On-line Novelty Detection Using the Kalman Filter and Extreme Value Theory

Hyoung-joo Lee, Stephen J. Roberts
Department of Engineering Science, University of Oxford
Parks Road, OX1 3PJ, Oxford, UK
hjlee@robots.ox.ac.uk, sjrob@robots.ox.ac.uk

Abstract

Novelty detection is concerned with identifying abnormal system behaviours and abrupt changes from one regime to another. This paper proposes an on-line (causal) novelty detection method capable of detecting both outliers and regime change points in sequential time-series data. Our approach is based on a Kalman filter in order to model time-series data and extreme value theory is used to compute a novelty measure in a principled manner. The proposed approach is shown to be effective via experiments on several real-world data sets.

1 Introduction

The goal of novelty detection is to identify abnormal system behaviours which are not consistent with the "normal" (or dominant) state of a system [4]. Such an approach has been of great importance in various domains where a severe imbalance exists between numbers of examples in differing classes or in cases in which the novel "class" cannot be uniquely defined but system "normality" can be. Examples include condition monitoring, identity verification, financial data analysis and biomedical signal processing.

For many practical applications, in particular in time-series domains such as finance, a novelty takes place for two different reasons. Firstly, an *outlier* arises when an observation deviates from a static norm [2]. A novelty can also be oserved when the norm itself changes dynamically, i.e. at a *change point* [3]. These two types of novelty, although closely related, have hitherto been approached separately.

This paper proposes an on-line novelty detection method capable of detecting both outliers and change points in time-series data. The Kalman filter [5] is employed to estimate dynamically the probability density of the target data. It is assumed that a deviation in the *observation space* indicates an outlier whereas one in the *state space* (which defines the dynamical model of the data) represents a change point between dynamic regimes. In order to derive a more principled and intuitive novelty measure from the estimated density, extreme value theory, in the semi-analytic framework described in [7], is adopted. The proposed approach is shown to be effective through experiments on artificial and real-world time-series data sets.

2 Proposed Approach

This section presents the main idea of the proposed approach. In particular, we firstly describe the Kalman filter multivariate autoregression model [6] via which a (multivariate Gaussian) probability density over the target variables can be obtained. Secondly, we introduce extreme value theory (EVT) which provides our novelty measure. We can model the dynamics of the state space using the Kalman filter, and detect changes of the underlying system as well as outliers in the observation sequence. On-line detection is possible since sequential modelling is sufficient for the Kalman filter. Furthermore, EVT enables us to obtain a novelty measure and determine a threshold in a more principled manner rather than just thresholding a probability density.

2.1 Kalman filter autoregression

Suppose a data set $\{\mathbf{y}_t\}_{t=1}^T$, where $\mathbf{y}_t = [y_{t,1}, \ldots, y_{t,d}]^\top \in \mathbb{R}^d$. In order to model the data, a Kalman filter multivariate autoregression model [6] is considered as follows:

$$\mathbf{a}_t = \mathbf{a}_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}(0, \mathbf{W}_t), \tag{1}$$

$$\mathbf{y}_t = \mathbf{H}_t \mathbf{a}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(0, \mathbf{V}_t). \tag{2}$$

With the state transition matrix assumed to be identity (i.e. we do not believe there to be systematic trends in the data in these examples), the state variable, $\mathbf{a}_t \in \mathbb{R}^k$, follows a Gaussian diffusion process with Markovian dynamics. The observation matrix, $\mathbf{H}_t \in \mathbb{R}^{d\times k}$ consists of the past observations of the time-series in these examples (though there is no reason why this could not be any observable data), i.e. $\mathbf{H}_t = \left[1, y^1_{t-1}, \ldots, y^d_{t-1}, \ldots, y^1_{t-p}, \ldots, y^d_{t-p}\right] \otimes \mathbf{I}_d$, where p is an AR model order, $\otimes$ is the Kronecker product, and $\mathbf{I}_d$ is a $(d \times d)$ identity matrix. Hence, eqs.(1-2) specify a system where, at time t, the state is a set of time-varying AR coefficients and the observation is estimated by a multivariate AR model with a model order of p.

Without going into details, estimating the variables, $\mathbf{a}_t, \mathbf{w}_t$, and $\mathbf{v}_t$, is identical to the standard Kalman filter, alternating prediction and update. The estimated probability density functions can be written as follows, repectively,

$$p(\hat{\mathbf{a}}_t) = \mathcal{N}(\hat{\mathbf{a}}_{t-1}, \mathbf{P}_t), \tag{3}$$
$$p(\mathbf{y}_t) = \mathcal{N}(\mathbf{H}_t\hat{\mathbf{a}}_{t-1}, \mathbf{S}_t). \tag{4}$$

The estimated covariances are, respectively, $\mathbf{P}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)(\mathbf{P}_{t-1} + \mathbf{W}_t)$ and $\mathbf{S}_t = \mathbf{V}_t + \mathbf{H}_t(\mathbf{P}_t + \mathbf{W}_t)\mathbf{H}_t^\top$, where $\mathbf{K}_t = (\mathbf{P}_{t-1} + \mathbf{W}_t)\mathbf{H}_t\mathbf{V}_t^{-1}$ is the Kalman gain.

The above probability densities can be used as novelty measures with an appropriate thresholding. Low densities in the observation space correspond to outliers which deviate from the system norm. On the other hand, low densities in the state space suggest that the statistics of the underlying system are changing i.e. potential dynamic regime changes are taking place. To find appropriate thresholds in the absence of labeled data we use the notion of extreme values, as detailed below.

2.2 Extreme value theory (EVT)

When novelty detection is performed based on estimation of a probability density function, a constant threshold is typically used such that a data point is considered novel if its probability density is lower than the threshold. This is equivalent to implicitly assuming a uniform density for the novel class and then carrying out discriminant analysis. However, it can be problematic since the uniform density is generally not an appropriate distribution. In addition, determining the threshold is data-dependent. For example, a threshold of 0.1 could be perfect for some data set, but it could be a disaster for another, especially in a high-dimensional space.

Extreme value theory (EVT) can be an alternative [7]. Suppose a set, $\mathbf{Z}_m = \{z_i\}_{i=1}^m$ with m *i.i.d* random variables drawn from a target distribution, $\mathcal{D}$. The extreme value probability (EVP) of z is a probability of z being the largest value in the set, i.e. $P_{EV}(z|\mathbf{Z}_m) = P(\max(\mathbf{Z}_m) \leq z)$. By the Fisher and Tippett theorem, an EVP is expressed as the generalised extreme value distribution which is generally not easy to evaluate. However, if the target distribution is assumed to be the one-sided standard Gaussian, i.e. $\mathcal{D} = |\mathcal{N}(0,1)|$, the EVP is analytically obtained in the form of the Gumbel distribution:

$$P_{EV}(z|\mathbf{Z}_m) = \exp\left[-\exp\left(-\frac{z-\mu_m}{\sigma_m}\right)\right], \tag{5}$$

where $\mu_m = (2\ln m)^{1/2} - \frac{\ln(\ln m)+2\pi}{2(2\ln m)^{1/2}}$ and $\sigma = (2\ln m)^{-1/2}$. Note that the "norming" parameters, μ_m and σ_m, depend only on the number of data, m, in the set. We use a notation $P_{EV}(z|m)$ for $P_{EV}(z|\mathbf{Z}_m)$ throughout this paper.

Assuming the Gaussian noise processes for the state and observation variables in the Kalman filter model, eq.(5) can be applied directly to our novelty detection approach. A Mahalanobis distance,

$$\sqrt{(\mathbf{x}-\mathbf{m}_x)^\top \mathbf{C}_x^{-1}(\mathbf{x}-\mathbf{m}_x)}, \tag{6}$$

is considered to be distributed as $|\mathcal{N}(0,1)|$ if $\mathbf{x}$ is Gaussian. Therefore, we can obtain EVPs of $\mathbf{a}_t$ and $\mathbf{y}_t$ as in eq.(5) by replacing $\{\mathbf{x}, \mathbf{m}_x, \mathbf{C}_x\}$ in eq.(6) with $\{\hat{\mathbf{a}}_t, \hat{\mathbf{a}}_{t-1}, \mathbf{P}_t\}$ in eq.(3) and $\{\mathbf{y}_t, \mathbf{H}_t\hat{\mathbf{a}}_{t-1}, \mathbf{S}_t\}$ in eq.(4), respectively.

In order to evaluate an EVP as in eq.(5), the number of data, m, should be specified. We adopt the concept of *run length* which means time since the last outlier or change point [1]. It is assumed that the run length is reset to 1 if a novelty arises or added by 1 if no novelty is observed. Hence, at time t, the run length, ℓ_t, stochastically has binary values,

$$\ell_t = \begin{cases} 1, & \text{with } \Pr = P_{EV}(t-1), \\ \ell_{t-1}+1, & \text{with } \Pr = 1 - P_{EV}(t-1), \end{cases}$$

where $P_{EV}(t)$ can be either $P_{EV}(\mathbf{a}_t)$ or $P_{EV}(\mathbf{y}_t)$, and $\ell_1 = 1$. In turn, we have the run length probability,

$$P(\ell_t = m) = \begin{cases} P_{EV}(t-1), & \text{for } m = 1, \\ (1 - P_{EV}(t-1))\, P(\ell_{t-1} = m-1), & \text{for } m \geq 2, \end{cases}$$

where $P(\ell_1 = 1) = 1$. As we are not directly interested in m, we obtain the EVP by marginal-

ising out the run length,

$$P_{EV}(t) = \sum_{m=1}^{t} P_{EV}(t|\ell_t = m)P(\ell_t = m).$$

The EVP can be used as a novelty measure; a novelty flag is raised if it exceeds some threshold, θ_{EV}. That is, an outlier is detected if $P_{EV}(\mathbf{a}_t) \geq \theta_{EV}$, or a change point if $P_{EV}(\mathbf{y}_t) \geq \theta_{EV}$. In this paper, we set $\theta_{EV} = 0.95$. Using an EVP as a novelty measure has a few advantages over thresholding the probability density. First, it makes more theoretical sense without making an inappropriate assumption of a uniform density over the novel class. Second, it is more interpretable since a EVP is more recognisable with a spike-like shape when a novelty arises, as will be shown in the next section. Third, it is in the form of probability, so setting the threshold at a reasonable value, e.g. 0.95, works for most data sets. On the other hand, a probability density has different scales for different data sets. The threshold can be determined only after examining the density for the whole data set, which is impossible and undesirable for novelty detection.

3 Experimental Results

This section presents experimental results on three data sets: an artificial data set, the Well-Log data set, and the Dow Jones data set. Although all of them are one dimensional for illustration purposes, extending to multi-dimensional data is straightforward. For each data set, a simple AR model was applied to a "run-in" period of first 100 time steps to choose the model order p and the initial noises, $\mathbf{W}_0$ and $\mathbf{V}_0$. All the Kalman filter AR models were first-order, i.e. $p = 1$ except for the well-log where it was second-order.

Figure 1 shows the results on the artificial data set which was generated by alternating first-order and zeroth-order random walks (upper). The change points between different orders were chosen to be at $t = 250, 500$, and 750 while "one-off" outliers were generated at $t = 200, 400, 600$ and 800. The observations and the estimated AR coefficients are shown in the upper and the lower panels, respectively. The vertical lines indicate novelties detected by our novelty detector. In the observation space (upper), the detector identified all four outliers and two change points out of three. However, it failed to detect a change point at $t = 250$. In the state space (lower), on the other hand, it identified all four outliers and all three change points while it raised one false alarm at around $t = 720$. Considering the similar

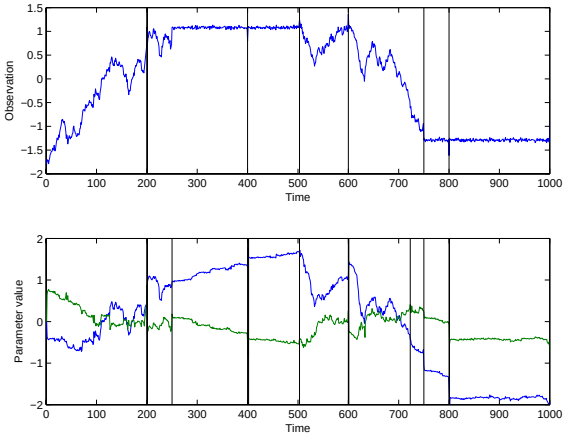

Figure 1. Artificial data set

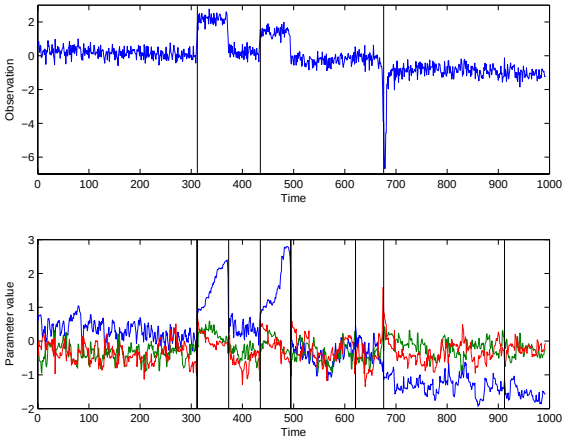

Figure 2. Well-Log data set

results in both spaces, we suppose that distinction between outliers and change points is vague in this data set.

The Well-Log data set contains measurements of nuclear magnetic response while drilling a well. Figure 2 shows the results from a latter part, about 1,000 time steps. The observation sequence has some variations in baseline, and a deep valley. In the observation space (upper), our detector picked up two deviations in baseline and the valley, but not two returns to the norm. In the state space (lower), it detected all the baseline variations and the valley. It caught two other points as novelties at $t = 620$ and 910. Looking at them closely, there are subtle changes at those points in frequency and magnitude of the signal.

Figure 3 shows the results on the daily returns of the Dow Jones Industrial Average from 1972 to 1975. Despite no evident outlier or change point, there were some events much relevant to the market, i.e. the dashed vertical lines: (A) conviction of Nixon's former aides (30 Jan 1973), (B) beginning of the OPEC embargo (19 Oct 1973), and (C) resignation of Nixon (09 Aug 1974). The novelty detector did not notice any outliers in the

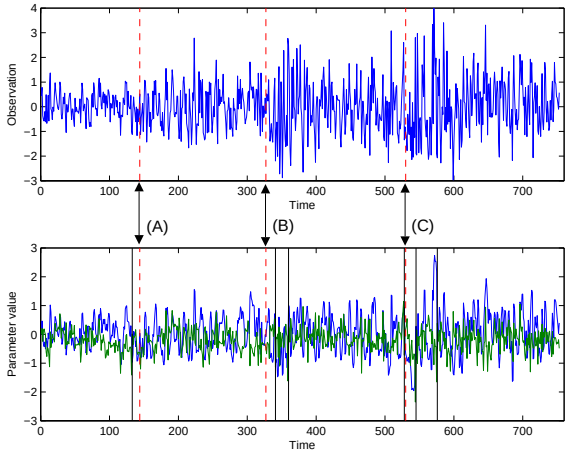

Figure 3. Dow Jones data set

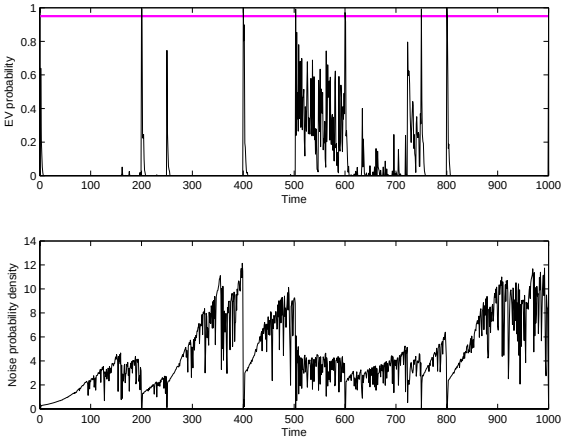

Figure 4. EVP and noise density

observation space. However, it did identify some change points in the state space. It did not *predict* the first change point or the conviction event, but it recognised some subtle changes in the market which had happened a little before the event. It detected the second change point a little late and the last one accurately.

Figure 4 compares the EVP (upper) and the noise density (lower) in the observation space of the artificial data set to illustrate practical usefulness of EVT. The thick horizontal line in the upper panel indicates the EVP threshold, $\theta_{EV} = 0.95$. The EVP is much easier to interpret, since it has spikes where it has to and is relatively flat elsewhere. The noise density oscillates over the whole period. Furthermore, while the EVP is in a form of a probability, the noise density is not with the range from 0 to roughly 12. The maximum values of the noise density were approximately 1 and 0.5 for the Well-Log and the Dow Jones data set, respectively. As a result, even though novelties could be detected by the noise density, the threshold should be optimised for a specific data set, which is difficult and undesirable for novelty detection.

4 Conclusion

This paper proposes an on-line novelty detection approach for time-series data, by employing a combined Kalman filter and extreme value theory framework. Working both in the observation and state space, we can detect not only outliers where observations are clear outliers, but also change points where the underlying system changes dynamically. Experiments on three data sets have shown promising results.

A few future research directions need to be addressed. First, as we have employed the Kalman filter, we assume linear transformations and Gaussian noises. In order to deal with general situations with non-linear transformations or non-Gaussian noises, we may need to consider particle filters or generalized extreme value theory. Second, we fixed the model order for each dataset. In a more general scenario, however, it may change over time. The dynamic model order can make estimation more accurate, and more importantly, the change in the model order can be regarded as a change point. Finally, we need to evaluate novelty detection performance for time-series data more quantitatively.

Acknowledgments

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-357-D00276).

References

[1] R. P. Adams and D. J. C. MacKay. Bayesian online changepoint detection. Technical report, Inference Group, University of Cambridge, 2007. Available as `http://arxiv.org/abs/0710.3742`.

[2] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley and Sons, New York, USA, 1994.

[3] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[4] M. Markou and S. Singh. Novelty detection: a review - part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[5] R. J. Meinhold and N. D. Singpurwalla. Understanding the Kalman filter. *The American Statistician*, 37(2):123–127, 1983.

[6] W. Penny and S. J. Roberts. Dynamic models for nonstationary signal segmentation. *Computers and Biomedical Research*, 32(6):483–502, 1999.

[7] S. J. Roberts. Novelty detection using extreme value statistics. *IEE Proceedings - Vision, Image and Signal Processing*, 146(3):124–129, 1999.