Data-Driven Anomaly Detection based on a Bias Change Model ⋆

André Carvalho Bittencourt * **and Thomas B. Schön** **

* *Department of Electrical Engineering, Linköping University, Sweden*
andrecb@isy.liu.se
** *Department of Information Technology, Uppsala University, Sweden*
thomas.schon@it.uu.se

Abstract: This paper proposes off-line and on-line data-driven approaches to anomaly detection based on generalized likelihood ratio tests for a bias change model. The procedure is divided into two steps. Assuming availability of a nominal dataset, a nonparametric density estimate is obtained in the first step, prior to the test. Second, the unknown bias change is estimated from test data. Based on the expectation maximization (EM) algorithm, batch and sequential maximum likelihood estimators of the bias change are derived for the case where the density estimate is given by a Gaussian mixture. Asymptotic expressions for the probabilities of error are suggested based on available results. Real world experiments illustrate the approach.

Keywords: Fault detection and diagnosis; Nonparametric methods; Estimation and filtering

1. INTRODUCTION

In anomaly detection, the main objective is to determine whether observations conform to expected (normal) behavior or not (i.e. an anomaly). Anomaly detection appears in a variety of applications, such as condition monitoring of machines, fraud detection, intrusion detection, etc. Chandola et al. (2009) provide a survey of anomaly detection. A factor that distinguishes anomaly detection to related detection problems is the lack of knowledge of the anomaly. This is a rather common situation, e.g. in condition monitoring and fault detection. A mathematical model is a common description of the available knowledge. However, it may be difficult to determine such model a priori in some applications. A more common situation is perhaps that it is possible to collect measurements (data) under normal conditions. This nominal dataset contains relevant information about the conforming behavior and it is possible to infer an anomaly based only on nominal data.

Examples of data-driven approaches to anomaly detection are one-class classification algorithms, e.g. Devroye and Wise (1980); Schölkopf et al. (2001), where a boundary region in the observation space is determined from a nominal dataset. Fresh observations falling outside this region are classified as anomalies. A shortcoming with such an approach is that all knowledge about the normal behavior is summarized by a region in the observation space. For instance, this approach would fail to recognize that if observations consistently fall in a low probability region of the support, it is more likely that an anomaly is present. An alternative is to estimate a model of the measurements density based on the nominal data. In this case, anomalies can be detected based on the probability that test data has under the estimated density model. Since it is often difficult to determine the family of distributions, mixture models are commonly used, e.g. Agarwal (2007), as well as nonparametric estimates (Desforges et al., 1998; Yeung and Chow, 2002). A shortcoming with approaches based on a model solely for the normal behavior is that it is not possible to provide an estimate of how certain the test is in the presence of an anomaly. This type of information is however often important in practice to support decisions of recovery actions.

With the possibility of determining probabilistic models for both the normal and abnormal behaviors, anomaly detection can be seen as a hypothesis testing problem (HTP). In a HTP, it is possible to quantify the decision uncertainties since probabilistic models are defined for the entire problem. In a binary HTP, the null hypothesis $\mathcal{H}^0$ describes the nominal behavior and the alternative hypothesis $\mathcal{H}^1$ describes the abnormal behavior. The hypotheses are described by the statistical behavior of the measurements $\mathbf{y} \in \mathbb{R}^d$ under each hypothesis,

$$\mathcal{H}^0 : \mathbf{y} \sim p^0(\mathbf{y}), \quad \mathcal{H}^1 : \mathbf{y} \sim p^1(\mathbf{y}). \tag{1}$$

When the hypotheses densities, $p^0(\mathbf{y})$ and $p^1(\mathbf{y})$, are given or when their family of parametric distributions are known, there are well-established statistical tests based on *likelihood ratios*, i.e. $\Lambda(\mathbf{y}) \triangleq p^1(\mathbf{y})/p^0(\mathbf{y})$, (Neyman and Pearson, 1933; Wald, 1945).

An approach to overcome the lack of knowledge for the anomaly is to define it as a change *relative to nominal*. In this manner, the available knowledge about the nominal behavior can be used to test for an anomaly. Here, a **bias** (location) change is considered, i.e. the density for the alternative hypothesis is written as $p^1(\mathbf{y}) = p^0(\mathbf{y} - \boldsymbol{\Delta})$, for a bias change $\boldsymbol{\Delta}$. Using this model, this article aims at providing an approach for anomaly detection that without

⋆ This work was supported by ABB and the Vinnova Industry Excellence Center LINK-SIC at Linköping University.

requiring specification of a density function and based only on availability of a nominal dataset,

- is flexible and can be used for different problems,
- can provide estimates of the decision uncertainties,
- requires only minimal and meaningful specification parameters from the user.

This is achieved with a two step approach. First, the nominal dataset is used to find a nonparametric estimate of the density function for $\mathcal{H}^0$, denoted $\widehat{p}^0(\mathbf{y})$. In the second step, incoming test measurements are used to find a maximum likelihood estimate $\widehat{\boldsymbol{\Delta}}$ of the unknown bias change. These estimates are used to define the approximate model

$$\mathcal{H}^0 : \mathbf{y}_i \sim \widehat{p}^0(\mathbf{y}), \quad \mathcal{H}^1 : \mathbf{y}_i \sim \widehat{p}^1(\mathbf{y}|\widehat{\boldsymbol{\Delta}}) = \widehat{p}^0(\mathbf{y} - \widehat{\boldsymbol{\Delta}}), \quad (2)$$

which is tested based on a generalized likelihood ratio test (GLRT) assuming this model to be true. Both on-line and off-line tests are devised.

The presentation is organized as follows, Sec. 2 presents the bias change model and reviews the GLRT. Sec. 3 presents the approaches used to find the estimate $\widehat{p}^0(\mathbf{y})$ based on a nominal dataset. The resulting density model will be a finite mixture distribution. Sec. 4 defines maximum likelihood estimators for $\boldsymbol{\Delta}$ based on the Expectation Maximization algorithm. Algorithms are derived for mixtures of multivariate Gaussian distributions. The use of GLRTs based on the approximate models (2) is illustrated in Sec. 5 through real data examples followed by concluding remarks.

2. THE BIAS CHANGE MODEL AND GLRT

The assumption that an anomaly will appear as a bias change from nominal gives the following hypotheses

$$\mathcal{H}^0 : \mathbf{y} \sim p^0(\mathbf{y}), \quad \mathcal{H}^1 : \mathbf{y} \sim p^1(\mathbf{y}|\boldsymbol{\Delta}) = p^0(\mathbf{y} - \boldsymbol{\Delta}), \quad (3)$$

for the unknown bias vector $\boldsymbol{\Delta}$. The expected value of $\mathbf{y}$ under $\mathcal{H}^1$ can be written as

$$\mathbb{E}_{p^1}[\mathbf{y}] = \mathbb{E}_{p^0}[\mathbf{y}] + \boldsymbol{\Delta}, \quad (4)$$

i.e., it changes the mean of $\mathbf{y}$ by $\boldsymbol{\Delta}$. This model is easy to interpret and bias changes are often considered when detecting anomalies, e.g. in the literature of fault diagnosis (Isermann, 2006). The model describes situations where the data is shifted in the observation space. The parameter $\boldsymbol{\Delta}$ also carries valuable information about the problem. E.g. if Σ is the density covariance, then $\boldsymbol{\Delta}\Sigma^{-1}\boldsymbol{\Delta}^T$ measures the significance of the change relative to the density volume, similar to a signal to noise ratio.

For N independent and identically distributed (i.i.d.) measurements $\mathbf{Y}_N = [\mathbf{y}_1, \cdots, \mathbf{y}_N]$, the objective is to decide whether $\mathbf{Y}_N$ belongs to $\mathcal{H}^0$ or $\mathcal{H}^1$ in (3). This can be done with a generalized likelihood ratio test (GLRT)

$$\begin{aligned} \log \Lambda_N(\mathbf{Y}_N) &= \log \frac{\max_{\boldsymbol{\Delta}} p_N^0(\mathbf{Y}_N - \boldsymbol{\Delta})}{p_N^0(\mathbf{Y}_N)} \\ &= \sum_{j=1}^{N} \log \frac{p^0(\mathbf{y}_j - \widehat{\boldsymbol{\Delta}}_N)}{p^0(\mathbf{y}_j)} \underset{\mathcal{H}^0}{\overset{\mathcal{H}^1}{\gtrless}} \eta, \end{aligned} \quad (5)$$

where $\widehat{\boldsymbol{\Delta}}_N$ is a maximum likelihood (ML) estimate of the unknown bias. The above notation means that $\mathcal{H}_0$ is chosen if the test statistic $\log \Lambda_N(\mathbf{Y}_N)$ is smaller than the threshold η otherwise $\mathcal{H}_1$ is chosen. This test can be extended to the sequential case by considering

$$\log \Lambda_n(\mathbf{y}_n) = \sum_{j=1}^{n} \log \frac{p^0(\mathbf{y}_j - \widehat{\boldsymbol{\Delta}}_j)}{p^0(\mathbf{y}_j)} \underset{\mathcal{H}^0}{\overset{\mathcal{H}^1}{\gtrless}} \eta, \quad (6)$$

where $\widehat{\boldsymbol{\Delta}}_j$ is a ML estimate computed sequentially.

Associated to any test is the probability of deciding incorrectly for $\mathcal{H}^0$, denoted β, and the probability of deciding incorrectly for $\mathcal{H}^1$, denoted α. For a GLRT they are given by $\beta = \int_{-\infty}^{\eta} \log \Lambda(\mathbf{y}|\mathcal{H}^1) d\mathbf{y}$ and $\alpha = \int_{\eta}^{\infty} \log \Lambda(\mathbf{y}|\mathcal{H}^0) d\mathbf{y}$. While in general no analytical solution is available, they can in principle be found based on Monte Carlo techniques. This is however difficult, in particular for the sequential case, when it will depend on the sequence $\widehat{\boldsymbol{\Delta}}_n$. An alternative is to find α and β based on the asymptotic behavior of the GLRT statistic. The asymptotic behavior of the test statistic is given by (Mackay, 2003, Appendices 6A-C)

$$2 \log \Lambda(\mathbf{y}|\mathcal{H}^0) \overset{\text{as.}}{\sim} \mathcal{X}_d^2, \quad (7a)$$

$$2 \log \Lambda(\mathbf{y}|\mathcal{H}^1) \overset{\text{as.}}{\sim} \mathcal{X}_d'^2(\lambda(\boldsymbol{\Delta}^1)), \; \lambda(\boldsymbol{\Delta}) \triangleq \boldsymbol{\Delta}^T \mathbf{F}(0) \boldsymbol{\Delta} \quad (7b)$$

where $\boldsymbol{\Delta}^1$ is the true parameter under $\mathcal{H}^1$, $\mathcal{X}_d^2$ is the chi-square distribution with d degrees of freedom, $\mathcal{X}_d'^2(\lambda)$ is the non-central chi-square with non-centrality parameter λ and $\mathbf{F}(\mathbf{0})$ is the Fisher information for $\boldsymbol{\Delta}$ evaluated at zero. This result is valid whenever $\widehat{\boldsymbol{\Delta}}$ tends to the true value $\boldsymbol{\Delta}^1$. Since under $\mathcal{H}^0$ the asymptotic behavior of the test statistic does not depend on unknowns, a threshold can be found from (7b) for a desired value of α. An estimate of β can also be computed based on the maximum likelihood estimate $\widehat{\boldsymbol{\Delta}}$. For a desired level α' this is summarized in the equations below

$$\eta(\alpha') = \inf_{\eta} \left\{ \int_{\eta}^{\infty} \mathcal{X}_d^2 \geq \alpha' \right\}, \; \beta(\alpha') = \int_{-\infty}^{\eta(\alpha')} \mathcal{X}_d'^2(\lambda(\widehat{\boldsymbol{\Delta}})). \quad (8)$$

To apply the GLRT for the bias change model, the unknown density $p^0(\mathbf{y})$ is needed. In a practical setup, it is often common to introduce assumptions on the data distribution, the Gaussian model being a common choice. Although the Gaussian model gives statistical tests that can be conveniently described by sufficient statistics (Van Trees and Kristine, 2013), it is clear that there will be situations where this model is a poor description of $\mathcal{H}^0$. In this paper, no assumption is forced about $\mathcal{H}^0$, instead, all knowledge is considered to be contained in a nominal dataset and data-driven approaches are sought.

3. NONPARAMETRIC DENSITY ESTIMATORS

A nominal dataset $\mathbf{Y}_{N_0}^0$ with N_0 i.i.d. observations from $\mathcal{H}^0$ is used to find a nonparametric density estimate $\widehat{p}^0(\mathbf{y})$. The density model will take the form of a finite mixture model

$$\widehat{p}^0(\mathbf{y}) = \sum_{k \in \mathcal{K}} \pi_k \, \kappa(\mathbf{y}; \mathbf{y}_k^0, \mathbf{h}), \; \sum_{k \in \mathcal{K}} \pi_k = 1, \; \pi_k > 0 \quad (9)$$

where $\mathcal{K}$ is an index set with cardinality $|\mathcal{K}| = K \leq N_0$, $\kappa(\cdot)$ is a kernel function satisfying $\kappa(\mathbf{y}) \geq 0$ and $\int \kappa(\mathbf{y}) \, \mathrm{d}\mathbf{y} = 1$. The bandwidth $\mathbf{h} \in \mathbb{R}^d$ is fixed and the weighting coefficients $\{\pi_k\}$ are found according to the chosen density estimator. Two nonparametric density estimators are presented next.

3.1 Kernel Density Estimator

The first type of estimator considered is a so called kernel density estimator (KDE), or Parzen estimator. The KDE based on the nominal dataset $\mathbf{Y}_{N_0}^0$ is given by a finite mixture model (9) with

$$\mathcal{K} = \{1, 2, \ldots, N_0\}, \quad \pi_k = \tfrac{|(\mathbf{h})|^{-1/2}}{N_0}, \tag{10a}$$

$$\kappa(\mathbf{y}; \mathbf{y}_k^0, \mathbf{h}) = \kappa\left(S(\mathbf{h})^{-1/2}\left(\mathbf{y} - \mathbf{y}_k^0\right)\right), \tag{10b}$$

where $S(\mathbf{h})$ is a positive definite scaling matrix. The KDE model has as many components as data points and the coefficients $\{\pi_k\}$ are fixed and identical. As shown by Parzen (1962); Cacoullos (1966), this estimator is consistent and asymptotically unbiased. The KDE method requires specification of the bandwidth $\mathbf{h}$. There are several approaches reported in the literature for bandwidth selection (Jones et al., 1996b,a). Here, a diagonal $S(\mathbf{h})$ will be considered with bandwidth elements chosen using Silverman's rule of thumb (Silverman, 1986),

$$S(\mathbf{h}) = \operatorname{diag}(\mathbf{h}), \quad \sqrt{h_j} = \tfrac{4}{d+2}^{\frac{1}{d+4}} N_0^{\frac{-1}{d+4}} \widehat{\sigma}_j, \tag{11}$$

for $j = \{1, \ldots, d\}$ and where $\widehat{\sigma}_j$ is an estimate of the data standard deviation over the jth dimension.

Besides requiring storage of the entire dataset, performing inference with a KDE will become computationally intensive when N_0 is large. An alternative is to consider reduced mixture models, with $K \ll N_0$ components. When the number of components K is fixed, it is possible to find maximum likelihood estimates for the parameters using, e.g., the EM algorithm (Dempster et al., 1977). A disadvantage with such an approach is that the number of components K must be pre-specified.

3.2 A Sparse Density Estimator

An alternative will be considered here based on the generalized cross entropy (GCE) method presented by Botev and Kroese (2011), which does not require specification of K or $\mathbf{h}$. For a dataset $\mathbf{Y}_{N_0}^0$, the estimate is given as

$$\widehat{p}^0(\mathbf{y}) = \sum_{k \in \mathcal{K}} \lambda_k^* \kappa(\mathbf{y}; \mathbf{y}_k^0, \mathbf{h}^*). \tag{12}$$

with $\mathcal{K} = \{1, \ldots, N_0\}$ and where the bandwidth h^* and weights λ_k^* are given by

$$(\mathbf{h}^*, \boldsymbol{\lambda}^*) = \Big\{(\mathbf{h}, \boldsymbol{\lambda}) : \mathbf{1}^T \boldsymbol{\lambda}(\mathbf{h}) = 1,$$

$$\boldsymbol{\lambda}(\mathbf{h}) = \arg\min_{\boldsymbol{\lambda} \geq 0} \boldsymbol{\lambda}^T C(\mathbf{h}) \boldsymbol{\lambda} - \boldsymbol{\lambda}^T \widehat{\phi}_i(\mathbf{h})\Big\}, \tag{13a}$$

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

$$\widehat{\phi}_i(\mathbf{h}) = \frac{1}{N_0 - 1} \sum_{j \neq i} \kappa(\mathbf{y}_j^0; \mathbf{y}_i^0, \mathbf{h}), \tag{13b}$$

$$C_{ij}(\mathbf{h}) = \int_{\mathbb{R}^d} \kappa(\mathbf{y}; \mathbf{y}_i^0, \mathbf{h}) \kappa(\mathbf{y}; \mathbf{y}_j^0, \mathbf{h}) \, \mathrm{d}\mathbf{y}, \tag{13c}$$

and $C(\mathbf{h}) \in \mathbb{R}^{N_0 \times N_0}$ is positive definite by construction.

This approach is algorithmically similar to the support vector density estimator by Vapnik and Mukherjee (2000), in which the condition $\mathbf{1}^T \boldsymbol{\lambda}(\mathbf{h}) = 1$ is included as a constraint in the QP and $\mathbf{h}$ is pre-specified. As noted by Botev and Kroese (2011), the QP in (13a) is closely related to the support vector regression problem with an ϵ-insensitive error function, see e.g. Bishop (2006), and most elements in $\boldsymbol{\lambda}^*$ will usually be close to zero.

To remove small components in $\boldsymbol{\lambda}^*$, a pruning approach is suggested here. Let $\boldsymbol{\lambda}^*$ be ordered as $\lambda_1^* \leq \lambda_2^* \leq \ldots \lambda_{N_0}^*$, the ϵ approximation of (12) is written by replacing $\mathcal{K}$ and λ_k^* in (12) with $\mathcal{K}_\epsilon$ and π_k^* respectively, where

$$\mathcal{K}_\epsilon : \Big\{k : \sum_{j=1}^{k} \lambda_j^* \geq \epsilon, 1 \leq k \leq N_0\Big\}, \; \pi_k^* \triangleq \frac{\lambda_k^*}{\sum_{j \in \mathcal{K}_\epsilon} \lambda_j^*}, \tag{14}$$

and $|\mathcal{K}_\epsilon| = K$ will typically be much smaller than the number of data samples N_0.

The GCE method requires solution of $C_{ij}(\mathbf{h})$ in (13c), which is not always analytically tractable. For the Gaussian case, i.e. $\kappa(\mathbf{y}; \mathbf{y}_k^0; S(\mathbf{h})) = \mathcal{N}(\mathbf{y}; \mathbf{y}_k^0, S(\mathbf{h}))$, it can be shown from completion of the squares that

$$C_{ij}(\mathbf{h}) = \mathcal{N}(\mathbf{y}_i^0; \mathbf{y}_j^0, 2S(\mathbf{h})). \tag{15}$$

To avoid solving (13a) for a d-dimensional $\mathbf{h}$, a simplification is made which considers a scalar bandwidth h applied to a diagonal covariance estimate, i.e.

$$S(h) = h\widehat{\Sigma}, \quad \text{where } \widehat{\Sigma}_{ij} = \begin{cases} 0, \text{if } i \neq j \\ \widehat{\sigma}_i^2, \text{if } i = j \end{cases} \tag{16}$$

where $i, j \in \{1, \ldots, d\}$. In this manner, only one bandwidth parameter needs to be found and different scaling is allowed for the different dimensions. The resulting problem (13a) is solved by addressing the nonlinear least squares

$$h^* = \arg\min_h \left(\mathbf{1}^T \boldsymbol{\lambda}(h) - 1\right)^2, \tag{17}$$

where $\boldsymbol{\lambda}(h)$ is the solution to the QP in (13a) with $C(\mathbf{h})$ found from (15) and $S(\mathbf{h})$ given by (16). The optimal weights are $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}(h^*)$ and the model of (12) is approximated as in (14).

4. ESTIMATING THE BIAS CHANGE

For $\widehat{p}^0(\mathbf{y})$ achieved using either the KDE or the GCE methods, the model for the alternative hypothesis in (2) can be written as the finite mixture

$$\widehat{p}^1(\mathbf{y}|\boldsymbol{\Delta}) = \widehat{p}^0(\mathbf{y} - \boldsymbol{\Delta}) = \sum_{k=1}^{K} \pi_k \kappa_k(\mathbf{y} - \boldsymbol{\Delta}), \tag{18}$$

where $\kappa_k(\mathbf{y}) \triangleq \kappa(\mathbf{y}; \mathbf{y}_k^0, \mathbf{h})$. The objective of this section is to derive maximum likelihood estimators of $\boldsymbol{\Delta}$ in (18) based on a measurement batch $\mathbf{Y}_N$ and on a measurement sequence $\{\mathbf{y}_n\}$. First, notice that for a mixture density $p(\mathbf{y})$, $\mathbb{E}_p[\mathbf{y}] = \sum_{k=1}^{K} \pi_k \mathbb{E}_{\kappa_k}[\mathbf{y}]$. Using this relation with (4) an estimate of Δ can be computed based on $\mathbf{Y}_N$ from the sample estimate

$$\widehat{\boldsymbol{\Delta}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i - E_{\widehat{p}^0}[\mathbf{y}] = \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i - \sum_{k=1}^{K} \pi_k \mathbb{E}_{\kappa_k}[\mathbf{y}] \tag{19}$$

This estimate is asymptotically unbiased. However, for a given sample $\mathbf{Y}_N$, it does not necessarily maximizes the likelihood function (e.g. if the density has multiple modes) and an alternative is needed. It is well known that direct optimization of the likelihood function in mixture models is problematic (Bishop, 2006). For mixtures, the Expectation Maximization (EM) algorithm can be used to provide maximum likelihood estimates.

4.1 Off-line Estimation using EM

The EM algorithm (Dempster et al., 1977), is a two step iterative procedure for finding maximum likelihood parameter estimates in probabilistic models involving latent variables. Let $\mathbf{X}$ and $\mathbf{Y}$ denote latent and measured variables, with joint distribution $p(\mathbf{Y},\mathbf{X}|\boldsymbol{\theta})$ governed by the parameter vector $\boldsymbol{\theta}$ and let

$$\begin{aligned}\mathcal{Q}\left(\boldsymbol{\theta},\boldsymbol{\theta}'\right) &\triangleq \int \ln p(\mathbf{Y},\mathbf{X}|\boldsymbol{\theta}) p(\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}')\,\mathrm{d}\mathbf{X} \\ &= \mathbb{E}_{\boldsymbol{\theta}'}[\ln p(\mathbf{Y},\mathbf{X}|\boldsymbol{\theta})|\mathbf{Y}].\end{aligned} \tag{20}$$

For iterates $\boldsymbol{\theta}_i$, the expectation (20) is computed for $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}_{i-1})$ in the E-step. In the M-step, the resulting $\mathcal{Q}$-function is maximized w.r.t. $\boldsymbol{\theta}$ to update the iterate $\boldsymbol{\theta}_i$. The steps are repeated until a convergence criterion is satisfied. The EM algorithm guarantees that the iterates satisfy $p(\mathbf{Y}|\boldsymbol{\theta}_i) \geq p(\mathbf{Y}|\boldsymbol{\theta}_{i-1})$ and therefore they eventually converge to a stationary point of the likelihood function. For a measurement batch $\mathbf{Y}_N$, the estimate achieved after convergence of the algorithm is denoted as $\widehat{\boldsymbol{\theta}}_N$.

As previously noted, the model (18) can be interpreted as a mixture model, where the parameter $\boldsymbol{\theta}=\boldsymbol{\Delta}$ is common to all mixture components. Hence, the model (18) can be written as $\widehat{p}^1(\mathbf{y}|\theta)=\sum_{k=1}^K \pi_k\kappa_k(\mathbf{y}|\theta)$. Introducing a discrete latent variable to denote which of the K components that generated a certain measurement $\mathbf{y}_n$, it is possible to show that the E-step amounts to (Bishop, 2006)

$$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta'}) = \sum_{n=1}^N\sum_{k=1}^K \zeta_{nk}(\boldsymbol{\theta}')\log\pi_k\kappa_k(\mathbf{Y}_N|\boldsymbol{\theta}), \tag{21}$$

$$\zeta_{nk}(\boldsymbol{\theta}') \triangleq \frac{\pi_k\kappa_k(\mathbf{Y}_N|\boldsymbol{\theta}')}{\sum_{j=1}^K \pi_j\kappa_j(\mathbf{Y}_N|\boldsymbol{\theta}')}. \tag{22}$$

The solution to the M-step depends on the form of the kernel function and on how the unknown parameters enter this function. Explicit solutions are given next for the Gaussian mixture model (GMM) with

$$\boldsymbol{\theta}=\boldsymbol{\Delta}, \quad \kappa_k(\mathbf{y}|\boldsymbol{\theta})=\kappa_k(\mathbf{y}-\boldsymbol{\Delta})=\mathcal{N}(\mathbf{y}-\boldsymbol{\Delta};\mathbf{y}_k^0,S),. \tag{23}$$

The M-step can be found explicitly by finding the solution to $\frac{\partial}{\partial\boldsymbol{\Delta}}\mathcal{Q}\left(\boldsymbol{\Delta},\boldsymbol{\Delta}'\right)=0$. This gradient is given by

$$\sum_{n=1}^N\sum_{k=1}^K \zeta_{nk}(\boldsymbol{\Delta}')\left[\frac{\partial}{\partial\boldsymbol{\Delta}}\log\kappa_k(\boldsymbol{x}_n-\boldsymbol{\Delta})\right]$$

the term in brackets simplifies to $S^{-1}[(\mathbf{y}_n-\mathbf{y}_k^0)-\boldsymbol{\Delta}]$ giving

$$\begin{aligned}\boldsymbol{\Delta} &= \frac{\sum_{n=1}^N\sum_{k=1}^K \zeta_{nk}(\boldsymbol{\Delta}')\left(\mathbf{y}_n-\mathbf{y}_k^0\right)}{\sum_{n=1}^N\sum_{k=1}^K \zeta_{nk}(\boldsymbol{\Delta}')} \\ &= \frac{1}{N}\sum_{n=1}^N\sum_{k=1}^K \zeta_{nk}(\boldsymbol{\Delta}')\left(\mathbf{y}_n-\mathbf{y}_k^0\right)\end{aligned} \tag{24}$$

where the last step follows since $\sum_{k=1}^K\zeta_{nk}(\boldsymbol{\Delta}')=1$. The resulting iterates $\boldsymbol{\Delta}_i$ produced from the EM algorithm are given in Algorithm 1 for a convergence criterion based on $\|\boldsymbol{\Delta}_i-\boldsymbol{\Delta}_{i-1}\|_2^2$. The algorithm can be initialized using (19), which for the GMM gives

$$\boldsymbol{\Delta}_0=\frac{1}{N}\sum_{i=1}\mathbf{y}_i-\sum_{k=1}^K\pi_k\mathbf{y}_k^0. \tag{25}$$

Algorithm 1 Batch EM for bias change in GMM

Set $i=1$, $\boldsymbol{\Delta}_{i-1}$ as (19) and $\epsilon>0$.
repeat
  E-Step: compute $\zeta_{nk}(\boldsymbol{\Delta}_{i-1})$ in (22)
  M-step: set $\boldsymbol{\Delta}_i$ according to (24)
until $\|\boldsymbol{\Delta}_i-\boldsymbol{\Delta}_{i-1}\|_2^2\leq\epsilon$
return $\widehat{\boldsymbol{\Delta}}_N=\boldsymbol{\Delta}_i$ {Return the estimate}

Algorithm 2 Sequential EM for bias change in GMM

Set $n=1$, $\widehat{\boldsymbol{\Delta}}_{n-1}$, $\gamma_0\in(0,1)$ and $\rho\in(\frac{1}{2},1]$
for all incoming $\mathbf{y}_n$ **do**
  E-Step: compute $\zeta_{nk}(\widehat{\boldsymbol{\Delta}}_{n-1})$ in (22)
  M-step: set $\gamma_n=\gamma_0 n^{-\rho}$ and $\widehat{\boldsymbol{\Delta}}_n$ according to (27)
end for

4.2 On-line Estimation using a Sequential version of EM

To evaluate the E-step in the EM algorithm, all measurements in $\mathbf{Y}$ must be available and the EM algorithm is therefore an off-line approach. A sequential version of EM was suggested in Cappé and Moulines (2009), based on a stochastic approximation of the E-step according to,

$$\widetilde{\mathcal{Q}}_n(\boldsymbol{\theta}) = \gamma_n\mathbb{E}_{\widehat{\boldsymbol{\theta}}_{n-1}}[\ln p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})|\mathbf{y}_n]+(1-\gamma_n)\widetilde{\mathcal{Q}}_{n-1}(\boldsymbol{\theta}), \tag{26}$$

where γ_n is the step-size, controlling the adaptation rate to incoming measurements. The M-step is unchanged and the estimate $\widehat{\boldsymbol{\theta}}_n$ is taken as the maximum of the $\widetilde{\mathcal{Q}}$-function. Consistency and convergence rate for the estimator (26) are studied in Cappé and Moulines (2009). For consistency, γ_n must be chosen such that $0<\gamma_n<1$, $\sum_{j=1}^\infty\gamma_j=\infty$ and $\sum_{j=1}^\infty\gamma_j^2<\infty$. To satisfy these conditions, the authors suggest the use of $\gamma_n=\gamma_0 n^{-\rho}$ for $\gamma_0\in(0,1)$ and $\rho\in(\frac{1}{2},1]$. The particular choice $\gamma_0=\rho=1$ is equivalent to the recursion of Equation 12 in Titterington (1984). For mixture models, (26) follows as

$$\widetilde{\mathcal{Q}}(\boldsymbol{\theta}) = \gamma_n\sum_{k=1}^K\zeta_{nk}(\widehat{\boldsymbol{\theta}}_{n-1})\log\pi_k\kappa_k(\mathbf{y}_n|\boldsymbol{\theta}) + (1-\gamma_n)\widetilde{\mathcal{Q}}_{n-1}(\boldsymbol{\theta})$$

where $\zeta_{nk}(\cdot)$ is evaluated at the previous estimate $\widehat{\boldsymbol{\theta}}_{n-1}$. For a GMM, a recursive solution to the M-step can be found. Starting with $\widetilde{\mathcal{Q}}_0(\boldsymbol{\Delta})=0$ and an initial $\widehat{\boldsymbol{\Delta}}_0$, direct maximization of $\widetilde{\mathcal{Q}}_1(\boldsymbol{\Delta}),\widetilde{\mathcal{Q}}_2(\boldsymbol{\Delta}),\ldots,\widetilde{\mathcal{Q}}_n(\boldsymbol{\Delta})$, for a sequence $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n$ gives

$$\begin{aligned}\widehat{\boldsymbol{\Delta}}_1 &= \sum_{k=1}^K\zeta_{1k}(\widehat{\boldsymbol{\Delta}}_0)(\mathbf{y}_1-\mathbf{y}_k^0), \\ \widehat{\boldsymbol{\Delta}}_2 &= \gamma_2\sum_{k=1}^K\zeta_{2k}(\widehat{\boldsymbol{\Delta}}_1)(\mathbf{y}_2-\mathbf{y}_k^0)+(1-\gamma_2)\widehat{\boldsymbol{\Delta}}_0,\ \cdots, \\ \widehat{\boldsymbol{\Delta}}_n &= \gamma_n\sum_{k=1}^K\zeta_{nk}(\widehat{\boldsymbol{\Delta}}_{n-1})(\mathbf{y}_n-\mathbf{y}_k^0)+(1-\gamma_n)\widehat{\boldsymbol{\Delta}}_{n-1}.\end{aligned} \tag{27}$$

Recursion (27) gives rise to Algorithm 2, which produces an estimate $\widehat{\boldsymbol{\Delta}}_n$ at each new measurement $\mathbf{y}_n$. Notice that the computational complexity of these algorithms are directly proportional to the number of kernels K. Therefore, the use of sparse models, such as the ones given by the GCE method, gives the advantage of a reduced computation load.

5. ILLUSTRATIVE EXAMPLES

5.1 Off-line Detection of Eruption Increase

The Old Faithful geyser dataset (Azzalini and Bowman, 1990) is considered here to illustrate the methods for the off-line multivariate case. The dataset contains 272 measurements with $d=2$ dimensions representing the registered length of the geyser's eruptions and the time in between them (both in minutes). A fraction $N_0 = 222$ of the measurements are used to estimate a density for the nominal model $\widehat{p}^0(\mathbf{y})$. Three different models are used:

- a Gaussian with parameters given from the standard maximum likelihood equations,
- a nonparametric model given by the KDE with a Gaussian kernel and bandwidth found using (11),
- a nonparametric model given by the GCE with a Gaussian kernel and an $\epsilon = 10^{-8}$ approximation.

The measurements $\mathbf{Y}_{N_0}$ are shown in Fig. 1(a) together with contour lines for the density models. The components chosen for the GCE model are also shown in Fig. 1(a) with a colormap relating to the weights π^*. With $K=32$, the GCE compresses the dataset by a factor of 86%. The GCE is also richer in details and with a tighter support compared to the KDE and Gaussian models.

A bias change is considered to illustrate the situation where the length of eruptions is increased by half a minute and the interval between them is reduced by 2 minutes, i.e. $\boldsymbol{\Delta} = [0.5, -2]^T$. These values are added to the $N=50$ remaining measurements, which can be seen in Fig.1(a). Using these abnormal measurements $\mathbf{Y}_N$, $\boldsymbol{\Delta}$ is estimated for the three different models. For the Gaussian model, a standard maximum likelihood estimate is used. The estimates for the GCE and KDE models are based on Algorithm 1 with initial values chosen $\boldsymbol{\Delta}_0 = [0,0]^T$ for a comparison. The iterates $\boldsymbol{\Delta}_i$ are shown in Fig. 1(b) as a function of iterations. Due to the sparsity of the GCE, $\widehat{\boldsymbol{\Delta}}_N$ is computed 40 times faster compared to the one given by the KDE. After convergence of the iterates, the GLRT test statistic $\log \widehat{\Lambda}_N(\mathbf{Y}_N)$ is computed for the different models, the values are 9.18, 21.71 and 83.71 for the Gaussian, KDE and GCE models respectively. Based on the asymptotic expressions (8), a threshold $\eta(0.01) = 4.60$ is found. All tests can detect the change, although the one based on the GCE gives a much clearer response.

5.2 On-line Wear Detection in an Industrial Robot

By processing torque measurements collected from an industrial robot joint, a scalar quantity y is generated to infer the mechanical condition of the joint gearbox (Bittencourt et al., 2012). The generated quantity y is positive and remains close to zero under normal conditions, deviating otherwise to indicate an anomaly. The data processing used in the generation of y makes it difficult to determine its distribution function. From this application, it is however possible to collect nominal measurements before the application of the test. Based on $N_0 = 45$ nominal samples, the same three models of Sec. 5.1 are considered. The resulting models and histogram of $Y^0_{N_0}$ can be seen in Fig. 2(a). The measurements distribution is multimodal and asymmetric, which makes the Gaussian model a poor representation of the measurements.

Using these models, the objective is to detect a wear fault appearing from $n=17$ in a sequence $\{y_n\}$ with $1 < n < 23$. Maximum likelihood estimates of a bias change Δ are computed sequentially for each model. For the Gaussian model, the standard maximum likelihood estimate is used and for the KDE and GCE models, Algorithm 2 is used with $\gamma_0 = 0.6$ and $\rho = 1$. The measurements $\{y_n\}$ and the different estimates $\widehat{\Delta}_n$ are shown in Fig. 2(b). Up to $n = 17$, y_n has values smaller than the mean for the Gaussian model making the estimate to deviate.

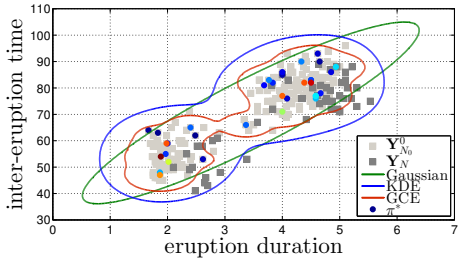
The resulting models are used in GLRTs of the form (6) which are shown in Fig. 2(c). The threshold $\eta(0.01)=3.32$ and $\beta(0.01)$ are found based on (8) where the Fisher information matrix is computed numerically, these are also shown in Fig. 2(c). For the GCE, hypothesis $\mathcal{H}^1$ is chosen for $n \geq 18$ and $\beta < 0.1$ for $n \geq 20$. For the KDE, $\mathcal{H}^1$ is chosen for $n \geq 20$ and $\beta < 0.6$ for $n \geq 21$. For the Gaussian model, alarms are generated in the interval $9 \leq n \leq 17$ and for $n \geq 20$, and β is always larger than 0.5 (notice that the GLRT is never reset when a change is detected).

In this application, an early detection is very important to allow for condition based maintenance, giving enough time to perform maintenance. To decide for maintenance actions, it is also very important to have few false alarms and to give a consistent estimate of the detection error β. In this application, the test obtained using the GCE model presented the best compromise for these requirements.

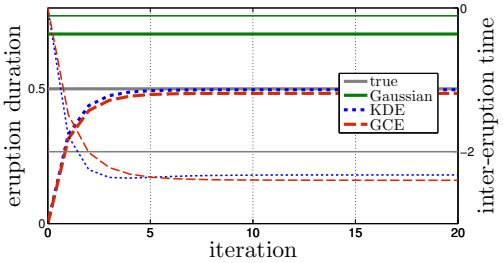
6. CONCLUSIONS

This paper proposed a two step approach for anomaly detection using a bias change model and GLRTs. In the first step, a model for the normality is found based on a nominal dataset. Nonparametric density estimates are used which give high flexibility since specification of a density function is not needed. In the second step, maximum likelihood estimates of a bias change are computed using the EM algorithm. The use of a sparse density model can considerably reduce the computations needed for the estimators. The density model and bias change estimate are then used in GLRTs to decide for presence of an anomaly. Both off-line and on-line cases are considered and the approach only requires availability of nominal data and minimal/meaningful specification in terms of a desired probability of false alarm (to find a threshold). Using asymptotic expressions for the GLRT, it is also possible to give estimates of the uncertainties associated with the decision, which are important to support higher level decisions. The efficacy of the approaches was illustrated in real data examples to detect an increase of eruptions in a geyser and a wear fault in an industrial robot joint. The results achieved show clear improvements compared to tests based on a Gaussian assumption.

Currently, the decision errors are estimated based on asymptotic expressions which may differ for a finite number of measurements. In this direction, it would be interesting to study approaches to provide estimates for the small sample behavior of the error probabilities. This could possibly lead to the derivation of adaptive thresholds.

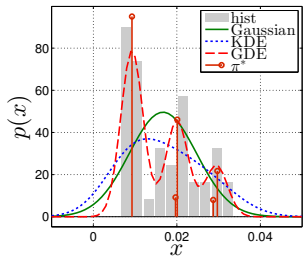


(a) Estimates of $p^0(\mathbf{y})$ and measurements.

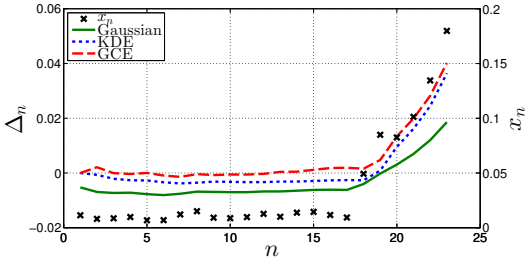


(b) Estimates of eruption duration (thick) and inter-eruption interval (thin). Notice the different axes.

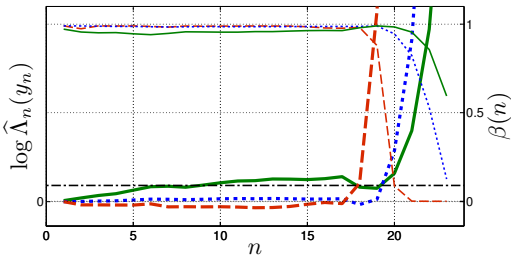
Fig. 1. GLRT for detection of eruptions increase in a geyser dataset. Notice how the test measurements $\mathbf{Y}_N$ in Fig. 1(a) overlaps with the support for the nominal models. Despite this, a detection is achieved with any of the models.



(a) Estimation of $p^0(x)$.



(b) Test data sequence and estimates $\widehat{\Delta}_n$.



(c) GLRT statistic (thick) and $\beta(0.01)$ (thin). Notice the different axes.

Fig. 2. GLRT for detection of abnormalities in the gearbox of a robot joint. Notice the prompt response of the test achieved with the GCE model compared to the KDE and Gaussian.

REFERENCES

D. Agarwal. Detecting anomalies in cross-classified streams: a Bayesian approach. *Knowledge and information systems*, 11(1): 29–44, 2007.

A. Azzalini and A. W. Bowman. A look at some data on the old faithful geyser. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(3):pp. 357–365, 1990.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, USA, 1st edition, 2006.

A. C. Bittencourt, K. Saarinen, and S. Sander-Tavallaey. A data-driven method for monitoring systems that operate repetitively - applications to wear monitoring in an industrial robot joint. In *Proc. of the 8th IFAC SAFEPROCESS*, volume 8, Mexico City, Mexico, 2012.

Z. Botev and D. Kroese. The generalized cross entropy method, with applications to probability density estimation. *Methodology and Computing in Applied Probability*, 13:1–27, 2011.

T. Cacoullos. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18:179–189, 1966.

O. Cappé and E. Moulines. On-line expectation-maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):pp. 1–38, 1977.

M. Desforges, P. Jacob, and J. Cooper. Applications of probability density estimation to the detection of abnormal conditions in engineering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 212(8):687–703, 1998.

L. Devroye and G. L. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, 1980.

R. Isermann. *Fault-Diagnosis Systems - An Introduction from Fault Detection to Fault Tolerance*. Springer, 2006.

C. Jones, J. Marron, and S. Sheather. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*, (11):337–381, 1996a.

M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996b.

D. J. C. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 1st edition, June 2003.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:pp. 289–337, 1933.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):pp. 1065–1076, 1962.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall/CRC, 1986.

D. M. Titterington. Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):pp. 257–267, 1984.

H. L. Van Trees and L. B. Kristine. *Detection, Estimation and Modulation Theory, Part I*. Wiley, New York, USA, 2nd edition, 2013.

V. N. Vapnik and S. Mukherjee. Support vector method for multivariate density estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 659–665. MIT Press, 2000.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):pp. 117–186, 1945.

D.-Y. Yeung and C. Chow. Parzen-window network intrusion detectors. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 4, pages 385–388. IEEE, 2002.