Data-Driven Anomaly Detection using a Bias Change Model

André Carvalho Bittencourt, Thomas Schön

Introduction

Testing for a Bias Change We model an anomaly as a bias change relative from nominal according to

$$\mathcal{H}^0: \ \mathbf{x} \sim p^0(\mathbf{x}), \quad \mathcal{H}^1: \ \mathbf{x} \sim p^1(\mathbf{x}) = p^0(\mathbf{x} - \boldsymbol{\Delta}).$$

An advantage with this viewpoint is that it is possible to **assess the decision errors** since both hypotheses are defined. For an unknown change $\boldsymbol{\Delta}$, the Generalized Likelihood Ratio Test (GLRT) can be used to determine presence of an anomaly

$$\log \Lambda_N(\mathbf{X}_N) = \log \frac{\max_{\boldsymbol{\Delta}} p_N^0(\mathbf{X}_N - \boldsymbol{\Delta})}{p_N^0(\mathbf{X}_N)} = \sum_{j=1}^{N} \log \frac{p^0(\mathbf{x}_j - \widehat{\boldsymbol{\Delta}}_N)}{p^0(\mathbf{x}_j)} \underset{\mathcal{H}^0}{\overset{\mathcal{H}^1}{\gtrless}} \eta,$$

and reads, accept $\mathcal{H}_0$ if $\Lambda_N(\mathbf{X}_N) < \eta$ otherwise choose $\mathcal{H}_1$. The asymptotic behavior of the test statistic is given as [M. Kay, (2009)]

$$2\log \Lambda(\mathbf{X}|\mathcal{H}^0) \overset{\text{asymp.}}{\sim} \mathcal{X}_d^2, \quad 2\log \Lambda(\mathbf{X}|\mathcal{H}^1) \overset{\text{asymp.}}{\sim} \mathcal{X}_d'^2(\boldsymbol{\Delta}^T \mathbf{F}(\mathbf{0}) \boldsymbol{\Delta})$$

from which it is possible to find a threshold for a desired probability of false P_f and an estimate of P_d using $\widehat{\boldsymbol{\Delta}}_N$. The GLRT is extended to the **sequential** case by considering the test

$$\log \Lambda_n(\mathbf{X}_n) = \sum_{j=1}^{n} \log \frac{p^0(\mathbf{x}_j - \widehat{\boldsymbol{\Delta}}_j)}{p^0(\mathbf{x}_j)} \underset{\mathcal{H}^0}{\overset{\mathcal{H}^1}{\lessgtr}} \eta$$

where $\widehat{\boldsymbol{\Delta}}_j$ is a maximum likelihood estimate found sequentially. The threshold and P_d are also found using the asymp. expressions and $\widehat{\boldsymbol{\Delta}}_n$.

Problem Description and Approach

The density $p^0(\mathbf{x})$ and change $\boldsymbol{\Delta}$ are considered unknown and there is only availability of a **nominal dataset** $\mathbf{X}_{N_0}^0$. In the **first step**, an estimate $\widehat{p}^0(\mathbf{x})$ for $\mathcal{H}^0$ is computed from $\mathbf{X}_{N_0}^0$ with a kernel density estimator as

$$\widehat{p}^0(\mathbf{x}) = \sum_{k \in \mathcal{K}} \pi_k \, \kappa(\mathbf{x}; \mathbf{x}_k^0, \mathbf{h}), \quad \sum_{k \in \mathcal{K}} \pi_k = 1, \quad \pi_k > 0, \quad |\mathcal{K}| = K \leq N_0.$$

A **Gaussian kernel** is used. In the **second step**, a maximum likelihood estimate $\widehat{\boldsymbol{\Delta}}$ of the unknown change is found. The estimate $\widehat{p}^0(\mathbf{x})$ and $\widehat{\boldsymbol{\Delta}}$ are used to define the approximate models

$$\mathcal{H}^0: \ \mathbf{x}_i \sim \widehat{p}^0(\mathbf{x}), \quad \mathcal{H}^1: \ \mathbf{x}_i \sim \widehat{p}^1(\mathbf{x}|\widehat{\boldsymbol{\Delta}}) = \widehat{p}^0(\mathbf{x} - \widehat{\boldsymbol{\Delta}})$$

which are used in a GLRT assuming it is true.

First Step - A Sparse KDE

We use a sparse density estimator which does not require specification of the bandwidth $\mathbf{h}$ or number of components K in the mixture. For the dataset $\mathbf{X}_{N_0}^0$, the generalized cross entropy method [Botev, (2011)] gives the estimate $\widehat{p}^0(\mathbf{x}) = \sum_{k \in \mathcal{K}_\epsilon} \pi_k^* \kappa(\mathbf{x}; \mathbf{x}_k^0, \mathbf{h}^*)$ with $(\mathbf{h}^*, \boldsymbol{\lambda}^*)$ given by

$$\left\{(\mathbf{h}, \boldsymbol{\lambda}) : \mathbf{1}^T \boldsymbol{\lambda}(\mathbf{h}) = 1, \ \boldsymbol{\lambda}(\mathbf{h}) = \arg\min_{\boldsymbol{\lambda} \geq 0} \ \boldsymbol{\lambda}^T C(\mathbf{h}) \boldsymbol{\lambda} - \boldsymbol{\lambda}^T \widehat{\phi}_i(\mathbf{h})\right\}$$

where the quadratic program for $\boldsymbol{\lambda}(\mathbf{h})$ is defined by

$$\widehat{\phi}_i(\mathbf{h}) = \frac{1}{N_0 - 1} \sum_{j \neq i} \kappa(\mathbf{x}_j^0; \mathbf{x}_i^0, \mathbf{h}), \ C_{ij}(\mathbf{h}) = \int_{\mathbb{R}^d} \kappa(\mathbf{x}; \mathbf{x}_i^0, \mathbf{h}) \kappa(\mathbf{x}; \mathbf{x}_j^0, \mathbf{h}) \, \mathrm{d}\mathbf{x},$$

and $C(\mathbf{h}) \in \mathbb{R}^{N_0 \times N_0}$ is positive definite by construction. For a Gaussian kernel C_{ij} can be found analytically. Small components in $\boldsymbol{\lambda}^*$ are removed using a **pruning** approach. Let $\boldsymbol{\lambda}^*$ be ordered as $\lambda_1^* \leq \lambda_2^* \leq \ldots \leq \lambda_{N_0}^*$, the ϵ approximation is written as $\widehat{p}^0(\mathbf{x}) = \sum_{k \in \mathcal{K}_\epsilon} \pi_k^* \kappa(\mathbf{x}; \mathbf{x}_k^0, \mathbf{h}^*)$ with

$$\pi_k^* \triangleq \frac{\lambda_k^*}{\sum_{j \in \mathcal{K}_\epsilon} \lambda_j^*}, \qquad \mathcal{K}_\epsilon : \{k : \sum_{j=1}^{k} \lambda_j^* \geq \epsilon, 1 \leq k \leq N_0\},$$

where $|\mathcal{K}_\epsilon| = K$ will typically be much less than N_0.

Second Step - Estimate of Δ using EM

The approximate model for $\mathcal{H}^1$ can be written as

$$\widehat{p}^1(\mathbf{x}|\boldsymbol{\Delta}) = \widehat{p}^0(\mathbf{x} - \boldsymbol{\Delta}) = \sum_{k=1}^{K} \pi_k \kappa_k(\mathbf{x} - \boldsymbol{\Delta}), \quad \kappa_k(\mathbf{x}) \triangleq \kappa(\mathbf{x}; \mathbf{x}_k^0, \mathbf{h})$$

which corresponds to a **mixture model** with the parameter $\boldsymbol{\Delta}$ common to each component. An estimate of $\boldsymbol{\Delta}$ is found using the Expectation Maximization (EM) algorithm. For a Gaussian kernel, an estimate $\widehat{\boldsymbol{\Delta}}_N$ is found after convergence of the iterates $\boldsymbol{\Delta}_i$ below

$$\boldsymbol{\Delta}_{i+1} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \zeta_{nk}(\boldsymbol{\Delta}_i) \left(\mathbf{x}_n - \mathbf{x}_k^0\right), \ \zeta_{nk}(\boldsymbol{\Delta}_i) \triangleq \frac{\pi_k \kappa_k(\mathbf{x}_n|\boldsymbol{\Delta}_i)}{\sum_{j=1}^{K} \pi_j \kappa_j(\mathbf{x}_n|\boldsymbol{\Delta}_i)}$$

A sequential estimate $\widehat{\boldsymbol{\Delta}}_n$ can be found based on a stochastic approximation [Cappé, (2009)]

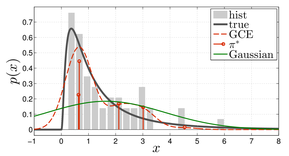
$$\widehat{\boldsymbol{\Delta}}_n = \gamma_n \sum_{k=1}^{K} \zeta_{nk}(\widehat{\boldsymbol{\Delta}}_{n-1}) \left(\mathbf{x}_n - \mathbf{x}_k^0\right) + (1 - \gamma_n) \widehat{\boldsymbol{\Delta}}_{n-1}$$

Note that a sparse density model (small K) can considerably **reduce the amount of computations** needed to find the estimates.

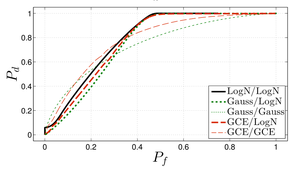
Miss Specification of Hypotheses Densities

Consider the problem of detecting a bias change of size $\Delta = 1$ in a log-normal distribution, i.e.

$$\begin{aligned} \mathcal{H}^0: \quad & p^0(x) = \log \mathcal{N}(x; 0, 1) \\ \mathcal{H}^1: \quad & p^1(x) = p^0(x - \Delta) \\ & \quad = \log \mathcal{N}(x - \Delta; 0, 1). \end{aligned}$$

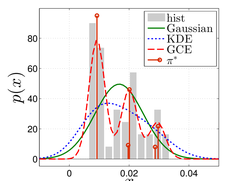


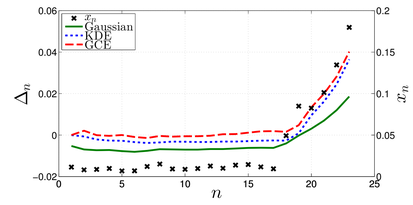
A nominal dataset is used with $N_0 = 50$ and Δ is assumed known. We apply a likelihood ratio test for 1 sample based on the true density (optimal), a Gaussian assumption and the GCE estimate.

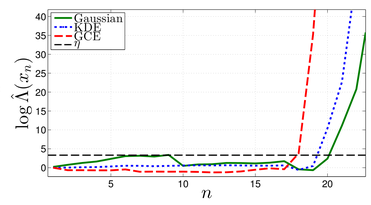


Sequential Anomaly Detection in an IRB

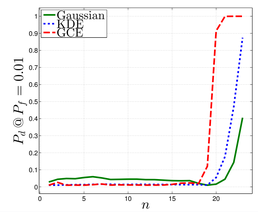
By processing torque data collected from an industrial robot joint, a scalar quantity x is generated to infer the mechanical condition of the joint gearbox. The data processing used in the generation of x makes it difficult to determine its distribution function.







$N_0 = 45$ nominal samples are used and three density models are considered. Δ is estimated sequentially in the data and GLRTs are applied. The threshold and P_d are found using the asymp. expressions for $P_f = 1\%$.



Summary

Data-driven approaches for anomaly detection using a bias change model.

- Only requires availability of nominal data
- Use of a kernel density estimate gives flexibility
- The bias change is found using EM algorithm
- The sparse density estimate used reduces amount of computations
- Approximate GLRTs used for both batch and sequential cases
- Choice of threshold, P_f and P_d estimated from asymp. expressions