Gaussian Process Subset Scanning for Anomalous Pattern Detection in Non-iid Data

William Herlands Carnegie Mellon University
Edward McFowland III University of Minnesota
Andrew G. Wilson Cornell University
Daniel B. Neill Carnegie Mellon University

Abstract

Identifying anomalous patterns in real-world data is essential for understanding where, when, and how systems deviate from their expected dynamics. Yet methods that separately consider the anomalousness of each individual data point have low detection power for subtle, emerging irregularities. Additionally, recent detection techniques based on subset scanning make strong independence assumptions and suffer degraded performance in correlated data. We introduce methods for identifying anomalous patterns in non-iid data by combining Gaussian processes with novel log-likelihood ratio statistic and subset scanning techniques. Our approaches are powerful, interpretable, and can integrate information across multiple data streams. We illustrate their performance on numeric simulations and three open source spatiotemporal datasets of opioid overdose deaths, 311 calls, and storm reports.

1 Introduction

Anomalous pattern detection is the task of identifying subsets of data points that systematically differ from the underlying model. Identifying anomalous patterns in real-world data is critical for understanding how people and systems deviate from expected behavior. In the spatiotemporal domain, timely identification of such patterns can allow for effective interventions. For example, detecting anomalous increases in opioid deaths can enable health care workers to effectively target overdose prevention programs. Similarly, patterns of increased 311 calls can help cities to better target services and allocate resources.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

To detect these anomalous patterns, we will address three key challenges. First, real-world data is extremely complex with non-trivial correlations across space, time, and other features. Treating data points as iid ignores important covariance structure and will substantially overestimate the anomalousness of detected patterns. Second, an event of interest often affects multiple nearby points. Simply considering how anomalous is each individual point loses power to detect subtle anomalies. Third, anomalous patterns are often irregularly shaped or discontiguous due to latent demographic or geographic features. Searching for these complex patterns is important for precision and detection power, yet exhaustive methods are computationally intractable and may result in overfitting.

A sensible approach to this problem is model-based anomaly detection, where a distribution is fit to model "regular" data. Points with a low likelihood under this distribution are identified as anomalous (Chandola et al., 2009; Hodge and Austin, 2004). To address the complex correlations in real-world systems, Gaussian processes (GPs) provide a natural means of learning covariance structure from data. However, GP anomaly detection has been typically used to classify *individual* points as outliers (Smith et al., 2014; Kowalska and Peel, 2012; Stegle et al., 2008). Such approaches have difficulty when confronted with subtle anomalies, where each individual data point may seem to conform to the underlying distribution, yet when taken as a group, they form a collectively anomalous pattern. Thus anomalous pattern detection is a conceptually and statistically different problem than anomaly or outlier detection.

A few recent GP models consider anomalous intervals (Reece et al., 2015) and sophisticated change points (Saatçi et al., 2010; Herlands et al., 2016) to detect intervals of anomalous points. However, these methods (the first two of which are applied exclusively to one-dimensional data) are limited to contiguous intervals in the input domain and cannot model the irregularly shaped anomalies we expect in complex data. Cheng et al. (2015) recently developed an anomalous

pattern detection technique for spatiotemporal data. However, this approach requires a corpus of anomaly-free training data, can only detect contiguous anomalous patterns, and is specific to video data.

In the statistics literature, spatial and subset scanning methods are commonly used to identify collectively anomalous subsets of data (Kulldorff, 1997; Neill, 2012). By combining information across a subset of data elements, they generate a strong signal of anomalous behavior. These approaches compute a log-likelihood ratio (LLR) of subsets being drawn from a null or anomalous distribution. The LLR is a powerful statistic that measures how much evidence exists in the data to conclude if the subset exhibits abnormal behavior (Kulldorff, 1997; Neill et al., 2005). A core challenge of subset scanning is searching through the $O(2^n)$ possible subsets of n data elements (Neill and Moore, 2004; Agarwal et al., 2006; Duczmal et al., 2007; Wu et al., 2009). Neill (2012) shows that certain LLR statistics satisfying a linear-time subset scanning (LTSS) property can be optimized in $O(n \log n)$ by ordering points according to a particular "priority function" and evaluating only n of the 2^n subsets. However, LTSS assumes that we can compute the contributions of individual points to the LLR. This is possible only when assuming that data is uncorrelated under the null (Neill, 2009), yet when applied to non-iid data this independence assumption would result in substantial false positive rates, as correlated fluctuations will be mistaken for anomalous movements.

1.1 Contributions

In this paper we introduce novel techniques for identifying anomalous patterns in non-iid data. Our methods are powerful and interpretable. By combining naturally interpretable GPs with localized anomalous patterns we can describe the "regular" data dynamics as well as quantify and corroborate anomalous regions with domain experts. Our main contributions are:

1. Combining GP modeling with subset scanning for powerful and interpretable detection of anomalous patterns in highly correlated data.
2. Proposing a new likelihood ratio statistic and subset scan technique for correlated data that do not assume conditional independence.
3. Performing hold-out GP inference while computing our new likelihood ratio statistic conditioned on GP hyperparameters, to avoid corrupting the null model with anomalies.
4. Developing two novel, principled approaches to the NP-hard problem of searching for the most anomalous subset, through a new iterative method and an application of the Generalized Rayleigh Quotient respectively.
5. We demonstrate our methods on numeric simulations, opioid-related deaths, 311 calls for service data, and multiple streams of sewer flooding reports and tree damage reports, illustrating interpretable and policy-relevant results.

The paper proceeds as follows: §2 provides background on GPs. §3 introduces a novel log-likelihood ratio statistic for non-iid data. §4 details the Gaussian Process Neighborhood Scan (GPNS) and the Gaussian Process Subset Scan (GPSS). Experimental results on numerical and real data are presented in §5.

2 Gaussian Processes

Consider data, (x, y), where $x = \{x_1, \ldots, x_n\}, x_i \in \mathbb{R}^D$, are inputs or covariates, and $y = \{y_1, \ldots, y_n\}$, $y_i \in \mathbb{R}$ are outputs or response variables indexed by x. We assume that y is generated from x by a latent function with a GP prior. In particular, $y = f(x) + \epsilon$, $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$

A GP is a nonparametric prior over functions completely specified by mean and covariance functions. The mean function, $m(x) = \mathbb{E}[f(x)]$, is the prior expectation of $f(x)$. The covariance function is given by $k(x, x') = \text{cov}(f(x), f(x'))$ (Rasmussen, 2006).

In this paper we use three important properties of GPs. First, we can draw samples from a GP prior since conditional on GP hyperparameters any finite collection of function values is distributed $\mathcal{N}(m(x), k(x, x))$

Second, if a function has a Gaussian noise model, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, then conditional on hyperparameters and data (x, y), we can derive a closed form expression for the predictive distribution of $f(x^*)$,

$$f(x^*)|x, y, x^* \sim \mathcal{N}\Big(k(x^*, x)[k(x, x) + \sigma_\epsilon^2 I]^{-1} y, \\ k(x^*, x^*) - k(x^*, x)[k(x, x) + \sigma_\epsilon^2 I]^{-1} k(x, x^*)\Big) \quad (1)$$

Third, GP hyperparameters, θ, can be learned by maximum likelihood optimization. While naively this requires $O(n^3)$ computations, we use scalable GP learning in the structured kernel inference (SKI) framework (Wilson and Nickisch, 2015) for $O(n)$ scalability.

3 LLR statistic for non-iid data

Considering (x, y) as defined in §2, we are interested in anomalous patterns that systematically differ from the

underlying data distribution. We frame this search as an LLR comparison between a null model of "regular" behavior and an alternative model of "anomalous" behavior. A single latent GP defines both models. Subsets of data with the highest LLR scores are identified as the most anomalous.

Using a GP as the foundational modeling technique enables us to learn complex covariance structure and seamlessly extend to high dimensions as well as missing data. GPs are also naturally interpretable, which can provide insight about the "regular" data dynamics.

Consider a given subset of data points defined by the binary weighting vector w, where $w_i = 1$ if (x_i, y_i) is included in the subset and $w_i = 0$ if excluded. Our null model, H_0, assumes that all points (regardless of w_i) are drawn from a function with a GP prior: $y = f(x)+\epsilon$, where $f(x) \sim GP(\theta_0)$ and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$. Our alternative model, $H_1(w)$, assumes that $y_i = f(x_i) + \epsilon$ for $w_i = 0$, and $y_i = g(f(x_i), \theta_1) + \epsilon$ for $w_i = 1$, where $g(\cdot)$ is any function of the latent GP.

Here we focus on the case of a mean shift, $g(f(x), \theta_1) = f(x)+\beta$, $\beta \in \mathbb{R}^1$. The covariance structure remains the same in the null and alternative models. This allows us to efficiently compute the posterior mean vector μ and covariance matrix Σ through GP inference, where $y \sim \mathcal{N}(\mu, \Sigma)$ under H_0, and $y \sim \mathcal{N}(\mu + \beta w, \Sigma)$ under $H_1(w)$. For posterior μ and Σ we condition on all data outside the subset of points represented by w, ensuring that null model estimates are not corrupted by anomalous observations. However, since anomalies are assumed to be rare, their influence on parameter estimation is minimal. Therefore we use all (x, y) for GP learning of the parameters of the null model θ_0.

We concentrate on mean changes since many real world cases concern anomalous levels of a quantity. Increases in localized drug overdoses, crime, and calls for city service are all mean shifts of great importance. Methods for identifying arbitrary changes in distribution – while able to detect other sorts of patterns – have reduced power to detect such mean shifts, due to more diffuse inductive biases. Persistent changes in covariance structure are typically considered change-points and require substantial data in both regimes as opposed to the localized anomalous patterns we detect.

To measure how anomalous is a subset defined by w, we compute the generalized log-likelihood ratio, $LLR(w) = \max_\beta LLR(w \mid \beta)$, where:

$$LLR(w \mid \beta) = \log \frac{\texttt{MNPDF}(y - \beta w \mid \mu, \Sigma)}{\texttt{MNPDF}(y \mid \mu, \Sigma)} \tag{2}$$

Here `MNPDF` is the multivariate normal probability density function. The most anomalous subset, w^*, is

$$\begin{aligned} w^* &= \underset{w}{\operatorname{argmax}}\, LLR(w) \\ &= \underset{w}{\operatorname{argmax}} \max_\beta -\frac{\beta^2}{2} w^T E w + \beta w^T E(y - \mu) \end{aligned} \tag{3}$$

where $E = \Sigma^{-1}$ for notational brevity. Conditional on w, the MLE $\beta^* = \arg\max_\beta LLR(w)$ can be calculated in closed form, $\beta^* = [w^T E(y - \mu)]/[w^T E w]$ (see supplementary material). Nevertheless, maximizing $LLR(w)$ is an NP-complete Integer Quadratic Program (Del Pia et al., 2014), so an optimal solution requires exponential-time computation. Note that the LTSS condition for a log linear-time subset search described in Neill (2012) does not apply, since it requires independent data with a diagonal covariance matrix.

3.1 Randomization testing

Given a method for finding anomalous subsets, the following randomization testing procedure determines an α-level significance threshold for $LLR(w)$ conditional on the parameters of the null model:

1. Repeatedly draw $y^{(r)} \sim GP(\theta_0)$, at the same covariates, x, as the real data for $r = 1...R$.
2. Scan over $(x, y^{(r)})$ with the chosen subset searching method. For each randomization r save the most anomalous LLR value, $LLR(w^{*,(r)})$.
3. Determine an α-level threshold for significance based on the $(1 - \alpha)$ quantile of the R maximum LLR values, above which any $LLR(w)$ from the original scan is considered statistically significant.

4 Efficient subset scanning

Having defined the LLR scan statistic to evaluate how anomalous is a given subset, we must now decide over which subsets to scan. Unconstrained optimization over $O(2^n)$ subsets is computationally infeasible for an exhaustive search. Additionally, an unconstrained search may return an unrelated set of points, reducing interpretability and increasing the potential for overfitting. Anomalous events in human data, such as drug usage and requests for government services, often affect multiple nearby points. Thus we assume that anomalous points are near one another. For example, in spatiotemporal data we assume that anomalous points are clustered in space and time. Following Neill (2012), we define the local "k-neighborhood" of each data point, consisting of that point and its $k-1$ nearest neighbors, for some k. We propose two approaches for using these neighborhoods to identify anomalous patterns: Gaussian Process Neighborhood Scan (GPNS) and Gaussian Process Subset Scan (GPSS).

4.1 GP Neighborhood Scan (GPNS)

Given a maximum neighborhood size k_{max}, GPNS searches over the $O(nk_{max})$ local neighborhoods consisting of the k-neighborhood for each point where $k = \{1, 2, \ldots, k_{max}\}$. Where neighborhoods are defined by Euclidean distance, the set of search regions are circular in shape. For each neighborhood, $(x^{(n)}, y^{(n)})$, we obtain posterior μ and Σ conditional on θ_0 and points $(x^{(-n)}, y^{(-n)})$. We then compute $LLR(w)$ for the neighborhood where $w = \vec{1}$, i.e., we evaluate the alternative hypothesis of the entire neighborhood being anomalous. GPNS pseudocode is presented in Alg. 1.

Algorithm 1: GPNS

for $k = 1 : k_{max}$ **do**
  for (x_i, y_i), $i = 1 : n$ **do**
    Define k-neighborhood, $n^{(k,i)}$, and infer (μ, Σ);
    Set $w^{(k,i)} = \vec{1} \in \{0, 1\}^k$;
    Compute β^* given $w^{(k,i)}$;
    Compute $LLR(w^{(k,i)})$;
  end
end
Choose $n^* = \text{argmax}_{n^{(k,i)}} LLR_{n^{(k,i)}}$;
Randomization testing for significance;

4.2 GP Subset Scan (GPSS)

While GPNS simplifies the exponential search, it requires constraining assumptions about the shape of neighborhoods and is only able to discover contiguous, spherical anomalous patterns. While there are approaches to increase the variety of neighborhood shapes without substantially degrading computational efficiency (Kulldorff et al., 2006; Neill and Moore, 2004; Kulldorff, 2001), these methods still require strict specification of potential anomalies. Such foreknowledge is unrealistic in real-world applications where natural boundaries, demographics, and stochastic effects lead to irregularly-shaped patterns. In such cases GPNS has reduced detection and explanatory power.

To flexibly detect irregularly-shaped patterns, GPSS conducts an unconstrained search for the most anomalous subset within neighborhoods of fixed size k. Specifically, we identify the subset of points $(x^{(s)}, y^{(s)}) \subseteq (x^{(n)}, y^{(n)})$ that maximize the LLR within each neighborhood. This allows us to identify highly irregular and even non-contiguous anomalous patterns. By restricting the search within a local neighborhood, we ensure that the identified patterns are coherent and interpretable. GPSS requires evaluating $O(n)$ neighborhoods, as presented in Alg. 2.

Algorithm 2: GPSS

Fix k at some size;
for (x_i, y_i), $i = 1 : n$ **do**
  Define k-neighborhood, $n^{(i)}$, and infer (μ,Σ);
  Approximate the optimal subset, $s^{(i)} \subseteq n^{(i)}$;
  Set each $w_j^{(i)} = 1(j \in s^{(i)})$;
  Compute β^* given $w^{(i)}$;
  Compute $LLR(w^{(i)})$;
end
Choose $s^* = \text{argmax}_{s^{(i)}} LLR_{s^{(i)}}$;
Randomization testing for significance;

Unfortunately, this procedure requires finding $w \in \{0, 1\}^k$ that maximizes the LLR of a subset within the neighborhood, $\text{argmax}_w \text{-}\frac{1}{2} w^T \beta E w \beta + w^T \beta E(y^{(n)}\text{-}\mu)$. This is still an Integer Quadratic Program, whose optimal solution is intractable even for moderately sized neighborhoods. Instead, below we formulate three approaches for finding approximate solutions.

4.2.1 β_{MAX} for conditionally optimal subset

Due to the full rank covariance matrix, we are unable to disentangle the individual contributions from each point to the LLR. However, if we condition on some subset of points, w, we are able to compute the conditional contribution of each point. First, note that conditional on w we can decompose w^* from Eq. 3 into a sum over each of the m points in the neighborhood

$$w^T \beta E(y^{(n)} - \mu) - \frac{1}{2} w^T \beta E w \beta$$
$$= \sum_i w_i \Big[\beta \big(E(y^{(n)} - \mu) \big)_i - \frac{1}{2} \big(\sum_{j \neq i} w_j E_{j,i} + E_{i,i} \big) \beta^2 \Big] \quad (4)$$

The contribution of point (x_i, y_i) to the LLR is the difference in LLR between $w_i = 0$ and $w_i = 1$. Due to the outer and inner sums, the change in the LLR is:

$$\beta \big(E(y^{(n)} - \mu) \big)_i - \frac{1}{2} \big(\sum_{j \neq i} 2 w_j E_{j,i} + E_{i,i} \big) \beta^2 \quad (5)$$

To maximize the LLR a point is only added to the subset if its contribution is positive. By setting Eq. 5 to zero we can compute β_{MAX_i}, the maximum β value for which to include point (x_i, y_i).

$$\beta_{MAX_i} = \Big[2 \big(E(y^{(n)}\text{-}\mu) \big)_i \Big] / \Big[\sum_{j \neq i} 2 w_j E_{j,i} + E_{i,i} \Big] \quad (6)$$

As proved in Speakman et al. (2016), we obtain the conditional optimal subset by using β_{MAX} as a priority function, ranking each data point by β_{MAX_i}, and

iteratively compute the score function for subsets including each additional point. This yields a log linear search over data points. Such an approach identifies the most anomalous subset with a positive mean shift. To find the most anomalous subset with a negative mean shift we simply rank data points by $-\beta_{MAX_i}$

Since the derivation of β_{MAX} is conditional on a subset w, we obtain the *conditional* optimal subset. In order to approximate an optimal solution we iteratively compute the conditional optimal subset beginning with a null subset, $w = \vec{0}$. This requires $O(\ell k \log(k))$ computation for some ℓ number of iterations. Pseudo-code for this algorithm can be found in the supplement.

For a diagonal Σ, β_{MAX} orders points according to $2(y_i^{(n)} - \mu_i)$, which is equivalent to the LTSS priority function for an independent Gaussian subset scan (Speakman et al., 2016). Thus β_{MAX} approach identifies the optimal subset in the independent case and is conditionally optimal in the dependent case.

4.2.2 Generalized Rayleigh Quotient method

We consider an alternative optimization approach to obtain an approximately optimal subset. Consider plugging the MLE solution, β^*, into w^* from Eq. 3,

$$w^* = \underset{w}{\text{argmax}} \Big[w^T \big(E(y^{(n)}\text{-}\mu)(y^{(n)}\text{-}\mu)^T E \big) w \Big] / \Big[w^T \big(2E \big) w \Big] \tag{7}$$

If we relax w such that $w \in \mathbb{R}^m$, this can be re-written as the generalized Rayleigh quotient, $(w^T A w)/(w^T B w)$, where $A = E(y^{(n)} - \mu)(y^{(n)} - \mu)^T E$, and $B = 2E$. Note that A is a symmetric matrix and B is a Hermitian positive-definite matrix. Taking the Cholesky decomposition $B = LL^T$, the generalized Rayleigh quotient can be written as a Rayleigh quotient (Yu et al., 2013), $R(A', w') = (w'^T A' w')/(w'^T w')$, where $A' = L^{-1} A L^{T^{-1}}$ and $w' = L^T w$. The maximum w' of the Rayleigh quotient, $w'_{max} = \text{argmax}_{w'} R(A', w') = \text{argmax}_{w'}(w'^T A w')/(w'^T w') = v^{(max)}$, is the largest eigenvector of A'. Since we defined $w' = L^T w$, then the maximum $w_{max} = L^{T^{-1}} v^{(max)}$ is the relaxed solution to our original optimization problem from Eq. 7.

Although w_{max} has non-integer elements, the ordering of the elements of this eigenvector corresponds to the importance of the data points in the neighborhood. Thus we scan over the ordered elements of w_{max}, iteratively adding each to the subset. Maximizing $LLR(w)$ over this linear number of subsets provides an approximate solution to the constrained integer program.

4.2.3 Forward stepwise optimization

A third approximation approach uses a greedy forward stepwise algorithm that iteratively sets one element $w_i = 1$ such that the objective is minimized in each iteration. Once the objective cannot be further minimized the optimization is terminated, thereby providing a greedy optimal solution. For a neighborhood of size k, the stepwise approach may require up to k iterations, evaluating $O(k)$ subsets at each iteration for a total of $O(k^2)$ computations.

4.3 Efficient Multi-Stream Search

Often we are interested in searching for anomalous patterns across multiple dimensions, or streams, of data. For example, anomalous patterns of damaged trees and sewer flooding can help localize severe storm damage. Multi-stream search can enhance the signal of subtle anomalies that affect multiple streams, and reduce false positive detections when perturbations in a single stream are not important to the application.

In principle, GPNS and GPSS can handle multiple streams by stacking the data from each stream and adding a final dimension to indicate from which stream the data came. Yet naive GP inference requires $O(n^3)$ complexity, so repeatedly concatenating data from multiple streams quickly leads to scalability issues. On the other hand, Kronecker-based scalability require a kernel that is multiplicatively decomposable over the input dimensions (Saatçi, 2012). This implies that the prior correlation structure is the same over all data dimensions except for the stream indicator. For example, Kronecker structure in spatiotemporal settings constrains streams to have the same prior spatiotemporal correlations. This assumption is overly restrictive for the complex data in which we are interested.

Instead, we learn independent GPs for each stream of data and then scan over neighborhoods in the data jointly for all streams. Posteriors for each stream are independently inferred from the associated GP. Thus for streams $s = 1, ..., S$, the posterior distribution for subset scanning contains a block diagonal covariance,

$$\mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_S \end{bmatrix}, \begin{bmatrix} \Sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_S \end{bmatrix} \right)$$

In this manner each stream can flexibly learn different prior covariance structures while still ensuring scalability equivalent to single-stream GPNS and GPSS. The one drawback of this approach is that inter-stream covariance information is not exploited for GP inference.

5 Experiments

We evaluate GPNS and GPSS using numeric simulations and three urban spatiotemporal datasets. We compare the methods against a number of competitive baseline algorithms from contemporary literature. First, we compare to an independent Gaussian subset scan, a state of the art anomalous pattern detection algorithm (Neill, 2009, 2012). Additionally, we compare against a standard GP anomaly detection approach (Kowalska and Peel, 2012; Stegle et al., 2008), in which we use the posterior distribution of the null GP model θ_0 regressed over the entire dataset to classify points beyond a given level-α significance threshold as anomalies. While all GP methods in this paper are agnostic to kernel choice, an RBF kernel and linear mean function were used for all experiments.

Although anomalous pattern detection is a distinct problem from outlier or anomalous point detection, we also compare against two commonly used outlier detection techniques: a one-class SVM (Schölkopf et al., 2001) and robust multivariate outlier detection using the Mahalanobis distance (Rousseeuw and Leroy, 2005; Rousseeuw and Van Zomeren, 1990).

Additional real data analyses are in the supplement.

5.1 Numeric experiments

For each numeric test, baseline data is drawn from a 2D GP (Rasmussen and Nickisch, 2016). Multiplicative anomalies of arbitrary shape are injected by scaling randomly sampled points, within a randomly chosen neighborhood, by a factor of ≥ 1. (Note that this simulation does not correspond to our method's assumption of an additive mean shift.) The most anomalous subset is computed using GPSS methods and baseline approaches. For the baseline GP approach and one-class SVM we provide additional information (the true percentage of the anomalous data) in order to determine their threshold levels.

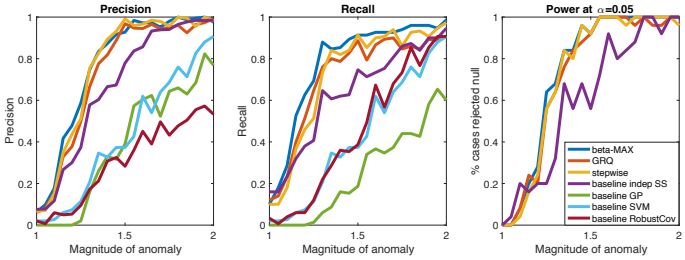

Figure 1: Precision, recall, and power at $\alpha = 0.05$ for GPSS methods and baseline anomaly detection approaches. The three GPSS methods dominate in all cases with the β_{MAX} performing best overall.

Varying the multiplicative factor between 1 and 2 we compute the average precision and recall in Fig. 1 over 50 tests in a 400 point grid for each multiplicative factor. Randomization testing ($\alpha = .05$) is performed for each synthetic test to determine the score threshold for significance. For precision and recall, truly anomalous points are "positive" and all other data is "negative." The GPSS approaches dominate all other methods for nearly the entire test range, with β_{MAX} performing best overall.

Additionally, for each test we use an exhaustive search to find the subset with the highest LLR. The ratios of the LLR of approximate GPSS solutions to $LLR(w^*)$ are shown in Fig. 2. Note that all approximation methods are relatively close to the optimal value. While the β_{MAX} approach dominates at large magnitudes, the GRQ dominates at small magnitudes and achieves a relatively stable ratio across all tests.

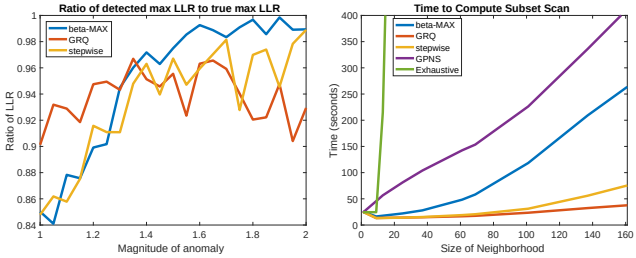

Figure 2: Numeric tests of GPNS and GPSS compared to exhaustive evaluation of $LLR(w^*)$. Left plot: ratio of maximum LLR identified by GPSS to true maximum LLR. Right plot: run time.

To test the methods' scalability we vary the maximum neighborhood size and measure run time. In Fig. 2 we compare GPSS, GPNS, and an exhaustive search for the optimal subset. The exhaustive search quickly becomes computationally intractable. Despite the added flexibility, GPSS is faster than GPNS because GP posterior inference is performed for fewer neighborhoods.

We consider the effect of the density of anomalies on GPSS and GPNS where "density" is defined by the proportion of anomalous points in the true subset (Fig. 3). While the stepwise method is competitive with the β_{MAX} and GRQ approaches at low densities, its precision and recall drop off steeply at high densities. Additionally, in relatively low density anomalies, where the anomalous shapes may be highly irregular, GPNS has substantially reduced precision and recall.

5.2 Urban opioid overdose deaths

A recent United States opioid epidemic has garnered national attention (US Department of Health and Hu-

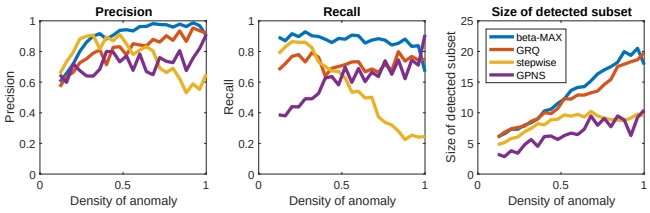

Figure 3: Precision, recall, and size of detected subset for GPSS and GPNS methods over subsets of varying density within a neighborhood.

man Services, 2016). We study monthly opioid overdose deaths in New York from 1999-2015 (US CDC, 2017). Data is provided at a county level for Manhattan, Brooklyn, Queens, the Bronx, Nassau County, and Suffolk County. Data is missing for some months in different counties. We apply GPSS and baseline approaches jointly to data across all time, latitude, and longitude, with randomization testing at $\alpha = 0.05$.

All three GPSS approaches identify two statistically significant anomalous patterns. While precise points

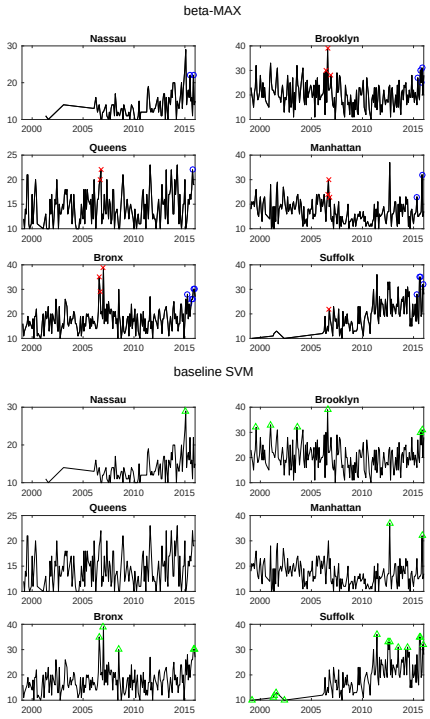

Figure 4: Monthly opioid overdose deaths in New York from 1999-2015. Top plot depicts the two statistically significant anomalies detected by β_{MAX}. Bottom plot depicts points detected by the one-class SVM.

selected by the methods differ slightly, Fig. 4 depicts the two anomalous regions discovered by β_{MAX} in blue circles and red crosses. With the exception of the independent subset scan, the baseline methods failed to discover a coherent anomalous pattern. Instead they selected individual points across space and time. For example, see results from the one-class SVM in Fig. 4.

The anomalies detected by GPSS correspond to important public health events. The blue circles at the end of 2015 indicate a surge in opioid deaths corresponding to a well known plague of fentanyl-related deaths in NYC (City of New York Office of the Mayor, 2017). The anomaly denoted by red crosses in 2006 is particularly interesting since it indicates a spike in opioid deaths immediately preceding the introduction of community training programs to administer a life-saving naloxone drug. This may indicate a surge in fatalities that was cut short by making naloxone more widely available and educating communities in its use.

5.3 Manhattan 311 requests

New York City's 311 system enables residents to request government services. We consider a local public health event that occurred on 01/22/16 in upper Manhattan. On that day, local news reported that residents were concerned due to brown tap water (Pichardo, 2016; CBS New York, 2016b). Detecting the extent of the residents' concerns is important to help identify and mitigate public health risks.

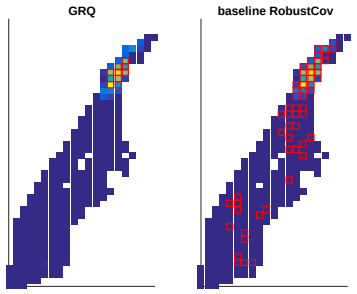

Figure 5: GPSS and robust covariance results for daily 311 requests in Manhattan on 01/22/16. Red squares indicate detected anomalies.

We consider daily 311 requests in Manhattan for the month of January 2016, aggregated over a 0.08 mile^2 grid (City of New York, 2017). We apply GPSS methods and baseline approaches with neighborhoods of up to 15 points. All GPSS methods identified an anomalous pattern around the locations and time of the water discoloration event. Baseline methods tended to substantially overestimate the anomaly's extent in both space and time. These results from January 22 are represented by the GRQ and the Robust baselines in Fig. 5. Blue and yellow squares indicate low and high volume of reports, respectively. Red squares indicate the top anomalous regions discovered by each method.

Ground truth does not exist for these hyper-local events so we cannot compute precision and recall. However, 311 requests have labeled types, although we used aggregated 311 calls as our data inputs. For each method we compute the ratio of water-related 311 calls to non-water-related calls in the detected anomalies. This "water signal-to-noise" ratio, listed in Table 1, indicates how precisely each method identified regions associated with many water-related requests. The entire dataset has a water signal-to-noise of 0.07.

Table 1: Signal-to-noise ratio of water-related 311 calls to non-water-related 311 calls for all methods.

Model	Signal-to-Noise
GRQ	7.22
Stepwise	7.22
β_{MAX}	7.22
Independent SS	7.06
Baseline GP	0.44
One-class SVM	0.23
RobustCov	0.12

5.4 Multi-stream: trees and sewers

Using the multi-stream procedure from §4.3, we consider 311 reports of damaged trees and sewer issues. Both streams indicate weather-related issues: damaged trees indicate high winds while sewer calls indicate substantial precipitation. Together, these data identify areas with dangerous post-storm conditions. Each complaint type is fit with an independent GP and the entire data is scanned jointly for anomalies.

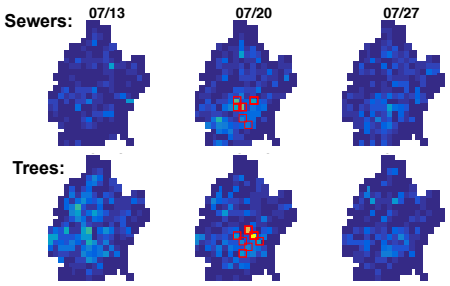

Figure 6: 311 calls for damaged trees and sewer issues from 2016 in Brooklyn. Red squares indicate the top anomalies discovered by the β_{max} approach.

We analyze data in Brooklyn aggregated weekly over a 0.08 mile^2 grid (City of New York, 2017). We conduct analyses for 2016 and 2010 with results depicted in Figs. 6 and 7. The number of sewer reports (per week, per cell) are plotted on top, and damaged tree reports on bottom. Red squares indicate the top anomalous regions discovered using the β_{max} approach.

The most anomalous regions in 2016 were all concentrated during the week of July 20th when a significant summer storm felled trees and flooded sewers, thus jointly affecting both data streams (CBS New York, 2016a). Conversely, although the week of July 13th experienced elevated reports of felled trees no anomalous region is detected since there is no corresponding increase in sewer flooding. This demonstrates how multi-stream search may help to regulate GPSS.

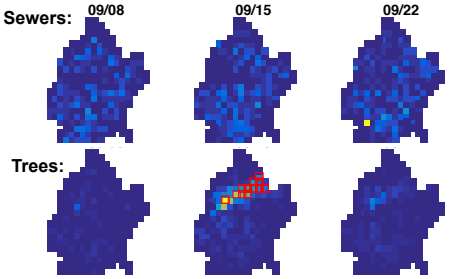

Figure 7: 311 calls for damaged trees and sewer issues from 2010 in Brooklyn. Red squares indicate the top anomalies discovered by the β_{max} approach.

The most anomalous regions in 2010 were all concentrated during the week of September 15th when an urban tornado cut through Brooklyn (AccuWeather, 2013). Unlike the 2016 results, these anomalies only occurred in reports of damaged trees. Also note the lone yellow square in the sewer data of September 22. Though the square indicates elevated number of calls, GPSS does not consider it anomalous since it does not represent a systematic shift in space and time.

6 Conclusions

We develop two GP-based subset scanning approaches to accurately and efficiently detect anomalous patterns in complex, highly correlated data. The results of GPNS and GPSS are coherent, powerful, and interpretable. While the simpler GPNS method may be sufficient for circular clusters, GPSS provides additional flexibility to accurately identify irregular cluster shapes. Unlike individual anomaly detection methods, the spatial locality enforced by our methods ensures coherent explanation of detected anomalous patterns. As the 311 and opioid applications demonstrate, our approaches can be used for studying and informing policy decisions. Future work could integrate categorical variables or extend to Student-t processes (Shah et al., 2014) for dealing with heavy-tailed noise.

This work was supported by NSF awards GRFP DGE-1252522 and IIS-0953330, the RK Mellon Foundation, and NCSU Laboratory for Analytical Sciences.

References

AccuWeather. Photos: Tornadoes ransacked Brooklyn, Queens in 2010, 2013. URL https://www.accuweather.com/en/weather-news/nyc-tornado-anniversary-2010/17830410.

D. Agarwal, A. McGregor, J. M. Phillips, S. Venkatasubramanian, and Z. Zhu. Spatial scan statistics: approximations and performance study. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 24–33, 2006.

CBS New York. Severe storms follow extreme heat for tri-state area, 2016a. URL http://newyork.cbslocal.com/2016/07/25/tri-state-extreme-heat/.

CBS New York. Discoloration in water rankles upper Manhattan residents, 2016b. URL http://newyork.cbslocal.com/2016/01/22/upper-manhattan-water/.

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

Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2909–2917, 2015.

City of New York. New York City open data, 2017. URL https://opendata.cityofnewyork.us/.

City of New York Office of the Mayor. HealingNYC: Preventing overdoses, saving lives, 2017. URL http://www1.nyc.gov/assets/home/downloads/pdf/reports/2017/HealingNYC-Report.pdf.

Alberto Del Pia, Santanu S Dey, and Marco Molinaro. Mixed-integer quadratic programming is in NP. *Mathematical Programming*, pages 1–16, 2014.

L. Duczmal, A. L. Cancado, R. H. Takahashi, and L. F. Bessegato. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics and Data Analysis*, 52:43–52, 2007.

William Herlands, Andrew Wilson, Hannes Nickisch, Seth Flaxman, Daniel Neill, Wilbert van Panhuis, and Eric Xing. Scalable Gaussian processes for characterizing multidimensional change surfaces. *AISTATS*, 2016.

Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22 (2):85–126, 2004.

Kira Kowalska and Leto Peel. Maritime anomaly detection using Gaussian process active learning. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 1164–1171. IEEE, 2012.

Martin Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496, 1997.

Martin Kulldorff. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):61–72, 2001.

Martin Kulldorff, Lan Huang, Linda Pickle, and Luiz Duczmal. An elliptic spatial scan statistic. *Statistics in Medicine*, 25(22):3929–3943, 2006.

D. B. Neill. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting*, 25:498–517, 2009.

D. B. Neill, A. W. Moore, M. R. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 218–227, 2005.

Daniel B Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, 2012.

Daniel B Neill and Andrew W Moore. Rapid detection of significant spatial clusters. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 256–265. ACM, 2004.

Carolina Pichardo. Brown water pours from faucets uptown after maintenance work, DEP says, 2016. URL https://www.dnainfo.com/new-york/20160122/washington-heights/brown-water-pours-from-faucets-uptown-after-maintenance-work-dep-says.

Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

Carl Edward Rasmussen and Hannes Nickisch. GPML Matlab code version 4.0, 2016. URL http://www.gaussianprocess.org/gpml/code/.

Steven Reece, Roman Garnett, Michael Osborne, and Stephen Roberts. Anomaly detection and removal using non-stationary Gaussian processes. *arXiv preprint arXiv:1507.00566*, 2015.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 589. John Wiley & Sons, 2005.

Peter J Rousseeuw and Bert C Van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639, 1990.

Yunus Saatçi. *Scalable inference for structured Gaussian process models*. PhD thesis, University of Cambridge, 2012.

Yunus Saatçi, Ryan D Turner, and Carl E Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 927–934, 2010.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

Amar Shah, Andrew Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to gaussian processes. In *Artificial Intelligence and Statistics*, pages 877–885, 2014.

Mark Smith, Steven Reece, Stephen Roberts, Ioannis Psorakis, and Iead Rezek. Maritime abnormality detection using Gaussian processes. *Knowledge and Information Systems*, 38(3):717–741, 2014.

Skyler Speakman, Sriram Somanchi, Edward McFowland III, and Daniel B Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics*, 25(2):382–404, 2016.

Oliver Stegle, Sebastian V Fallert, David JC MacKay, and Søren Brage. Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering*, 55(9):2143–2151, 2008.

US CDC. WONDER database, 2017. URL https://wonder.cdc.gov/.

US Department of Health and Human Services. The opioid epidemic: By the numbers, 2016. URL https://www.hhs.gov/sites/default/files/Factsheet-opioids-061516.pdf.

Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *International Conference on Machine Learning*, pages 1775–1784, 2015.

M. Wu, X. Song, C. Jermaine, S. Ranka, and J. Gums. A LRT framework for fast spatial anomaly detection. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 887–896, 2009.

Shi Yu, Léon-Charles Tranchevent, Bart De Moor, and Yves Moreau. *Kernel-based data fusion for machine learning*. Springer, 2013.

Supplementary Material

Gaussian Process Subset Scanning for Anomalous Pattern Detection in Non-iid Data

William Herlands Carnegie Mellon University
Edward McFowland III University of Minnesota
Andrew G. Wilson Cornell University
Daniel B. Neill Carnegie Mellon University

1 Alternative model MLE

Given data, (x, y), we can determine the optimal mean shift, β^* through maximum likelihood estimation as shown below. Let μ, Σ be the posterior mean and covariance of the null model in the domain of x, and denote $E = \Sigma^{-1}$ for brevity.

$$\begin{aligned}\beta^* &= \max_{\beta}\Big((2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}\exp(-\frac{1}{2}(y-w\beta-\mu)^T \\ &\qquad E(y-w\beta-\mu))\Big) \\ &= \max_{\beta} -\frac{1}{2}(y-w\beta-\mu)^T E(y-w\beta-\mu) \\ &= \max_{\beta}(y-\mu)^T Ew\beta - \frac{1}{2}(w\beta)^T E(w\beta)\end{aligned} \tag{1}$$

We take the derivative with respect to β and set it to zero

$$\begin{aligned}\frac{\delta LLR(w)}{\delta\beta} &= (y-\mu)^T Ew - (w\beta^*)^T E(w) = 0 \\ &\Rightarrow (w\beta^*)^T E(w) = (y-\mu)^T Ew \\ &\Rightarrow \beta^* = \frac{w^T E(y-\mu)}{w^T Ew}\end{aligned} \tag{2}$$

2 Iterative β_{MAX} algorithm to approximate optimal subset

Since the derivation of β_{MAX_i} is conditional on a subset w, we obtain the *conditional* optimal subset. In order to approximate an optimal solution we use iteratively compute the conditional optimal subset beginning with a null subset, $w = \vec{0}$. This is an $O(\ell k \log(k))$ algorithm for some ℓ number of iterations, where k is the size of the neighborhood. Pseudo-code is depicted in Alg. 1.

Algorithm 1: Iterative β_{MAX_i} algorithm

Result: Highest scoring subset w^*
Initialize $w = \vec{0}$;
for $l = 1 : \ell$ **do**
  Compute $\beta_{MAX_i}\ \forall i$ conditioned on the current value of w;
  Find highest scoring subset, $w^{(l)}$, using a linear search over sorted β_{MAX_i};
  Compute $LLR(w^{(l)})$;
  Set $w = w^{(l)}$;
end
Choose $w^* = \arg\max_{w^{(l)}} LLR(w^{(l)})$

3 Constrained β_{MAX} optimization over blocks

Although we focus on unconstrained subsets searching within neighborhoods, real world applications sometimes require a more constrained optimization. For example, in spatiotemporal phenomena it is often useful to consider anomalous patterns that are nearby in space and contiguous over time. We can enforce such constraints by predefining mutually exclusive blocks of points, $(x^{(B)}, y^{(B)}) \subseteq (x^{(n)}, y^{(n)})$ where points in a block must all either be included in, or excluded from, a subset.

When considering blocks of points we can compute the total contribution from all points in the block, though we must also account for additional off-diagonal terms in E due to the blocking of data points. Following the derivation in Section 4.2.1 of the main paper, we can derive the β_{MAXb} for each block,

$$\beta_{MAX_B} = \sum_{i\in B}\frac{2\big(E(y^{(n)}-\mu)\big)_i}{\big(\sum_{j\notin B}2w_jE_{j,i} + E_{i,i} + \sum_{k\in B}E_{k,i}\big)} \tag{3}$$

This can be used in a lightly modified version of Algorithm 1 where the β_{MAXb} of blocks, not individual points, is iteratively computed.

4 School Absenteeism

Public schools in New York City record and publish daily student attendance (NYC Department of Education, 2017). Given the importance of education on future outcomes there is tremendous interest in understanding patterns of school absenteeism. We consider public school attendance data in Manhattan for the 2015-2016 school year. The data is messy, with missing entries and non-uniform placement of school locations. We aggregate data at weekly level and remove the last four weeks of the school year since they contain known high absenteeism rates that are not of interest to Department of Education officials.

We apply GPSS methods and baseline approaches with neighborhoods of up to ten local schools. All GPSS methods identified an anomaly around January to February 2016 concentrated on West Side of Manhattan. The results from GRQ around the time of the detected anomaly are presented in Fig. 1. Each dot represents a school location, with yellow dots indicating high attendance and blue dots indicating low attendance. The space-time locations of schools in the top ten anomalous subsets are bordered in red.

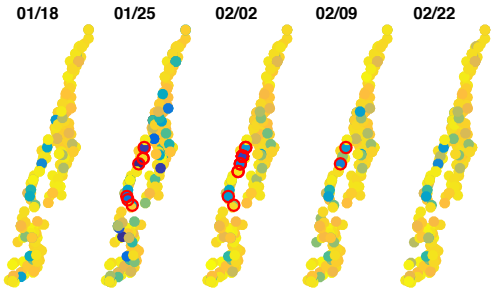

Figure 1: School absenteeism results from Manhattan using GRQ. Each dot represents a school location, with yellow dots indicating high attendance and blue dots indicating low attendance. The space-time locations of schools in the top ten anomalous subsets are bordered in red.

The detected anomalies correspond to a category five blizzard which may have disrupted teachers and students from attending school even though no snow day closings were reported at the time. Further research is required to understand why the West Side of Manhattan differed systematically from the rest of the borough. Baseline anomaly detection methods did not identify a coherent anomaly and instead detected anomalies throughout the year.

References

NYC Department of Education. Data about schools, 2017. URL http://schools.nyc.gov/AboutUs/schools/data/Attendance.htm.