Computational Statistics and Data Analysis 56 (2012) 3276–3285

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust analysis of default intensity

Tiziano Bellini, Marco Riani *

Università di Parma, Via Kennedy 6, I-43100, Parma, Italy

ARTICLE INFO

Article history:
Available online 3 April 2011

Keywords:
Forward search
Outlier
Kalman filter
Term structure of default intensity

ABSTRACT

The problem of robust estimation and multivariate outlier detection of the term structure of default intensity is considered. Both the multivariate Vasicek and CIR models, embedding the Kalman filter algorithm in a forward search context, are used to estimate default intensity. The focus is not on the estimation of credit models including jumps, but on the automatic detection of masked multiple outliers in multivariate time series. Both simulated and real market credit spread time series are analyzed. In order to make inference on outliers, confidence envelopes which are virtually independent of the estimated parameters are introduced. The output is not only a unique default intensity term structure curve, as often used in the financial literature, but a robust confidence interval within which default intensity is likely to stay.

1. Introduction

Pricing default risk has received a lot of attention from both practitioners and researchers (see for example Amendola et al., 2006; Fuertes and Kalotychou, 2006; Amendola et al., 2008). If, on the one hand, the standard theoretical paradigm is the contingent claims of Merton (1974), who explicitly linked the risk of a firm's default to the variability in the firm's asset value, on the other hand, the intensity approach has become very popular and widely used in practice.

The intensity approach has been pursued considering that at each instant there is some probability that a firm defaults on its obligations. Both this probability and the recovery rate, in the event of default, may vary stochastically through time. These stochastic processes determine the price of credit risk and, although they are not formally linked to the firm's asset value, there are presumably some underlying relations. This approach has been described as a reduced form model (Duffie and Singleton, 1999) and is used throughout this paper.

Our research is focused on the definition of a procedure to detect masked multiple outliers or jumps in multivariate time series and is applied to the term structure of default intensity. With regards to intensity estimation, we do not concentrate on jump diffusion models, which have been extensively used in the most recent literature (Eckner, 2009), since our goal is to detect jumps or masked multiple outliers allowing the researcher to analyze, free from masking and swamping problems, the impact that each unit, outlier or not, exerts on the estimated model. We exploit both multifactor Vasicek and CIR models which constitute the most popular paradigms in the context of term structure analysis. In order to estimate model parameters we exploit the Kalman filter representation described by Duan and Simonato (1999).

Many approaches have been developed in statistics to detect atypical observations. Given that traditional deletion methods, due to the well known masking effect, may not lead to the identification of the contaminated observations, the forward search was proposed, originally in linear and nonlinear regression by Atkinson and Riani (2000), as a powerful general method for detecting multiple masked outliers and for determining their effect on inferences about models fitted to data. In the forward search the evolution of residuals, parameter estimates and inferences is monitored as the subset size

* Corresponding author. Tel.: +39 0521902473; fax: +39 0521902375.
E-mail addresses: tiziano.bellini@unipr.it (T. Bellini), mriani@unipr.it (M. Riani).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.03.007

increases. Results are presented as forward plots which show the evolution of the quantities of interest as a function of the subset size.

In this paper we combine different lines of research in the context of term structure of default intensity and, for the first time in the literature, we extend the forward search to multivariate Kalman filter time series analysis. According to our knowledge this is the first attempt in the literature to try to estimate in a robust way default intensity. What makes the forward search particularly appealing over other robust methods based on high breakdown point estimators in the context of default intensity is the fact that we do not need to add additional iterative procedures which may lead to non-convergence or to find local minima, but we can simply use traditional maximum likelihood estimators and monitor how these estimates evolve as the subset increases.

We begin our data analysis on a simulated uncontaminated dataset, then we introduce contaminants and, stressing the need to find theoretical boundaries for the inference on outliers, we check the effectiveness of our procedure exploiting envelopes obtained through Monte Carlo simulations. As additional contribution to time series analysis, we consider the fast procedure to obtain envelopes from order statistics (Riani et al., 2009). The final step of this research is to apply our framework to real market time series. Our procedure allows us to obtain not only a term structure of default intensity, but also a robust confidence interval in which intensity is likely to stay.

The paper is organized as follows. In Section 2 we describe the mechanics of default intensity estimation through the Kalman filter. In Section 3 we describe the forward search for multivariate time series analysis. In Section 4 we apply our technique to simulated uncontaminated and contaminated credit spread time series. In Section 5 we extend our approach to real time series. Section 6 contains concluding remarks and directions for future research.

2. Default intensity estimation through the Kalman filter

In the financial literature, it is well known that, in a continuously compounded environment, at time 0, the price of a default free zero coupon bond with maturity τ can be expressed as

$$V(\tau) = E\left[\mathrm{e}^{-\int_0^\tau R(u)\mathrm{d}u}\right], \tag{1}$$

where $R(\cdot)$ is the risk free rate of interest and the expectation is taken over the possible paths of $R(\cdot)$.

According to Duffie and Singleton (1999), the reduced form pricing for a defaultable zero coupon bond $\tilde{V}(\tau)$ can be expressed as

$$\tilde{V}(\tau) = E\left[\mathrm{e}^{-\int_0^\tau \tilde{R}(u)\mathrm{d}u}\right] = E\left[\mathrm{e}^{-\int_0^\tau [R(u)+\lambda(u)]\mathrm{d}u}\right], \tag{2}$$

where $\lambda(\cdot)$ denotes the intensity of default.

Considering that the difference between the price of a defaultable zero coupon bond $\tilde{V}(\cdot)$ and a default free zero coupon bond $V(\cdot)$ is due to the difference in their interest rates $\tilde{R}(\cdot)$ and $R(\cdot)$, following Geyer et al. (2001), we argue that default intensity can be modeled through credit spreads $Y(\tau)$ obtained as follows

$$Y(\tau) = \tilde{R}(\tau) - R(\tau). \tag{3}$$

There is a clear analogy between conventional models for the risk free interest rate term structure and the above described structure. Given the convenience offered by analytical solution of the above equation and given that our goal is to detect atypical observations, we avoid the direct inclusion of jumps into the model and we assume that the instantaneous differential default intensity $\mathrm{d}Y(t)$, in the Vasicek setting (Vasicek, 1977), can be modeled as follows

$$\mathrm{d}Y(t) = k(\theta - Y(t))\mathrm{d}t + \sigma \mathrm{d}W(t), \tag{4}$$

or, using the CIR specification (Cox et al., 1985), as follows

$$\mathrm{d}Y(t) = k(\theta - Y(t))\mathrm{d}t + \sigma\sqrt{Y(t)}\mathrm{d}W(t), \tag{5}$$

where $W(t)$ is a Wiener process, $\theta \geq 0$ is the long run average of instantaneous default intensity, $k \geq 0$ is the mean reverting rate at which the process returns to its long run mean and $\sigma \geq 0$ is the volatility parameter of the process.

Extending the approach of Duan and Simonato (1999) to credit spreads, we can state that

$$Y(\tau) = \frac{1}{\tau}\left[B(\cdot)Y(t) - \ln(A(\cdot))\right], \tag{6}$$

where $A(\cdot)$ and $B(\cdot)$ contain the parameters θ, k, σ and the risk premium parameter ν. The precise functional form of $A(\cdot)$ and $B(\cdot)$ will be detailed when we describe the Kalman filter algorithm.

The stochastic differential equations above mentioned are specified in continuous time, but focusing on daily credit spreads y_t and concentrating on some nodes (maturity dates) of the term structure of default intensity, we need to describe their evolution in a discrete time setting. In general, inference from discrete time observations can be based on an approximation to the continuous-time likelihood, replacing Lebesgue integrals and Itô integrals by Riemann-Itô sums

provided the observation times, as in our case, are closely spaced (see for example Yoshida, 1990). On the other hand, if the time between observations is bounded away from 0 one should adjust the score function or, starting from the continuous-time likelihood function, should modify appropriately the estimating equations as suggested in Bibby and Sørensen (1995). Other alternative strategies are based on the Euler or Milstein scheme where, according to Bolder (2001), we can actually solve the stochastic differential equations for y_t and then discretize this solution. Following this latter scheme, once we do the discretization, we conform our notation to that usually employed in state space analysis (Koopman et al., 1999) to indicate the discretized measurement equation y_t and the transition equation α_t.

The general state space form is applied to a multivariate time series y_t where, for each time $t = 1, \ldots, T$, we consider N observations. N is the number of nodes of the intensity curve on which credit spreads are observed.

The standard Kalman filter provides a recursive algorithm for computing the minimum mean squared error estimator of α_t conditional on $y_1, \ldots, y_{t-1}$, i.e.

$$a_{t|t-1} = E(\alpha_t|y_1, \ldots, y_{t-1}) = T_t a_{t-1} + d_t,$$

and its mean squared error (MSE)

$$MSE(a_{t|t-1}) = E[(a_{t|t-1} - \alpha_t)(a_{t|t-1} - \alpha_t)'|y_1, \ldots, y_{t-1}] = P_{t|t-1}.$$

More specifically, the Kalman filter is the set of recursions

$$\begin{array}{ll} \upsilon_t = y_t - c_t - Z_t a_{t|t-1} & F_t = Z_t P_{t|t-1} Z_t' + G_t G_t' \\ q_t = q_{t-1} + \upsilon_t' F_t^{-1} \upsilon_t & K_t = (T_t P_{t|t-1} Z_t' + H_t G_t') F_t^{-1} \\ a_{t+1|t} = T_t a_{t|t-1} + d_t + K_t \upsilon_t & P_{t+1|t} = T_t P_{t|t-1} T_t' + H_t H_t' - K_t F_t K_t', \end{array} \tag{7}$$

with $q_0 = 0$. The filter innovations (one step ahead prediction errors) are indicated by υ_t and their variance by $F_t = \mathrm{var}(\upsilon_t) = \mathrm{var}\{y_t - E(y_t|y_1, \ldots, y_{t-1})\}$. These two quantities form the necessary ingredients for the computation of the likelihood

$$l(\hat{\Theta}) = -0.5NT \cdot \ln(2\pi) - 0.5 \sum_{t=1}^{T} \left(\ln |F_t| + \upsilon_t' F_t^{-1} \upsilon_t \right). \tag{8}$$

The multivariate state space form for Vasicek and CIR models with J factors has the following measurement equation

$$\underbrace{\begin{pmatrix} y_t(\tau_1) \\ \vdots \\ y_t(\tau_N) \end{pmatrix}}_{y_t} = \underbrace{\begin{pmatrix} \frac{B_1(\tau_1)}{\tau_1} & \cdots & \frac{B_J(\tau_1)}{\tau_1} \\ \vdots & \ddots & \vdots \\ \frac{B_1(\tau_N)}{\tau_N} & \cdots & \frac{B_J(\tau_N)}{\tau_N} \end{pmatrix}}_{Z_t} \underbrace{\begin{pmatrix} \alpha_{t,1} \\ \vdots \\ \alpha_{t,J} \end{pmatrix}}_{\alpha_t} - \underbrace{\begin{pmatrix} \frac{A(\tau_1)}{\tau_1} \\ \vdots \\ \frac{A(\tau_N)}{\tau_N} \end{pmatrix}}_{c_t} + \underbrace{\begin{pmatrix} \eta_{t,1} \\ \vdots \\ \eta_{t,N} \end{pmatrix}}_{G_t \epsilon_t}, \tag{9}$$

where $\tau_{(\cdot)}$ denotes the maturity (node of the term structure), and the functional forms for $A(\cdot)$ and $B(\cdot)$ can be formulated, for the Vasicek model, as

$$A(\tau) = \sum_{j=1}^{J} \left(\frac{\gamma_j (B_j(\tau) - \tau)}{k_j^2} - \frac{\sigma_j^2 B_j^2(\tau)}{4k_j} \right), \tag{10}$$

$$B_j(\tau) = \frac{1}{k_j}[1 - \mathrm{e}^{-k_j \tau}], \tag{11}$$

where

$$\gamma_j = k_j^2 \left(\theta_j - \frac{\sigma_j \nu_j}{k_j} \right) - \frac{\sigma_j^2}{2}. \tag{12}$$

On the other hand, for the CIR model we have

$$A(\tau) = \sum_{j=1}^{J} \ln \left(\frac{2\gamma_j \mathrm{e}^{\frac{(\gamma_j + k_j + \nu_j)\tau}{2}}}{(\gamma_j + k_j + \nu_j)(\mathrm{e}^{\gamma_j \tau} - 1) + 2\gamma_j} \right)^{\frac{2k_j \theta_j}{\sigma_j^2}}, \tag{13}$$

$$B_j(\tau) = \frac{2(\mathrm{e}^{\gamma_j \tau} - 1)}{(\gamma_j + k_j + \nu_j)(\mathrm{e}^{\gamma_j \tau} - 1) + 2\gamma_j}, \tag{14}$$

where

$$\gamma_j = \sqrt{(k_j + \nu_j)^2 + 2\sigma_j^2}. \tag{15}$$

For both models, $\sigma_j \geq 0$ is the volatility parameter of the process for the jth factor. The meaning of the other parameters k_j, θ_j and ν_j, referring to each of the J factors, has already been described below Eq. (6). A diagonal covariance structure for the errors in the measurement equation $G_t G_t'$ is assumed in order to ensure the identification of model parameters and reduce the complexity of Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization.

The transition equation is as follows

$$\underbrace{\begin{pmatrix}\alpha_{t+1,1}\\ \vdots \\ \alpha_{t+1,J}\end{pmatrix}}_{\alpha_{t+1}} = \underbrace{\begin{pmatrix} e^{-k_1\Delta t} & 0 & \cdots & 0 \\ 0 & e^{-k_2\Delta t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & e^{-k_J\Delta t}\end{pmatrix}}_{T_t} \underbrace{\begin{pmatrix}\alpha_{t,1}\\ \vdots \\ \alpha_{t,J}\end{pmatrix}}_{\alpha_t} + \underbrace{\begin{pmatrix}\theta_1(1-e^{-k_1\Delta t})\\ \vdots \\ \theta_J(1-e^{-k_J\Delta t})\end{pmatrix}}_{d_t} + \underbrace{\begin{pmatrix}\psi_{t,1}\\ \vdots \\ \psi_{t,J}\end{pmatrix}}_{H_t\epsilon_t}, \tag{16}$$

where Δ_t, for daily observations, is conventionally assumed to be $1/250$.

According to the Vasicek model specification, $H_t H_t'$ is assumed to be diagonal, with jth diagonal element

$$H_t H_t'(j) = \frac{\sigma_j^2}{2k_j}(1 - e^{-2k_j\Delta t}). \tag{17}$$

In the case of the CIR model this matrix depends on a state space process with jth diagonal element

$$H_t H_t'(j) = \frac{\theta_j\sigma_j^2}{2k_j}\left(1 - e^{-k_j\Delta t}\right)^2 + \frac{\sigma_j^2}{k_j}\left(e^{-k_j\Delta t} - e^{-2k_j\Delta t}\right)\alpha_{t-1,j}. \tag{18}$$

In the next section we show how to apply the forward search to the above described Kalman filter.

3. The forward search for the multivariate Kalman filter

The forward search is made up of the following three main steps: initialization, progression and monitoring. The first task is to find the appropriate starting subset of observations which is free from multiple masked outliers. In the time series context the initial subset can be chosen among q blocks of contiguous observations of a predefined dimension b. More precisely, to find the initial subset, we perform the search over all possible blocks and we choose the one that is considered most compact according to a certain distance. It is interesting to notice that the choice of the initial subset does not dramatically influence the search and, in particular, it does not affect the final steps where the most important findings of the analysis are concentrated. The second step is the way we progress in the search. At each step we rank units according to a specified distance continuing until all units are included in the subset. The third task is to monitor some suitable quantities along the search. In what follows we describe how these steps are performed.

1. Division of the dataset into q blocks. We split our time series in blocks and, in order to retain the dependence structure, we include the first observation as the first unit of each initial subset. As pointed out in Riani (2004), the choice of the number of blocks does not substantially effect the procedure. We define the number of units of each block according to the pragmatic rule $b \approx \sqrt{T}$.
2. Likelihood estimation. We estimate the vector $\hat{\Theta}_{S_m}$, where m indicates the subset size, computing the likelihood of Eq. (8). We perform the Kalman filter estimation by using the Kalman gain if units belong to the subset, while we do not use the Kalman gain for the other units (Harvey, 1989). For the initial subset we carry out the same estimation for each S_b obtained from the division in q blocks above mentioned.
3. Squared Mahalanobis distances. After estimating the parameter vector $\hat{\Theta}_{S_m}$, we compute the squared Mahalanobis distances for each unit t as follows
$$d_{t,S_m} = \upsilon_t' \Sigma_{S_m}^{-1} \upsilon_t, \tag{19}$$
where the matrix $\Sigma_{S_m}^{-1}$ is estimated as the inverse of
$$\frac{\Sigma_{t\in S_m}\upsilon_t\upsilon_t'}{m-1}. \tag{20}$$
We calculate distances for all T units, but in order to estimate υ_t, we apply the Kalman gain only to the units belonging to the subset.
4. Initial subset S_{b*}. For each of the q blocks, we compute the median squared Mahalanobis distance. We choose as initial b-dimensional subset, S_{b*}, the one with the lowest median. This is a generalization of the least median of squares criterion in regression (Rousseeuw, 1984). Without loss of generality, instead of taking the median squared Mahalanobis distance we could have taken a trimmed sum. This would have led to a generalization of the LTS estimator to time series (Rousseeuw and Van Driessen, 1999).

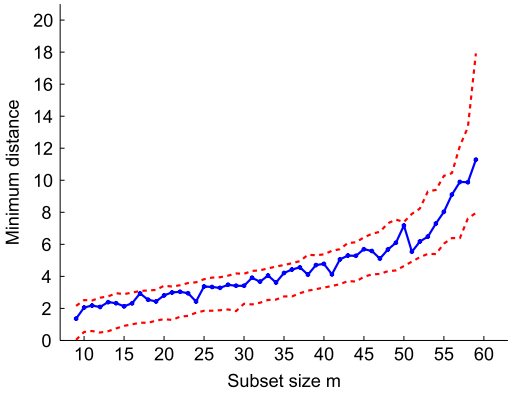

Fig. 1. Minimum distance outside the subset $d_{\min}(m)$ for the simulated credit spreads. The fact that $d_{\min}(m)$ increases almost monotonically as the size of the subset increases and reaches its maximum at the end of the search is an indication of the absence of masking. The dotted lines give a 95% confidence interval.

5. Subset S_{b+1}. We add to S_{b*} the unit with the lowest squared Mahalanobis distance obtained considering the parameter vector estimated on S_{b*}.
6. Progressing in the search, subset $S_{m>b+1}$. S_m is made up by the m units with the lowest squared Mahalanobis distances. In order to compute these distances, we exploit the parameter vector $\hat{\Theta}_{S_{m-1}}$ estimated considering the $m-1$ units belonging to S_{m-1}. Given that in progressing the search from $m-1$ to m more than one unit can join the subset, for each m we completely rerun the Kalman filter.

At each step m of the search we monitor the minimum distance of units not belonging to the subset $d_{\min}(m)$ and we exploit envelopes to make inference on outliers (Riani and Atkinson, 2007).

In the next section we apply our forward search framework to a simulated dataset explaining how to compute $d_{\min}(m)$ and the envelopes.

4. The forward search applied to simulated credit spreads

We apply the state space representation described in the previous sections to generate $N = 4$ paths of 60 units each, of the Vasicek one factor model with parameter vector $\Theta_{sim} = (0.03, 0.6, 0.01, -0.002, 0.0081, 0.0090, 0.0012, 0.0091)'$. We chose this set of parameters because they can be representative of real time series. Proceeding in this way we obtain four time series corresponding to daily credit spreads at maturities: 3, 5, 7 and 10 years. We stress that we consider so few time series because only on them are there actively traded interest rates. Starting from them, we are able to estimate model parameters which allow us to obtain, through Eq. (6), the intensity curve for all maturities τ.

Following the search as described in the previous section, we concentrate, first of all, on monitoring the distances d_{t,S_m} for each unit along the search. The goal pursued through this analysis is to highlight influential units (trajectories which change considerably during the search), masked multiple outliers or jumps (set of trajectories which have a very similar atypical behavior along the search, but at the end are completely mixed with the remaining observations) and isolated outliers (trajectories very different from the majority of the observations).

Starting from $d_{t,S_m} = \upsilon_t' \Sigma_{S_m}^{-1} \upsilon_t$, we focus on the minimum distance of units outside the subset

$$d_{\min}(m) = \min[d_{t,S_m}], \quad t \notin S_m. \tag{21}$$

As in standard regression (Atkinson and Riani, 2000), this distance increases progressively along the search and its slope is higher in the last few steps. The monitoring of $d_{\min}(m)$ helps to have an idea about potential outliers. We expect distances at the end of the search to be large, but this definition of *large* is subjective, without reference to any null distribution. For this reason, we extend to our setting the envelope procedure for outlier detection introduced by Riani et al. (2009) for i.i.d. observations. The idea underlying the generation of envelopes is to find theoretical boundaries for the quantities monitored during the forward search, that allow for simultaneous inference on outliers. We try to achieve this goal by generating a large number of paths and verifying whether observations stay within bounds obtained from such generations.

Using the parameter vector Θ_{sim}, we generate paths for each subset size m obtaining a distribution of $d_{\min}(m)$ from which we can extract percentiles. Lower and upper envelopes are the collection of these point-wise values which we use as a threshold for the detection of outliers.

Fig. 1 shows the monitoring of the minimum distance outside the subset $d_{\min}(m)$ (solid line) together with a 95% confidence interval (dotted lines). In this case the trajectory is well inside the envelopes along the search and indicates the absence of atypical observations. In the presence of an isolated outlier we expect to find a trajectory which lies inside the envelopes and goes out in the final step. On the other hand, in the presence of multiple outliers (as we shall see

Table 1
Vector of Θ Vasicek one factor model parameters to simulate envelopes.

	θ	k	σ	ν	$\sigma_{G_t G'_t,1}$	$\sigma_{G_t G'_t,2}$	$\sigma_{G_t G'_t,3}$	$\sigma_{G_t G'_t,4}$
Θ_{sim}	0.03	0.6	0.01	−0.002	0.0081	0.0090	0.0012	0.0091
Θ_2	0.10	0.5	0.3	−0.001	0.0705	0.0510	0.0014	0.0196
Θ_3	0.07	0.2	0.001	−0.2	0.0826	0.9587	0.0751	0.6502

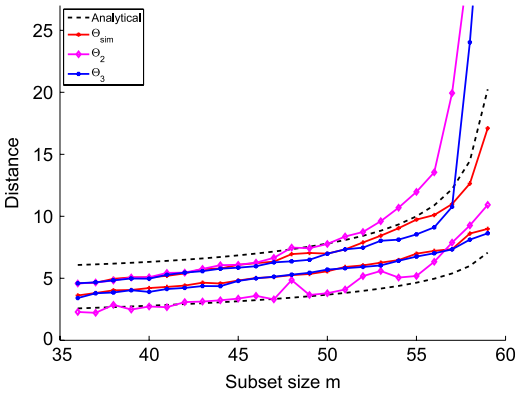

Fig. 2. Envelopes at confidence level 95% starting from different parameter vectors. Analytical envelopes (based on the order statistics approximation) share substantially the same shape as envelopes obtained through Monte Carlo simulations generated from different parameter sets.

in the next section with a real data example), we expect to observe a trajectory of the minimum distance which lies above the envelopes in the final part of the search but, due to the masking effect, goes back inside in the final step. If there is a misspecification in the model, or for example if the errors are non-normal, we expect that the trajectory of the minimum deletion residual is persistently above the upper threshold. Finally, a misspecification combined with outliers is likely to produce a trajectory which is persistently above the confidence levels and at the end is characterized by upward jumps.

It is interesting to remark that the shape of the envelopes obtained through the above mentioned procedure is virtually unaffected by the parameter vector Θ from which we simulate. In other words, starting from different parameter vectors, such as those given in Table 1, we obtain the same envelope structure as is shown in Fig. 2.

In addition, considering that the above mentioned Monte Carlo procedure is time consuming, we also tried to exploit the order statistic approximation proposed by Riani et al. (2009) for squared Mahalanobis distances in the presence of i.i.d observations. Fig. 2 shows that this technique in the case of time series data can be used just to have a rough approximation. For this reason, in what follows, we rely on envelopes obtained through Monte Carlo simulations.

In order to check whether our procedure is really capable to detect outliers, we contaminate units 25 to 27 in 1 and 2 year credit spread dataset. We perform the whole forward search on the contaminated dataset. The top left panel of Fig. 3 shows the monitoring of the trajectories associated with the distances d_{t,S_m}. This panel clearly shows that the trajectories associated with the contaminated units are very different from those of the other units. In order to establish whether these units can be considered atypical, we plot on the top right panel the curve of $d_{\min}(m)$ together with a confidence envelope based on all the observations. This panel shows that even if the final value of $d_{\min}(m)$ is inside the envelope at the final step its trajectory is well above the upper confidence envelope in the previous steps. In order to understand how many outliers are present in the data we need to resuperimpose the confidence bands using a reduced number of observations. The bottom left panel of Fig. 3 shows that if we consider an envelope based on $T - 3$ units, the trajectory of the minimum deletion residual always lies inside the confidence band. The bottom right panel of Fig. 3 shows that if we consider an envelope based on $T - 2$ units the curve starts going outside the envelope in the final step. In other words, when the first contaminated unit enters the subset, there is a huge increase in $d_{\min}(m)$. This leads us to declare three outliers.

It is interesting to notice that the most important findings of our procedure are concentrated on the last steps of the search. We carried out the analysis even without considering the block procedure described in previous sections and, apart from the very beginning of the search, we obtained the same results as above. In addition, we carried out the same analysis introducing contaminations at the beginning, at the end and in other positions of the time series and we obtained the same findings. We notice that the forward search applied to this simulated dataset takes roughly 40 s with a *Pentium V* computer and each simulation to construct the envelopes takes the same time.

In order to verify the effectiveness of our framework in a setting which is not obtained exploiting a simulation algorithm based on i.i.d. realizations, in the next section we apply the forward search to real market financial time series.

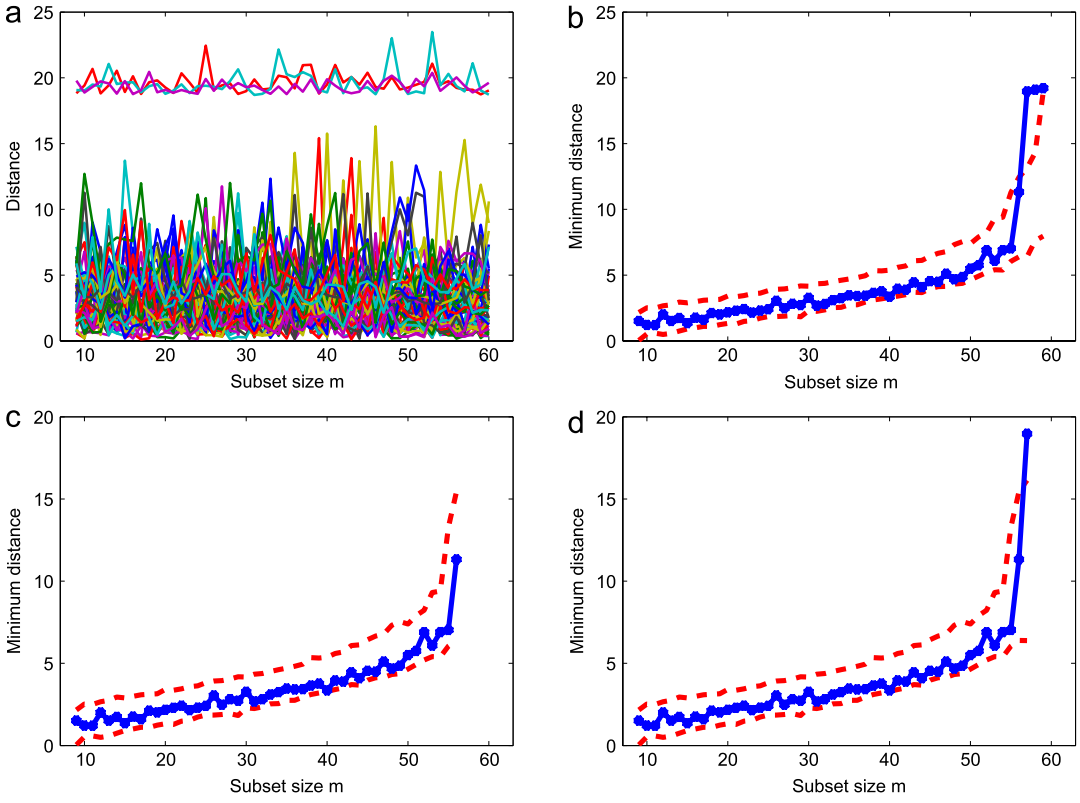

Fig. 3. Forward analysis of the contaminated data. (a) Monitoring of the distances d_{t,S_m}. There are three trajectories which seem to be very different from the others. (b) Monitoring of $d_{\min}(m)$ with an envelope based on T observations. (c) Monitoring of $d_{\min}(m)$ using an envelope based on $T-3$ observations. (d) Monitoring of $d_{\min}(m)$ using an envelope based on $T-2$ observations. When the first contaminated unit enters the subset, there is a huge increase in $d_{\min}(m)$.

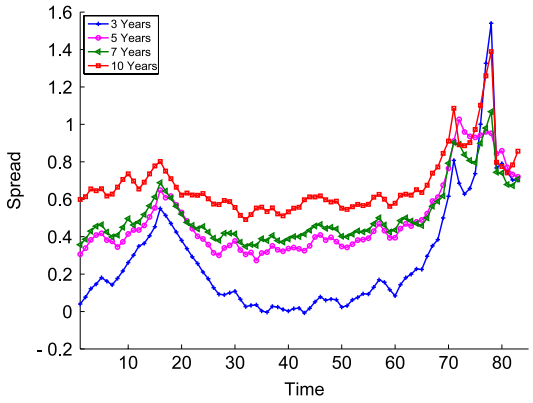

Fig. 4. Spanish daily credit spreads from January 2010 to April 2010. Spreads are computed as difference between government bond interest rates for maturities 3, 5, 7, 10 years and swap rates.

5. The forward search applied to Spanish credit spreads

In this section we consider the time series of Spanish daily credit spreads from January 2010 to April 2010 as plotted in Fig. 4. These spreads are obtained as difference between government bond interest rates for maturities 3, 5, 7, 10 years and swap rates.

Fig. 4 shows that at the end of the time series there are some upward and downward peaks which are probably due to rumors about the Spanish financial situation after the Greek crisis. Given that outlier detection improves forecasting, it is interesting to verify whether some units can be considered as atypical and make appropriate inferences.

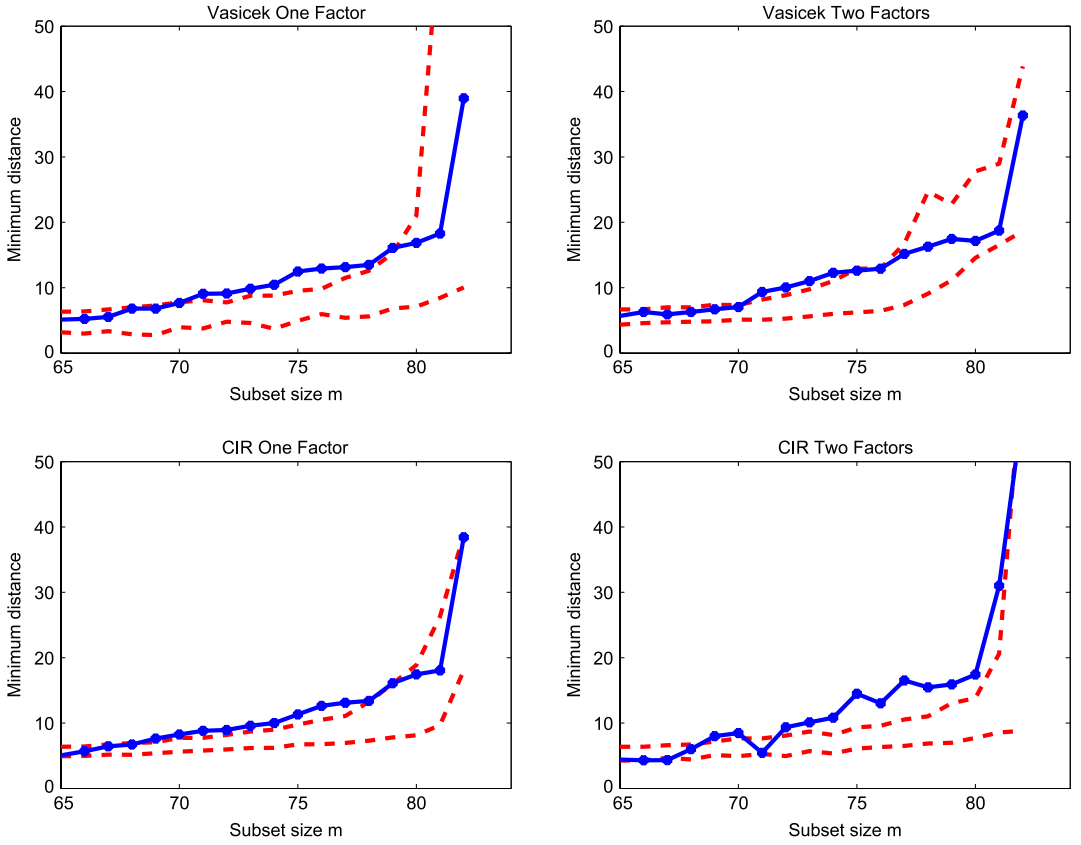

Fig. 5. Monitoring of $d_{\min}(m)$ for Spanish credit spreads for Vasicek and CIR, one and two factor models, with 95% confidence bands. The observed curve of minimum squared Mahalanobis distance lies outside the upper 95% confidence level in the last part of the search but goes back inside the envelopes in the final steps showing a *masking* phenomenon.

Fig. 5 shows the monitoring of the minimum squared Mahalanobis distance for Vasicek and CIR, one or two factor models. This figure shows that the trajectory of $d_{\min}(m)$ crosses the upper envelope in a persistent way in the last part of the search, but in the final steps $d_{\min}(m)$ stays within the envelopes. This phenomenon is known as *masking*. One of the main advantages of the forward search is the effectiveness in detecting *masked* units. In order to further verify whether units which cross envelopes can be considered outliers, we superimpose envelopes as we have described in the previous Section, confirming what we have stated above.

Starting from Eq. (6) we can draw credit intensity curves by considering the parameter vector $\hat{\Theta}_{S_{T-p^*}}$ where p^* is the number of units declared as outliers. Furthermore, we can exploit forward search parameter estimates up to step $T - p^*$ to build up lower and upper frontiers for the intensity credit curve. At each step of the search we have a vector of parameter estimates. Then, we can use these estimates to obtain credit curves for all maturities τ (we concentrate on maturities: $\{1, \ldots, 10\}$). We repeat this procedure for all steps of the search obtaining a set of curves. For each annual node of the curve we extract lower and upper values obtaining lower and upper boundaries as the collection of these point-wise values as shown in Fig. 6 (Vasicek model), and in Fig. 7 (CIR model). At the same time, as shown through box plots, parameters estimated along the search allow to show the dispersion of intensity at each maturity τ.

In both Figs. 6 and 7 the shape of the term structure is aligned to the relationship among credit spreads which, apart from *masked* atypical observations, is increasing as function of maturity τ.

As correctly pointed out by a referee, the suggested approach can be applied *on line* in the sense that when a new observation arises, the new trajectory of the minimum deletion residual can be recalculated taking advantage in an efficient way of the envelopes previously calculated. An observation that can be flagged permanently as an outlier after this has become evident at a certain significance level.

6. Concluding remarks

In this paper we combine different lines of research in the context of term structure of default intensity. For the first time in the literature, relying on squared Mahalanobis distance as a measure to rank observations, we extend the forward search to multivariate time series based on the Kalman filter analysis. We apply our general framework to the special issue of estimating default intensity exploiting Vasicek and CIR one and two factor models.

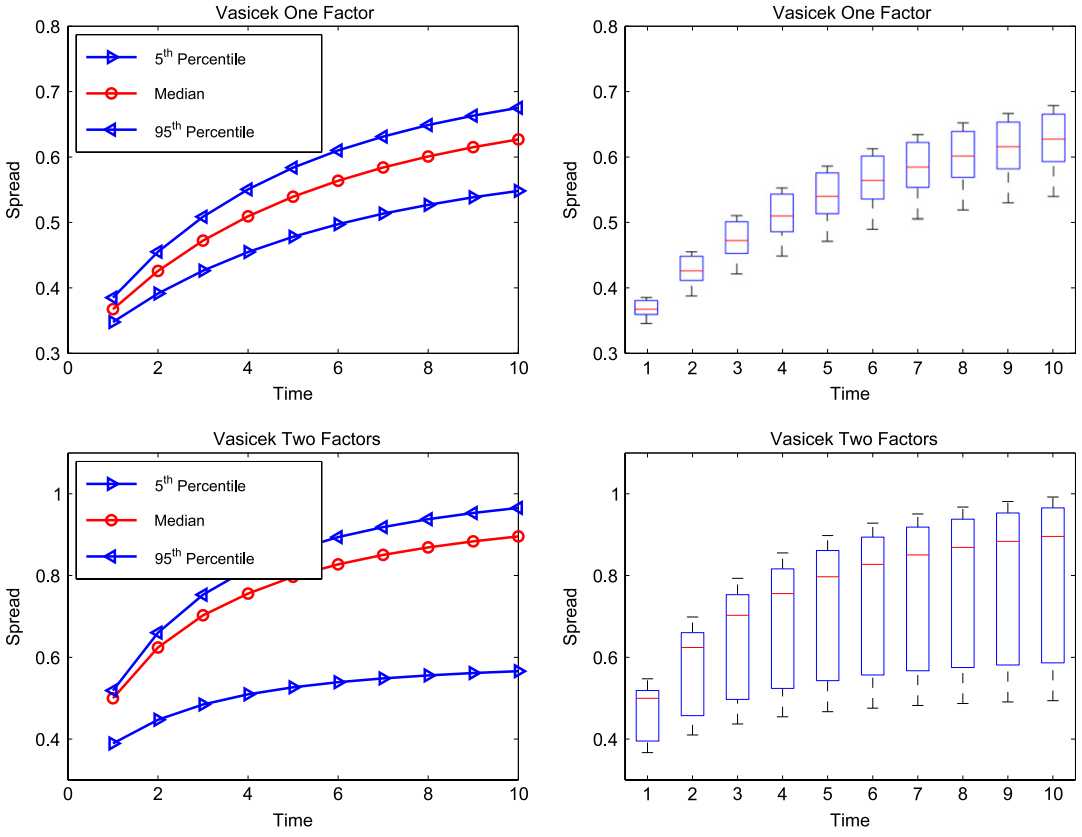

Fig. 6. Spanish term structure of default intensity: Vasicek one and two factor models. In the top left panel Vasicek 1 lower, median and upper percentiles (of the default intensity curve obtained considering the parameter vectors estimated along the search) are compared. In the top right panel box plots for maturities $\tau = \{1, \ldots, 10\}$ are presented. The bottom plots show the same analysis for the Vasicek two factor model.

In order to find distributional boundaries for inferences on outliers we introduce a procedure for computing envelopes. This procedure is based on random simulations of the state space model. We show that these envelopes are virtually independent of the simulation parameter set and we compare them to the analytical envelopes obtained through the order statistics procedure of Riani et al. (2009).

We apply the forward search to both simulated and real market datasets. We concentrate on the simulated environment in order to analyze the main features of our framework. We start from an uncontaminated setting, then we introduce contaminations to verify whether our approach is effective in detecting atypical observations. Then, we apply the forward search to Spanish credit spread time series, concentrating on the period between January 2010 and April 2010 (after the Greek financial crisis). Envelope analysis allows us to detect atypical units highlighting influential masked observations.

Considering parameter estimates at each step of the search, we build up not only a unique default intensity term structure, as is usual in the literature, but also an interval within which default intensity is likely to stay.

This work is the first step in multivariate time series forward search analysis. Further research needs to be devoted to deal with more sophisticated models and a greater effort will be required to obtain *universal* techniques for constructing envelopes without resorting to Monte Carlo simulations.

Notes and acknowledgments

The code has been written in Matlab and can be downloaded together with the Spanish dataset from the web pages of the authors http://www.riani.it or http://economia.unipr.it/docenti/belliniT. Furthermore, from the web address http://www.riani.it/matlab it is possible to download the FSDA toolbox, a collection of routines which implement the forward search both in regression and multivariate analysis in an exploratory data analysis content or in a inferential way for automatic outlier detection.

Last but not least, we are grateful to the Associate Editor, to two anonymous referees and to Professor Anthony C. Atkinson for very constructive suggestions on earlier drafts.

Our work was supported by the grant "Nuove metodologie robuste per l'analisi di dati complessi of the Ministero dell'Università e della Ricerca - Progetti di Ricerca di Interesse Nazionale 2008".

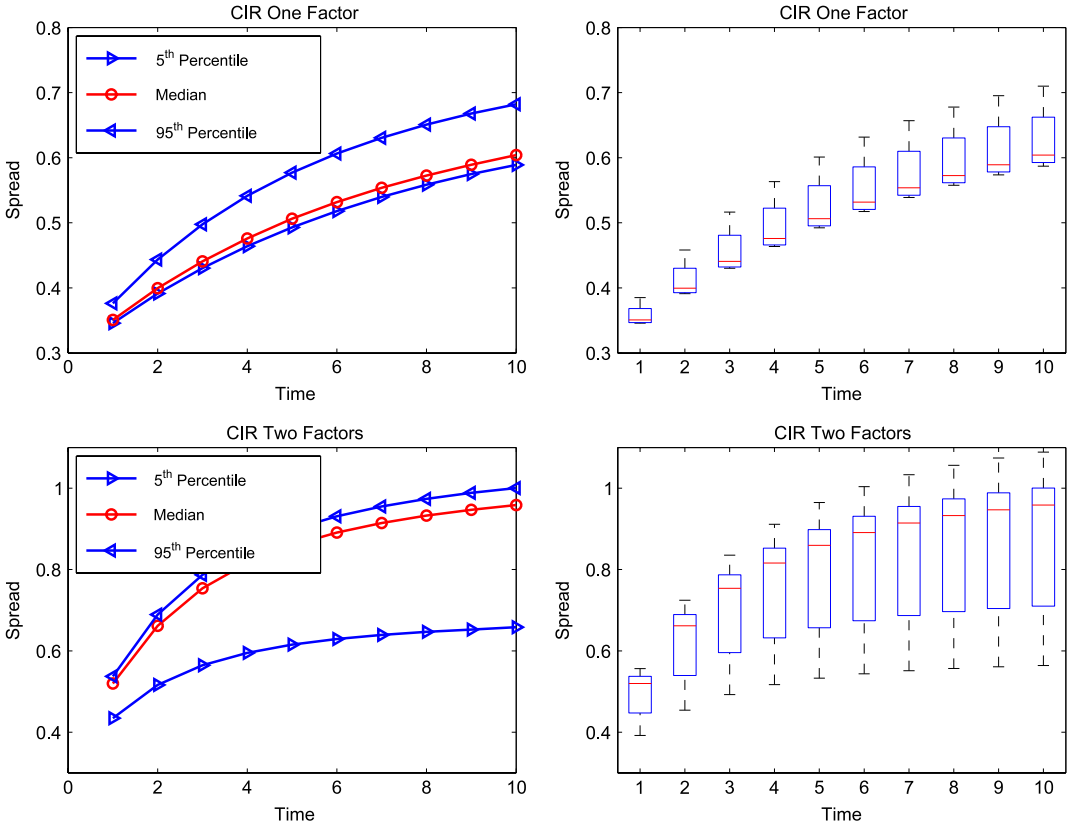

Fig. 7. Spanish term structure of default intensity: CIR one and two factor models. In the top left panel CIR one lower, median and upper percentiles (of the default intensity curve obtained considering the parameter vectors estimated along the search) are compared. In the top right panel box plots for maturities $\tau = \{1, \ldots, 10\}$ are presented. The bottom plots show the same analysis for the CIR two factor model.

Appendix. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2011.03.007.

References

Amendola, A., Francq, C., Koopman, S., 2006. Nonlinear Modelling and Financial Econometrics. Computational Statistics and Data Analysis 51, 2115–2117. (special issue).
Amendola, A., Belsley, D., Kontoghiorghes, E., van Dijk, H., Omori, Y., Zivot, E., 2008. Statistical and Computational Methods in Finance. Computational Statistics and Data Analysis 52, 2842–2845. (special issue).
Atkinson, A.C., Riani, M., 2000. Robust Diagnostic Regression Analysis. Springer - Verlag, New York.
Bibby, B., Sørensen, M., 1995. Martingale estimation functions for discretely observed diffusion processes. Bernoulli 1, 17–39.
Bolder, D., 2001. Affine term structure models: Theory and implementation. Working Paper, Bank of Canada.http://www.bankofcanada.ca/en/res/wp/2001/wp01-15a.pdf.
Cox, J., Ingersoll, J., Ross, S., 1985. A theory of the term structure of interest rates. Econometrica 53, 385–407.
Duan, J., Simonato, J., 1999. Estimating and testing exponential-affine term structure models by Kalman filter. Review of Quantitative Finance and Accounting 13, 111–135.
Duffie, D., Singleton, K.J., 1999. Modeling term structures of defaultable bonds. Review of Financial Studies 12, 687–720.
Eckner, A., 2009. Computational techniques for basic affine models of portfolio credit risk. Journal of Computational Finance 100, 1–35.
Fuertes, A., Kalotychou, E., 2006. Early warning systems for sovereign debt crises: The role of heterogeneity. Computational Statistics and Data Analysis 51, 1420–1441.
Geyer, A., Kossmeier, S., Pichler, S., 2001. Empirical analysis of European government yield spreads. Working Paper, University of Vienna.
Harvey, A.C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.
Koopman, S.J., Shephard, N., Doornik, J., 1999. Statistical algorithms for models in state space form using ssfpack 2.2. Econometrics Journal 2, 113–166.
Merton, R., 1974. On the pricing of corporate debt: the risk structure of interest rates. Journal of Finance 29 (2), 449–470.
Riani, M., 2004. Extension of the forward search to time series. Studies in Nonlinear Dynamics and Econometrics 8 (2), Article 2.
Riani, M., Atkinson, A.C., 2007. Fast calibrations of the forward search for testing multiple outliers in regression. Advances in Data Analysis and Classification 1, 123–141.
Riani, M., Atkinson, A.C., Cerioli, A., 2009. Finding an unknown number of multivariate outliers. Journal of the Royal Statistical Society, Series B 71, 201–221.
Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American Statistical Society 79, 871–880.
Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.
Vasicek, O., 1977. An equilibrium characterization of the term structure. Journal of Financial Economics 5, 177–188.
Yoshida, N., 1990. Estimation for diffusion processes from discrete observations. Journal of Multivariate Analysis 41, 220–242.