ST 337 / ST 405:
Bayesian Forecasting and Intervention

Dr Bärbel Finkenstädt Rand[1], Department of Statistics, University of Warwick

[1]B.F.Finkenstadt@warwick.ac.uk.

CONTENTS

CHAPTER 4

FORECASTING SYSTEMS

Usually, the forecasting performance of any model deteriorates through time. Means of incorporating external information or even new parameters must be allowed for, in order to maintain this performance. This interaction is called a forecasting system. In order to asses the performance of the model one has to monitor it and, if necessary, intervene.

4.1 Management by exception

Data is processed using a routine statistical model and its forecasts accepted by a decision procedure unless exceptional circumstances arise. Exceptions come up in two general ways:

Feed-Forward This kind of intervention occurs when information received by experts anticipates an unexpected major change (e.g. in an inventory system, we know that the price of a certain product will increase in two months time; so, we expect an exceptional increase in the stock of that product) . This may lead to qualitative, quantitative or even to conceptual model adjustments. These interventions are *anticipatory* in nature.

Feedback This intervention responds to suspected model inadequacies signalled by the model monitor. These may relate to a large error, over-forecasting, increased volatility, ... These interventions are *corrective* in nature.

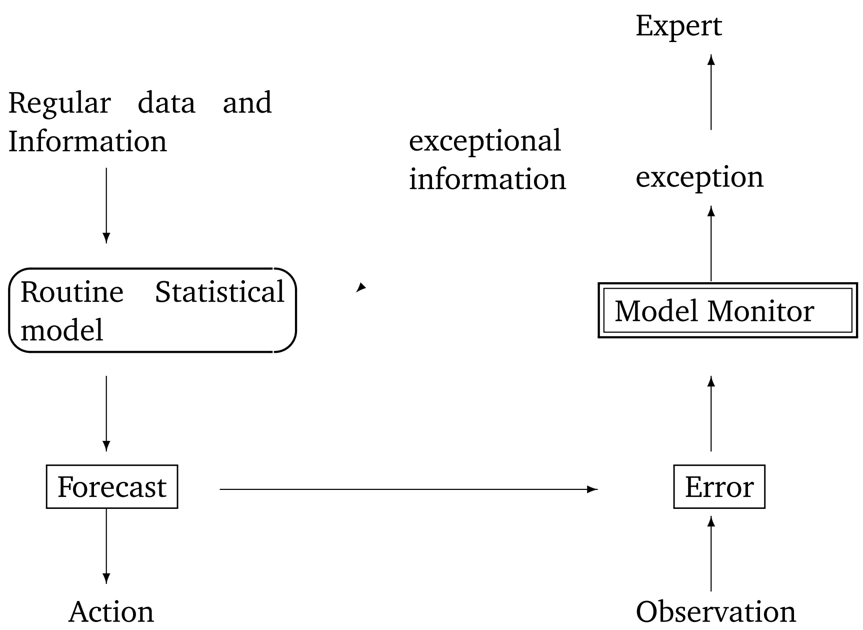

Figure 4.1. *A forecasting system operates according to the principles of management by exception.*

4.2 Monitoring

Adequacy of a model by itself is difficult to asses and it is often better to compare the performance of alternative models. As Figure 4.1 illustrates, we will keep the routine model until an exception appears and then, depending on its type, the model will be revised, altered or changed. So, we first have to set up a monitoring scheme capable to detect such exceptions.

Generally, there are two main ways of statistical testing:

- Relative testing: does one model outperform another? This requires more than one model and is often done through Bayes factors (see later).
- Diagnostic (absolute) testing: here we simply contrast observed and expected behaviour on the basis of the model assumptions.

Often, we start with diagnostic testing, and once we have detected a problem, we need to think about alternative models. A good starting point for diagnostic testing are the forecast errors (these are actually not residuals, but mean residuals; remember that $y_t \mid D_{t-1}$ has an entire distribution centred over f_t).

4.2.1 Error analysis

Forecasting errors (mean residuals) measure the retrospective departure of the model from the data. Of paramount importance in a forecasting system are the one-step ahead forecasting errors, $e_t = y_t - f_t$. Under the DLM assumptions (with unknown variance), the forecast distribution is $y_t \mid D_{t-1} \sim \mathrm{St}\left(y_t \mid f_t, q_t, n_{t-1}\right)$, thus the error has a predictive distribution

$$e_t \mid D_{t-1} \sim \mathrm{St}\left(e_t \mid 0, q_t, n_{t-1}\right) \;;$$

and therefore, $u_t = e_t/\sqrt{q_t}$ follow a standard Student t distribution, with n_{t-1} d.f.

Thus, from the model assumptions, u_t has mean zero, variance equal to $\frac{n_{t-1}}{n_{t_1}-2}$ which is approximately one (we are assuming t is not too small), and is serially uncorrelated.

Plotting u_t over time is usually the best starting point to look for such deviations from the assumptions in the sequence as:

i.– Individual, extreme errors. Possibly due to outlier observations.

ii.– Groups or patches. A sequence of consecutive errors of the same sign suggests either a local drift away from zero location, or the development of positive correlation between errors.

iii.– Patches of errors with alternated signs suggest negative correlation between them.

iv.– Patches of error with large absolute value, without any clear pattern.

Isolated, individual large errors suggest the occurrence of an outlier. Typically, the best strategy in this cases is just to ignore the observation, make $D_t = D_{t-1}$, and flag a signal in the monitoring scheme indicating that an outlier was observed at time t.

Moderate positive (negative) errors may not be large enough in order to trigger the outlier alert. However, a sequence of consecutive such errors clearly indicates that the model is systematically under (over) forecasting. The cumulative sum (CUSUM) of the last l_t forecast errors,

$$E_t = \sum_{r=0}^{l_t-1} e_{t-r}$$

provides us with a means of determining when such a patch should activate the monitoring alert.

4.2.2 Bayes factors

For Bayesians, the natural way to express uncertainty is through probability theory. Thus, if we are (as it is usually the case) uncertain about which sampling model to choose, it is natural to use a probability perspective. We can simply treat the unknown model index as a random quantity, over which we need to specify a prior (just like any other unknown parameter).

Therefore, if we have a set of possible models, $M_1, \dots, M_r$, say, all possibly of different dimensions and with different parameter vectors, $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_r$, we need to specify our prior probabilities on the models, i.e.

$$\left\{\mathrm{P}[M_1], \dots, \mathrm{P}[M_r]\right\} \qquad \text{such that} \qquad \sum_{i=1}^{r} \mathrm{P}[M_i] = 1$$

(indicating that we are only considering the r models in our set). If we consider two models, say M_i and M_j, the **prior odds** of these models will be $\mathrm{P}[M_i]/\mathrm{P}[M_j]$, denoting the prior probability ratio between these two models.

By simply using Bayes Theorem, we can express the **posterior odds**, the ratio of posterior probabilities between two alternative models, based on observations $\boldsymbol{y} = \{y_1, \dots, y_T\}'$ as

$$\frac{\mathrm{P}[M_i \mid \boldsymbol{y}]}{\mathrm{P}[M_j \mid \boldsymbol{y}]} = \frac{\mathrm{P}[M_i]\, f\left(\boldsymbol{y} \mid M_i\right)}{\mathrm{P}[M_j]\, f\left(\boldsymbol{y} \mid M_j\right)}\,, \tag{4.1}$$

where $f\left(\boldsymbol{y} \mid M_i\right)$ is the so-called **marginal likelihood**, given by

$$f\left(\boldsymbol{y} \mid M_i\right) = \int_{\Theta_i} f\left(\boldsymbol{y} \mid \boldsymbol{\theta}_i, M_i\right) f\left(\boldsymbol{\theta}_i \mid M_i\right)\, \mathrm{d}\boldsymbol{\theta}_i\,,$$

defining the prior distribution for the parameters in M_i through $f\left(\boldsymbol{\theta}_i \mid M_i\right)$ (a pdf).

The ratio of marginal likelihoods $B_{ij} = f\left(\boldsymbol{y} \mid M_i\right)/f\left(\boldsymbol{y} \mid M_j\right)$, is called the **Bayes factor** and indicates how much the data favours M_i over M_j. If $B_{ij} > 1$ the data favours M_i and the posterior odds between M_i and M_j will exceed the prior odds.

From (4.1), posterior odds between any two alternative models are equal to the prior odds times the Bayes factor. Now, if we entertain r models, a posteriori we have

$$\mathrm{P}[M_i \mid \boldsymbol{y}] = \frac{\mathrm{P}[M_i]\, f\left(\boldsymbol{y} \mid M_i\right)}{\sum_{k=1}^{r} \mathrm{P}[M_k]\, f\left(\boldsymbol{y} \mid M_k\right)}\,.$$

These probabilities can be used for **model choice**, e.g. choosing the model with highest posterior probability; or even **model averaging**, which formally deals with model uncertainty, by averaging (for instance, predictions) over models using posterior model probabilities as weights.

Often we use prior odds equal to unity and the interest, thus, focusses on the Bayes factors. In the next subsection we see how Bayes factors can be used in model monitoring for DLM's.

4.2.3 Bayes factors for DLM's

When setting up a DLM one has to decide, among other things, on the order of the polynomial component, the number of harmonics to include, if covariates are available, how many and which of them should be used, ... Usually, more than one model may seem sensible for a time series. Assume that M_1 and M_2 are two competing models, in order to decide which one is better for the data available at time t, one can calculate the corresponding Bayes factor

$$B_t^{12} = \frac{f\left(\boldsymbol{y}_t \mid \boldsymbol{D}_{t-1}, M_1\right)}{f\left(\boldsymbol{y}_t \mid \boldsymbol{D}_{t-1}, M_2\right)}.$$

The Bayes factor of M_1 versus M_2, based on the sequence of k consecutive observations $\{y_t, y_{t-1}, \ldots, y_{t-k+1}\}$ is given by

$$B_t(k) = \prod_{j=t-k+1}^{t} B_j = \frac{f\left(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-k+1} \mid D_{t-k}, M_1\right)}{f\left(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-k+1} \mid D_{t-k}, M_2\right)}$$

One key feature of the Bayes factor is that evidence in favour of M_1 accumulates multiplicatively as data are processed:

$$B_t(k) = B_t B_{t-1}(k-1) \qquad k = 2, \ldots, t.$$

Alternatively, on the log scale, evidence accumulates additively.

The basic idea would then be to select a threshold value $B^\star$ and flag a signal in favour of using M_2 instead of M_1, whenever the cumulated Bayes factor surpasses the threshold. One such a scheme is as follows:

Let

$$L_t = \min_{1 \le k \le t} B_t(k),$$

with $L_1 = B_1$. The quantities L_t are updated sequentially by

$$L_t = B_t \min\{1, L_{t-1}\} \qquad t > 1.$$

The minimum at time t is taken at $k = \ell_t$, with $L_t = B_t(\ell_t)$, where the *run-length*, ℓ_t, is sequentially updated by

$$\ell_t = \begin{cases} 1 + \ell_{t-1} & L_{t-1} < 1 \\ 1 & L_{t-1} \ge 1 \end{cases}.$$

The run-length,ℓ_t, is the number of recent consecutive observations that lead to evidence against M_1 in favour of the alternative model, M_2. As ℓ_t grows, evidence against M_1 builds up, and L_t drops rapidly. Only if $\ell_t = 1$, i.e. the previous observation favoured M_1, do we compare models only on the evidence in the current observation.

Once L_t drops below a certain (preset) threshold, we will typically need to take action. This could be by discarding y_t as an outlier, if $\ell_t = 1$; or intervening in the model if $\ell_t > 1$, as explained in the next subsection.

4.3 Intervention

Once an exception is signaled and/or exceptional information is received, the forecasting system updates, revises or changes the routine statistical model in order to accommodate it. As said in the first section of this chapter, there are two general way of intervention, we start with

4.3.1 Feed-Forward Intervention

Outliers

The most common strategy to follow when a single atypical observation is seen is simply to ignore it. The rationale behind this action is that such an observation does not bring new, relevant information into the analysis, so $D_t = D_{t-1}$ and the posterior for time t is just the prior, $\boldsymbol{\theta}_t \mid D_t \sim \mathrm{N}\left(\boldsymbol{\theta}_t \mid \boldsymbol{m}_t, C_t\right)$ where $m_t = a_t$ and $C_t = R_t$.

Additional evolution noise

When new external information about the circumstances affecting $\boldsymbol{y}_t$ is received, one way to proceed is to alter the elements of the variance matrix, R_t, in order to reflect increased uncertainty about all the parameters of the model. Alternatively, if the information is such that the expert can discriminate which elements are more likely to be affected, only the corresponding (block) elements of R_t might be altered, leaving the rest unchanged. Likewise, trend, seasonal, regression and other components can be selectively altered in order to reflect expected changes in their future behaviour.

All these interventions can be accommodated within the DLM framework, such that the resulting model is still a DLM and so, updating and forecast equation forms remain the same.

More in particular, we wish to adjust the system before observing y_t, so the relevant information set is D_{t-1}. Normally we would have the following distribution for the state vector:

$$\boldsymbol{\theta}_t | D_{t-1} \sim \mathrm{N}\left(\boldsymbol{\theta}_t \mid \boldsymbol{a}_t, R_t\right)$$

which, in the presence of the information leading to the intervention, I_t, is changed to, say,

$$\boldsymbol{\theta}_t | I_t, D_{t-1} \sim \mathrm{N}\left(\boldsymbol{\theta}_t \mid \tilde{\boldsymbol{m}}_t, \tilde{C}_t\right).$$

There could be an anticipated shift in the mean of $\boldsymbol{\theta}$ (which can correspond to level or slope) and/or an increased uncertainty.

In the usual situation, we have

$$\boldsymbol{\theta}_t = G_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$$

leading to the usual updates $\boldsymbol{a}_t = G_t \boldsymbol{m}_{t-1}$ and $R_t = G_t C_{t-1} G_t' + W_t$. By now adding an independent noise term, say, $\boldsymbol{\zeta}_t \sim \mathrm{N}\left(\boldsymbol{\zeta}_t \mid \boldsymbol{h}_t, H_t\right)$, we can change the system equation to

$$\boldsymbol{\theta}_t = G_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t + \boldsymbol{\zeta}_t$$

leading to the intervention updates

$$\tilde{\boldsymbol{m}}_t = G_t \boldsymbol{m}_{t-1} + \boldsymbol{h}_t = \boldsymbol{a}_t + \boldsymbol{h}_t$$

and

$$\tilde{C}_t = G_t C_{t-1} G_t' + W_t + H_t = R_t + H_t.$$

Note that we can not accommodate a decrease in uncertainty, since H_t is positive definite. A more general transformation of the system equation can deal with such cases, however, as explained in exercise sheet 4.

4.3.2 Feedback Intervention

Once a monitoring scheme has been set up, handling of detected exceptions must be done according to the forecasting system described above. Of course, some automatic responses can be implemented to deal with the most common situations; however, careful analysis of less common circumstances is an integral part of the system.

If an exception occurs on the basis of a single observation (i.e. with $\ell_t = 1$), then it might be that the best response is to treat y_t as an outlier (but this is not easy to identify if we have no further observations beyond y_t).

In general, a change in the circumstances brings an increase in the variance of the system, thus feedback intervention in this case may be achieved through

$$\boldsymbol{\theta}_t = G_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t + \boldsymbol{\xi}_t$$

where $\boldsymbol{\omega}_t$ is the usual evolution error term, independent of the additional, automatic intervention noise term, $\xi_t \sim \mathrm{N}\left(\xi_t \mid 0, H_t\right)$. The variance matrix, H_t, should be specified to reflect the increase in uncertainty brought up by the detected exception; in particular, if no element is to be singled out, one option is to take $H_t = (c-1)W_t$, with $c > 1$, which inflates the overall variance of $\boldsymbol{\theta}_t$ by a factor of c.

Another option, frequently used in practice is to combine the above programme with the discount factors technique; decreasing the discount factor of the appropriate model block to reflect the increase of uncertainty.

Example 1 (NHS prescriptions data).
To illustrate the concepts above we will analyse some NHS data. The data consists of deseasonalised NHS monthly prescriptions (in thousands) from Jan/68 to Jan/72, depicted in the upper part of Figure 4.2. There we can see that the series has two different stable levels, one before June/69 around 220 and another afterwards of around 195. Thus, we choose a first order polynomial trend discount DLM $\{1, 1, V, W_t\}$:

$$\begin{array}{lll} \text{obs. eq.} & Y_t = \mu_t + \nu_t & \nu_t \sim \mathrm{N}\left(\nu_t \mid 0, \lambda^{-1}\right) \\ \text{sys. eq.} & \mu_t = \mu_{t-1} + \omega_t & \omega_t \sim \mathrm{N}\left(\omega_t \mid 0, W_t\right) \\ \text{Precision} & \lambda \sim \mathrm{Ga}\left(\lambda \mid \frac{n_t}{2}, \frac{n_t s_t}{2}\right) & \end{array}$$

Together with the prior settings $\mu_0 \sim \mathrm{N}\left(\mu_0 \mid 220, 25\right)$ with a discount factor of $\delta_t = 0.9$ and for the precision $n_0 = 4$ and $s_0 = 49$.

As an alternative to the formal Bayes factor monitoring of Subsection 4.2.3, an automatic monitoring scheme, based on CUSUM's is implemented. So every time that the CUSUM, $S_{t,m}$, of the standardised forecast errors, u_t,

$$S_{t,m} = \left| \sum_{i=0}^{m-1} u_{t-i} \right|$$

exceeds a predetermined threshold, the monitor will flag up a signal.

From historical records we know that:

(*i*) A rise in prescription charges in June 69 resulted in a sustained fall in the underlying demand for prescriptions and

(*ii*) an influenza epidemic in December 1970 resulted in a large number of prescriptions for that month, but did not affect the underlying demand. ◁

In first instance, we will ignore this additional information and let the model run as described above (i.e. a feedback intervention analysis).

The monitoring scheme will signal at June/69, and our intervention will consist of changing the level mean to 200 and its variance to 25. The effect of this intervention can be seen in the top panel of Figure 4.2: the level is immediately set to 200 (actually, this may be a bit of a cheat; it might have been better to only increase the variance).

A second signal is issued by the scheme in December/70. We will mark this as an outlier, i.e. it will be ignored for the next updates, but the run-length will be incremented. The result of this intervention can be visualised again in the top panel of Figure 4.2: the observation corresponding to Dec/70 is deleted and the forecast for the next period is the same as that for the previous ($D_{Dec/70} = D_{Nov/70}$).

The lower panel of Figure 4.2 depicts the series in terms of standard deviations. This visualisation is useful as we would expect that roughly 95% of the observations fall within two standard deviations, as is the case.

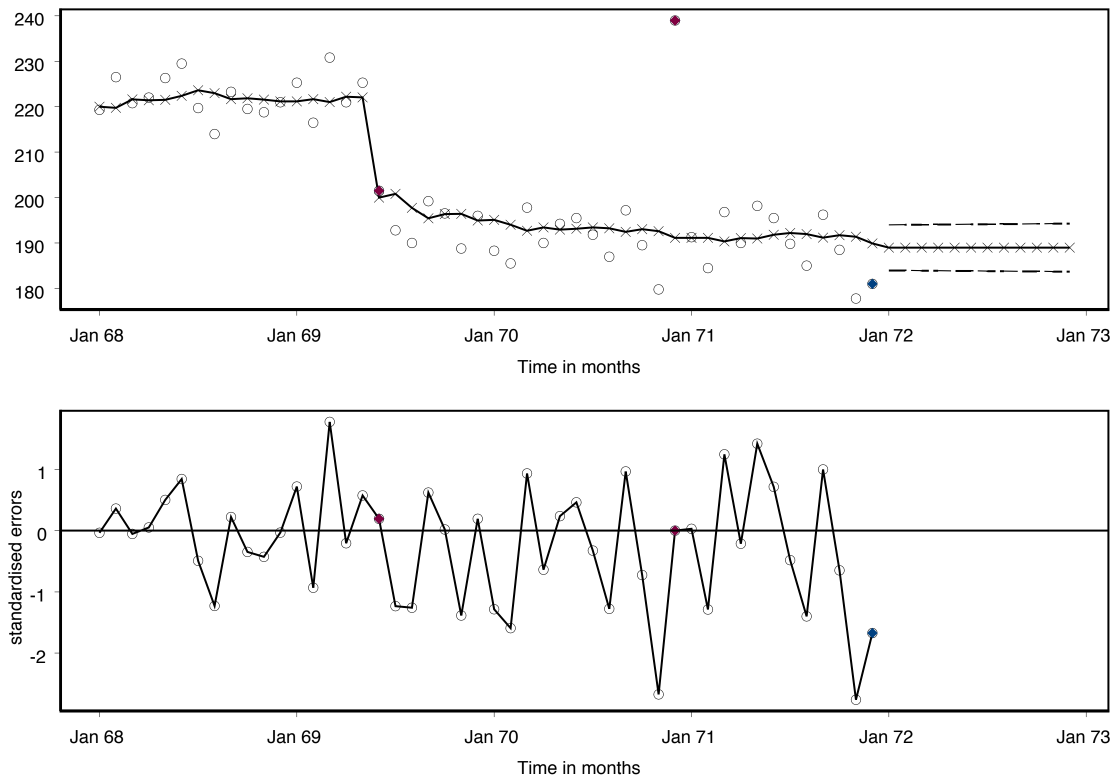

Figure 4.2. *NHS prescription example, forecasts with feedback intervention.*

A second option is to analyse the series using the available information from the beginning; i.e. a feed-forward intervention. As at $t = 1$ we already know that two exceptions occur at $t = 18$ and $t = 36$, we can include this in our model and intervene at those points. Again, our intervention will consist of setting the level mean to 200 and the variance to 25, at June/69 and to flag Dec/70 as an outlier.

The results of this modelling are illustrated in Figure 4.3. These plots are almost identical to the previous ones, the main difference is that in the later interventions are programmed to be included at $t = 18, 36$; while in the former, interventions will take place iff the monitoring scheme triggers a flag.

To illustrate this point, let us change now the threshold that the automatic CUSUM scheme must exceed to signal an exception, and increase it by 15%. The result of this is that no monitor signal is issued at the second from last observation, this can be seen in Figure 4.4

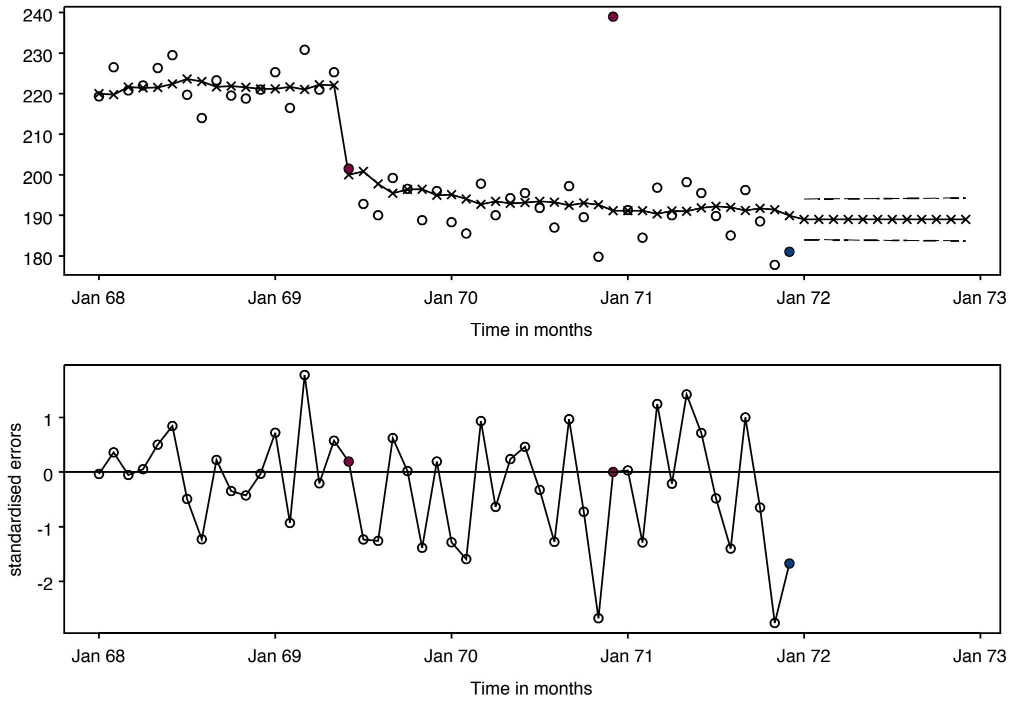

Figure 4.3. *NHS prescription example, forecasts with feedforward intervention.*

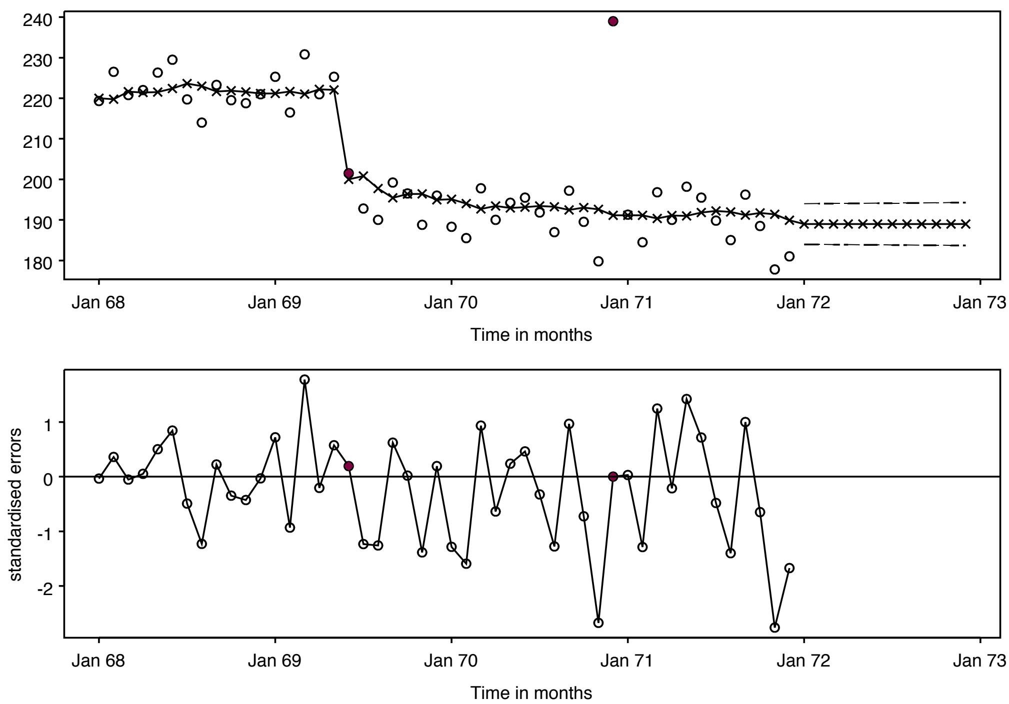

Figure 4.4. *NHS prescription example, forecasts with feedforward intervention and increased* CUSUM *threshold.*