Bayesian Robustness
IMS Lecture Notes - Monograph Series (1996) Volume 29

NONPARAMETRIC SPECICATION OF ERROR TERMS IN DYNAMIC MODELS [1]

BY FABIO CORRADI AND FABRIZIA MEALLI
Università di Firenze

In this paper the first order polynomial dynamic model is considered introducing a nonparametric specification of the error terms, using mixtures of Dirichlet processes. In order to make inference about the relevant parameters of the model the Gibbs Sampling approach is used. The approach is suitable to cope with features like outliers and changes in level, both for prediction and detection purposes, showing some characteristics of robustness due to the memory of the processes. An example is shown using artificial data.

1. Introduction. From their original formulation, the state space models for nonstationary time series have been widely used and largely improved.

In the Bayesian approach, advances trying to remove the normality assumption for the error terms can be recognized in West et al. (1985) and, more recently, Carlin et al. (1992) introduced the Gibbs sampling approach to the estimation of a multivariate nonnormal, nonlinear state space model, covering a wide variety of possible distributions. Besides these attempts and in order to robustify the estimation of the state parameters, Meinhold and Singpurwalla (1989) specified the distributions of the error terms and the state parameters as (multivariate) Student-t, being able to cope with outlying observations. In order to avoid strict assumptions on the error terms, a possible solution can be found in the nonparametric approach; in the Bayesian framework, Ferguson (1973) introduced the Dirichlet process (DP), later leading also to a solution for the nonparametric density estimation problem via Mixtures of Dirichlet processes (MDP) (Antoniak, 1974; Ferguson, 1983). Since substantial computational difficulties arise also for rather small amount of data, a Gibbs sampling solution was proposed by Escobar (1988) and Escobar and West (1995).

In this paper we merge the two approaches, producing a simple univariate first-order polynomial model with error terms specified in a nonparametric way.

As an important special case of nonnormal error terms, we focus our attention on modelling discrepant observations. We will show how the uncertainty about the distribution of the error terms allows to sensibly process such observations; these can represent different features of the data, which must be differently treated in the estimation of the state and of the other parameters of the model.

[1] This work was supported by MURST 60% funds. We would like to thank Mike West for useful suggestions. Responsability for errors remains ours.
Key Words:: Dynamic Models, Nonparametric Approach, Dirichlet Process, Outliers, Changes in Level.

In a retrospective assessment of the series it seems important to produce a robust analysis, while in forecasting all the detected sources of uncertainty must be taken into account. The approach proposed considers the possibility that an observation could belong to the prior predicted distribution or that it might be regarded as an outlier or, finally, as representative of a change in the level of the state. The nonparametric density estimation approach naturally assigns a probability to each of these features producing multimodal densities in the error terms distribution to be employed for forecasting; it will be shown that it also provides a robust estimation of the state and a tool to detect outliers and changes in level.

In a parametric setting, the idea to have different components in the distribution of the error terms can be recognized in multi-process models class II as described in West and Harrison (1989), even if, in our approach, the number, the type and the probability of the components must not be specified in advance.

2. The Model. Consider the first order polynomial dynamic linear model:

$$\begin{aligned} x_t &= x_{t-1} + w_t \\ y_t &= x_t + e_t, \end{aligned}$$

where, for $t = 1, \ldots, n$, x_t is the state parameter, y_t the observation, while w_t and e_t represent the state and the observation error terms respectively. The errors are assumed serially and mutually independent and independent of the initial state $x_0 \sim N(m_0, A_0)$.

Assuming normality for the error terms, $\forall t$ we have

$$\begin{aligned} (e_t|\mu_t^e, V_t^e) &\sim N(\mu_t^e, V_t^e), \quad \Pi_t^e = (\mu_t^e, V_t^e) \\ (w_t|\mu_t^w, v_t^w) &\sim N(\mu_t^w, V_t^w), \quad \Pi_t^w = (\mu_t^w, V_t^w). \end{aligned}$$

If we are unsure about the distribution of the parameters $\Pi's$ we can assume $\{\Pi_1^e, \ldots, \Pi_n^e\}$ and $\{\Pi_1^w, \ldots, \Pi_n^w\}$ as a sample coming from uncertain prior distributions G^e and G^w modelled as two bivariate independent Dirichlet processes, so that the predictive distributions of w_t and e_t result to be Dirichlet mixtures of normals. So we assume:

$$\begin{aligned} G^e(\Pi^e|m^e, B^e, S^e) &\sim \mathcal{D}(\alpha^e, G_0^e) \\ G^w(\Pi^w|m^w, B^w, S^w) &\sim \mathcal{D}(\alpha^w, G_0^w), \end{aligned}$$

being G_0 the expected distribution function of G, defined as

$$\begin{aligned} G_0^e &= N(\mu^e|m^e, B^e) IG(V^e|\frac{s^e}{2}, \frac{s^e S^e}{2}) \\ G_0^w &= N(\mu^w|m^w, B^w) IG(V^w|\frac{s^w}{2}, \frac{s^w S^w}{2}), \end{aligned}$$

and α being a positive scalar representing the concentration of the prior about its expectation. Specifying G_0 in a non-conjugate form overcomes some well known undesirable features of the conjugate one, allowing to separate the variance of the mean from the variance of the observations. As noted in West et al. (1994) this seems important when some influential observation might be present in data, which is crucial for our purposes.

The main results about the DP, for continous G_0, can be summarized as follows: assuming the Π_i known, the posterior for G is still a DP, i.e. $G(\Pi|\Pi_1,\ldots,\Pi_n) \sim \mathcal{D}(\alpha+n, G_n)$, where

$$(1) \qquad G_n(\Pi|\Pi_1,\ldots,\Pi_n) = \alpha a_n G_0(\Pi) + a_n \sum_{i=1}^{n} \delta_{\Pi_i}(\Pi),$$

$\delta_{\Pi_i}(\Pi)$ is the unit point mass at Π_i and $a_n = 1/(\alpha+n)$.

If n is large compared to α, the next value of Π is very likely coincident with one of the other n values of Π. Defining a k-configuration as a correspondence between a set $\{\Pi_i\}$, $i=1,\ldots,n$, and a set of distinct Π's $\{\Pi_j^*\}$, $j=1,\ldots,k$, $k \leq n$, being n_j the number of $\Pi_i = \Pi_j^*$, then, conditionally on a specified k-configuration, $G_n(\Pi)$ can be expressed by

$$(2) \qquad G_n(\Pi|\Pi_1,\ldots,\Pi_n) = \alpha a_n G_0(\Pi) + a_n \sum_{j=1}^{k} n_j \delta_{\Pi_j^*}(\Pi).$$

Further, defining $\Pi_{(i)} = (\Pi_1\ldots,\Pi_{i-1},\Pi_{i+1},\ldots,\Pi_n)$, the distribution of any Π_i conditional on $\Pi_{(i)}$ is

$$(3) \qquad p(\Pi_i|\Pi_{(i)}) = \alpha a_{n-1} G_0(\Pi_i) + a_{n-1} \sum_{j=1, i\neq j}^{n} \delta_{\Pi_j}(\Pi_i).$$

Of course, the values of the unobservable Π's are unknown but their distribution can be simulated using Gibbs sampling (Escobar and West, 1995). Since the $\Pi_{(i)}$ are unknown, as well as Π_i, equation (3) is just the full conditional prior of Π_i. According to an observational model $(y_i|\Pi_i)$ and a prior $G(\Pi) \sim \mathcal{D}(\alpha G_0)$, the required conditional posterior is:

$$(4) \quad p(\Pi_i|y_i,\Pi_{(i)}) = c^{-1} l(y_i|\Pi_i)(\alpha a_{n-1} G_0(\Pi_i) + a_{n-1} \sum_{j=1, i\neq j}^{n} \delta_{\Pi_j^*}(\Pi_i))$$

where $l(y_i|\Pi_i)$ is the likelihood for Π_i.

Now, consider again our model. All the introduced hyperparameters can be either directly specified or a learning procedure can be established. Having the Gibbs sampling solution in mind, so favoring the conditional conjugacy between the prior and the relevant likelihood, the hyperparameters

can be conveniently modelled as follows: $B^e \sim IG(\frac{t_0^e}{2}, \frac{R_0^e}{2})$, $m^e \sim N(m_0^e, A_0^e)$, $S^e \sim G(\frac{a_0^e}{2}, \frac{b_0^e}{2})$, $B^w \sim IG(\frac{t_0^w}{2}, \frac{R_0^w}{2})$, $m^w \sim N(m_0^w, A_0^w)$, $S^w \sim G(\frac{a_0^w}{2}, \frac{b_0^w}{2})$.

The model implies that the distribution of the state parameters is:

$$(x_t|x_{t-1}, \Pi_t^w) \sim N(x_{t-1} + \mu_t^w, V_t^w), \quad \forall t.$$

3. Computation. Finding the exact posterior of G^e and G^w, and then the predictive distribution for e and w, implies the consideration of all the possible configurations of the $\Pi^{*e} = \{\Pi_1^{*e}, \ldots, \Pi_{k^e}^{*e}\}$ and the $\Pi^{*w} = \{\Pi_1^{*w}, \ldots, \Pi_{k^w}^{*w}\}$ for each value of k^e and k^w and all the possible arrangements of the n Π's into k distinct Π^*'s.

The implementation of the Gibbs sampling scheme requires the definition of the full conditional posterior for all the parameters involved in the model, a set of starting values for x_t, for Π^e and Π^w, $\forall t$, a value for all the remaining parameters and, finally, a set of values for the hyperpriors $m_0, A_0, t_0, R_0, a_0, b_0$ for e and w.

Defining, $\forall t$, $E_t = y_t - x_t$ and $W_t = x_t - x_{t-1}$, we have to run M cycles of the scheme. Follow these steps:

1. Start the cycle sampling the Π^e's parameters: from (4), the required full conditional posterior is:

$$\begin{aligned} p(\Pi_t^e|\Pi_{(t)}^e, y_t, x_t) &\propto N(E_t|\mu_t^e, V_t^e)(\alpha a_{n-1} G_0(\Pi^e) + a_{n-1} \sum_{j=1}^{k^e} n_j^e \delta_{\Pi_j^{*e}}(\Pi_t^e)) \\ &= q_{0,t} G(\Pi_t^e) + \sum_{j=1}^{k^e} q_{j,t} \delta_{\Pi_j^{*e}}(\Pi_t^e), \end{aligned} \tag{5}$$

where n_j^e is reduced by 1 if Π_t^e belongs to the j-th component,

$$\begin{aligned} q_{0,t} &\propto \alpha a_{n-1} \int\int N(E_t|\mu^e, V^e) N(\mu^e|m^e, B^e) IG(V^e|\frac{s^e}{2}, \frac{s^e S^e}{2}) d\mu^e dV^e \\ &\propto \alpha a_{n-1} \int N(E_t|m^e, V^e + B^e) IG(V^e|\frac{s^e}{2}, \frac{s^e S^e}{2}) dV^e, \end{aligned} \tag{6}$$

$$q_{j,t} \propto n_j a_{n-1} N(E_t|\mu_j^{*e}, V_j^{*e}) \tag{7}$$

and

$$\begin{aligned} G(\Pi_t^e) &\propto N(E_t|\mu^e, V^e) G_0(\Pi^e) \\ &\propto N(E_t|\mu^e, V^e) N(\mu^e|m^e, B^e) IG(V^e|\frac{s^e}{2}, \frac{s^e S^e}{2}), \end{aligned}$$

so that,

$$(8) \qquad (\mu_t^e|E_t, V_t^e) \sim N(E_t - \frac{V_t^e}{B^e + V_t^e}(E_t - m^e), \frac{V_t^e B^e}{B^e + V_t^e}),$$

$$(9) \qquad (V_t^e|E_t, \mu_t^e) \sim IG(\frac{s^e + 1}{2}, \frac{s^e S^e + (E_t - \mu_t^e)^2}{2}).$$

Some comments are in order.

The non-conjugacy of G_0 produces some problems in the computation of $q_{0,t}$ and in sampling from $G(\Pi_t^e)$. Following West et al. (1994), the integral in (6) can be approximated by Monte Carlo integration, averaging over draws from the base prior G_0. In order to sample from the distribution $G(\Pi_t^e)$ we use (8) and (9). Since in (8) the value of V_t^e is unknown, we first sample μ_t^e conditional on a starting value for V_t^e and then V_t^e conditional on μ_t^e. Iterating these two draws establishes a Markov Chain, leading to an approximate draw from (μ_t^e, V_t^e).

In the Gibbs sampling scheme, $q_{0,t}$ represents the probability to sample from the $G(\Pi_t^e)$ and the $q_{j,t}$'s the probability that the actual Π_t^{*e} is coincident to the jth Π_j^{*e}, given the configuration. Drawing a variate from a multinomial distribution of specified parameters $q_{0,t}$ and $q_{j,t}$'s implies that if the variate is drawn from the state 0 a new Π_t^e is sampled using (8) and (9), otherwise $\Pi_t^e = \Pi_j^{*e}$, corresponding to the sampled jth state. Consider the configuration reached at each Gibbs run: this typically includes components representing errors with mean near zero plus, possibly, some others identifying outliers and changes in level of different size. The probability to allocate an error to each of these components, or to establish a new cluster, depends not only on their likelihoods, as in usual testing procedures, but also on the number of elements in each cluster and, more specifically, on their relative frequency with respect to the total number of observations. The result is that, for instance, a sufficiently large group of outliers even of small size can be globally detected because of their number, while a relative small group of errors of larger size might be supposed to belong to the zero mean component, taking into account that, with a large amount of observations, few of them could come from the tails.

2. After a complete sample from $p(\Pi_t^e|\Pi_{(t)}^e, y_t, x_t)\ \forall t$, the full conditional posteriors can also be provided for the hyperparameters m^e, B^e, S^e. Applying Bayes theorem we have:

$$(10) \qquad p(m^e|\mu^{*e}, B^e) \propto \prod_{j=1}^{k^e} N(\mu_j^{*e}|m^e, B^e) N(m^e|m_0^e, A^e)$$

$$= N\left(\frac{\sum_{j=1}^{k^e}\mu_j^{*e}}{k^e}\frac{A^e}{A^e+\frac{B^e}{k^e}} + m_0^e\frac{\frac{B^e}{k^e}}{A^e+\frac{B^e}{k^e}}, \frac{A^eB^e}{k^eA^e+B^e}\right)$$

$$p(B^e|\mu^{*e}, m^e) \propto \prod_{j=1}^{k^e} N(\mu_j^{*e}|m^e, B^e) IG(B^{*e}|\frac{t_0^e}{2}, \frac{R_0^e}{2})$$

$$= IG\left(\frac{t_0^e+k^e}{2}, \frac{R_0^e+\sum_{j=1}^{k^e}(\mu_j^{*e}-m^e)^2}{2}\right) \tag{11}$$

$$p(S^e|V^{*e}) \propto \prod_{j=1}^{k^e} IG(V_j^{*e}|\frac{s^e}{2}, \frac{s^eS^e}{2}) G(S^e|\frac{a_0^e}{2}, \frac{b_0^e}{2})$$

$$= G\left(\frac{a_0^e+k^es^e}{2}, \frac{b_0^e+s^e\sum_{j=1}^{k^e}\frac{1}{V_j^{*e}}}{2}\right). \tag{12}$$

3. Now the contribution of each Gibbs run can be incorporated in the reconstruction of the predictive density for the observational errors. Since:

$$\begin{aligned}(e|\mathbf{y}, \mathbf{x}, \Pi^e, m^e, B^e, S^e) &\sim \alpha a_n N(m_0^e, V^e+B^e) \\ &+ a_n \sum_{j=1}^{k_{(m)}^e} n_{j(m)}^e N(\mu_{j(m)}^{*e}, V_{j(m)}^{*e})\end{aligned} \tag{13}$$

the total reconstruction is obtained averaging (13) over all the performed M runs. Note that, at each run, a different configuration may arise by simulating the relevant conditionals in the solution of the posterior $p(\Pi^e|D)$, $D = \{y_1, \ldots, y_n\}$.

4. The same approach followed in steps 2 and 3 can be applied to solve the problem of the reconstruction of the predictive density for w, simply replacing E_t with W_t and, of course, all the relevant distributions.

5. Consider the state parameters' full conditional posterior distribution

$$p(x_t|x_{(t)}\mathbf{y}, \Pi^e, \Pi^w) \propto l(y_t|x_t, \Pi_t^e) p(x_t|x_{t-1}, \Pi_t^w) p(x_{t+1}|x_t, \Pi_{t+1}^w)$$

where $(y_t|x_t, \Pi_t^e) \sim N(x_t+\mu_t^e, V_t^e)$, $(x_t|x_{t-1}, \Pi_t^w) \sim N(x_{t-1}+\mu_t^w, V_t^w)$, $(x_{t+1}|x_t, \Pi_{t+1}^w) \sim N(x_t+\mu_{t+1}^w, V_{t+1}^w)$, so that the full conditional, for any $1 < t < n$, is

$$(x_t|x_{(t)}\mathbf{y}, \Pi^e, \Pi^w) \sim N\left(\frac{\frac{y_t-\mu_t^e}{V_t^e}+\frac{x_{t-1}+\mu_t^w}{V_t^w}+\frac{x_{t+1}-\mu_{t+1}^w}{V_{t+1}^w}}{\frac{1}{V_t^e}+\frac{1}{V_t^w}+\frac{1}{V_{t+1}^w}}, \frac{1}{\frac{1}{V_t^e}+\frac{1}{V_t^w}+\frac{1}{V_{t+1}^w}}\right),$$

with some minor differences for the two endpoints $t = 1$ and $t = n$.

4. Data analysis example. An artificial data set is considered to provide some insights of the model capabilities. A time series of length $n = 100$ was generated where the state level was drawn from a $N(40, 1)$ and the state and the observational errors were randomly sampled from two mixtures of normals as follows:

$$\begin{aligned} e &\sim 0.84N(0,2) + 0.06N(-15,2) + 0.06N(12,2) \\ w &\sim 0.84N(0,1) + 0.06N(8,1). + 0.06N(-10,1) \end{aligned}$$

To obtain the posterior distribution for all the parameters involved in the model a starting value for each parameter and the specification of the hyperpriors are required. Since in this example we consider an artificial data set, there is not a genuine prior information available for determining the hyperpriors, so that vague prior distributions were provided to the procedure. For the B's the variances of the mean of G_0^e and G_0^w, we specified their expected values such that outliers and changes in level in the range [-30, +30] can be sampled with a reasonable support ($t_0^e = t_0^w = 2$, $R_0^e = R_0^w = 200$). For the S', the parameters controlling the prior means of the variance of each components, the hyperparameters were chosen setting a rather vague prior with expected value equal to the variance of the generated errors ($a_0^e = a_0^w = 1$, $b_0^e = 0.5$, $b_0^w = 1$).

Suggestions about the values of the $\alpha's$ can be obtained from the prior distribution of k, derived by Antoniak (1974), taking as prior information that a range of 2-5 components are expected. For $2 \leq k \leq 5$, the computed values give support to prior values of α in the range $0.3 - 0.6$, so we set $\alpha^e = \alpha^w = 0.5$. Further, to start the iteration scheme, the x_t were initialized at 40; $\forall t$, the initial level of the series; all the $\mu's$ were initialized at 0, and all the V's had starting values sampled from their $IG(\frac{s}{2}, \frac{sS}{2})$ distributions, taking $s = 1$ so that the prior is given a weight of one observation.

In Figure 1 the series is shown, superimposing the levels at which the series was generated and the posterior mean of the state.

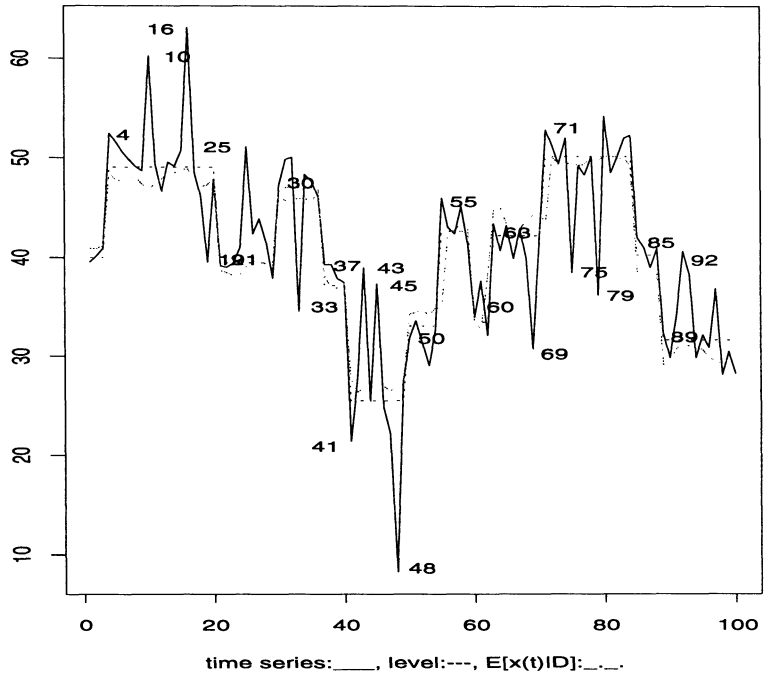

FIGURE 1: *Time series, non-stochastic level and E[x(t)|D)]*

We can observe how the estimated mean of the state closely follows the true level. The procedure discriminates between the outliers, which are identified as can be seen from their low influence on the state, and the changes in level, which are properly taken into account as the quick changes in level show.

In Figure 2 the simulated error mixture densities are superimposed to the original ones, giving a rather clear idea of the number and the size of the main features of the errors. A very moderate shrinking is present, showing how the filtering mechanism works when the errors are represented by a mixture of different distributions, each having a smaller variance with respect to an unique distribution with heavier tails, as suggested in usual robust analysis. Such overall reconstructions of the densities are important for forecasting purposes, as they allow to include the detected characteristics of the series in the forecasts; nevertheless, since only artificial data are considered, this analysis was not performed and will be undertaken in a forthcoming paper.

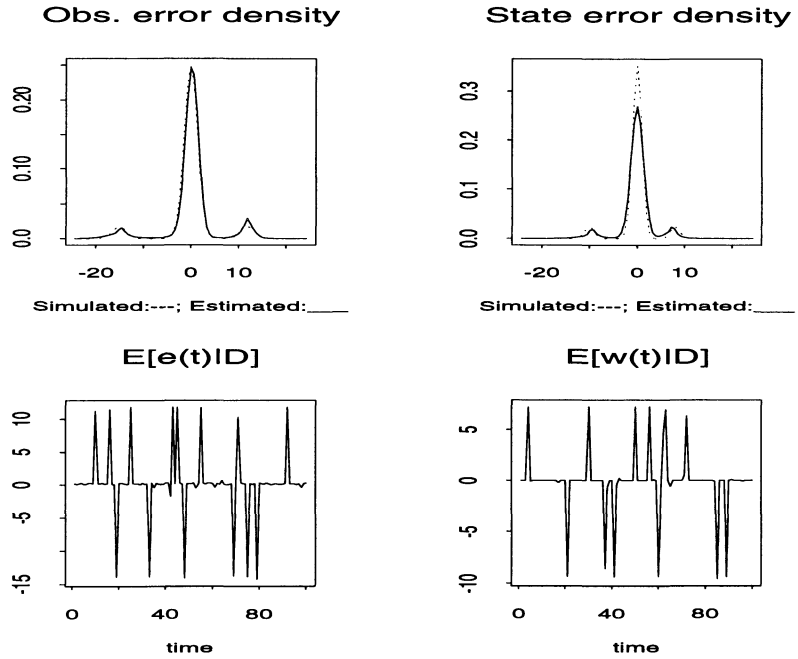

FIGURE 2: *Reconstruction of the error densities*

In the same picture the Monte Carlo posterior means of E_t and W_t give a summary of the error detection procedure: note the correspondence between the outliers and changes in level highlighted in figure 1 and the non-zero posterior means. A more detailed analysis of the errors at each time can be conducted, considering the availability, at each run of the Gibbs scheme, of the component at which each e_t and w_t is assigned. In this respect the Monte Carlo reconstruction of each observation error density can be obtained by

$$(e_t|D) \approx \frac{1}{M} \sum_{m=1}^{M} N(\mu^e_{t,m}, V^e_{t,m}) \qquad \forall t, (t = 1, ..., n),$$

and similarly for the state errors. Such densities can be used to single out the features of each state and observation error. Simply looking at some characteristics of the obtained densities, like the number and the position of the modes and the overall variability, some classifications can be proposed. Of course some more formal classification procedure could be used, although visual evidence seems to be clear enough in many situations.

In Figure 3 some typical examples of the obtained results are reported. In the first two pictures an outlier and a change in level are clearly detected while in the third one the presence of two modes makes the detection uncertain.

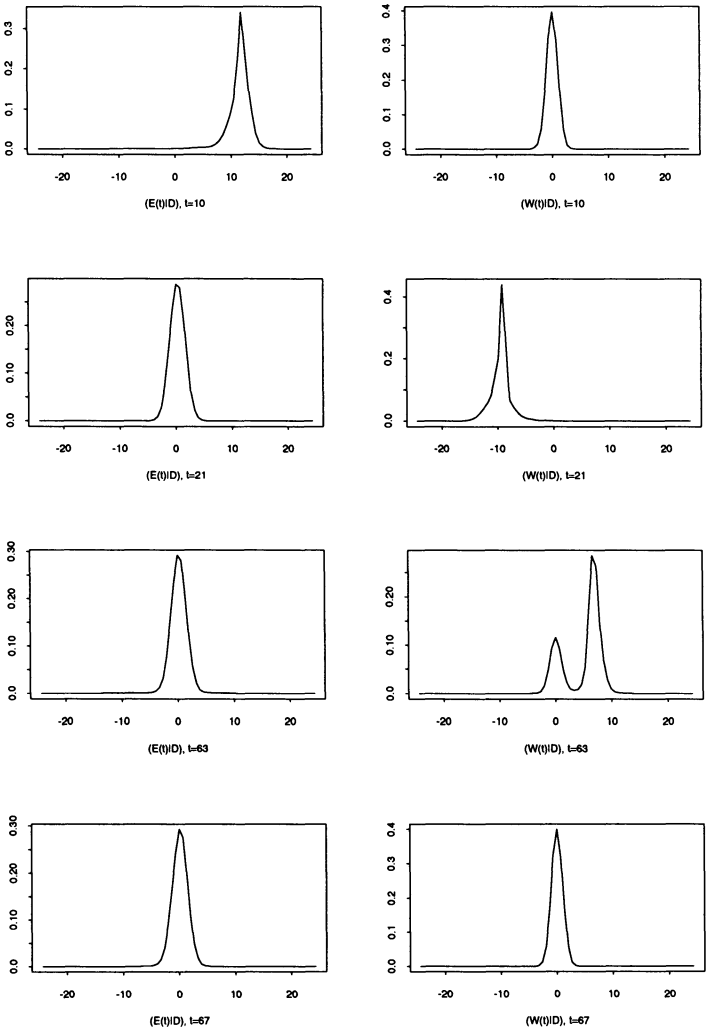

FIGURE 3: *Examples of error densities reconstruction*

The last line of the picture shows a situation where neither outliers nor changes in level can be singled out. According to this classification, the detected features can be compared with the data generation process and the results of the performed analysis are summarized in Table 1. Considering together the correctly and the uncertainly classified cases, the procedure provides correct indications for more than 90% of the errors.

TABLE 1 - *Classification of the observation and state errors*

	Zero mean errors	Outliers	Changes in level
Correctly classified	71	12	8
Uncertainly classified	2	0	1
Misclassified	3	0	3
Total	76	12	12

5. Concluding remarks. The approach proposed in this paper must be regarded as a retrospective assessment of the most relevant characteristics of a time series modelled by a first-order polynomial DLM.

In this respect, the memory of the Dirichlet processes, used in modelling the parameters of the error terms, allows to globally take into account the main features of the series producing a new approach to the detection of outliers and changes in level.

Directions for future research include the extension of the approach to models involving a more complex parametrization and the use of the posterior distribution of the errors as an assessment of the components of multiprocess models to be employed for forecasting purposes.

REFERENCES

ANTONIAK, C. (1974). Mixtures of Dirichlet Processes with Application to Nonparametric Problems. *The Annals of Statistics* **2** 1152-1174.

CARLIN B.P., POLSON N.G., STOFFER D.S. (1992). A Monte Carlo Approach to Nonnormal and Nonlinear State Space Modeling. *Journal of the American Statistical Association* **87** 493-500.

ESCOBAR, M.D. (1988). *Estimating the Means of several normal populations by estimating the distribution of the means.* Unpublished Ph.D. thesis, Yale University.

ESCOBAR, M.D., WEST, M. (1995). Bayesian Prediction and Density Estimation. *Journal of the American Statistical Association.* (to appear).

FERGUSON T.S. (1973). A Bayesian Analysis of some Nonparametric Problems. *The Annals of Statistics* **1** 209-230.

FERGUSON T.S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, (Hrizvi and J. Rustagi, eds.) New York, Academic Press.

MEINHOLD R.J., SINGPURWALLA N.B., (1989). Robustification of Kalman Filter Models. *Journal of the American Statistical Association* **84** 479-486.

WEST M., HARRISON P.J., MIGON H.S. (1985). Dynamic Generalized Linear Model. *Journal of the American Statistical Association* **82** 1032-104.

WEST M., HARRISON P.J. (1989). *Bayesian Forecasting and Dynamic Models.* Springer Verlag, New York.

WEST M., MUELLER P., ESCOBAR M.D. (1994). Hierarchical Priors and Mixtures Models, With Application in Regression and Density Estimation. In *Aspects of Uncertainty*, (P.R. Freeman and A.F.M. Smith, eds.), John Wiley and Sons Ltd.

DIPARTIMENTO DI STATISTICA
UNIVERSITÀ DI FIRENZE