Anomaly detection in homogenous populations: a sparse multiple kernel-based regularization method

Tianshi Chen, Martin S. Andersen, Alessandro Chiuso, Gianluigi Pillonetto and Lennart Ljung

Abstract—A problem of anomaly detection in homogenous populations consisting of linear stable systems is studied. The recently introduced sparse multiple kernel based regularization method is applied to solve the problem. A common problem with the existing regularization methods is that there lacks an efficient and systematic way to tune the involved regularization parameters. In contrast, the hyper-parameters (some of them can be interpreted as regularization parameters) involved in the proposed method are tuned in an automatic way, and in fact estimated by using the empirical Bayes method. What's more, both the parameter and hyper-parameter estimation problems can be cast as convex and sequential convex optimization problems. It is possible to derive scalable solutions to both the parameter and hyper-parameter estimation problems and thus provide a scalable solution to the anomaly detection.

I. INTRODUCTION

One of the main reasons for the interests in anomaly detection in homogenous populations is that problems of this kind naturally arise in many applications, e.g., the flight operations quality assurance (FOQA) system, [1], [2], [3], and the anomaly detection in eye-tracking data for quantification of motor symptoms in Parkinson’s disease [4], [5]. In this paper, we study the anomaly detection in homogenous populations consisting of \( N_d \) linear dynamical systems. Normally, all \( N_d \) systems are supposed to have the same nominal dynamics. Now assume that \( k \ll N_d \) systems have different dynamics from the majority of the population, and moreover, it is unknown what the nominal dynamics is, what the number \( k \) is, and which \( k \) systems that have different dynamics. The goal is to detect the \( k \) systems (anomalies) based on a collection of data from systems in the population.

The existing methods in [1], [2], [3] are linear regression models based regularized least squares methods. In [1], [2], \( l_2 \) regularization method is used and there are a couple of regularization parameters and a threshold that have to be chosen carefully in order to single out the anomalies. However, it is not clear how to tune them and they are chosen empirically to give the best results; see [1]. In [3], a sum-of-norm regularized method is used and there is only one regularization parameter that requires to be tuned. So the tuning in [3] is easier than [1], [2]. Nevertheless, the issue on how to tune the regularization parameter in an efficient and systematic way is not addressed either and it is chosen by heuristic methods and by trial and errors. Another problem with the method in [3] is that it is essentially based on least squares estimates of linear regression models and classical model structure selection methods. As have been shown in [6], [7], [8], [9], [10], the maximum likelihood/prediction error method (ML/PEM) equipped with the classical structure selection methods like Akaike’s information criterion (AIC), Bayesian information criterion (BIC) and cross validation sometimes behaves not perfectly for short data. As such the accuracy of the method in [3] could be improved by using methods more suitable for handling short data.

In this paper, the sparse multiple kernel-based regularization method (SMKRM) recently introduced in [9] is applied to tackle the anomaly detection problem. It is a variant of the kernel-based regularization method in [6], [7], [8] with multiple kernels used instead of single kernels, which in many cases can provide more accurate and more robust model estimates for short data. The use of multiple kernel leads to a couple of features. One feature relevant to the topic here is that the estimation of the hyper-parameters by maximizing the marginal likelihood favors sparse optimal hyper-parameters, which enables SMKRM to tackle various structure detection problems, e.g., the sparse dynamic network identification and the segmentation of linear systems; see [9]. Here, we show that SMKRM can very well handle the anomaly detection problem. In contrast with [1], [2], [3], the hyper-parameters are tuned in an automatic way by the marginal likelihood maximization method, which corresponds to the empirical Bayes method in Bayesian learning. What’s more, both the parameter and hyper-parameter estimation problem can be cast as convex and sequential convex optimization problems. We have also developed scalable solution to both the parameter and hyper-parameter estimation problem and thus provided a scalable solution to the anomaly detection problem.

II. PROBLEM STATEMENT

Assume that the population of interest consists of \( N_s \) linear stable causal systems, and these \( N_s \) systems are described by

\[
y_i(t) = G_{i,0}(q)u_i(t) + v_i(t), \quad i = 1, \ldots, N_s, t = 1, \ldots, N_d \tag{1}
\]

where \( t \) is the time instant, \( N_d \) is the number of collected data, \( i \) is the system index, \( q \) is the time shift operator, \( G_{i,0}(q) \) is the unknown transfer function (model) describing the relation...
from the input \( u_i(t) \in \mathbb{R} \) to the measured output \( y_i(t) \in \mathbb{R} \). Further, assume that for \( i = 1, \cdots, N_s \), \( G_i(0)(q) \) is stable and casual, and the disturbance \( v_i(t) \) is white Gaussian distributed with mean zero and unknown variance \( \sigma^2_i \) and moreover, independent of \( v_j(t) \) for \( i \neq j \) and \( j = 1, \cdots, N_s \).

We will in the following say that the population behaves normally if \( G_1(0)(q) = \cdots = G_{N_s}(0)(q) = G_0(0)(q) \), where \( G_0(0)(q) \) is the unknown nominal model of the population. Conversely, if any system has a model deviating from the nominal one \( G_0(0)(q) \), we will say that the population behaves abnormally and consider that system as an anomaly. In this paper, we focus on the situation where the population behaves normally due to a minority of anomalous systems in the population. More specifically, further assume that there are \( k \) anomalies where \( k \ll N_s \), but \( k \) is unknown and it is also unknown which \( k \) systems are anomalies. Our goal is to detect the \( k \) anomalies from the population based on the collected data \( Z = \{ y_i(t), u_i(t), t = 1, \cdots, N_d, i = 1, \cdots, N_s \} \).

Remark 2.1: For convenience, we only consider the case where \( v_i(t) \) is white. For the case where \( v_i(t) \) is a filtered white noise, i.e., \( v_i(t) = H_i(0)(q)e_i(t) \), where \( H_i(0)(q) \) is both stable and inversely stable and \( e_i(t) \) is white Gaussian distributed with mean zero and unknown noise variance, can be handled in a similar way; see e.g. [10].

III. LINEAR REGRESSION MODELS AND SUM OF NORM REGULARIZATION METHOD

A. Linear regression models and Least squares method

The transfer function \( G_i(0)(q) \) can be written as

\[
G_i(0)(q) = \sum_{k=1}^{+\infty} g_{i,k}^0 q^{-k}, \quad i = 0, 1, \cdots, N_s
\]

where the coefficients \( g_{i,k}^0, k = 1, 2, \cdots, +\infty \) form the impulse response of \( G_i(0)(q) \). Since \( G_i(0)(q) \) is stable, the impulse response of \( G_i(0)(q) \) decays exponentially. It is often enough to truncate the infinite impulse response of \( G_i(0)(q) \) to a finite one: treat \( g_{i,k}^0 = 0 \) for \( k > n \) where \( n > 0 \) is a sufficiently large positive integer, and to assume \( G_i(0)(q) \) can be described by a finite impulse response (FIR) model

\[
G_i(q, \theta_{i,0}) = \sum_{k=1}^{N_s} g_{i,k}^0 q^{-k}
\]

where the FIR model parameter \( \theta_{i,0} \) is defined as

\[
\theta_{i,0} = [g_{i,1}^0, g_{i,2}^0, \cdots, g_{i,N_s}^0]^T
\]

Since \( \theta_{i,0}, i = 0, 1, \cdots, N_s \) are unknown, they have to be estimated from the given data \( Z \). With (3), a model of system (1) can be described by

\[
y_i(t) = \sum_{k=1}^{n} g_{i,k} u_i(t-k) + v_i(t), t = 1, \cdots, N_d, i = 1, \cdots, N_s
\]

which can be further written as a linear regression model

\[
Y_i = \Phi_i \theta_i + V_i, \quad i = 1, \cdots, N_s
\]

In (5) and (6), \( g_{i,k} \) and \( \theta_i \) are, for clarity, used to differentiate from the true quantities \( g_{i,k}^0 \) and \( \theta_{i,0} \), respectively, and in (6), the \( j \)th row of \( Y_i, V_i \in \mathbb{R}^{N_d} \) and \( \Phi_i \in \mathbb{R}^{N_d \times n} \) are \( y_j(i), v_j(i) \), and \( [u_j(i-1) \cdots u_j(i-n)] \), respectively. From (6), if \( \Phi_i \) is full column rank, a least squares (LS) estimate of \( \theta_i \) can be obtained by solving

\[
\hat{\theta}_i = \arg \min_{\theta_i} \| Y_i - \Phi_i \theta_i \|^2 = (\Phi_i^T \Phi_i)^{-1} \Phi_i^T Y_i
\]

where \( \| \cdot \| \) denotes the Euclidean norm.

Remark 3.1: For \( y_i(t), t = 1, \cdots, N_d \), we need to know \( u_i(t), t = 0, \cdots, t-n \) in order to form (6), which however are not known. This issue can be handled in different ways, see e.g., [11] for general discussions and [12] for specific discussions in the context of kernel-based regularization method. Here, we will follow the suggestion in [12] and treat the unknown \( u_i(t), t = 0, \cdots, t-n \), as zeros. It is worth to note that this treatment has already been used in [6], [7] before.

B. Classical model structure selection methods and sum of norm regularization method

The number of available data, in practice, is finite and sometimes, can be small due to time and/or cost reasons. In this case, other methods should be used instead. The anomaly detection problem is essentially a model structure selection problem. Let \( \tilde{k} \) denote the upper bound of the number of anomalies in the population. There are two levels of model structure selection:

1) For each \( k = 1, \cdots, \tilde{k} \), assume that there are \( k \) anomalies in the population and the goal is to determine which \( k \) systems in the population are anomalies;

2) With the result of 1), the goal is to determine the number of anomalies in the population among \( k = 1, \cdots, \tilde{k} \).

As is well known, a traditional method to tackle the model structure selection problem of dynamical systems is AIC, BIC, cross validation and other statistical tests; see e.g. [11]. For level 1), since all possible model structures (hypotheses) have the same model complexity in terms of the same number of anomalies, applying AIC, BIC or the like is equivalent to choose the model structure with the one that gives the smallest data fit, or equivalently the largest likelihood. However, the number of model structures (hypotheses) needed to be checked may be very large and thus this traditional method may be prohibitive to apply; see [3]. As a result, a convex relaxation method based on sum-of-norm regularization method is proposed in [3]:

\[
\min_{\theta_i, i=0, \cdots, N_s} \sum_{i=1}^{N_s} \| Y_i - \Phi_i \theta_i \|^2 + \lambda \sum_{i=1}^{N_s} \| \theta_i - \theta \|^2
\]

where \( \| \cdot \| \) is defined as \( \| \cdot \| / \sigma^2 \) and the regularization parameter \( \lambda > 0 \) is a tuning parameter to balance the tradeoff between the data fit and the number of anomalies. There exists \( \lambda_{\text{max}} > 0 \) such that for any \( \lambda \geq \lambda_{\text{max}} \), all estimates \( \theta_1, \cdots, \theta_{N_s} \) are same so there is no anomaly. As we gradually decrease \( \lambda \), the number of anomalies gradually increase from 0 to 1, \cdots, \tilde{k}. In level 1), we thus have obtained the \( k \) anomalies for each \( k = 1, \cdots, \tilde{k} \) by solving (8) for suitable \( \lambda \). For level 2), we need to determine a suitable \( k \). This is again a model structure selection problem however
with different model complexity in terms of the number of anomalies. Since there are only \( \bar{k} \) (relatively small) model structures (hypotheses) to check, the classical model structure selection methods can be used, e.g. AIC, BIC, etc.

There are two major problems for (8). First, as have been examined in [6], [7], [8], [9], [10], the maximum likelihood/prediction error method (ML/PEM) equipped with classical structure selection methods like AIC, BIC and cross validation, sometimes behaves not perfectly for short data. As such the accuracy of the method in [3] could be improved by using methods more suitable for handling short data. Second, there is no efficient and systematic way to tune the regularization parameter \( \lambda \) in (8). In order to complete the task to single out \( k = 1, \cdots, \bar{k} \) possible anomalies, the current method is to first find \( \lambda_{\text{max}} \) and then gradually reduce \( \lambda \) by trial and errors.

IV. SPARSE MULTIPLE KERNEL BASED REGULARIZATION METHOD

The anomaly detection problem is handled here by using a kernel based regularization method:

\[
\hat{\theta}_i = \arg \min_{\theta_i, i = 0, \cdots, N_s} \sum_{i = 1}^{N_s} \left\{ \|Y_i - \Phi_i \theta_i \|^2 \sigma_i^2 + (\theta_i - \theta_0)^T P(\alpha_i)^{-1} (\theta_i - \theta_0) \right\}
\]

where for each \( i = 0, \cdots, N_s \), \( P(\alpha_i) \geq 0 \) is a kernel (matrix), \( \alpha_i \) is the associated hyper-parameter, and \( \hat{\theta}_i \) should be interpreted as the estimate of the true model parameter \( \theta_i \).

If for some \( i = 1, \cdots, N_s \), \( P(\alpha_i) = 0 \) and thus \( \hat{\theta}_i = \theta_0 \), which indicates the \( i \)th system is not an anomaly. On the contrary, if for some \( i = 1, \cdots, N_s \), \( P(\alpha_i) \neq 0 \) and thus \( \hat{\theta}_i \neq \theta_0 \), which indicates that the \( i \)th system is an anomaly. In this way, the anomaly detection problem is converted to a problem of finding a suitable sparse pattern of hyper-parameters. The design of the kernel structure of \( P(\alpha) \) and the estimation method of the hyper-parameter \( \alpha \) play a key role for the success of the anomaly detection. They should collaborate to give group sparse hyper-parameters among \( \alpha_i \), \( i = 1, \cdots, N_s \) and in turn, \( P(\alpha_i) = 0 \) for certain \( i = 1, \cdots, N_s \). We find the multiple kernel structure and the empirical Bayes method for hyper-parameter estimation fits our need perfectly.

A. Multiple kernel

The kernel \( P(\alpha) \) represents our prior knowledge on the linear stable and casual system. Three families of single kernel structures have been proposed in [6], [8], [12] including the stable spline (SS) kernel, the diagonal/correlated (DC) kernels and the state-space model induced kernels, e.g.,

\[
\text{SS} \quad P_{s,k}^{\alpha} (\beta) = \begin{cases} \frac{3k^2}{2} (1 - \frac{k}{2})^2 & k \geq j \\ \frac{3k^2}{2} (k - \frac{k}{2})^2 & k < j \end{cases}, \quad \beta = \lambda 
\]

\[
\text{DC} \quad P_{d,k}^{\beta} (\lambda) = \lambda^{(k+j)/2} \rho^{k-j}, \quad \beta = [\lambda \rho]^T
\]

where \( \beta \) is the hyper-parameter with \( 0 \leq \lambda < 1, |\rho| \leq 1 \).

These single kernel structures are good for modeling of systems with not very complicated dynamics, but not convenient for structure detection problems. Hence we proposed in [9] the use of multiple kernel for modeling of systems with complicated dynamics and for structure detection problems. Multiple kernels are conic combinations of fixed kernels and have the combination coefficients as hyper-parameters:

\[
P(\alpha) = \sum_{j=1}^{m} c_{ij} P_j(\alpha), \quad \alpha = [c_{i1}, \cdots, c_{im}]^T
\]

where \( P_j > 0, i = 1, \cdots, m \), are fixed kernels and \( c_{ij} \geq 0, j = 1, \cdots, m \), are the hyper-parameters. The fixed kernels \( P_j \) can be constructed in different ways. Here, we mainly consider the way to construct \( P_j \) as instances of single kernels (10). In addition, assume without loss of generality each predefined kernel \( P_j \) can be decomposed as

\[
P_j = L_j L_j^T, \quad L_j \in \mathbb{R}^{n \times n}, \quad j = 1, \cdots, m
\]

B. Hyper-parameter estimation

The empirical Bayes method is used to estimate the hyper-parameter \( \alpha_i, i = 1, \cdots, N_s \). We first need to embed the regularization term \( (\theta_i - \theta_0)^T P(\alpha_i)^{-1} (\theta_i - \theta_0) + \beta^T P(\alpha_0)^{-1} \theta_0 \) in (9) in Bayesian framework.

Assume that \( v(t) \) in (1) is independent of \( \theta_i, i = 0, \cdots, N_s \), and moreover, \( \theta_0 \) is Gaussian distributed as

\[
p(\theta_0) = \mathcal{N}(0, P(\alpha_0))
\]

and \( \theta_i, i = 1, \cdots, N_s \) conditioned on \( \theta_0 \) are independent with each other and moreover, Gaussian distributed as

\[
p(\theta_i | \theta_0) = \mathcal{N}(\theta_i | \theta_0, P(\alpha_i)), \quad i = 1, \cdots, N_s
\]

For now we assume \( \alpha_0, \alpha_i, \sigma^2_i, i = 1, \cdots, N_s \) are known and consider the MAP (maximum a posteriori) estimation:

\[
\max_{\theta_i, i = 0, 1, \cdots, N_s} p(\theta_0, \theta_1, \cdots, \theta_N | Y_1, \cdots, Y_N)
\]

\[
= \max_{\theta_i, i = 0, 1, \cdots, N_s} \Pi_{i=1}^{N_s} p(Y_i | \theta_i) \Pi_{i=1}^{N_s} p(\theta_i | \theta_0) p(\theta_0)
\]

Due to the assumption on the disturbance \( v(t) \), \( i = 1, \cdots, N_s \) in (1), \( Y_i \) conditioned on \( \theta_i \), \( i = 1, \cdots, N_s \), are independent with each other and Gaussian distributed as

\[
p(Y_i | \theta_i) = \mathcal{N}(\Phi_i \theta_i, \sigma_i^2 I_{N_i})
\]

Taking into account (16), (13) and (14), the MAP estimation problem (15) takes exactly the form of (9).

Now consider the hyper-parameter estimation. The empirical Bayes method is actually a maximum likelihood method for estimating the hyper-parameters in the prior distribution and it is also named marginal likelihood (or type II maximum likelihood) maximization method:

\[
\max_{\alpha_0, \alpha_i, \sigma^2_i, i = 1, \cdots, N_s} p(Y_1, \cdots, Y_N | \alpha_0, \alpha_i, \sigma^2_i, i = 1, \cdots, N_s)
\]

Note that

\[
p(Y_1, \cdots, Y_N | \alpha_0, \alpha_i, \sigma^2_i, i = 1, \cdots, N_s)
\]

\[
= \int \Pi_{i=1}^{N_s} p(Y_i | \theta_i, \sigma^2_i) p(\theta_i | \theta_0, \alpha_i) p(\theta_0 | \alpha_0) d\theta_0 d\theta_1 \cdots d\theta_N
\]
Further from (16), (13) and (14), we have

\[
p(Y_1, \cdots, Y_{N_s} | \alpha_0, \alpha_i, \sigma_i^2, i = 1, \cdots, N_s) = \mathcal{N} \left( 0, \begin{bmatrix} \Phi_1^T & \Phi_2^T & \cdots & \Phi_{N_s}^T \end{bmatrix} \Sigma \begin{bmatrix} \Phi_1 \Phi_2 \cdots \Phi_{N_s} \end{bmatrix} \right) \tag{18}
\]

where \( \text{diag}(A_1, A_2) \) is a block diagonal matrix with matrices \( A_1 \) and \( A_2 \) on the main diagonal.

**Remark 4.1:** The noise variance \( \sigma_i^2, i = 1, \cdots, N_s \), are assumed to be equal and moreover, known in [3]. Here, they are treated as “hyper-parameters” and estimated together with the other hyper-parameters \( \alpha_i, i = 0, 1, \cdots, N_s \) with the marginal likelihood maximization method.

Eq. (18) implies that the marginal likelihood maximization problem (17) can be put into the following form:

\[
\begin{align*}
& \text{minimize} \quad \mathcal{L} = Y^T \left( \sum_{i=1}^{p} x_i B_i B_i^T \right)^{-1} Y + \log \det \left( \sum_{i=1}^{p} x_i B_i B_i^T \right) \\
& \quad \text{where } p = (N_s + 1)m + N_s,
\end{align*}
\]

\[
Y = \begin{bmatrix} Y_1^T & \cdots & Y_{N_s}^T \end{bmatrix}^T
\]

\[
x = \begin{bmatrix} \alpha_0^T & \alpha_1^T & \sigma_1^2 & \cdots & \alpha_{N_s}^T & \sigma_{N_s}^2 \end{bmatrix}^T
\]

\[
C = \left\{ x \in \mathbb{R}^p \mid x_i \geq 0, i = 1, \cdots, m, \right. \\
\left. x_{k(m+k-1)+i} \geq 0, x_{(k+1)m+k} > 0, i = 1, \cdots, m, k = 1, \cdots, N_s, \right. \\
\left. i = 1, \cdots, m, k = 1, \cdots, N_s \right\}
\]

\[
B_i = \begin{bmatrix} \Phi_1^T & \cdots & \Phi_{N_s}^T \end{bmatrix} L_i, \quad i = 1, \cdots, m
\]

where \( L_i \) is a lower triangular matrix with \( L_{ii} = \sum_{j=1}^{m} \phi_{ij} \phi_{ij} \).

The multiple kernel (11) and the marginal likelihood maximization method (18), or equivalently (19), collaborate together to yield a feature that is the key for the anomaly detection. As shown in [13, Theorem 1], [9, Theorem 4.1, Remark 4.1], the optimization problem (19) favors sparse \( x \), that is, the marginal likelihood maximization problem (18) favors sparse hyper-parameters \( \alpha_i, i = 0, 1, \cdots, N_s \). This is the reason why the proposed sparse multiple kernel-based regularization method is capable to solve the anomaly detection problem in homogenous populations.

### V. SCALABLE SOLUTION TO THE ANOMALY DETECTION

The optimization problem (19) has \((N_s + 1)m + N_s\) variables and (9) has \((N_s + 1)n\) variables. Clearly, if \( N_s \) is large, both of the two optimization problems can be prohibitive to solve on a single computer. Therefore, it is interesting and important to investigate scalable solutions, which break the large problem into a number of small problems that can be handled by cheap computing facilities.

#### A. SCALABLE LOCALLY OPTIMAL SOLUTION TO (19)

Although (19) is nonconvex, it is a difference of convex function programming problem and its locally optimal solution can be obtained efficiently by using sequential convex optimization techniques, e.g., the majorization minimization (MM) algorithm [14]. In the following, we first recall the MM algorithm [9, Algorithm 3.1] and then develop its scalable version in the context of anomaly detection.

The optimization problem (19) can be rewritten as

\[
\begin{align*}
& \text{minimize } f(x) \\
& \quad \text{subject to } \|w_i\|^2 \leq 2x_i z_i, x_i \geq 0, z_i \geq 0, i = 1, \cdots, p,
\end{align*}
\]

\[
Y = \sum_{i=1}^{p} w_i \gamma_i
\]

where \( w = [w_1^T \cdots w_p^T]^T \) with \( w_i \in \mathbb{R}^n \) for \( i = 1, \cdots, m \), \( w_i \in \mathbb{R}^n \) for \( i = km + k - 1 + j, j = 1, \cdots, m \), \( k = 1, \cdots, N_y \), and \( w_i \in \mathbb{R}^n \) for \( i = (k+1)m+k, k = 1, \cdots, N_y \).

The gradient \( \nabla h(x^{[q]}) \) in (25) has to be computed and given at each iteration. We will discuss how to compute \( \nabla h(x^{[q]}) \) in a scalable way in the next subsection. For now we assume that \( \nabla h(x^{[q]}) \) is given and focus on how to derive a scalable...
solution to (25) based on ADMM. It is worth noting that (25) is almost separable except the equality constraint in (25). As can be seen from (20d), the matrices $B_i$, $i = m+1, \ldots, p$ are highly structured and in fact are block sparse. Making use of this structural property can lead to scalable solutions.

First, we group the matrices $B_i$ and vectors $w_i$, $i = 1, \ldots, p$ in the following way

$$B_{[0]} = [B_1 \cdots B_m], B_{[i]} = [B_{(m+1)i} \cdots B_{(i+1)m}],$$
$$w_{[0]} = [w_1^T \cdots w_m^T]^T, w_{[i]} = [w_{(m+1)i}^T \cdots w_{(i+1)m+i-1}^T]^T, \quad i = 1, \ldots, N_s,$$

where $B_{[0]} \in \mathbb{R}^{N_0 \times nm}$, $w_{[0]} \in \mathbb{R}^{nm}$ and $B_{[i]} \in \mathbb{R}^{N_i \times (nm+N_d)}$, $w_{[i]} \in \mathbb{R}^{nm+N_d}$, $i = 1, \ldots, N_s$. It follows from (20d) that $B_{[i]}$, $i = 0, \ldots, N_s$, can be partitioned as:

$$B_{[i]} = \begin{bmatrix} B_{0i} & \vdots & B_{Ni} \end{bmatrix}, \quad i = 1, \ldots, N_s,$$

where $B_{0i} \in \mathbb{R}^{N_i \times mm}$, $B_{1i} \in \mathbb{R}^{N_i \times (nm+N_d)}$, $i = 1, \ldots, N_s$ are defined in an obvious way. Then the optimization problem (25) can be rewritten as

$$\min_{x, w} \left\{ 2(1^T z) - \nabla h(x^0) x + \sum_{i=1}^{p} \tilde{I}_{x_i \geq 0} (x_i) + \tilde{I}_{z_i \geq 0} (z_i) \right\} + I_{|w_{[i]}|^2 \leq 2\eta} (x_i, z_i, w_{[i]}) + \sum_{i=1}^{N_s} \tilde{I}_{B_{0i}w_{[0i]} + B_{1i}w_{[i]} = Y_{[i]}(w_{[0i]}, w_{[i]}]} (28)$$

where $I(\cdot)$ denotes an indicator function, e.g. $I_{x_i \geq 0}(x_i)$ means that $\tilde{I}_{x_i \geq 0}(x_i) = 0$ for $x_i \geq 0$ and $\tilde{I}_{x_i \geq 0}(x_i) = \infty$ otherwise.

The coupling in (28) is due to the last term. Now define

$$\tilde{x} = [x^T \ z^T \ w]^T$$
$$\tilde{y} = [\tilde{x}^T \ \tilde{z}^T \ \delta_0^T \ \cdots \ \delta_{N_s}^T \ \mu_1^T \ \cdots \ \mu_{N_s}^T]^T$$

where $\tilde{x}, \tilde{z} \in \mathbb{R}^p$ and $\delta_i \in \mathbb{R}^{nm}$, $\mu_i \in \mathbb{R}^{nm+N_d}$, $i = 1, \ldots, N_s$. Then the optimization problem (28) can be rewritten as:

$$\min_{\tilde{x}, \tilde{y}} G_1(\tilde{x}) + G_2(\tilde{y})$$

subject to $\tilde{A}\tilde{x} = \tilde{y}$

where

$$G_1(\tilde{x}) = 2(1^T z) - \nabla h(x^0) x + \sum_{i=1}^{p} \tilde{I}_{x_i \geq 0} (x_i) + \tilde{I}_{z_i \geq 0} (z_i)$$
$$+ \tilde{I}_{|w_{[i]}|^2 \leq 2\eta} (x_i, z_i, w_{[i]}) (31a)$$
$$G_2(\tilde{y}) = \sum_{i=1}^{N_s} I_{B_{0i} \delta_i + B_{1i} \mu_i = Y_{[i]}}(\delta_i, \mu_i) (31b)$$

$$\tilde{A} = \begin{bmatrix} I_{2p \times 2p} & 0_{2p \times nm} & 0_{2p \times (nm+N_d)} \ 0_{nmN_i \times 2p} & I_{nm} & 0_{nmN_i \times (nm+N_d)} \ 0_{N_i(nm+N_d) \times 2p} & 0_{N_i(nm+N_d) \times nm} & I_{N_i(nm+N_d)} \end{bmatrix} (31c)$$

Introducing a Lagrange multiplier vector $\tilde{\eta} \in \mathbb{R}^{2p+(2m+N_d)N_s}$ corresponding to the equality constraint $\tilde{A}\tilde{x} = \tilde{y}$ yields the augmented Lagrangian function of (30):

$$L_{\tilde{\rho}}(\tilde{x}, \tilde{y}, \tilde{\eta}) = G_1(\tilde{x}) + G_2(\tilde{y}) + \tilde{\eta}^T(\tilde{A}\tilde{x} - \tilde{y}) + \frac{(\tilde{\rho}/2)||\tilde{A}\tilde{x} - \tilde{y})^2}{(32)}$$

where $\tilde{\rho} > 0$. Then applying ADMM yields the following iterative solution of (30):

$$\tilde{x}^{(k+1)} = \arg\min_{\tilde{x}} [G_1(\tilde{x}) + \left(\frac{\tilde{\eta}^{(k)})}{\tilde{\rho}}\tilde{A}\tilde{x}^{(k)}\right)^T]$$
$$\tilde{y}^{(k+1)} = \arg\min_{\tilde{y}} [G_2(\tilde{y}) - \left(\frac{\tilde{\eta}^{(k)})}{\tilde{\rho}}\tilde{y}^{(k)}\right)^T]$$

where $\tilde{\rho} > 0$, and the initial vectors $\tilde{\eta}^{(0)}$ and $\tilde{y}^{(0)}$ are arbitrary.

It is interesting to see that the first two optimization problems with respect to $\tilde{x}$ and $\tilde{y}$ in (33) can still be casted as conic optimization problems with rotated quadratic cone constraints. More specifically, for the first optimization problem in (33), it is equivalent to

$$\min_{\tilde{x}} \left\{ 2(1^T z) - \nabla h(x^0) x + \left(\frac{\tilde{\eta}^{(k)})}{\tilde{\rho}}\tilde{A}\tilde{x}^{(k)}\right)^T \right\}$$

subject to $x_i \geq 0, z_i \geq 0, ||w_{[0]}||^2 \leq 2\eta, i = 1, \ldots, p$

$$||w_{[i]}||^2 - \left(\frac{\tilde{\eta}^{(k)})}{\tilde{\rho}}\right)^2 \leq l_{i+1, i}, i = 1, \ldots, N_s$$

where $l = [l_1 \cdots l_{N_s}]^T$. For the second optimization problem in (33), it is equivalent to

$$\min_{\tilde{y}} -\left(\frac{\tilde{\eta}^{(k)})}{\tilde{\rho}}\tilde{y}^{(k)}\right)^T \tilde{y} + \left(\frac{\tilde{\rho}}{2}\right)I^T \tilde{y}$$

subject to $B_{0i} \delta_i + B_{1i} \mu_i = Y_{[i]}, i = 1, \ldots, N_s$

$$||w_{[0]}||^2 - \left(\frac{\tilde{\eta}^{(k)})}{\tilde{\rho}}\right)^2 \leq l_{i}, ||w_{[i]}||^2 - \left(\frac{\tilde{\eta}^{(k)})}{\tilde{\rho}}\right)^2 \leq l_{N_s+i}, i = 1, \ldots, N_s$$

What’s more, it is worth noting that

$$\tilde{B}_{0i} = [\tilde{\Phi}_{i} L_{i1} \cdots \tilde{\Phi}_{i} L_{im}], \tilde{B}_{1i} = [\tilde{\Phi}_{i} L_{i1} \cdots \tilde{\Phi}_{i} L_{im}, l_{i}),(34)$$

which only depends on the data collected on the $i$th system. Therefore, the computations in (33) are separable with respect to the $i$th system, $i = 1, \ldots, N_s$. We thus have obtained an ADMM based scalable solution to the optimization problem (25). The following theorem guarantees that the scalable solution (33) converges to the solution of the original optimization problem (25).

**Theorem 5.1:** Consider (25). The sequences $\{\tilde{x}^{(k)}\}$, $\{\tilde{y}^{(k)}\}$ and $\{\tilde{\eta}^{(k)}\}$ generated by the ADMM based scalable solution (33) for any $\tilde{\rho} > 0$, initial vectors $\tilde{x}^{(0)}$ and $\tilde{y}^{(0)}$, converge. Moreover, the sequence $\{\tilde{x}^{(k)}\}$ converges to the optimal solution of the original problem (25).

**Proof:** Since the optimal solution set of problem (30) is nonempty and $\tilde{A}^T \tilde{A}$ is invertible, the conclusion follows by [17, Ch. 3, Prop. 4.2].

From (22) and (20d), $\Sigma(x)$ is a sum of low rank matrix and a block diagonal matrix. By exploring this structure, it is possible to compute $\nabla h(x)$ in a scalable way. Sometimes, it is also interesting to know the estimate $\tilde{\theta}_i$, $i = 0, 1, \ldots, N_s$ in (9). Since (9) is a simple $l_2$ regularized least squares problem, there exist a couple of scalable solutions to this problem. The details are omitted due to the space limitation.
VI. NUMERICAL SIMULATION

A. Test systems and data sets

Test systems: We first generate 11 generic systems in the same way in [8], [9]. In particular, a continuous-time system of 30th order is first generated using the command m=rss(30) in MATLAB. The continuous-time system m is then sampled at 3 times of its bandwidth to yield the corresponding discrete-time system md using the following commands in MATLAB: bw=bandwidth(m); f = bw*3*2*pi; md=c2d(m,1/f,’zoh’). If all poles of md are within the circle with center at the origin and radius 0.95, set the feedthrough matrix of md to 0 and save it as one generic system.

Test data sets: Randomly select 1 generic system as the nominal system and generate the first 40 data sets based on this generic system as follows. Simulate the nominal system with a white Gaussian input signal with unit variance, get the noise-free output, and then add to the noise free output a white Gaussian noise. The signal-to-noise ratio (the ratio between the variance of the noise free output and that of the additive noise) is randomly chosen from [1, 5]. We treat the remaining 10 generic systems as anomalies and for each system we generate 1 data set in a similar way as the first 40 data sets. So in total there are 50 data sets (Nt = 50): the first 40 correspond to normal systems and the last 10 correspond to anomalies. Each data set contains 40 data points (Nd = 40). All data sets are collected after getting rid of initial conditions effect.

B. Simulation setup and results

To test the proposed approach, the fixed kernels Pj in (11) are constructed based on the TC (tuned correlated) kernel

\[ TC(\beta) = \min(\lambda^T, \lambda^j), \beta = \lambda \]  

which is a special case of the DC kernel (10b) and also known as the first order stable spline kernel [7]. In particular, the following grid of \( \beta \) with \( \beta = \lambda = 0.7 : 0.02 : 0.95 \) is used to compute 11 fixed kernels \( P_j \) in (11), i.e., \( m = 11 \). The FIR model order in (3) is set to 40, i.e., \( n = 40 \).

We then compute \( \alpha_i \) the estimate of the hyper-parameter \( \alpha_i \) in (11) by solving the problem (19) based on (25) or (33). Clearly, if for some \( i \), \( \sum_{j=1}^{11} \hat{c}_{i,j} \) is nonzero, it implies that the corresponding \( i \)th system is an anomaly. Fig. 1 shows the profile of \( \sum_{j=1}^{11} \hat{c}_{i,j}, i = 1, \cdots, 50 \). If \( \sum_{j=1}^{11} \hat{c}_{i,j} \) is nonzero for certain \( i \), it implies that the corresponding \( i \)th system is an anomaly.

VII. CONCLUSION AND FUTURE WORKS

In this paper, we have studied the problem of anomaly detection in homogenous populations by the recently introduced sparse multiple kernel based regularization method. The problem is solved by converting it to a problem of finding a suitable sparse pattern of the hyper-parameters, whose estimation is handled automatically by the empirical Bayes method, i.e. the marginal likelihood maximization method. In the future, it is interesting to investigate and give theoretical comparisons on if utilizing the kernels suitable for system identification can give more accurate decisions than the sum-of-norm regularization method in [3].

REFERENCES


