Diss. ETH No. 16063

State Space Methods for Robust Estimation in AR-Models With External Regressors

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Doctor of Mathematics

presented by
CHRISTIAN SANGIORGIO
Dipl. Math. ETH
born June 20, 1973
citizen of Balerna TI

accepted on the recommendation of
Prof. Dr. H. R. Künsch, examiner
Prof. Dr. F. Hampel, co-examiner
Prof. Dr. W. Stahel, co-examiner

2005

Summary

Regression models with time series as both response and explanatory variables play an important role in several fields of science. The easiest time series regression model is built with independent $\mathcal{N}(0, \sigma^2)$ distributed errors. But these error assumptions are not fulfilled in many real situations. In reality, correlation in the random errors is present and/or some outliers are observable.
Many methods have been developed to take into consideration these features. The random errors have been modeled using an AR or ARIMA process. Robust methods have been developed for time series autoregressions, generalizing robust methods applied in the usual regression context. But the proposed procedures have only been partially adapted to the time series situation and a simultaneous treatment of regression models with outliers and correlated errors has been missing.

A promising idea to handle both correlated errors and outliers is to write the time series regression models in state space form. State space models constitute a large and flexible class of models. They consist of an equation which describes the dynamics of the observation sequence (Y_t) using an unobserved first order Markov process (X_t) (the *state process*). For example, the popular Gaussian ARMA model is equivalent to a linear Gaussian state space model.
Parameter estimation and inference for state space models rely on the assessment of the distributions of the unknown states X_t given the observed values $Y_1, \ldots, Y_s$. Three cases are distinguished: the prediction $(t > s)$, the filtering $(t = s)$ and the smoothing $(t < s)$ *state distributions*. General recursive methods for calculating the respective state space distributions have been derived, but the formulae contain convolution integrals. Thus, an easy and closed form for the results is possible only under strict assumptions (Gaussian distributions for both the er-

rors and the starting X_0). But, the resulting methods are not robust and therefore outliers can have disastrous effects. Unfortunately, it is not easy to robustify these methods. On the other hand, it is possible to choose other error distributions than the Gaussian one to model the outliers. Then, approximations are needed in order to compute the state distributions. A new approach which works well also with high dimensional states X_t consists of approximating the state distributions by samples generated using Monte Carlo methods.

In the present thesis, we consider a time series regression model and assume linearity in the error term. Both auto-regressive correlated errors and observation outliers (*additive outliers*) are considered and no constraints are set on the relationship between response and explanatory time series. Moreover, missing values in the response series and/or in the explanatory series are allowed.

The resulting state distributions are computed approximately using Monte Carlo techniques. Special care is used to develop recursive algorithms which are both fast and reliable. In fact, these characteristics are a prerequisite for the estimation of the unknown parameters in the considered time series regression model.

The estimation issue is solved by applying the maximum likelihood method. The resulting estimates are robust thanks to the assumed error distributions. In addition, the maximum likelihood method permits to compute approximate confidence intervals by the usual likelihood procedures. The difficulty with this approach is that the likelihood function cannot be computed in closed form and thus it has to be approximated using again Monte Carlo methods. The developed methods are illustrated with some examples.

Finally, we consider the robustness of filter and smoother distributions to analyse how reliable the developed methods are.

Riassunto

In parecchi campi della scienza si incontrano modelli di regressione con variabili indipendenti e variabile dipendente date da serie temporali. Il modello più semplice in questi casi assume errori indipendenti e distribuiti normalmente con valore atteso zero e varianza σ^2. Purtroppo queste assunzioni per l'errore sono disattese in molte applicazioni reali. Infatti gli errori sono correlati e/o presentano valori anomali ("outliers").
Parecchi metodi sono stati sviluppati per tenere in considerazione queste caratteristiche. Per esempio gli errori sono stati modellati usando processi AR o ARIMA. Inoltre metodi robusti sono stati sviluppati per serie temporali con errori autocorrelati generalizzando i metodi robusti usati nella regressione. Ma le procedure proposte sono state solo parzialmente adattate alle caratteristiche delle serie temporali. L'analisi simultanea di modelli di regressione con errori correlati e valori anomali è ancora agli inizi.

Un' idea promettente per trattare sia gli errori correlati che i valori anomali è di scrivere le regressioni tra serie temporali usando modelli di stato ("state space models"). Questi ultimi costituiscono una classe ampia e flessibile di modelli. Consistono in un' equazione che descrive la dinamica delle osservazioni (Y_t) usando un processo markoviano (X_t) di ordine uno che non è osservabile direttamente (il cosiddetto *processo di stato*). Per esempio il modello ARMA con errori gaussiani, attualmente molto in voga, è equivalente a un modello lineare di stato con errori gaussiani.
Il problema maggiore nell'utilizzare i modelli di stato è dato dal calcolare le distribuzioni degli stati X_t conoscendo le osservazioni $Y_1, \ldots, Y_s$. Tre casi vengono distinti: la distribuzione di stato per la previsione $(t > s)$, per il filtro $(t = s)$ e per la ricostruzione a posteriori ("smoothing",

$t < s$). Per calcolare queste distribuzioni sono state sviluppate ricorsioni generali che però contengono convoluzioni. Così ricorsioni facili e in forma chiusa possono essere derivate solo nel caso in cui gli errori e la distribuzione iniziale X_0 siano date dalla distribuzione di Gauss. Queste ricorsioni però non sono robuste e così valori anomali nelle osservazioni possono avere un effetto disastroso. Purtroppo non è facile rendere robuste le ricorsioni trovate. D'altra parte, distribuzioni di errori diverse da quella di Gauss possono essere usate per modellare i valori anomali ma con lo svantaggio di dover usare delle approssimazioni per poter calcolare le distribuzioni di stato. Un nuovo approccio che funziona pure per stati X_t con dimensione elevata consiste nell'approssimare le distribuzioni di stato con un campione generato usando metodi di Monte Carlo.

Nella presente tesi consideriamo un modello di regressione tra serie temporali con linearità nei termini d'errore. Sia errori autocorrelati che valori anomali additivi sono presi in considerazione. Inoltre nessuna restrizione viene posta alla funzione che lega la serie temporale dipendente alle serie temporali indipendenti. Alcuni valori delle serie temporali dipendente e/o indipendenti possono mancare.
Le distribuzioni di stato che ne derivano sono calcolate approssimativamente usando metodi di Monte Carlo. Particolare cura è usata per sviluppare algoritmi ricorsivi che siano allo stesso tempo veloci e affidabili. Infatti queste due caratteristiche sono un prerequisito fondamentale per svillupare i metodi di stima dei parametri sconosciuti nel modello considerato.
La stima in sè viene trovata applicando il metodo della massima verosimiglianza ("maximum likelihood method"). Le stime risultanti sono robuste grazie alle distribuzioni dell'errore scelte. Inoltre, il metodo della massima verosimiglianza permette di trovare intervalli di confidenza approssimativi usando le tecniche sviluppate per questo metodo. La difficoltà in questo approccio sta nel fatto che la funzione di probabilità ("likelihood function") non può essere espressa in forma chiusa per il modello considerato e così deve essere approssimata usando nuovamente metodi di Monte Carlo. I metodi sviluppati vengono poi illustrati tramite alcuni esempi.
Da ultimo, consideriamo la robustezza delle distribuzioni di filtro ed "smoothing" trovate per analizzare quanto affidabili siano i metodi sviluppati.