The Annals of Applied Statistics
2022, Vol. 16, No. 2, 721–743
https://doi.org/10.1214/21-AOAS1508

SCALABLE CHANGE-POINT AND ANOMALY DETECTION IN CROSS-CORRELATED DATA WITH AN APPLICATION TO CONDITION MONITORING

BY MARTIN TVETEN[1,a], IDRIS A. ECKLEY[2,b] AND PAUL FEARNHEAD[2,c]

[1]*Department of Mathematics, University of Oslo,* [a]*tveten@nr.no*
[2]*Mathematics and Statistics, Lancaster University,* [b]*i.eckley@lancaster.ac.uk,* [c]*p.fearnhead@lancaster.ac.uk*

Motivated by a condition monitoring application arising from subsea engineering, we derive a novel, scalable approach to detecting anomalous mean structure in a subset of correlated multivariate time series. Given the need to analyse such series efficiently, we explore a computationally efficient approximation of the maximum likelihood solution to the resulting modelling framework and develop a new dynamic programming algorithm for solving the resulting binary quadratic programme when the precision matrix of the time series at any given time point is banded. Through a comprehensive simulation study we show that the resulting methods perform favorably compared to competing methods, both in the anomaly and change detection settings, even when the sparsity structure of the precision matrix estimate is misspecified. We also demonstrate its ability to correctly detect faulty time periods of a pump within the motivating application.

1. Introduction. Modern machinery can be perplexingly complicated and interlinked. The interruption of one machine may cause downtime of a whole operation, in addition to a repair being both costly, time consuming and arduous. This has spawned an enormous interest in (remote) condition monitoring of industrial equipment to detect deviations from normal operation such that optimal uptime can be achieved and impending faults discovered before they occur. Overviews of condition monitoring techniques for different equipment exist for pump turbines (Egusquiza et al. (2015)), wind turbines (Tchakoua et al. (2014)) and audio and vibration signals (Henriquez et al. (2014)), among others. A common theme is the decision problem of when the machinery is running abnormally—a problem that lends itself well to statistical change-point analysis.

The current work is motivated by a problem of detecting time intervals (segments) of suboptimal operation of an industrial process pump. We will refer to these segments as "anomalies" or "segments," because they correspond to deviations from some predefined baseline pump behaviour. The pump is equipped with sensors that measure temperatures and pressures over time at various locations. Other operational variables, such as the flow rate and volume fractions for the different fluids being pumped, are also recorded. If present, the aim is to estimate the start- and end point of anomalies as well as indicate which variables are anomalous. This is useful information to the operators of the pump to pinpoint the source of historical problems and learn from them. Another reason for performing such an analysis is to create a clean reference data set that can be used to train a model of the equipment's baseline behaviour before deploying the method for online condition monitoring. The particular data set we consider contains four anomalies that have been manually labelled by engineers familiar with this data, based on retrospectively looking for signs in the data of degrading performance.

The starting point of our methodology is to assume that, during normal operation of the pump, the data follows a baseline stationary distribution and during suboptimal operation, the

Received October 2020; revised June 2021.
Key words and phrases. Anomaly, binary quadratic programme, change points, cross-correlation, outliers.

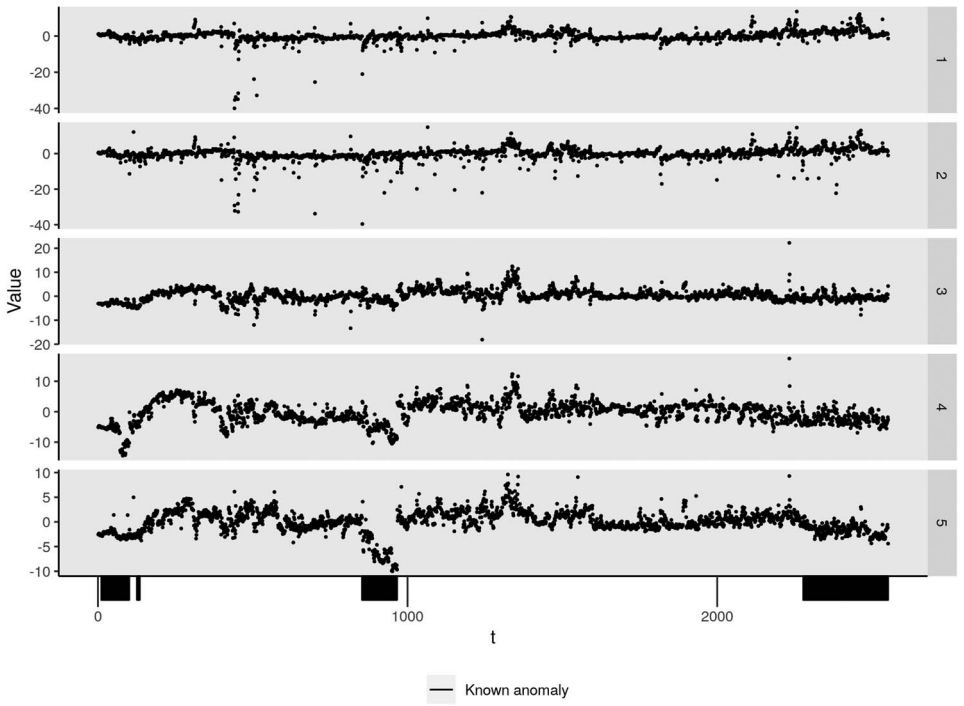


FIG. 1. *Pump data after preprocessing, with four known segments of suboptimal operation marked by black lines on the x-axis. The correlation between variables* 1 *and* 2 *is* 0.89, *and the pairwise correlations between variables* 3, 4 *and* 5 *are all above* 0.6.

mean of the distribution changes abruptly for some period of time, before it reverts back to the baseline mean. This is known as an *epidemic change-point* model in the literature (Kirch, Muhsal and Ombao (2015)), but in the presence of our application, we will refer to it as the *anomaly* model. A challenge with the pump data is that the mean changes as a consequence of what is being pumped and other operating conditions in addition to suboptimal operation. To decrease the dependence on the operating conditions and thus increase the signal from changes due to suboptimal operation, we divide the variables into sets of *state* variables and *monitoring* variables and regress the monitoring variables onto the state variables (similar to Klanderman et al. (2020)). The remaining five-variate time series of monitoring residuals are shown in Figure 1, where the known anomalies are marked on the time axis. Observe that the strength of the known anomalies vary as well as which variables seem to be affected. It is also apparent that the mean changes outside of the known anomalous segments. Detecting and estimating these segments is also important, as they may correspond to previously unknown anomalies or constitute data for which the current model between state and monitoring variables fit poorly and hence point to how it should be improved.

The pump data after preprocessing also exhibit strong cross-correlation, due to the proximity of the sensors to each other, with the correlation of variables 1 and 2 being 0.89 and the pairwise correlations between variables 3, 4 and 5 all being above 0.6. Most existing methods for detecting a change or anomaly in a subset of variables ignore cross-correlation (though see Wang and Samworth (2018)). If not accounted for, however, cross-correlation will hamper the detection of more subtle anomalies, as illustrated by the simulated example in Figure 2. The benefit of undertaking multivariate change-point detection is to borrow strength between variables to detect smaller changes than would be possible if each variable were considered separately. Including cross-correlation in the model, if sufficiently strong, will increase the power of detection. This is particularly true for sparse changes, an observation also made by Liu, Gao and Samworth (2021).

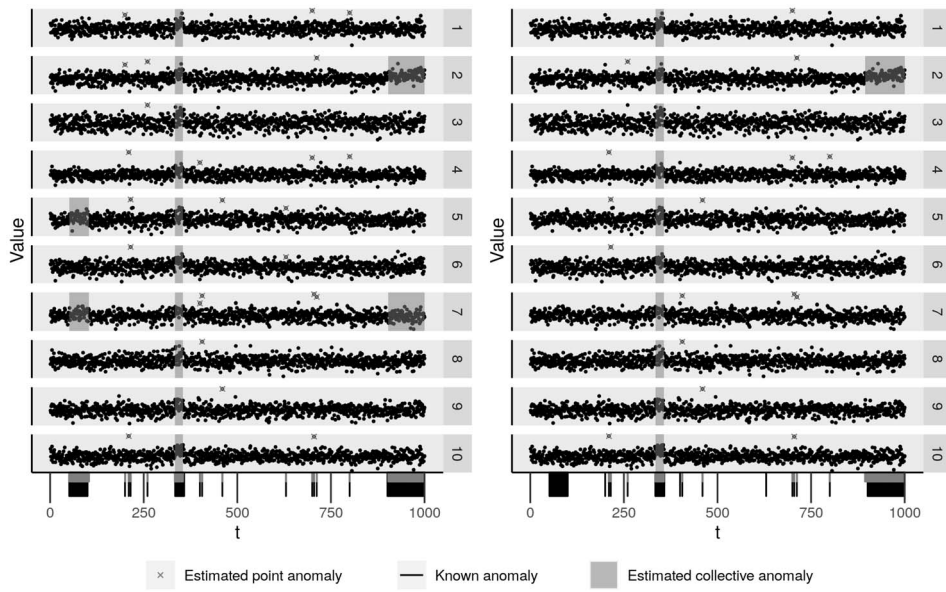


FIG. 2. *Modelling cross-correlation increases detection power for a fixed Type I error probability, especially for sparse changes. Both plots show the same set of* 1000 *simulated observations from a* 10*-variate Gaussian distribution with a global constant correlation of* 0.5*, containing three collective anomalies at* $t \in (50, 100]$, $(333, 358]$, $(900, 1000]$, *affecting the means of variables* $\{6, 10\}$, $\{1, \ldots, 10\}$ *and* $\{9\}$, *respectively, and* 12*-point anomalies affecting two random variables each. The left plot displays the estimates of collective and point anomalies of our method, which incorporates cross-correlations, while the right plot shows estimates when the method ignores cross-correlations. As both methods were tuned to achieve* 0.05 *probability of a false positive under the global correlation null model, the two sparse anomalies are not detected in the right plot as a trade-off with error control.*

Our main methodological contribution is to develop a novel test statistic based on a penalised cost approach for detecting multiple anomalies/epidemic changes in a subset of means of cross-correlated time series. The test is designed to be powerful for both sparse and dense alternatives as well as being computationally fast and scalable. This is crucial for our method also to be useful for anomaly detection problems of higher dimensionality than our process pump example. Anomalies are then detected by using the test within a PELT-type algorithm (Killick, Fearnhead and Eckley (2012)) to optimise exactly over all possible start- and end points of anomalies.

Through the work on making the method scalable, we derive an algorithm which may be of independent interest within combinatorial optimisation. Our test statistic is an approximation to the maximum likelihood solution of our problem, formulated as what is known as an unconstrained binary quadratic program (BQP). We show that such optimisation problems can be solved exactly by a dynamic programming algorithm scaling linearly in the number of variables, p, if the matrix in the quadratic part of the objective function is sparse in a banded fashion. In the anomaly detection problem this corresponds to having a banded precision matrix. We present a simple preprocessing step for obtaining a banded estimate of the precision matrix of our data and show empirically that detecting the anomalies, using such an estimate, leads to gains in power over methods that ignore cross-correlation, even when the banded assumption is incorrect.

A further challenge in many applications, such as the pump data of Figure 1, is the presence of outliers. If left unattended, it is well known that they will interfere with the detection of changes (Fearnhead and Rigaill (2019)). To handle outliers, we incorporate the distinction between point and collective anomalies, introduced in the CAPA (collective and point

anomalies) and MVCAPA (multivariate CAPA) methods of Fisch, Eckley and Fearnhead (2021a, 2021b). A point anomaly is defined as an anomalous segment of length one—a single anomalous observation—while a collective anomaly is an anomalous segment of length two or longer. This distinction enables the method to classify sporadic outliers as point anomalies rather than confusing them with a collective anomaly. We call our anomaly detection algorithm CAPA-CC, short for collective and point anomalies in cross-correlated data.

To the best of our knowledge, there are no other methods designed specifically for the multiple point and collective anomaly detection problem in multivariate, cross-correlated data with both sparse and dense anomalies. Current approaches to detect collective anomalies assume independence across series (Fisch, Eckley and Fearnhead (2021b), Jeng, Cai and Li (2013)). Alternatively, methods like Kirch, Muhsal and Ombao (2015) model correlated series but focus on detecting changes in the cross-correlation.

For the general change-point problem of a sparse or dense change in the mean, the literature is mostly concentrated on methods that either allow for sparse changes but assume cross-independence (Xie and Siegmund (2013), Jirak (2015), Cho and Fryzlewicz (2015), Cho (2016), Bardwell et al. (2019)) or allow cross-dependence but assume changes are dense (Horváth and Hušková (2012), Li et al. (2019), Bhattacharjee, Banerjee and Michailidis (2019), Westerlund (2019)). The inspect method of Wang and Samworth (2018) is a notable exception to this rule, as it is designed to estimate sparse changes in the mean of potentially cross-correlated data. Whilst general change-point methods can also be used for the anomaly detection problem, some power is expected to be lost, as there is no assumption of a shared baseline parameter.

The paper is organised as follows: We first describe the anomaly detection problem in detail in Section 2, before considering our solution in Section 3. Particular focus is put on the single collective anomaly case and our BQP solving algorithm for approximating the maximum likelihood solution. We then briefly describe how the same ideas can be applied to the general change-point detection problem in Section 4. In Section 5 we cover a useful strategy for robustly estimating the precision matrix with a given sparsity structure, and we suggest strategies for tuning our method. Section 6 contains an extensive simulation study for assessing the performance of our method. We conclude by presenting the analysis of the pump data in Section 7.

2. Problem description. Suppose we have n observations, $\{\mathbf{x}_t\}_{t=1}^n$, of p variables, $\mathbf{x}_t = (x_t^{(1)}, \ldots x_t^{(p)})$, where each $\mathbf{x}_t$ has mean, $\boldsymbol{\mu}_t$, and a common precision matrix, $\mathbf{Q}$, encoding the conditional dependence structure between the variables. Our interest is in detecting collective anomalies that are characterised by a change in the mean of the data.

In our anomaly detection problem, segments of the data will be considered anomalous if the mean, $\boldsymbol{\mu}_t$, is different from a baseline mean, $\boldsymbol{\mu}_0$. Let K be the number of collective anomalies, where the kth anomaly, for $k = 1, \ldots, K$, starts at observation $s_k + 1$, ends at observations e_k and affects the components in a subset $\mathbf{J}_k \subseteq [p]$. So, the model assumes that the mean vectors, $\boldsymbol{\mu}_t$, are given by

$$\mu_t^{(i)} = \begin{cases} \mu_1^{(i)} & \text{if } s_1 < t \leq e_1 \text{ and } i \in \mathbf{J}_1, \\ \quad\vdots & \\ \mu_K^{(i)} & \text{if } s_K < t \leq e_K \text{ and } i \in \mathbf{J}_K, \\ \mu_0^{(i)} & \text{otherwise,} \end{cases} \tag{1}$$

where $e_k \leq s_{k+1}$ such that no overlapping anomalous segments are allowed. To distinguish collective anomalies from point anomalies, which we will consider later, we make the assumption that collective anomalies are of length at least 2, that is, $e_k - s_k \geq 2$. The rationale

is that point anomalies, that is, anomalies that affect data at isolated time points are likely to be caused by different factors than collective anomalies. In our application, point anomalies may be due to sensor errors, whereas collective anomalies indicate underlying issues with the machinery. In some cases, one may also be given information about the minimum and maximum segment length of a collective anomaly, $l \geq 2$ and $l < M \leq n$, respectively, such that $l \leq e_k - s_k \leq M$ for all k.

Our aim is to infer the number of collective anomalies, K, as well as their locations within the data, $(s_k, e_k, \mathbf{J}_k)_{k=1}^K$, together with the anomalous means, $\boldsymbol{\mu}_k^{(i)}$ for $i \in \mathbf{J}_k$, in a computationally efficient manner.

During method development we assume that the baseline parameter, $\boldsymbol{\mu}_0$, and the precision matrix, $\mathbf{Q}$, is known. In practice, these will be estimated from the data, using robust statistical methods described in Section 5.1. Later, to enable quick computation, we will also assume that $\mathbf{Q}$, or an estimate of $\mathbf{Q}$, is sparse in a banded fashion. A sparse precision matrix corresponds to cases where only a few of the variables are conditionally dependent.

3. Detecting anomalies.

3.1. *A single collective anomaly.* In this section we consider the anomaly detection problem described in Section 2 for $K \leq 1$. Our approach is to model the data as being realisations of multivariate Gaussian random variables, independent over time, and to use a penalised likelihood approach to detect an anomaly.

We will use the following notation: For a p-vector $\mathbf{x}$ and set $\mathbf{J} \subseteq [p]$, $\mathbf{x}^{(\mathbf{J})} := (x^{(i)})_{i \in \mathbf{J}}$ and $\mathbf{x}(\mathbf{J}) := (x^{(i)} I\{i \in \mathbf{J}\})_{i=1}^p$, where $I\{i \in \mathbf{J}\})$ is the indicator function. For a matrix $\mathbf{X}$, $\mathbf{X}_{\mathbf{J},\mathbf{K}}$ denotes the submatrix of rows $\mathbf{J}$ and columns $\mathbf{K}$. Both $-\mathbf{J}$ and $\mathbf{J}^{\mathbf{c}}$ refer to the complement of a set $\mathbf{J}$. The k-subscripts enumerating the anomalies will be skipped when the referenced anomaly is clear from the context.

Define the cost of introducing an anomaly from time-point $s+1$ to e in variables $\mathbf{J}$ as twice the negative log-likelihood of multivariate Gaussian data

$$C\big(\mathbf{x}_{(s+1):e}, \boldsymbol{\mu}(\mathbf{J})\big) = \sum_{t=s+1}^{e} \big(\mathbf{x}_t - \boldsymbol{\mu}(\mathbf{J})\big)^{\mathsf{T}} \mathbf{Q} \big(\mathbf{x}_t - \boldsymbol{\mu}(\mathbf{J})\big), \tag{2}$$

where, for simplicity, we have dropped added constants. Now, for ease of presentation and without loss of generality, we assume $\boldsymbol{\mu}_0 = \mathbf{0}$. Then, the log-likelihood ratio statistic of the observations $\mathbf{x}_{(s+1):e}^{(\mathbf{J})}$ being anomalous is given by

$$S(s, e, \mathbf{J}) = C(\mathbf{x}_{(s+1):e}, \mathbf{0}) - \min_{\boldsymbol{\mu}(\mathbf{J})} C\big(\mathbf{x}_{(s+1):e}, \boldsymbol{\mu}(\mathbf{J})\big). \tag{3}$$

We refer to $S(s, e, \mathbf{J})$ as the *saving* realised by allowing the observations $\mathbf{x}_{(s+1):e}^{(\mathbf{J})}$ to have a different mean from $\mathbf{0}$. In a maximum likelihood spirit the aim is to maximise the savings $S(s, e, \mathbf{J})$ over start points s, end points e and subset $\mathbf{J}$, and infer the anomalous segment thereof. However, as we vary $\mathbf{J}$, we are optimising over differing numbers of means in the anomalous segment, and the savings will always increase, as we optimise over more parameters. One way of dealing with this is to introduce a penalty that is a function of the number of anomalous variables, $P(|\mathbf{J}|)$, and maximise the penalised savings instead. This gives us the following anomaly detection statistic:

$$S := \max_{l \leq s-e \leq M} S(s, e) := \max_{l \leq s-e \leq M} \max_{\mathbf{J}} \big[S(s, e, \mathbf{J}) - P\big(|\mathbf{J}|\big)\big]. \tag{4}$$

Recall that l and M are the minimum and maximum segment length, respectively. An anomaly is declared if (4) is positive and the maximising $(s, e, \mathbf{J})$ is a point estimate of the anomaly's position in the data.

Throughout this article we use a piecewise linear penalty function of the form

$$P(|\mathbf{J}|) = \min\big(\alpha_{\text{sparse}} + \beta|\mathbf{J}|, \alpha_{\text{dense}}\big) = \begin{cases} \alpha_{\text{sparse}} + \beta|\mathbf{J}|, & |\mathbf{J}| < k^*, \\ \alpha_{\text{dense}}, & |\mathbf{J}| \geq k^*, \end{cases} \tag{5}$$

where $k^* = (\alpha_{\text{dense}} - \alpha_{\text{sparse}})/\beta$. We will refer to $|\mathbf{J}| < k^*$ as being in the *sparse regime* and $|\mathbf{J}| \geq k^*$ as being in the *dense regime*. Such a penalty function ensures that our method can be powerful against both sparse and dense alternatives. In addition, we can apply the results from Fisch, Eckley and Fearnhead (2021b) where it is shown that, if our modelling assumptions are correct, setting $\alpha_{\text{dense}} = p + 2\sqrt{p\psi} + 2\psi$, $\alpha_{\text{sparse}} = 2\psi$ and $\beta = 2\log(p)$, for $\psi = \log(n)$, results in a false positive rate that tends to 0 as n grows. Furthermore, Fisch, Eckley and Fearnhead (2021b) show that scaling the penalty function (5) by a factor b is appropriate in many situations where the modelling assumptions do not hold, such as when there is dependence over time.

Note that $[p]$ is always the maximiser in the dense regime and that β is the additional penalty for adding an extra variable to the anomalous subset in the sparse regime. We will exploit these properties when deriving an efficient optimisation algorithm in Section 3.2.

To compute the anomaly detection statistic, S, we need the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\mu}}(\mathbf{J})$ of $\boldsymbol{\mu}(\mathbf{J})$, where the means of variables $j \in \mathbf{J}$ are allowed to vary freely while the others are restricted to 0. Optimising the multivariate Gaussian likelihood (2), with respect to such a subset restricted mean, results in the following MLE for the mean components in $\mathbf{J}$:

$$\hat{\boldsymbol{\mu}}^{(\mathbf{J})}_{(s+1):e} = \bar{\mathbf{x}}^{(\mathbf{J})}_{(s+1):e} + \mathbf{Q}^{-1}_{\mathbf{J},\mathbf{J}}\mathbf{Q}_{\mathbf{J},-\mathbf{J}}\bar{\mathbf{x}}^{(-\mathbf{J})}_{(s+1):e}. \tag{6}$$

The corresponding p-vector $\hat{\boldsymbol{\mu}}(\mathbf{J})$ is constructed by placing $\hat{\boldsymbol{\mu}}^{(\mathbf{J})}$ at indices $\mathbf{J}$ and zeroes elsewhere. Finally, putting the MLE back into the expression for the saving and suppressing the subscripts $(s+1):e$ to not clutter the display gives us that

$$S(s, e, \mathbf{J}) = (e-s)\big(2\bar{\mathbf{x}} - \hat{\boldsymbol{\mu}}(\mathbf{J})\big)^{\mathsf{T}}\mathbf{Q}\hat{\boldsymbol{\mu}}(\mathbf{J}). \tag{7}$$

Unfortunately, the complicated form of the MLE (6) means that the number of operations required for finding the exact maximum penalised saving over subsets $\mathbf{J}$ is $O(2^p)$. The optimisation problem is not only combinatorial but also nonlinear, and, as far as we know, there is no reformulation of the saving (7) that would make the problem notably more tractable. We thus opt for an approximation to the saving (7) to achieve scalability.

3.2. *Approximate savings for anomaly detection.* Our idea for a computationally efficient approximation of the subset-maximised penalised savings $S(s, e)$ is to replace the MLE in (7) with the subset-truncated sample mean,

$$\bar{\mathbf{x}}(\mathbf{J}) = \bar{\mathbf{x}} \circ \mathbf{u}, \tag{8}$$

where $\mathbf{u} = (I\{i \in \mathbf{J}\})_{i=1}^{p}$ and $\circ$ is the elementwise (Hadamard) product. That is, under the sparse regime we aim to maximise the approximate penalised saving,

$$\tilde{S}(s, e) := \max_{\mathbf{J}}\big[\tilde{S}(s, e, \mathbf{J}) - P(|\mathbf{J}|)\big] = \max_{\mathbf{J}}\big[(e-s)\big(2\bar{\mathbf{x}} - \bar{\mathbf{x}}(\mathbf{J})\big)^{\mathsf{T}}\mathbf{Q}\bar{\mathbf{x}}(\mathbf{J}) - \beta|\mathbf{J}|\big] - \alpha_{\text{sparse}}. \tag{9}$$

Under the dense regime the exact maximum is given by $S(s, e, [p]) - \alpha_{\text{dense}}$.

An important motivation for using $\bar{\mathbf{x}}(\mathbf{J})$ is that finding $\tilde{S}(s, e)$ corresponds to what is known as a *binary quadratic program* (BQP). The unconstrained version of such optimisation problems are of the form

$$\max_{\mathbf{u}\in\{0,1\}^p} \mathbf{u}^{\mathsf{T}}\mathbf{A}\mathbf{u} + \mathbf{u}^{\mathsf{T}}\mathbf{b} + c, \tag{10}$$

where $\mathbf{A}$ is a real, symmetric, $(p \times p)$-dimensional matrix, $\mathbf{b}$ is a real, p-dimensional vector and c is a real scalar. BQPs are NP-hard, in general (Garey and Johnson (1979)), even if $\mathbf{A}$ is positive or negative definite. If $\mathbf{A}$ is r-banded, however, we show that BQPs can be solved with $O(p2^r)$ operations. Proposition 1 confirms that $\max_{\mathbf{J}}[\tilde{S}(s,e,\mathbf{J}) - P(|\mathbf{J}|)]$ is indeed a BQP. The proof is given in Section A.1 of the Supplementary Material (Tveten, Eckley and Fearnhead (2022)).

PROPOSITION 1. *Let $\alpha, \beta \geq 0$, $\bar{\mathbf{x}} \in \mathbb{R}^p$ and $\bar{\mathbf{x}}(\mathbf{J}) = \mathbf{u} \circ \bar{\mathbf{x}}$, where $\mathbf{u}$ is a binary vector with 1 at positions $\mathbf{J}$ and 0 elsewhere. Then, solving*

$$\max_{\mathbf{J}}\big[(e-s)\big(2\bar{\mathbf{x}} - \bar{\mathbf{x}}(\mathbf{J})\big)^{\mathsf{T}}\mathbf{Q}\bar{\mathbf{x}}(\mathbf{J}) - \beta|\mathbf{J}|\big] - \alpha, \tag{11}$$

corresponds to a BQP with $\mathbf{A} = -(e-s)\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathsf{T}} \circ \mathbf{Q}$, $\mathbf{b} = 2(e-s)(\bar{\mathbf{x}} \circ \mathbf{Q}\bar{\mathbf{x}}) - \beta$ and $c = -\alpha$.

To explain the dynamic program (Algorithm 1) for solving the BQP when the precision matrix $\mathbf{Q}$, and hence $\mathbf{A}$, is r-banded; it is illustrative to consider the case of $r = 1$. The key idea is that if we cycle through the variables in turn, then the choice of which of the variables $d, \ldots, p$ are anomalous will depend on the variables $1, \ldots, d-1$ only through whether variable $d-1$ is anomalous or not. Thus, we can obtain a recursion by considering these two possibilities separately.

In the case of $r = 1$, the BQP for $\max_{\mathbf{J}}[\tilde{S}(s,e,\mathbf{J}) - P(|\mathbf{J}|)]$ is given by

$$\max_{\mathbf{u}\in\{0,1\}^p} \sum_{d=1}^{p}(b_d + A_{d,d})u_d + 2\sum_{d=2}^{p} A_{d,d-1}u_d u_{d-1} + c, \tag{12}$$

where $A_{d,i} = -(e-s)Q_{d,i}\bar{x}_d\bar{x}_i$ for $i = d, d-1$, $b_d = 2(e-s)\bar{x}_d\sum_{i=d-1}^{d+1} Q_{d,i}\bar{x}_i - \beta$ and $c = -\alpha$. Let $\tilde{S}_1(d)$ and $\tilde{S}_0(d)$ be the maximal approximate penalised savings of variables $1, \ldots, d \leq p$ conditional on variable d being anomalous ($u_d = 1$) or not ($u_d = 0$) for a fixed s and e. Moreover, we write $\tilde{S}_{(0,u)}(d)$ and $\tilde{S}_{(1,u)}(d)$ for $u = 0, 1$ when additionally conditioning on variable $d-1$ being 0 or 1. Then, by initialising from $\tilde{S}(0) := c$, $\tilde{S}_0(1) = \tilde{S}(0)$ and $\tilde{S}_1(1) = \tilde{S}(0) + b_1 + A_{1,1}$, the following two-stage recursion holds for $d = 2, \ldots, p$:

$$\begin{aligned}\tilde{S}_{(0,u)}(d) &= \tilde{S}_u(d-1),\\ \tilde{S}_{(1,u)}(d) &= \tilde{S}_u(d-1) + b_d + A_{d,d} + 2uA_{d,d-1}\end{aligned} \tag{13}$$

for $u = 0, 1$, and

$$\tilde{S}_u(d) = \max\big(\tilde{S}_{(u,0)}(d), \tilde{S}_{(u,1)}(d)\big) \tag{14}$$

such that $\max(\tilde{S}_0(p), \tilde{S}_1(p)) = \max_{\mathbf{J}}[\tilde{S}(s,e,\mathbf{J}) - P(|\mathbf{J}|)]$ when $r = 1$. Note that the computational complexity of finding the optimum in this case is only $O(p)$.

To extend the recursion to more general precision matrices, observe that the dynamic program given by (13) and (14) can be described by an unbalanced binary tree (Figure 3). Initialisation occurs at levels 0 and 1 of the tree. Thereafter, two selected nodes at level $d-1$ grow children nodes, according to (13), before two of the four nodes at level d are selected as parents for the next level by the max operation in (14). The path from the maximum node at the final level back to the root encodes the optimal $\mathbf{u}$. In the following we will refer to the vector of 0's and 1's along the path from a certain node back up to the root as the "position" of a node.

By using the tree description, it is easier to generalise the algorithm to any neighbourhood structure of each variable d. When $r = 1$, we only have to consider the two options of variable $d-1$ being 0 or 1 at every step d, whereas for a general band, we have to consider all

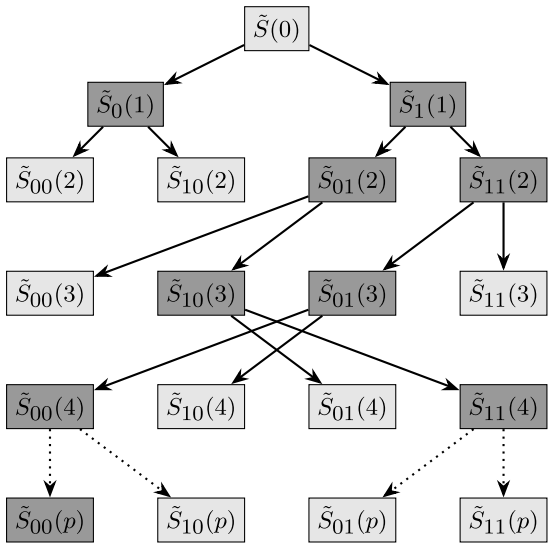


FIG. 3. *The unbalanced binary tree structure of the dynamic program for solving* (11) *for* 1*-banded* $\mathbf{Q}$ *and fictitious data. At each level, children nodes are grown, conditional on the value of the parent node, corresponding to variable/level* d *being anomalous* ($u_d = 1$) *or not anomalous* ($u_d = 0$). *The darker nodes are the selected parents of the nodes on the next level. Note that in this* 1*-banded example* (*a*) *one parent amongst* $\tilde{S}_{00}(d)$ *and* $\tilde{S}_{01}(d)$ *is always selected, assuming* $u_d = 0$, *and similarly* (*b*) *one parent amongst* $\tilde{S}_{10}(d)$ *and* $\tilde{S}_{11}(d)$ *is always selected, assuming* $u_d = 1$. *Observe that the maximum value to the BQP in this example is* $\tilde{S}_{00}(p)$, *with "position"* $\mathbf{u} = (1, 1, 0, 0, \ldots, 0)$.

combinations of variables $d - r, \ldots, d - 1$ being 0 or 1. A further adaptation to the precision matrix at hand can be made by excluding those variables among $d - r, \ldots, d - 1$ that will never be visited again, at each step d. To be precise, let us define the neighbours of variable d by $N_d := \{i : A_{d,i} \neq 0\}$, and the potential lower neighbours of d by $P_d^< := \{\max(1, d - r), \ldots, d - 1\}$ for $d \geq 2$ and $P_1^< := \varnothing$. At each step d we have to condition on all 0-1-combinations of the variables in

$$M_d := P_d^< \setminus \left(\bigcup_{i=d}^{d+r} N_i\right)^{\mathsf{c}} = P_d^< \cap \left(\bigcup_{i=d}^{d+r} N_i\right). \tag{15}$$

We call the variables in M_d the *extended neighbours* of d; see Figure 4 for an example of how the M_d's are constructed.

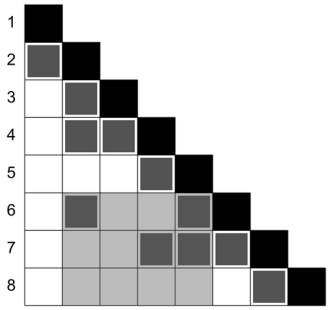


FIG. 4. *An example* 4*-banded* $\mathbf{A}$ *matrix where the diagonal is black, other nonzero elements are dark grey and zero-elements are white. The transparent, grey region illustrates how the extended neighbours of* $d = 6$ *are found; the column indices of the grey region correspond to* $P_6^< = \{2, 3, 4, 5\}$, *but variable* 3 *can be excluded, as it is not in any of the coming neighbourhoods, making* $M_6 = \{2, 4, 5\}$. *The other extended neighbourhoods in this example are* $M_1 = \varnothing$, $M_2 = \{1\}$, $M_3 = \{2\}$, $M_4 = \{2, 3\}$, $M_5 = \{2, 4\}$, $M_7 = \{4, 5, 6\}$ *and* $M_8 = \{7\}$.

To accomodate more complicated neighbourhood structures, we have to extend the scalar indicators u needed when $r = 1$, to vector indicators $\mathbf{u}_d \in \{0, 1\}^{|M_d|}$ that give us the position of a node in the tree relative to M_d. I.e., $\mathbf{u}_d$ tells us which extended neighbours of d are on (1) or off (0). At each level d, all $2^{|M_d|}$ possible on-off-combinations must be conditioned on, resulting in $2^{|M_d|+1}$ recursive updates, given by

$$\begin{aligned} \tilde{S}_{(0,\mathbf{u}_d)}(d) &= \tilde{S}_{\mathbf{u}_d}(d-1), \\ \tilde{S}_{(1,\mathbf{u}_d)}(d) &= \tilde{S}_{\mathbf{u}_d}(d-1) + b_d + A_{d,d} + 2\mathbf{u}_d^{\mathsf{T}}\mathbf{A}_{d,M_d}, \end{aligned} \tag{16}$$

where $(0, \mathbf{u}_d)$ and $(1, \mathbf{u}_d)$ indicates the positions of the 0-child and 1-child nodes relative to M_d. All these children nodes constitute the nodes at level d, and we will refer to them as $\{\tilde{S}(d)\}$.

The parent-selecting step in the general case also becomes more complex since the extended neighbourhoods can evolve in many different ways. To explain this step in detail, we use the notation position$(\tilde{S}(d))$ to refer to the 0-1-vector that gives the position of a given node in our binary tree representation of the algorithm. For example, position$(\tilde{S}_{10}(4)) = (1, 1, 0, 1)$ in Figure 3. Now, the parent for each $\mathbf{u}_d$ are determined by maximising over the variables that will never be visited again,

$$\tilde{S}_{\mathbf{u}_d}(d-1) = \max_{\mathbf{v}\in\mathbf{V}} \tilde{S}_{\mathbf{v}}(d-1), \tag{17}$$

where $\mathbf{V} = \{\mathbf{v} \in \text{positions}(\{\tilde{S}(d-1)\}) : \mathbf{v}^{(M_d)} = \mathbf{u}_d\}$ is the set of positions at level $d-1$ that match the on-off pattern indicated by $\mathbf{u}_d$ relative to M_d.

The final procedure is summarised in Algorithms 1 and 2. Note that we also keep track of the minimum number of anomalous variables at each level d through the term $\underline{k}$. In this way the recursions can be stopped as soon as the anomaly is guaranteed to lie in the dense regime. For an r-banded matrix the computational complexity is bounded by $O(\sum_{d=1}^{p} 2^{|M_d|}) \leq O(p2^r)$, and, if the anomaly is estimated as dense, the number of operations may be substantially less.

Algorithm 1 Dynamic programming BQP solver for banded matrices

Input: $\mathbf{A}$, $\mathbf{b}$, c, $\{M_d\}_{d=1}^{p}$, k^*

$d = 1$, $\underline{k} = 0$, $\tilde{S}(0) = c$.
while $d \leq p$ and $\underline{k} \leq k^*$ **do**
  for $\mathbf{u}_d \in \{0, 1\}^{|M_d|}$ **do**
    $\mathbf{V} = \{\mathbf{v} \in \text{positions}(\{\tilde{S}(d-1)\}) : \mathbf{v}^{(M_d)} = \mathbf{u}_d\}$.
    $\tilde{S}_{\mathbf{u}_d}(d-1) = \max_{\mathbf{v}\in\mathbf{V}} \tilde{S}_{\mathbf{v}}(d-1)$.
    $\tilde{S}_{(0,\mathbf{u}_d)}(d) = \tilde{S}_{\mathbf{u}_d}(d-1)$.
    $\tilde{S}_{(1,\mathbf{u}_d)}(d) = \tilde{S}_{\mathbf{u}_d}(d-1) + b_d + A_{d,d} + 2\mathbf{u}_d^{\mathsf{T}}\mathbf{A}_{d,M_d}$.
  end for
  $\underline{k} = \min_{\mathbf{v}\in\text{positions}\{\tilde{S}(d)\}} \mathbf{v}^{\mathsf{T}}\mathbf{1}$.
  $d = d + 1$.
end while
$\tilde{\mathbf{J}} = \text{argmax}\{\tilde{S}(p)\}$.
$\tilde{S} = \max\{\tilde{S}(p)\}$.
return: $\tilde{S}$, $\tilde{\mathbf{J}}$.

Algorithm 2 The approximate penalised saving for anomaly detection used in CAPA-CC

Input: $\bar{\mathbf{x}}$, $\mathbf{Q}$, $\{M_d\}_{d=1}^p$, β, α_{sparse}, α_{dense}, k^*, e, s.

$\mathbf{A} = -(e-s)\bar{\mathbf{x}}\bar{\mathbf{x}}^\intercal \circ \mathbf{Q}$.

$\mathbf{b} = 2(e-s)(\bar{\mathbf{x}} \circ \mathbf{Q}\bar{\mathbf{x}}) - \beta$.

$c = -\alpha_{\text{sparse}}$

$\tilde{S}, \tilde{\mathbf{J}}$ from Algorithm 1 with input $(\mathbf{A}, \mathbf{b}, c, \{M_d\}_{d=1}^p, k^*)$

$S = S(s, e, [p]) - \alpha_{\text{dense}}$.

if $\tilde{S} \geq S$ **return:** $\tilde{S}, \tilde{\mathbf{J}}$.

else return: S, $[p]$.

3.3. *Properties of the approximation.* Our main evaluation of the approximation's performance is done through simulations, where in Section C.1 of the Supplementary Material (Tveten, Eckley and Fearnhead (2022)) we demonstrate that the approximation and the MLE give almost equal results for low p. Some properties regarding how $\tilde{S}(s, e)$ compares to $S(s, e)$, however, can be derived theoretically.

First, under the dense penalty regime the approximate MLE is equal to the MLE because the optimal $\mathbf{J}$ is $[p]$ in both cases, making $\hat{\boldsymbol{\mu}}(\mathbf{J}) = \bar{\mathbf{x}}$. Thus, we are only approximating the savings under the sparse penalty regime.

Second, $\tilde{S}(s, e) \leq S(s, e)$ for all start- and end points s and e. This follows by definition of the MLE which is present in $S(s, e)$; $\hat{\boldsymbol{\mu}}(\mathbf{J})$ is the minimiser in (3), and, consequently, no other estimator can make the saving larger. Using the approximation will, therefore, not increase the probability of falsely detecting anomalies. The only effect it may have is a reduction in power.

In addition to the lower bound of 0 on the approximation error, Proposition 2 gives an upper bound which is useful for distilling what drives a potential decrease in performance. The proof is given in Section A.2 in the Supplementary Material (Tveten, Eckley and Fearnhead (2022)).

PROPOSITION 2. *Let* $\mathbf{W}(\mathbf{J})$ *be the matrix where* $\mathbf{W}(\mathbf{J})_{\mathbf{J},-\mathbf{J}} = \mathbf{Q}_{\mathbf{J},\mathbf{J}}^{-1}\mathbf{Q}_{\mathbf{J},-\mathbf{J}}$ *and is* 0 *elsewhere, and* $\hat{\mathbf{J}} = \operatorname{argmax}_{\mathbf{J}}[S(s, e, \mathbf{J}) - P(|\mathbf{J}|)]$. *Then, the following bound on the approximation error holds for all* $s < e$:

$$0 \leq S(s, e) - \tilde{S}(s, e) \leq (e-s)\lambda_{\max}\big(\mathbf{Q}\mathbf{W}(\hat{\mathbf{J}})\big)\big\|\bar{\mathbf{x}}_{(s+1):e}\big(\hat{\mathbf{J}}^{\mathbf{c}}\big)\big\|^2. \tag{18}$$

The right-hand side of (18) suggests that the relative approximation error will be largest for sparse anomalies in strongly correlated data, as this is the situation when $\|\bar{\mathbf{x}}_{(s+1):e}(\hat{\mathbf{J}}^{\mathbf{c}})\|^2$ is largest (see Section A.2 in the Supplementary Material (Tveten, Eckley and Fearnhead (2022))). The simulation results in Section C.1 of the Supplementary Material (Tveten, Eckley and Fearnhead (2022)) support this conclusion that the greatest difference in performance occurs when there is a sparse anomaly in strongly correlated data, although the difference is small in the tested settings.

3.4. *Multiple point and collective anomalies.* We can extend the described method for detecting a single collective anomaly to detecting multiple collective anomalies and also to allow for point anomalies within the baseline segments. To incorporate point anomalies, we follow the approach of Fisch, Eckley and Fearnhead (2021a, 2021b) by defining point anomalies as collective anomalies of length 1. Thus, the optimal approximate saving of a point anomaly at time t can be defined as

$$\tilde{S}'(t) = \max_{\mathbf{J}}\big[\tilde{S}(t, t, \mathbf{J}) - \beta'|\mathbf{J}|\big]. \tag{19}$$

In accordance with Fisch, Eckley and Fearnhead (2021b), we set $\beta' = 2\log p + 2\psi$, where $\psi = \log n$ as in Section 3.1. As for the collective anomaly penalty function, β' can be scaled by a constant factor b' to achieve appropriate error control.

We can now extend our penalised likelihood framework. The estimates for the collective anomalies, $\tilde{K}$ and $(\tilde{s}_k, \tilde{e}_k, \tilde{\mathbf{J}}_k)$ for $k = 1, \ldots, \tilde{K}$, and point anomalies, $\tilde{O}$ and $\tilde{\mathbf{J}}_t$ for $t \in \tilde{O}$, can then be obtained by minimising the penalised cost

$$\max_{K \in [\lfloor n/l \rfloor], s_k, e_k} \sum_{k=1}^{K} \tilde{S}(s_k, e_k) + \max_{O \subseteq [n]} \sum_{t \in O} \tilde{S}'(t), \tag{20}$$

subject to $\tilde{e}_k - \tilde{s}_k \geq l \geq 2$, $\tilde{e}_k \leq \tilde{s}_{k+1}$ and $(\bigcup_k [\tilde{s}_k + 1, \tilde{e}_k]) \cap \tilde{O} = \varnothing$.

The optimisation problem (20) can be solved exactly by a pruned dynamic program, using ideas from the PELT algorithm of Killick, Fearnhead and Eckley (2012). Defining $C(m)$ as the maximal penalised approximate savings for observations $\mathbf{x}_{1:m}$, the basis for our PELT algorithm is the following recursive relationship:

$$C(m) = \max\Big(C(m-1), \max_{0 \leq t \leq m-l} \big[C(t) + \tilde{S}(t, m)\big], C(m-1) + \tilde{S}'(t)\Big), \tag{21}$$

for $C(0) = 0$. The first term in the outer maximum corresponds to no anomaly at m, the second term to a collective anomaly ending at m and the third term to a point anomaly at m.

The computationally costly part of (21) is the maximisation over all possible starting-points t in the term for collective anomalies. Due to this term, the runtime of this dynamic program scales quadratically in n. If one specifies a maximum segment length M, however, the runtime is reduced to $O(Mn)$ at the risk of missing collective anomalies that are longer than M. The PELT algorithm is able to prune those t's in the term for the collective anomalies that can never be the maximisers, thus reducing computational cost whilst maintaining exactness. Proposition 3 gives a condition for when t can be pruned. The proof is given in Section A.3 of the Supplementary Material (Tveten, Eckley and Fearnhead (2022)).

PROPOSITION 3. *If there exists an $m \geq t - l$ such that*

$$C(t) + \tilde{S}(t, m) + \alpha_{dense} \leq C(m), \tag{22}$$

then, for all $m' \geq m + l$, $C(m') \geq C(t) + \tilde{S}(t, m')$.

Proposition 3 states that if (22) is true for some $m \geq t - l$, t can never be the optimal start-point of an anomaly for future times $m' \geq m + l$ and can, therefore, be skipped in the dynamic program. Killick, Fearnhead and Eckley (2012) show that if the number of change points increases linearly in n, then such a pruned dynamic program can scale linearly. In the worst case of no change points, however, the scaling is still quadratic in n.

Calculating $C(n)$ in (21) by PELT with savings computed from Algorithm 2 constitutes our CAPA-CC algorithm.

4. Relation to general change-point detection. So far, we have considered the anomaly detection problem which is a special case of the change-point detection problem. In the change-point model the changing parameter can change freely at every change point. The anomaly model restricts the change-point model by assuming there is a (known) baseline parameter the data reverts to at every other change point. In terms of the anomaly model (1), the change-point model is given by setting $e_k = s_{k+1}$ and $e_K = n$ and assuming all mean vectors to be unknown, including $\boldsymbol{\mu}_0$. In this section we highlight the benefits of making a distinction between changes and anomalies in light of our application and, briefly, describe how our method can be adapted to change-point detection in general.

In our application of condition monitoring a process pump as well as in other anomaly detection applications, the aim is to classify observations as either conforming to some baseline behaviour or being anomalous. Moreover, the majority of observations belong to the baseline group (or else the anomalies would not really be anomalies), and the remaining, anomalous observations may have any location and grouping within the data set (see Figure 1). The anomaly model adapts the change-point model to this setting by assuming that the baseline behaviour is characterised by a common stationary distribution, and that each anomaly is characterised by two (unknown) change points—one change from the baseline distribution to some other distribution and another change back to the baseline. In this way the model clearly distinguishes between which segments are in line with the baseline and those that are anomalous. In a general change-point model with K change points, on the other hand, observations are classified into $K+1$ distinct segments. These segments would subsequently have to be labelled as either baseline or anomalous by some additional rule if anomaly detection was the aim. In addition, the anomaly model enables borrowing strength across the entire data set for estimating the baseline distribution, rather than separately estimating each parameter between each anomaly, increasing the power of detecting anomalies.

If, however, a classical change-point analysis is of interest, the methodology described in Section 3 can be adapted. The overarching strategy in the corresponding change-point problem is to embed a test statistic for a single change point within binary segmentation or a related algorithm, such as wild binary segmentation (Fryzlewicz (2014)) or seeded binary segmentation (Kovács et al. (2020)). The test statistic for a single change point can be derived in a similar fashion as the test for a single anomaly given in (9). For a detailed derivation, algorithm and simulation results for the change-point detection problem, see the Supplementary Material (Tveten, Eckley and Fearnhead (2022)), Sections B and C.3.

5. Implementational details.

5.1. *Robustly estimating the mean and precision matrix.* To detect anomalies in practice, we need an estimate of $\mathbf{Q}$ and $\boldsymbol{\mu}_0$, as they are very rarely known a priori. We will use the median of each series $\mathbf{x}_{1:n}^{(i)}$ to estimate $\boldsymbol{\mu}_0^{(i)}$. To estimate $\mathbf{Q}$, we use a robust version of the GLASSO algorithm (Friedman, Hastie and Tibshirani (2008)). This algorithm takes as input an estimate of the covariance matrix, $\hat{\boldsymbol{\Sigma}}$, and an adjacency matrix $\mathbf{W}$. An estimate $\hat{\mathbf{Q}}(\mathbf{W})$ of $\mathbf{Q}$ is then computed by maximising the penalised log-likelihood

$$\log\det\Theta - \operatorname{tr}(\hat{\boldsymbol{\Sigma}}\Theta) - \|\Gamma \circ \Theta\|_1 \tag{23}$$

over nonnegative definite matrices Θ, where we define the entries of Γ to be $\gamma_{ij} = 0$ if $w_{ij} = 1$ or $i = j$ and $\gamma_{ij} = \infty$ otherwise. This can be seen as producing the closest estimate of $\mathbf{Q}$, based on $\hat{\boldsymbol{\Sigma}}^{-1}$, subject to the sparsity pattern imposed by $\mathbf{W}$. To compute $\hat{\mathbf{Q}}$ efficiently, we use the R package `glassoFast` (Sustik and Calderhead (2012)).

As input for $\hat{\boldsymbol{\Sigma}}$, we use an estimate, $\mathbf{S}$, of the covariance in the raw data that is robust to the presence of anomalies. Our robust estimator is constructed from the Gaussian rank correlation and the median absolute deviation estimator of the standard deviation, as suggested by Öllerer and Croux (2015). To be precise, let $\operatorname{mad}(\mathbf{x}^{(i)})$ be the median absolute deviation of all measurements of variable i, and

$$r_{\text{Gauss}}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) := r\big(\Phi^{-1}\big(R(\mathbf{x}^{(i)})/(n+1)\big), \Phi^{-1}\big(R(\mathbf{x}^{(j)})/(n+1)\big)\big) \tag{24}$$

be the Gaussian rank correlation between variables i and j, where r is the sample Pearson correlation and $R(\mathbf{x})$ is a vector of the ranks of each x_t within $\mathbf{x}$. Then, the robust pairwise covariances are estimated by

$$s_{ij} = \operatorname{mad}(\mathbf{x}_{1:n}^{(i)})\operatorname{mad}(\mathbf{x}_{1:n}^{(j)})r_{\text{Gauss}}(\mathbf{x}_{1:n}^{(i)}, \mathbf{x}_{1:n}^{(j)}). \tag{25}$$

5.2. *Tuning*. There are two primary tuning parameters in CAPA-CC: The adjacency matrix $\mathbf{W}$ in the precision matrix estimator of Section 5.1 and the scaling factor b in the penalty function for collective anomalies. This section contains guidelines for tuning them after some notes on the remaining tuning parameters.

The scaling factor for the point anomaly penalty, b', can be tuned separately from b if the application dictates it, but a reasonable default is to let $b' = b$ and tune both penalties simultaneously. The minimum and maximum segment lengths of collective anomalies, l and M, are tuning parameters solely for the convenience of speeding up computation if there is knowledge of such limits in an application. Otherwise, they default to $l = 2$ and $M = n$.

A number of different considerations can go into choosing $\mathbf{W}$. From a modelling perspective, selecting $\mathbf{W}$ corresponds to deciding on a model for the conditional independence structure; $w_{ij} = 0$ means variables are assumed to be conditionally independent, while $w_{ij} = 1$ means variables are conditionally dependent. For spatial data, for example, the choice of $\mathbf{W}$ is the same as choosing the neighbourhood structure in a conditional autoregressive model, where $w_{ij} = 1$ if and only if spatial region i is a neighbour of spatial region j. In our process pump example this would mean specifying which sensors are neighbours.

Computational considerations can also guide the choice of $\mathbf{W}$, however. As we have seen, CAPA-CC scales exponentially in the band of $\mathbf{Q}$. Hence, the band of $\mathbf{W}$ governs the runtime of our algorithms to a large extent. A reasonable default choice of $\mathbf{W}$ is, therefore, a low value of r in the r-banded adjacency matrix $\mathbf{W}(r)$, defined by

$$w_{ij} = \begin{cases} 1 & \text{if } 0 < |i - j| \leq r, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

In the simulations of the next section, we illustrate that good performance can be achieved even when specifying $\mathbf{W}$ to have a much narrower band than the true $\mathbf{Q}$.

In cases where the precision matrix is sparse but not banded, bandwidth reduction algorithms, such as the Cuthill–McKee algorithm (Cuthill and McKee (1969)) and the Gibbs–Poole–Stockmeyer algorithm (Lewis (1982)), can be a useful preprocessing step before running CAPA-CC.

Several strategies can also be employed for tuning the penalty scaling factor b. The first strategy requires a training set containing only baseline observations. This training set can either be used to estimate a model (e.g., Gaussian) of the baseline behaviour of the data or to constitute the empirical distribution of the baseline data. Anomaly-free data sets can then be sampled parametrically or nonparametrically from the baseline model to obtain bootstrap estimates $\hat{\alpha}$ of $\alpha = P(\hat{K} > 0 | K = 0)$ for a fixed b. A practitioner can thus select a target probability of false positives α and find b that meets this criterion within a selected interval of error, $\alpha \pm \delta$, and level of confidence governed by the number of repetitions used to calculate $\hat{\alpha}$ per b.

A second criterion is to find the smallest b such that a user-selected tolerable number of false alarms is raised in the training set. This strategy is much less computationally intensive, as it avoids the bootstrap sampling, but the error control hinges more strongly on how generalisable the training set is.

If there is no training set available, a third tuning strategy is to adjust b until a desired number of anomalies are output by CAPA-CC. As b is increased, the ordering of the anomalies in terms of significance will gradually be revealed. We explore the pump data set by this tuning strategy in Section 7.

6. Simulation study. We next turn to examine the power and estimation accuracy of CAPA-CC in a range of data settings. In almost all cases we test the robustness of the method

against an incorrectly specified adjacency matrix in the precision matrix estimate. We concentrate on the single anomaly setting first, before comparing several state-of-the-art methods in the multiple anomaly setting.

We have chosen a widely used one-parameter version of the *conditional autoregressive* (CAR) model, called the *row-standardised* CAR model, as our primary testbed (see, for instance, Ver Hoef, Hanks and Hooten (2018) for a concise introduction). This CAR model is given by

$$\mathbf{Q}_{\mathrm{CAR}}(\rho, \mathbf{W}) := \operatorname{diag}(\mathbf{W}\mathbf{1}) - \rho\mathbf{W}, \tag{27}$$

where $\mathbf{W}$ is an adjacency matrix as before. $\mathbf{Q}_{\mathrm{CAR}}$ is then standardised so that $\mathbf{Q}^{-1}$ becomes a correlation matrix, and we let $\boldsymbol{\mu}_0 = \mathbf{0}$ throughout. Conveniently, the sparsity structure of $\mathbf{Q}_{\mathrm{CAR}}$ follows directly from the design of $\mathbf{W}$. In our simulations we consider data with precision matrices corresponding to the r-banded neighbourhood structures given in (26) and regular lattice neighbourhood structures. To define the $m \times m$ lattice adjacency matrix, let (u, v) denote the coordinate of a node in the lattice for $0 \leq u, v \leq m$. The neighbourhood of (u, v) is considered to be $\{(u-1, v), (u+1, v), (u, v-1), (u, v+1)\}$. Coordinates are then enumerated by $i = (u-1)m + v$ such that the square lattice adjacency matrix $\mathbf{W}_{\mathrm{lat}}$ can be defined by $w_{ij} = 1$ if i and j are neighbours and 0 otherwise. For the sake of brevity, we also define $\mathbf{Q}_{\mathrm{lat}}(\rho) := \mathbf{Q}_{\mathrm{CAR}}(\rho, \mathbf{W}_{\mathrm{lat}})$ and $\mathbf{Q}(\rho, r) := \mathbf{Q}_{\mathrm{CAR}}(\rho, \mathbf{W}(r))$. In addition to the CAR models, we also test performance under the constant correlation model, given by

$$\mathbf{Q}_{\mathrm{con}}(\rho) := \left(\rho\mathbf{1}\mathbf{1}^{\mathsf{T}} + (1-\rho)\mathbf{I}\right)^{-1}. \tag{28}$$

Note that we use $\mathbf{W}^*$ to refer to the true adjacency matrix of the data.

If more than one series changes, the power of different methods may depend on how similarly each series change. To investigate this, we consider the following ways of simulating anomalous means, $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$: $\boldsymbol{\mu}_k^{(\mathbf{J}_k)} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{J}_k, \mathbf{J}_k})$, where $\boldsymbol{\Sigma}$ is the data covariance matrix and $\boldsymbol{\mu}_k^{(\mathbf{J}_k)} \sim N(\mathbf{0}, (\mathbf{Q}_{\mathrm{con}}(\rho))^{-1})$. We refer to anomalies being drawn from the former and latter classes, respectively, by $\boldsymbol{\mu}_{(\boldsymbol{\Sigma})}$ and $\boldsymbol{\mu}_{(\rho)}$. Note that $\rho = 0$ and $\rho = 1$ correspond to the special cases of the means being independent and equal for the changing variables, respectively. After sampling a mean vector, it is scaled by a constant to achieve a specific signal strength $\vartheta_k := \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_0\|_2 = \|\boldsymbol{\mu}_k\|_2$. Moreover, unless stated otherwise, we let $\mathbf{J}_k = \{1, 2, \ldots, J_k\}$, where $J_k \in [p]$ denotes the number of changing variables.

In all simulations the penalty functions or detection thresholds are tuned to achieve $\alpha = 0.05 \pm 0.02$ probability of false positives in data simulated from the appropriate true null distribution (see Section 5.2). One thousand and 500 bootstrap repetitions were used for each b to obtain $\hat{\alpha}$ for $p = 10$ and $p = 100$, respectively.

6.1. *Single anomaly detection.* To the best of our knowledge, there are no other statistical methods tailored for jointly detecting sparse and dense anomalies in correlated multivariate data. A comparison between methods for independent multivariate data was performed by Fisch, Eckley and Fearnhead (2021b), where their MVCAPA method was shown to generally outperform other competitors. Hence, we focus on comparing MVCAPA against a range of CAPA-CC scenarios, including various incorrectly specified versions, exploring the trade-offs between the two methods. We evaluate methods in terms of power to detect an anomaly of increasing signal strength and also assess the correctness of the estimated subset of anomalous variables, $\mathbf{J}$.

In the following, "Whiten + MVCAPA" means that the input to MVCAPA are the whitened observations $\mathbf{S}^{-1/2}\mathbf{x}_t$, where $\mathbf{S}$ is the robust covariance matrix estimate (25), whereas a plain "MVCAPA" takes the raw data $\mathbf{x}_t$. Note that Whiten + MVCAPA scrambles the sparsity structure of an anomalous mean such that the recovery of $\mathbf{J}$ is lost. It is, however, still interesting to include in the comparisons of detection power as no sparsity structure has

to be imposed on the covariance or precision matrix, in contrast to CAPA-CC. We, therefore, expect Whiten + MVCAPA to perform well when the precision matrix as well as the change is dense.

6.1.1. *Independence vs. dependence.* As the performance of the anomaly detection methods we consider ultimately hinges on the performance of a test statistic at each pair (s, e), we compare performance assuming that the location of the collective anomaly is known a priori. That is, we fix the collective anomaly at $(s, e) = (n/2, n/2 + 10)$ and compare the power of $\tilde{S}(s, e)$ with the corresponding test statistic, assuming cross-independence used within MVCAPA. In CAPA-CC we test using the true precision matrix $\mathbf{Q}$, an estimate based on the true adjacency structure $\hat{\mathbf{Q}}(\mathbf{W}^*)$, as well as misspecified banded adjacency structures with $r = 1, 2, 4$. The power at each point along the power curve is estimated from 1000 ($p = 10$) or 500 ($p = 100$) simulated datasets, and the same datasets were used for all methods. The full set of tested scenarios include all combinations of $\{(n, p), \mathbf{Q}, \rho, J, \boldsymbol{\mu}_{(\cdot)}\}$ for $(n, p) = (100, 10), (200, 100)$, $\mathbf{Q} = \mathbf{Q}(2), \mathbf{Q}_{\text{lat}}, \mathbf{Q}_{\text{con}}$, $\rho = 0.3, 0.5, 0.7, 0.9, 0.99$, $J = 1, \lfloor\sqrt{p}\rfloor, p$ and change classes $\boldsymbol{\mu}_{(\boldsymbol{\Sigma})}$, $\boldsymbol{\mu}_{(0)}$, $\boldsymbol{\mu}_{(0.8)}$, $\boldsymbol{\mu}_{(0.9)}$ and $\boldsymbol{\mu}_{(1)}$. In addition, we have also varied which series are anomalous for selected scenarios. Note that CAPA-CC($\mathbf{Q}$) represents the performance of an oracle method. For larger n relative to p, however, the difference between CAPA-CC($\mathbf{Q}$) and CAPA-CC($\hat{\mathbf{Q}}(\mathbf{W}^*)$) will decrease.

A first main finding, illustrated in Figure 5, is that, for detecting a single anomalous variable, incorporating correlations leads to higher power, even when misspecifying the structure of the precision matrix estimate. The stronger the correlation, the higher the gain in power. For a collection of densely correlated variables, even using a 1-banded estimate of the precision matrix leads to a big improvement in power for sparse anomalies, compared to MVCAPA (the bottom row of plots). It is somewhat surprising that Whiten + MVCAPA performs comparably to CAPA-CC in this setting of a very sparse change.

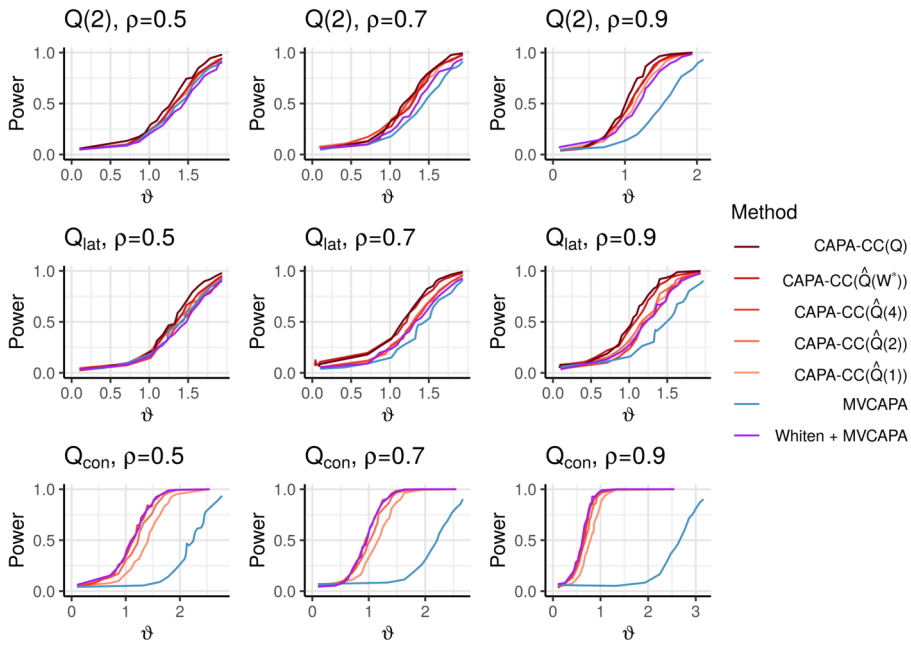
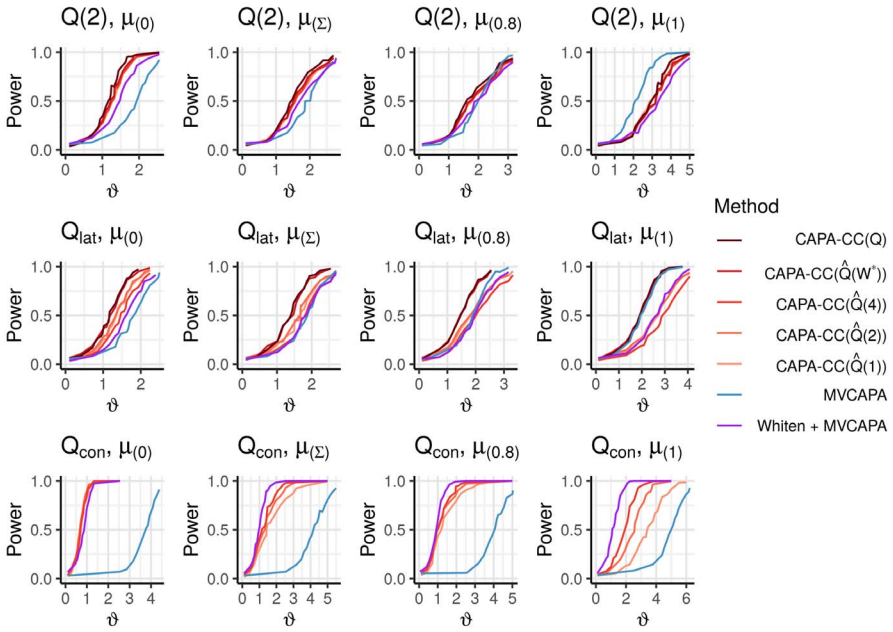
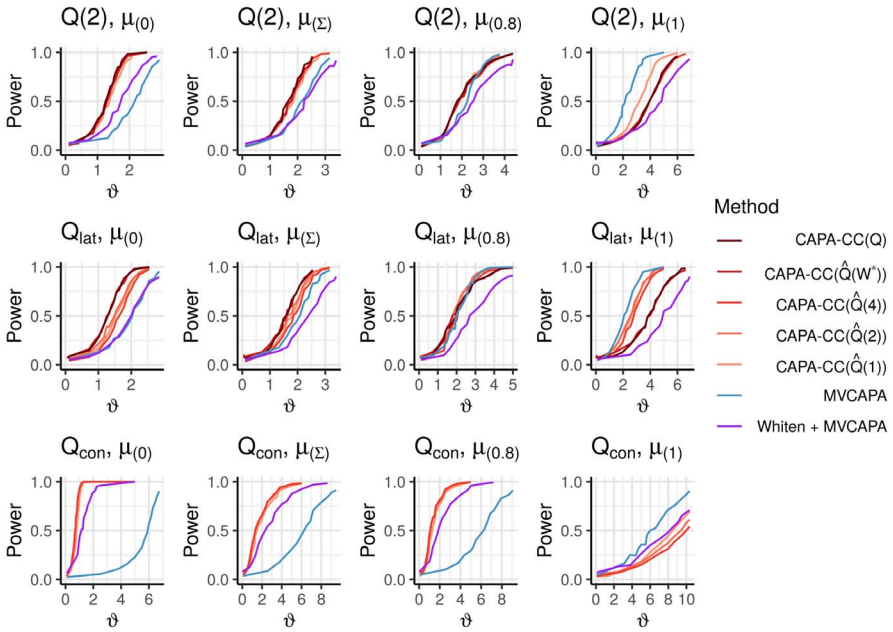


FIG. 5. *Power curves for correct and misspecified versions of CAPA-CC for a single known anomaly at* $(s, e) = (100, 110)$ *when* $J = 1$ *and* $p = 100$. *Results for* 2*-banded, lattice and globally constant correlation precision matrices are shown from top to bottom, with increasing* ρ *from left to right. Other parameters*: $n = 200$, $\alpha = 0.05$ *and* 500 *repetitions were used during tuning and for each point along the power curves.*



(a) $J = 10$.



(b) $J = 100$.

FIG. 6. *Power curves for a single known anomaly at* $(s, e) = (100, 110)$ *and* (*a*) $J = 10$ *and* (*b*) $J = 100$, *when* $p = 100$ *and* $\rho = 0.9$. *The methods are the same as in Figure* 5. *From left to right, the columns of plots show results for the anomalous means being sampled from* $N(\mathbf{0}, \mathbf{I})$, $N(\mathbf{0}, \mathbf{\Sigma})$, $N(\mathbf{0}, \mathbf{Q}_{\text{con}}^{-1}(0.8))$ *and* $\mu^{(i)} = \mu$ *for all* $i \in \mathbf{J}$ *in the right-most column. From top to bottom are results for* 2*-banded, lattice and global constant correlation data precision matrices. Other parameters*: $n = 200$, $\alpha = 0.05$ *and* 500 *repetitions per point along the power curves.*

The picture for more than one anomalous variable is more complex. Figure 6 displays the results for different precision matrices and classes of changes for $p = 100$ and $\rho = 0.9$ when (a) $J = 10$ and (b) $J = 100$. Observe that, for all precision matrices and J's (entire first col-

umn), CAPA-CC is superior for anomalous means sampled from the independent normal distribution ($\boldsymbol{\mu}_{(0)}$). This is also the case when the anomalous means are sampled from a normal distribution with the data correlation matrix ($\boldsymbol{\mu}_{(\boldsymbol{\Sigma})}$) (entire second column), with the exception of $J = 10$ and global constant correlation. The power of CAPA-CC decreases, however, when the anomalous means have very similar or equal values, as in the case of means being sampled from to $\boldsymbol{\mu}_{(0.8)}$ and $\boldsymbol{\mu}_{(1)}$. Surprisingly, for the special case of equally sized anomalous means and a banded or lattice precision matrix, MVCAPA is more powerful than using the true model for the precision in CAPA-CC($\mathbf{Q}$). For $J = 100$, this is also the case for equal changes in the global constant correlation model. The same phenomenon can be observed for other methods as well (see Section B in the Supplementary Material (Tveten, Eckley and Fearnhead (2022))), and we discuss it further in Section 8. As expected, Whiten + MVCAPA performs well for $\mathbf{Q}_{\text{con}}$ precision matrices, but the misspecified versions of CAPA-CC outperforms it when $J = 100$. For low values of ρ, we observe almost no difference between the different methods which is why we focus on $\rho \geq 0.5$. For higher values of ρ than 0.9, the gain from incorporating correlations in the method increases. For $p = 10$, the corresponding results look qualitatively similar; see Section C.2 of the Supplementary Material (Tveten, Eckley and Fearnhead (2022)) for more details.

6.1.2. *Variable selection.* Although CAPA-CC is not designed to estimate $\mathbf{J}$ consistently, it is worth investigating the behaviour of $\hat{\mathbf{J}}$ so that it is interpreted with sufficient caution. Note that we now use $\hat{\mathbf{J}}$ to refer to the output estimate of $\mathbf{J}$ for all algorithms. Also, recall that we let $J := |\mathbf{J}|$ and $\hat{J} := |\hat{\mathbf{J}}|$.

For $p = 10$ and 100, the precision and recall of $\hat{\mathbf{J}}$ from MVCAPA as well as both true and misspecified versions of CAPA-CC were compared in the single known anomaly setting, described in Section 6.1.1. We also included the exact ML method for $p = 10$. Whiten + MVCAPA is excluded from these simulations since the decorrelation transform breaks up the sparsity structure of the anomalies.

Under a 2-banded precision matrix model we see from Tables 3 and 4 in the Supplementary Material (Tveten, Eckley and Fearnhead (2022)) that both CAPA-CC and the exact ML method tend to have higher recall, but slightly lower precision, than MVCAPA. The reason for this is illustrated in Figure 7, where it can be observed that all the methods that incorporate cross-correlations overestimate J more frequently than MVCAPA. In particular, CAPA-CC more often estimates anomalies as dense. This effect is seen more clearly for $p = 100$ (Figure 20 in the Supplementary Material (Tveten, Eckley and Fearnhead (2022))), where estimating J becomes increasingly hard as J grows closer to the boundary k^* between sparse and dense changes. Moreover, we found that the estimated subset is quite sensitive to the scaling of the penalties relative to the signal strength ϑ. If a more accurate estimate of

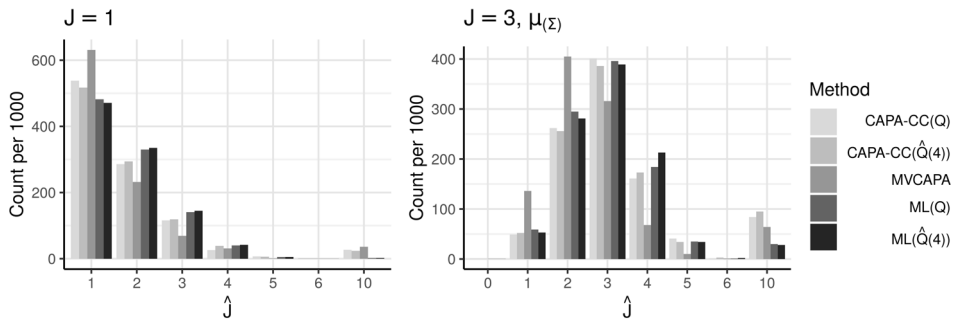


FIG. 7. *Estimated sizes of* $\mathbf{J}$ *for* $\mathbf{J} = \{1\}$ (*left*) *and* $\mathbf{J} = \{1, 2, 3\}$ (*right*) *when* $p = 10$ *and the location of the anomaly is assumed known. Other parameters*: $n = 100$, $\mathbf{Q} = \mathbf{Q}(2, 0.9)$, $s = 10$, $e = 20$, $\vartheta = 2$, $\boldsymbol{\mu}_{(\boldsymbol{\Sigma})}$, $\alpha = 0.005$.

$\mathbf{J}$ is desired, we thus recommend running a postprocessing step by optimising the penalised saving for each anomalous segment using only the sparse penalty regime.

6.2. *Multiple anomaly detection*. The simulation study is concluded by comparing the following methods in a multiple anomaly setting with and without point anomalies:

- CAPA-CC with a misspecified precision matrix $\hat{\mathbf{Q}}(4)$.
- MVCAPA and Whiten + MVCAPA.
- The inspect method of Wang and Samworth (2018). We test both the version assuming cross-independence implemented in the R-package `InspectChangepoint` as well as the version including cross-correlations discussed in their paper. To distinguish the two versions, we refer to the former as inspect($\mathbf{I}$) and the latter as inspect($\hat{\mathbf{Q}}$), where $\hat{\mathbf{Q}}$ is the inverse of the robust covariance matrix estimator (25).
- The group fused LARS method of Bleakley and Vert (2011), implemented in the R-package `jointseg`.

In addition, we tested the methods of Wang et al. (2020) and Safikhani and Shojaie (2020) for detecting changes in vector autoregressive models, but they were excluded due to poor computational scaling in p or n. For example, the method of Wang et al. (2020) with a maximum segment length of 100 takes around 13 minutes to complete on a single $p = 10$, $n = 1000$ data set on a typical computer, and the method of Safikhani and Shojaie (2020) scales exponentially in $\hat{K}$. The included methods are all tuned to a specific false positive probability α on data sets of size $\min(n, 200)$, except the group fused LARS, which uses the default model selection procedure of `jointseg` proposed in Bleakley and Vert (2011). To speed up computation, we set the maximum segment length of CAPA-CC and MVCAPA to $M = 100$

Performance is measured by the *adjusted rand index* (ARI; Hubert and Arabie (1985)) of classifying observations as either anomalous (point or collective) or baseline. The ARI measures the accuracy of the classification but adjusts for the sizes of the classes. It is, therefore, suitable in an unbalanced classification problem such as ours.

As inspect and the group fused LARS method are not made specifically for the anomaly setting, as opposed to MVCAPA and CAPA-CC, we do not expect them to be competitive. However, since they could be used for the purpose, we include them to measure the gain of using a dedicated anomaly detection method rather than a generic change-point detection method. Our heuristic for turning the change-point detection methods into an anomaly classifier is as follows: If the sample mean of an estimated segment has L_2 norm greater than 1, the observations within the segment are classified as anomalous, and if the L_2 norm is smaller than or equal to 1, they are classified as baseline. Adjacent segments, where both are classified as collective anomalies, by this rule are also merged to a single collective anomaly if the sign of $\sum_{j=1}^{p} \bar{x}^{(j)}_{(s+1):e}$ in each of the two segments is the same.

Also, note that we use a misspecified precision matrix in CAPA-CC since this is most realistic, but improved performance on the order of what can be seen in Figures 5 and 6 could be achieved by selecting the correct model.

Table 1 displays the results for $p = 100$, $n = 1000$ with three evenly spaced collective anomalies of lengths (30, 20, 10), different affected subsets, affected means sampled from $\boldsymbol{\mu}_{(\boldsymbol{\Sigma})}$ and $\vartheta = 1$ in signal strengths of sizes $\vartheta(1, 2, 3)$. The results are again generally favorable for CAPA-CC($\hat{\mathbf{Q}}(4)$), in particular, for the banded and lattice precision matrices, while Whiten + MVCAPA is slightly better for the global constant correlation matrix when point anomalies are absent. The group fused LARS and inspect($\mathbf{I}$) methods achieved approximately 0 ARI on all the tested scenarios, including the different signal strength parameters of $\vartheta = 1, 1.5, 2$.

TABLE 1

ARI of classifying baseline and anomalous observations when $p = 100$, $n = 1000$, $(\vartheta_k)_{k=1}^3 = (1, 2, 3)$, *the change class is* $\boldsymbol{\mu}_{(\boldsymbol{\Sigma})}$, $\{(s_k, e_k)\}_{k=1}^3 = \{(300, 330), (600, 620), (900, 910)\}$ *and* $\mathbf{J}_1 = \{1\}$, $\mathbf{J}_2 = \{1, \ldots, 10\}$, $\mathbf{J}_3 = \{1, \ldots, 10, 46, \ldots, 55, 91, \ldots, 100\}$, *based on* 100 *repetitions. Point anomalies are placed at* 10 *fixed locations, each randomly affecting a single variable with size sampled from* $N(0, 4 \log p)$. *The largest value for each data setting is given in bold. Note that the results for inspect*($\mathbf{I}$) *and the group fused LARS methods are excluded from the table since their ARIs are approximately* 0 *in all the tested scenarios*

$\mathbf{Q}$	ρ	Pt. anoms	CAPA-CC($\hat{\mathbf{Q}}(4)$)	W + MVCAPA	MVCAPA	inspect($\hat{\mathbf{Q}}$)
$\mathbf{Q}(2)$	0.5	–	**0.23**	0.09	0.20	0.05
$\mathbf{Q}(2)$	0.5	✓	**0.40**	0.25	0.37	0.01
$\mathbf{Q}(2)$	0.7	–	**0.34**	0.19	0.12	0.06
$\mathbf{Q}(2)$	0.7	✓	**0.43**	0.30	0.31	0.00
$\mathbf{Q}(2)$	0.9	–	**0.53**	0.43	0.05	0.13
$\mathbf{Q}(2)$	0.9	✓	**0.61**	0.46	0.26	0.03
$\mathbf{Q}_{\text{lat}}$	0.5	–	**0.21**	0.08	0.12	0.05
$\mathbf{Q}_{\text{lat}}$	0.5	✓	**0.29**	0.26	0.25	0.08
$\mathbf{Q}_{\text{lat}}$	0.7	–	**0.27**	0.21	0.13	0.05
$\mathbf{Q}_{\text{lat}}$	0.7	✓	**0.35**	0.31	0.25	0.10
$\mathbf{Q}_{\text{lat}}$	0.9	–	**0.34**	0.28	0.09	0.08
$\mathbf{Q}_{\text{lat}}$	0.9	✓	0.33	**0.42**	0.18	0.14
$\mathbf{Q}_{\text{con}}$	0.5	–	0.44	**0.52**	0.00	0.06
$\mathbf{Q}_{\text{con}}$	0.5	✓	**0.50**	0.49	0.11	0.03
$\mathbf{Q}_{\text{con}}$	0.7	–	0.60	**0.65**	0.00	0.08
$\mathbf{Q}_{\text{con}}$	0.7	✓	**0.66**	0.64	0.10	0.04
$\mathbf{Q}_{\text{con}}$	0.9	–	0.66	**0.82**	0.00	0.26
$\mathbf{Q}_{\text{con}}$	0.9	✓	0.71	**0.82**	0.09	0.10

The full set of multiple anomaly simulation results, covering anomalous means sampled from $\boldsymbol{\mu}_{(0)}$ and $\boldsymbol{\mu}_{(0.8)}$ in addition to $\boldsymbol{\mu}_{(\boldsymbol{\Sigma})}$, and $\vartheta = 1.5, 2$ in addition to $\vartheta = 1$, can be found in Section C.4 of the Supplementary Material (Tveten, Eckley and Fearnhead (2022)). The results for $\boldsymbol{\mu}_{(0.8)}$ are very similar to the results for $\boldsymbol{\mu}_{(\boldsymbol{\Sigma})}$, and the results for $\boldsymbol{\mu}_{(0)}$ are slightly more favorable for CAPA-CC compared to the other methods. As ϑ increases, the ARI of all methods increase, and the differences in performance decrease. In the scenarios with point anomalies when $\vartheta = 1.5$ and $\vartheta = 2$, a lot is gained by using CAPA-CC or (Whiten +) MVCAPA rather than inspect.

7. Pump data analysis. We now return to the problem of inferring anomalous segments and variables in the pump data described in the Introduction. Recall that the data was preprocessed by regressing a set of monitoring variables onto a set of state variables, such that we are left with five series of residuals to detect anomalies in (Figure 1). Some of the residuals are strongly correlated (Figure 8), suggesting that incorporating cross-correlations when modelling them is advantageous based on our simulation study.

Before running CAPA-CC on the pump data, the penalties must be tuned and input parameters selected. The tuning of the penalties accounts for all features in the data that we have not modelled, for example, autocorrelation, a nonstationary correlation matrix and trends in the data's mean not associated with segments of suboptimal operation. As we do not have training data guaranteed to only contain baseline observations, we instead tune the penalties such that a chosen number of the most significant anomalies are output, as discussed in Section 5.2. To test performance, we tune b such that the correct number of collective anomalies (four) are output to see how they align with the known ones. Since there are many outliers in the data set, we want to retain the default level of outlier-robustness and, therefore, keep

	1	2	3	4	5
1	1	0.89	0.08	-0.14	0.14
2		1	0.18	-0.02	0.13
3			1	0.78	0.64
4				1	0.67
5					1

FIG. 8. *The robustly estimated correlations* (*see* (25)) *of the pump data after preprocessing*.

the point anomaly scaling at 1, while adjusting b. This tuning procedure resulted in a scaling factor of $b = 11$. For the remaining inputs, we set $\mathbf{Q}$ to the inverse of the correlation matrix in Figure 8, a minimum segment length $l = 5$, and use no maximum segment length; see Section D in the Supplementary Material (Tveten, Eckley and Fearnhead (2022)) for results on the robustness to these choices of tuning parameters.

The final result is shown in Figure 9. Before interpreting the output, it is important to know that the start points of the known anomalies are more uncertain than the end points; the end point is the time where the pump was brought back to normal operation, whereas the start point has been set based on a retrospective analysis by the engineers. With this in mind, we observe that three out of four estimated collective anomalies are within three separate known anomalous segments, with the estimated end points being more accurate than the estimated start points. The short known anomaly from $t = 125$ to $t = 135$ is missed, as there is virtually no signal of it in the data. The estimated anomaly from $t = 1306$ to $t = 1362$, however, does not overlap with a known anomaly, but it clearly looks anomalous by eye. This segment

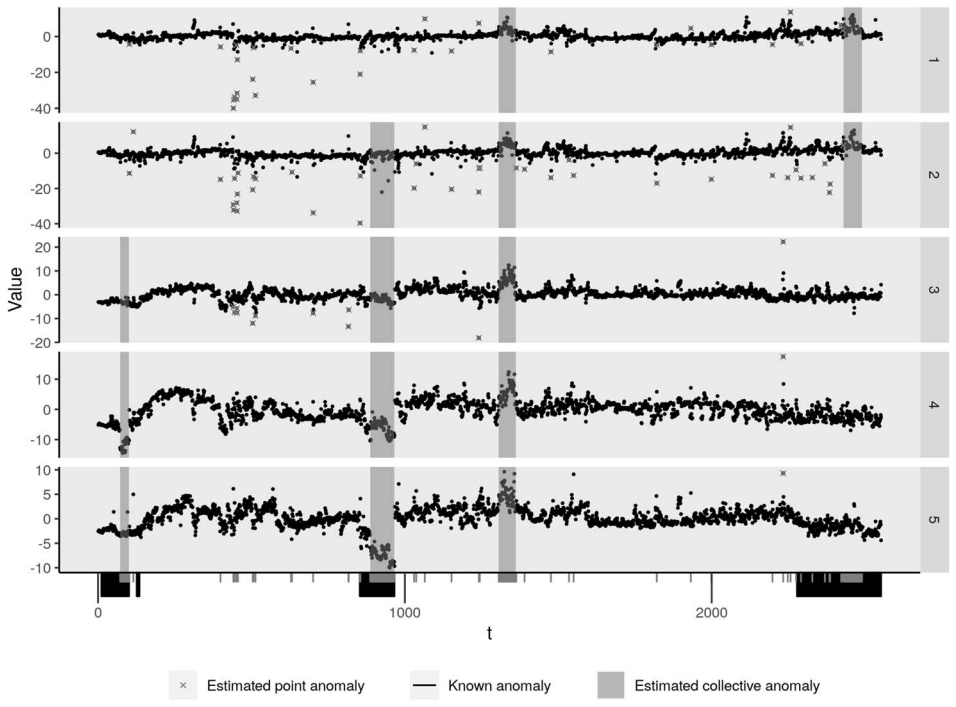


FIG. 9. *The four most significant estimated collective anomalies in the five residual times series derived from the pump data. Tuning parameters*: $b = 11$, $b_{\text{point}} = 1$, $l = 5$ *and* $M = n$.

is also of interest to detect since it may correspond to an unknown segment of suboptimal operation. If not, this segment points to a part of the data that fits our linear regression model poorly, indicating that a more sophisticated model might be in order if fewer false alarms are required. In general we expect that a better model for linking the state variables with the monitoring variables would improve the results even further because more of the trend in the mean not associated with the known anomalies would be absorbed by the model rather than leaking into the residuals.

In addition, notice the importance of including point anomalies in the analysis for this application. Rerunning CAPA-CC on the data without inferring point anomalies resulted in four additional false collective anomalies being inferred for $b = 11$.

8. Conclusions. In this article we have proposed computationally efficient penalised cost-based methods for detecting multiple sparse and dense anomalies or changes in the mean of cross-correlated data. In addition to estimating the locations of the anomalies/change points, the methods indicate which components are affected by a change. This is important to understand why and how changes or anomalies have occured. At the computational core of these methods lies a novel dynamic programming algorithm for solving banded unconstrained binary quadratic programs which approximate the Gaussian likelihood ratio test for a subset mean change.

The motivation of our methodological development comes from condition monitoring of an industrial process pump, where strong cross-correlations between spatially adjacent sensor measurements could be observed. Although several modelling assumptions were violated, three out of four known anomalies could be detected, with only one potential false alarm, when analysing the data with CAPA-CC. Even better results can be expected by using a more accurate model to remove trends not associated with anomalies. Also of interest for this application is being able to detect collective anomalies in real-time. The CAPA framework we have adopted has been shown to be able to be applied in online settings (Fisch, Bardwell and Eckley (2020)), and similar ideas could be used to produce a sequential version of CAPA-CC.

When assessing the method's performance empirically, special attention was paid to how incorporating cross-correlations in the model affected the results, compared to ignoring it as most existing methods do. We found that, for low to medium levels of dependence, there was almost no difference in power or estimation accuracy; for example, for $\rho < 0.5$ in the 2-banded and lattice precision matrices, and $\rho < 0.2$ for the constant correlation matrix, in the case of $p = 100$ variables. For increasingly stronger dependence above these levels, either in the form of a denser precision matrix or higher correlation parameter, the benefit of including cross-correlation in the model of the data grows in almost all tested cases.

The exception to this rule is connected to the somewhat surprising finding that the shape of the change in mean across variables influences the magnitude of the advantage of including cross-correlations quite strongly. In positively correlated data, changes that affect many series and are of very similar, or the same, size for each series can be harder to detect when including cross-correlations in the model. For example, in a model with strong positive correlations it is much harder to detect if a moderately large amount of variables changes by the same amount in the same direction, than if these variables changes by varying amounts in wildly varying directions. The intuition behind this is that in the former case, the change mimics the expected behaviour of the data, given the variables' strong positive dependence, while, in the latter, the change strongly violates the model's expectation. The model assuming independence, on the other hand, is completely agnostic to the shape of the changed mean vector. As a result, the benefits of including correlations in the model is small or, perhaps, even negative, if variables in the data is strongly dependent, and interest lies on detecting moderately sparse to dense and similarly changing variables.

Acknowledgments. We are grateful to OneSubsea for sharing their data with us and to Alex Fisch and Daniel Grose for helpful discussions.

Funding. Martin Tveten was supported by the Norwegian Research Council, project 237718 (Big Insight), while Idris A. Eckley and Paul Fearnhead were partly supported by EPSRC grant EP/N031938/1 (StatScale).

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/21-AOAS1508SUPPA; .pdf). Proofs of the propositions, additional comments to Proposition 2, derivation of the related change-point test, and detailed results from the simulation study, for both anomaly and change-point detection. More results on the pump data example for different choices of tuning parameters are also given.

Code (DOI: 10.1214/21-AOAS1508SUPPB; .zip). Efficient implementations of the CAPA-CC and CPT-CC algorithms as well as the code for reproducing the simulation study is available in the R package `capacc`, downloadable at https://github.com/Tveten/capacc. CAPA-CC will be included in a future version of the R package `anomaly` on CRAN, which contains the CAPA family of methods.

REFERENCES

BARDWELL, L., FEARNHEAD, P., ECKLEY, I. A., SMITH, S. and SPOTT, M. (2019). Most recent change-point detection in panel data. *Technometrics* **61** 88–98. MR3933661 https://doi.org/10.1080/00401706.2018.1438926

BHATTACHARJEE, M., BANERJEE, M. and MICHAILIDIS, G. (2019). Change point estimation in panel data with temporal and cross-sectional dependence. Preprint. Available at arXiv:1904.11101.

BLEAKLEY, K. and VERT, J.-P. (2011). The group fused Lasso for multiple change-point detection. Preprint. Available at arXiv:1106.4199.

CHO, H. (2016). Change-point detection in panel data via double CUSUM statistic. *Electron. J. Stat.* **10** 2000–2038. MR3522667 https://doi.org/10.1214/16-EJS1155

CHO, H. and FRYZLEWICZ, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 475–507. MR3310536 https://doi.org/10.1111/rssb.12079

CUTHILL, E. and MCKEE, J. (1969). Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the* 1969 24*th National Conference. ACM* '69 157–172. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/800195.805928

EGUSQUIZA, E., VALERO, C., VALENTIN, D., PRESAS, A. and RODRIGUEZ, C. G. (2015). Condition monitoring of pump-turbines. New challenges. *Measurement* **67** 151–163. https://doi.org/10.1016/j.measurement.2015.01.004

FEARNHEAD, P. and RIGAILL, G. (2019). Changepoint detection in the presence of outliers. *J. Amer. Statist. Assoc.* **114** 169–183. MR3941246 https://doi.org/10.1080/01621459.2017.1385466

FISCH, A. T. M., BARDWELL, L. and ECKLEY, I. A. (2020). Real time anomaly detection and categorisation. Preprint. Available at arXiv:2009.06670.

FISCH, A. T. M., ECKLEY, I. A. and FEARNHEAD, P. (2021a). A linear time method for the detection of point and collective anomalies. *Stat. Anal. Data Min.* To appear. Available at arXiv:1806.01947.

FISCH, A. T. M., ECKLEY, I. A. and FEARNHEAD, P. (2021b). Subset multivariate collective and point anomaly detection. *J. Comput. Graph. Statist.* 1–31. https://doi.org/10.1080/10618600.2021.1987257

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441. https://doi.org/10.1093/biostatistics/kxm045

FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. MR3269979 https://doi.org/10.1214/14-AOS1245

GAREY, M. R. and JOHNSON, D. S. (1979). *Computers and Intractability*: *A Guide to the Theory of NP-Completeness*. W. H. Freeman and Co., San Francisco, CA. MR0519066

HENRIQUEZ, P., ALONSO, J. B., FERRER, M. A. and TRAVIESO, C. M. (2014). Review of automatic fault diagnosis systems using audio and vibration signals. *IEEE Trans. Syst. Man Cybern. Syst.* **44** 642–652. https://doi.org/10.1109/TSMCC.2013.2257752

HORVÁTH, L. and HUŠKOVÁ, M. (2012). Change-point detection in panel data. *J. Time Series Anal.* **33** 631–648. MR2944843 https://doi.org/10.1111/j.1467-9892.2012.00796.x

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218. https://doi.org/10.1007/BF01908075

JENG, X. J., CAI, T. T. and LI, H. (2013). Simultaneous discovery of rare and common segment variants. *Biometrika* **100** 157–172. MR3034330 https://doi.org/10.1093/biomet/ass059

JIRAK, M. (2015). Uniform change point tests in high dimension. *Ann. Statist.* **43** 2451–2483. MR3405600 https://doi.org/10.1214/15-AOS1347

KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418 https://doi.org/10.1080/01621459.2012.737745

KIRCH, C., MUHSAL, B. and OMBAO, H. (2015). Detection of changes in multivariate time series with application to EEG data. *J. Amer. Statist. Assoc.* **110** 1197–1216. MR3420695 https://doi.org/10.1080/01621459.2014.957545

KLANDERMAN, M. C., NEWHART, K. B., CATH, T. Y. and HERING, A. S. (2020). Fault isolation for a complex decentralized waste water treatment facility. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 931–951. MR4133153 https://doi.org/10.1111/rssc.12429

KOVÁCS, S., LI, H., BÜHLMANN, P. and MUNK, A. (2020). Seeded binary segmentation: A general methodology for fast and optimal change point detection. Preprint. Available at arXiv:2002.06633.

LEWIS, J. G. (1982). Algorithm 582: The Gibbs–Poole–Stockmeyer and Gibbs–King algorithms for reordering sparse matrices. *ACM Trans. Math. Software* **8** 190–194. https://doi.org/10.1145/355993.355999

LI, J., XU, M., ZHONG, P.-S. and LI, L. (2019). Change point detection in the mean of high-dimensional time series data under dependence. Preprint. Available at arXiv:1903.07006.

LIU, H., GAO, C. and SAMWORTH, R. J. (2021). Minimax rates in sparse, high-dimensional change point detection. *Ann. Statist.* **49** 1081–1112. MR4255120 https://doi.org/10.1214/20-aos1994

ÖLLERER, V. and CROUX, C. (2015). Robust high-dimensional precision matrix estimation. In *Modern Nonparametric, Robust and Multivariate Methods* (K. Nordhausen and S. Taskinen, eds.) 325–350. Springer, Cham. MR3444335

SAFIKHANI, A. and SHOJAIE, A. (2020). Joint structural break detection and parameter estimation in high-dimensional nonstationary VAR models. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2020.1770097

SUSTIK, M. A. and CALDERHEAD, B. (2012). GLASSOFAST: An efficient GLASSO implementation. UTCS Technical Report **TR-12-29**.

TCHAKOUA, P., WAMKEUE, R., OUHROUCHE, M., SLAOUI-HASNAOUI, F., TAMEGHE, T. A. and EKEMB, G. (2014). Wind turbine condition monitoring: State-of-the-art review, new trends, and future challenges. *Energies* **7** 2595–2630. https://doi.org/10.3390/en7042595

TVETEN, M., ECKLEY, I. A. and FEARNHEAD, P. (2022). Supplement to "Scalable change-point and anomaly detection in cross-correlated data with an application to condition monitoring." https://doi.org/10.1214/21-AOAS1508SUPPA, https://doi.org/10.1214/21-AOAS1508SUPPB

VER HOEF, J. M., HANKS, E. M. and HOOTEN, M. B. (2018). On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spat. Stat.* **25** 68–85. MR3809256 https://doi.org/10.1016/j.spasta.2018.04.006

WANG, T. and SAMWORTH, R. J. (2018). High dimensional change point estimation via sparse projection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 57–83. MR3744712 https://doi.org/10.1111/rssb.12243

WANG, D., YU, Y., RINALDO, A. and WILLETT, R. (2020). Localizing changes in high-dimensional vector autoregressive processes. Preprint. Available at arXiv:1909.06359.

WESTERLUND, J. (2019). Common breaks in means for cross-correlated fixed-T panel data. *J. Time Series Anal.* **40** 248–255. MR3915529 https://doi.org/10.1111/jtsa.12407

XIE, Y. and SIEGMUND, D. (2013). Sequential multi-sensor change-point detection. *Ann. Statist.* **41** 670–692. MR3099117 https://doi.org/10.1214/13-AOS1094