# Adaptive Streaming Anomaly Analysis

**Zhao Xu**
NEC Laboratories Europe
Heidelberg, Germany

**Lorenzo von Ritter**
Technical University of Munich
Garching, Germany

**Kristian Kersting**
Technical University of Dortmund
Dortmund, Germany

## Abstract

Detecting anomalous activities from time series data is critical for enhancing availability and security of systems in many domains. Streaming data usually contains complex dynamic patterns, which complicates the learning process. In this paper, we present a nonparametric Bayesian method AOTS to help automating the model learning for anomaly detection in streaming time series. The method learns the dynamics of anomaly-contaminated time series with submodular optimization based kernel selection to effectively adapt to the data and identify potential anomalous events. Experiments on real data show encouraging results.

## 1 Introduction

With the rapid growth of Web, mobile and Internet of Things (IoT) systems, anomaly detection in time series [1, 4, 2] has attracted increasing interests to improve the availability, performance, and the overall service experience of the systems. In this paper, we propose a flexible nonparametric Bayesian method AOTS for anomaly detection in time series data. The method models time series with Student-t processes (TPs) [9] with time $t$ as predictors. The heavy-tailed distributions of TPs provide robustness against anomalies to better capture the normal patterns. Additionally modeling observations as a function of time $t$ allows for continuous time processes, and can resolve the problem of discretizing time into evenly spaced intervals. On the other hand, time series in the real-world systems usually involves complex dynamic patterns, which makes automatic kernel selection and construction critical in the learning process. Automatic Bayesian covariance discovery [7, 3] is very helpful for addressing this problem. We develop a kernel selection method based on submodular optimization to effectively adapt to the data and reduce manual efforts in model construction. Initial experiments on real data demonstrate promising results.

## 2 Automated Time Series Anomaly Detection

Assume that there is a time series $\mathbf{y} = \{y_1, y_2, \ldots\}$ of an infinite number of observations. We model the time series as a function $y_t = f(t)$ (shortened as $f_t$) with time $t$ as predictors. The modeling method can avoid introducing anomalies into predictors like the autoregression based methods do, thus potentially reduces the complexity of modeling anomalies in time series. Additionally the continuous time processes can resolve the problem of time discretization and introduce extra flexibility.

The function $f$ itself is unknown. Here we assume that $f$ is random and can be arbitrary mathematical form without requiring manual definition in advance. It will be drawn from a function space with a

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

robust nonparametric distribution, Student-t process (TP) [9]. Formally the generative process is:

$$\zeta|\nu, k \sim \mathcal{IWP}(\nu, k); \quad f|\mu, \zeta \sim \mathcal{GP}(\mu\ , (\nu-2)\zeta)$$

That is, we first draw a kernel function $\zeta$ from an inverse Wishart process $\mathcal{IWP}(\nu, k)$ with degrees of freedom $\nu > 2$ and base kernel $k$. Then the functions (i.e. the time series) are drawn from a Gaussian process $\mathcal{GP}(\mu, (\nu-2)\zeta)$ with the kernel $\zeta$. $\mu$ denotes mean function, and is often set as zero without loss of generality. The TP has one more level than the GP, thus the sampled time series can be more flexible to capture the complex patterns in time series.

The base kernel $k$ plays an important role, especially when dynamics of time series are complicated [7, 3, 8]. To automate model construction of Student-t process with less manual intervention, we develop a greedy forward-selection method inspired by submodular optimization [5, 6]. In particular, there are a finite set $\Omega$ of kernels (e.g. linear, squared exponential, periodic kernels) and their multiplicative compositions (e.g. linear $\times$ periodic). Here we only consider multiplication of two kernels due to the possible overfitting problem. Additionally we have a set function $z$: $2^{|\Omega|} \to \mathbb{R}$, which is the minimal negative log likelihood of the time series (with the optimal kernel parameters $\theta$). Then the model construction problem is cast as: select a subset $A \subset \Omega$ minimizing the negative log likelihood $\mathcal{NLL}$ of the observed time series. Note that, Student-t process is not closed under addition. Thus we incorporate all the selected kernels into the base kernel of the inverse Wishart process, rather than a sum of several individual Student-t processes. The selection method is shown as Alg. 1.

**Algorithm 1:** Model construction of the AOTS method

**Input :** $\Omega$ (candidate kernels), $\mathbf{y}$ (observed time series), $\tau$ (stop condition, default 0.01)
Initialization: $A = \emptyset$, K = None, k=None, r=1.0;
**while** $r > \tau$ **do**
  $A \leftarrow A \cup k$, $\Omega \leftarrow \Omega \setminus k$, $K \leftarrow K + k$ ;
  Find a kernel $k \in \Omega$ that offers minimal $\mathcal{NLL}$ together with the selected kernels $A$. Its hyperparameters $\theta$ are optimized with gradient descent method. The $\mathcal{NLL}$ and the derivatives are computed as follows:
  $\mathcal{NLL} = \frac{n}{2}\log(\nu-2) + \log\left(B\left(\frac{\nu}{2}, \frac{n}{2}\right)\right) + \frac{1}{2}\log\left(|K+k|\right) + \frac{\nu+n}{2}\log\left(1 + \frac{\mathbf{y}^T(K+k)^{-1}\mathbf{y}}{\nu-2}\right)$ ;
  $\frac{\partial}{\partial\theta_i}\mathcal{NLL} = -\frac{1}{2}Tr\left(\left(\frac{\nu+n}{\nu+\beta-2}\alpha\alpha^T - (K+k)^{-1}\right)\frac{\partial k}{\partial\theta_i}\right)$ ;
  $\frac{\partial}{\partial\tilde{\nu}}\mathcal{NLL} = (\nu-2)\left[\frac{n}{2(\nu-2)} - \psi\left(\frac{\nu+n}{2}\right) + \psi\left(\frac{\nu}{2}\right)\frac{1}{2}\log\left(1+\frac{\beta}{\nu-2}\right) - \frac{(\nu+n)\beta}{2(\nu-2)(\nu+\beta-2)}\right]$ ;
  $r = (\mathcal{NLL}^{(A)} - \mathcal{NLL}^{(A\cup k)})/\mathcal{NLL}^{(A)}$ ;
**Output :** Selected kernels $A$ and their hyperparameters

$K = LL^T$ (Cholesky decomposition), $\alpha = L^T \setminus (L \setminus \mathbf{y})$, and $\hat{\nu} = \log(\nu-2)$. $\psi$ is the digamma function.

Given the constructed model, we can estimate the normality of the time series in the near future to identify the anomalous events. Technically a predictive distribution is computed based on the learned model and the observed time series, then the anomaly is detected with its $z$ score. For efficient computation, we develop rank one update, as the time series is often observed sequentially. In particular, the predictive mean $m_*$ and variance $var_*$ at time $n+2$ given the new observation $y_{new}$ at time $n+1$ and previous observations $\mathbf{y}$ are computed as:

$$\tilde{\ell}_{new} = L \setminus \tilde{\mathbf{k}}_{new}; \quad \tilde{\ell}_{new,new} = \left(k_{new,new} + \sigma^2 - \tilde{\ell}_{new}^T\tilde{\ell}_{new}\right)^{1/2}$$

$$\tilde{a}_{new} = (y_{new} - \tilde{\ell}_{new}^T\mathbf{a})/\tilde{\ell}_{new,new}; \quad L \leftarrow \begin{bmatrix} L & 0 \\ \tilde{\ell}_{new}^T & \tilde{\ell}_{new,new} \end{bmatrix}; \quad \mathbf{a} \leftarrow \begin{bmatrix} \mathbf{a} \\ \tilde{a}_{new} \end{bmatrix}$$

$$\mathbf{b} = L \setminus \tilde{\mathbf{k}}_*; \quad \beta = \mathbf{a}^T\mathbf{a}; \quad m_* = \mathbf{b}^T\mathbf{a}; \quad var_* = \frac{\nu+\beta-2}{\nu+n-1}(k_{*,*} - \mathbf{b}^T\mathbf{b}),$$

where $\tilde{\mathbf{k}}_{new}$ is the covariance between $y_{new}$ and $\mathbf{y}$, and $\tilde{\mathbf{k}}_*$ is the covariance between $f_{n+2}$ and $[\mathbf{y}, y_{new}]^T$. $k_{new,new}$ and $k_{*,*}$ are variances of $f_{new}$ and $f_{n+2}$.

## 3 Experiments and Conclusion

We evaluate the AOTS method using the airline passenger data with randomly added outliers. The first 120 time steps are training data to learn the models, and the rest is used to test performance of anomaly detection. The experimental results are reasonable and illustrated as Fig. 1. The left panel shows the predicted time series and the detected anomalies. The middle panel is the iteratively selected kernels with the gradually decreasing negative log likelihood. One can find that the returns are diminishing. To give intuitions of the selected kernels, the right panel visualizes sample time series drawn from the corresponding Student-t processes. The results demonstrate the superiority of the proposed nonparametric Bayesian method with automatic kernel selection in anomaly detection.

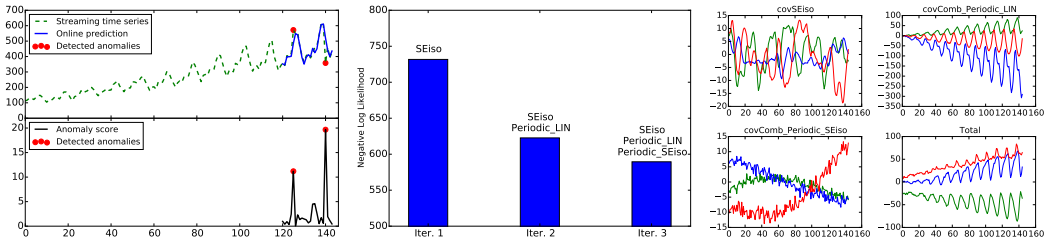


Figure 1: Detected anomalies from the airline passenger data.

**Acknowledgments** ZX was funded by the EU FP7 Project SMARTIE (contract no. 609062). KK acknowledges the support of the DFG Collaborative Research Center SFB 876, project B4.

## References

[1] Charu C. Aggarwal. *Outlier analysis*. Springer, 2013.

[2] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15:1–15:58, 2009.

[3] D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[4] M. Gupta, J. Gao, C. Aggarwal, and J. Han. Outlier detection for temporal data: a survey. *IEEE Transaction on Knowledge and Data Engineering*, 25(1), 2014.

[5] A. Krause, B. McMahan, C. Guestrin, and A. Gupta. Robust submodular observation selection. *Journal of Machine Learning Research (JMLR)*, 9:2761–2801, 2008.

[6] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research (JMLR)*, 9:235–284, 2008.

[7] J. R. Lloyd, D. Duvenaud, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Automatic construction and natural-language description of nonparametric regression models. In *AAAI*, 2014.

[8] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain. Gaussian processes for time series modelling. *Philosophical Transactions of the Royal Society of London A*, 371(1984), 2013.

[9] A. Shah, A. Wilson, and Z. Ghahramani. Student-t processes as alternatives to gaussian processes. In *Proceedings of AISTATS*, 2014.