{+'}\boldsymbol{F}_t^{-1}\boldsymbol{v}_t, \tag{6}$$

for $t = i, \ldots, T$ with starting conditions: $\boldsymbol{A}_i^+ = \boldsymbol{0}$, $\boldsymbol{S}_{i-1}^+ = \boldsymbol{0}$ and $\boldsymbol{s}_{i-1}^+ = \boldsymbol{0}$. This amounts to apply the KF to the intervention signature $I_t(i)\boldsymbol{I}$.