Fast Change Point Detection for Electricity Market Analysis

William Gu[*]
Jaesik Choi[*,‡]
Ming Gu[*,†]
Horst Simon[*]
Kesheng Wu[*]

[*]Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[†]Univeristy of California Berkeley, Berkeley, CA, USA
[‡]Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea

Lawrence Berkeley
National Laboratory

Fast Change Point Detection for Electricity Market Analysis

William Gu, Jaesik Choi, Ming Gu, Horst Simon, Kesheng Wu

September 8, 2013

Abstract

Electricity is a vital part of our daily life; therefore it is important to avoid irregularities such as the California Electricity Crisis of 2000 and 2001. In this work, we seek to predict anomalies using advanced machine learning algorithms, more specifically a Change Point Detection (CPD) algorithm on the electricity prices during the California Electricity Crisis. Such algorithms are effective, but computationally expensive when applied on a large amount of data. To address this challenge, we accelerate the Gaussian Process (GP) for 1-dimensional time series data. Since GP is at the core of many statistical learning techniques, this improvement could benefit many algorithms. In the specific Change Point Detection algorithm used in this study, we reduce the overall computational complexity from O(n^5) to O(n^2), where the amountized cost of solving a GP projet is O(1). Our efficient algorithm makes it possible to compute the Change Points using the hourly price data during the California Electricity Crisis. By comparing the detected Change Points with known events, we show that the Change Point Detection algorithm is indeed effective in detecting signals preceding major events.

Keywords: Change Point Detection, Gaussian Process, Semiseparable matrix, GPSS, BOCPD

1 Introduction

The California Electricity Crisis of 2000 and 2001 is reported to have cost the state's economy about 40 billion dollars [27]. From May 2000 to December 2001, the state experienced severe shortages in electric power caused by unusual weather, state deregulation policies, as well as illicit market manipulation by energy companies [9, 25]. Electricity prices skyrocketed by up to a factor of 800%, as depicted in Figure 1. To allow the market regulators and participants time to respond such irregularities, we aim to detect some leading indicators for such catastrophic events.

Previously, we have applied the similar idea of seeking leading indicators in the stock market [4]. The more general theme is to extract insight from massive amounts of data [11, 22]. In this spirit, we seek to develop an algorithm that is capable of detecting subtle signs of trouble from the available data about the electricity market. However, the detection algorithm used in the earlier study relies on the structure of the stock market that is not present in the electricity market [6]. In this work, we explore a class of machine learning techniques known as Change Point Detection algorithms [1, 2, 17].

Given a time series, Change Points are instances where the process producing the measurements undergoes abrupt and significant changes [2, Ch. 1]. Assuming the time series follows a certain generative model, the Change Point Detection (CPD) algorithms aim to identify changes in the parameters of the model or changes in the model itself. Given a time series such as the electricity market, the change points detected

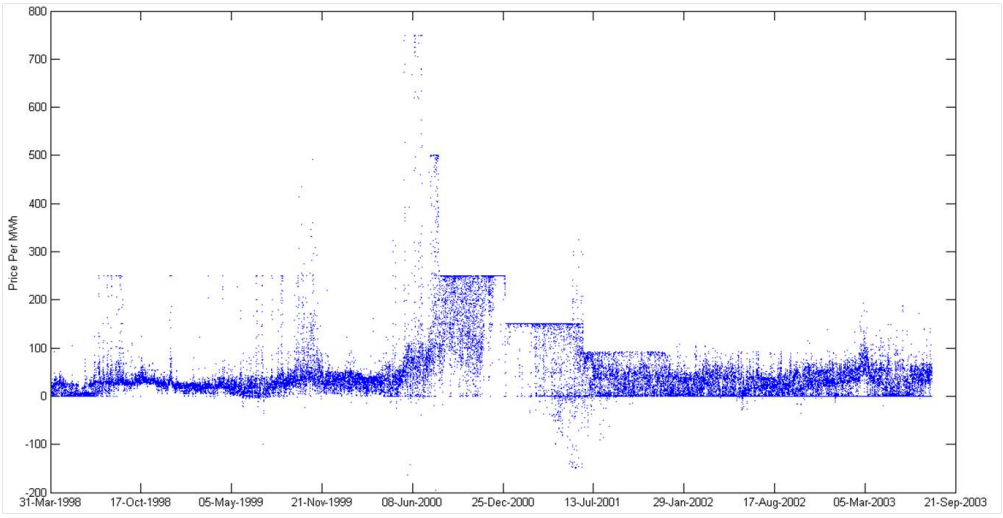

Figure 1: The historical CA ISO price in north California from April 1998 to July 2003.

could suggest changes in important factors affecting the electricity market. Correlating these change points with known events could be useful in understanding the operations of the electricity market and identifying anomalies. While CPD has been used in many applications including robotics and process control, CPD is especially relevant to financial time series, where risk resulting from parameter changes is often neglected in existing models [13, 19]. In this work, we choose to focus on one of the most effective CPDs, known as the Bayesian Online Change Point Detection (BOCPD) [19]. This method avoids the subtle pitfalls of most others and has been demonstrated to be able to detect true change points in nonstationary time series [19].

A notable challenge in using sophisticated statistical learning methods like CPD is that they are computationally expensive. Therefore, they are typically ill-suited for working with large amounts of data. Most existing studies on CPD use only hundreds of data points [19], whereas large time series from financial applications might have thousands or tens of thousands of data points. Our first objective in this work is to reduce the computational complexity of the BOCPD algorithm so it can deal with large data records.

In this work, we exercise the new algorithm with the California Spot Market electricity prices, known as the ISO (or CAISO) prices. They can be thought of as the difference between the actual price and the price set in the day-ahead market [5, 6], therefore, they can be negative in value. This makes it quite different from the prices of typical commodity. However, this does not present any additional difficulty to CPD algorithms. Various published reports [15, 7, 12, 27] have provided details of Enron manipulation schemes including oversubscribing congested transmission lines and causing artificial regional differences, creating uncertainty in the spot markets. We choose to study the ISO prices because these manipulations are more likely to be reflected as the irregularities in ISO prices. The specific data collection we use is from University of California Energy Institute [1]. This data set contains the electricity prices from 1998 to 2003. Since there is a significant amount of documented evidence surrounding the events during this time period, any change point our program might detect could be compared against information in literature. This makes the data particularly useful for studying Change Point Detection algorithms.

[1] `http://www.ucei.berkeley.edu/`

In the remainder of this paper, we provide a brief overview of the Gaussian Process in Section 2 and the BOCPD algorithm in Section 3. In Section 4, we present the techniques used to accelerate the Gaussian Process in the BOCPD algorithm. We first present the covariance matrix used in the Gaussian Process in a semi-separable form, and then describe a recursive solution procedure that produces n solutions in $O(n)$ time. We briefly describe the implementation of the BOCPD using the new Gaussian Process and measure its performance against another version using a well-known implementation of the Gaussian Process. These performance results are presented in Section 5, where we also describe how the detected change points are related to known events reported in the literature.

2 Gaussian Process

The Gaussian Process (GP) is a popular regression tool with many different uses [17, 21]. In this work, it is used as the core of a change point detection procedure. This section provides a brief overview about its computational complexity and its use in the change point detection procedure.

Formally, a Gaussian Process is a stochastic process x_t ($t \in T$), for which any finite linear combination of samples has a multivariate Gaussian distribution. GPs are nonparametric Bayesian, and can be considered as a nonparametric prior over functions [17]. At its core, GP is a stochastic process that assigns its input points to a Gaussian distribution and uses the Gaussian distribution to make predictions about new values. As a non-parametric model, GP makes no underlying assumptions about its inputs other than a specified mean function (m), which is usually set to zero ($m(x) = 0$), and a covariance function (κ) parameterized by a set of hyper-parameters. A popular choice is the set of Matérn covariance functions defined by [16]

$$\kappa(x, x') = \sigma^2 \frac{\Gamma(s+1)}{\Gamma(2s+1)} \sum_{i=0}^{s} \frac{(s+i)!}{i!(s-i)!} \left(\frac{\sqrt{8\nu} r}{\ell} \right)^{s-i} \mathbf{e}^{\left(-\frac{\sqrt{2\nu} r}{\ell}\right)}, \tag{1}$$

where $r = \|x - x'\|$; σ and ℓ are hyper-parameters; $\nu = s + 1/2$ is a half-integer; and $\Gamma(\cdot)$ is the gamma function. For $\nu = 3/2$, $\kappa(x, x')$ takes a simpler form

$$\kappa(x, x') = \sigma^2 \left(1 + \frac{\sqrt{3} r}{\ell} \right) \mathbf{e}^{\left(-\frac{\sqrt{3} r}{\ell}\right)}. \tag{2}$$

Assuming the availability of some noisy observations $y_1, \cdots, y_n$ of the dependent variable y at points $x_1, \cdots, x_n$, one can use GP regression to estimate the value of y at a new point x_{n+1}. Let σ_n be the standard deviation of the noise. If we define the covariance matrix as

$$K = \begin{pmatrix} \kappa(x_1, x'_1) & \kappa(x_1, x'_2) & \cdots & \kappa(x_1, x'_n) \\ \kappa(x_2, x'_1) & \kappa(x_2, x'_2) & \cdots & \kappa(x_2, x'_n) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(x_n, x'_1) & \kappa(x_n, x'_2) & \cdots & \kappa(x_n, x'_n) \end{pmatrix} + \sigma_n^2 I, \tag{3}$$

then the best estimate for y_{n+1} is

$$y_* = K_* K^{-1} (y_1, y_2, \cdots, y_n)^T \tag{4}$$

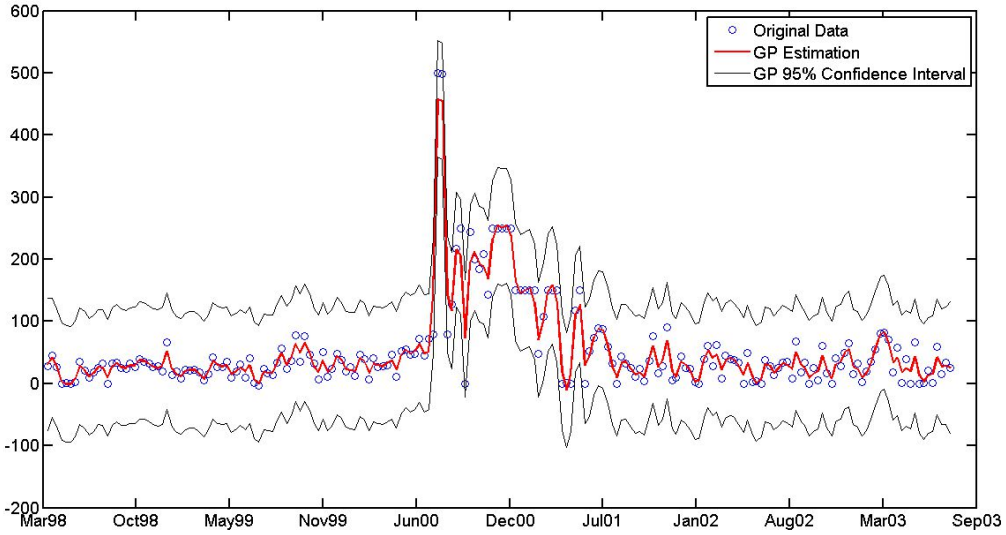

Figure 2: Gaussian Process

with variance

$$\mathbf{var}(y_*) = K_{**} - K_* K^{-1} K_*^T,$$

where $K_{**} = \kappa(x_{n+1}, x_{n+1}) + \sigma_n^2$ and

$$K_* = \left(\kappa(x_{n+1}, x_1'), \kappa(x_{n+1}, x_2'), \cdots, \kappa(x_{n+1}, x_n')\right).$$

As K^{-1} is involved, the above expressions typically require $O(n^3)$ operations and $O(n^2)$ memory to compute.

Although GP is computationally expensive, its non-parametric nature and its ability to provide a confidence interval allows it to adapt better to the changes in data than a typical parametric model could, thus yielding superior predictions. Figure 2 illustrates the GP approximating the data and a 95% confidence interval.

Figure 3 compares the GP regression with the less expensive, parametric ARIMA model; GPs smaller errors testify to its higher accuracy. Our primary objective in this work is to reduce the computational cost of GP while retaining its effectiveness.

3 Bayesian Online Change Point Detection

The Change Point Detection (CPD) [2, 3, 18] is an algorithm that detects changes in sequential data unders the assumption that the sequence data is composed of several runs. A run is best defined as the data of a specific time interval where the data fits a stochastic process without large deviations. In practice, it is not always clear how to split two consecutive runs. More generally, dividing a long sequence of data into runs in a challenging task. CPD algorithms generally work by estimating the length of the run (or run length) at every data point.

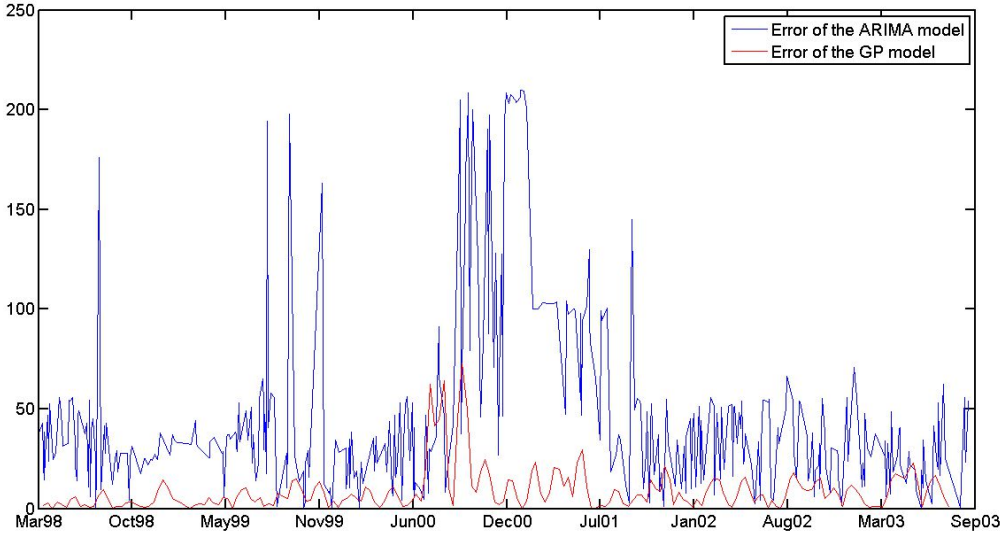

Figure 3: GP vs. ARIMA: GP generally has smaller errors.

Let r_t be the random variable representing the run length at time t. The goal of the CPD algorithm is to find the distribution of the random variable, $p(r_t)$. The Bayesian Online Change Point Detection (BOCPD) [19] finds the (distribution of) run length with a Bayesian update equation.

Given a sequence of data up to time $t-1$, $y_{1:t-1}$, and the distribution of run length $p(r_{t-1})$, the BOCPD algorithm predicts the run length r_t and data y_t as follow,

$$\begin{aligned} p(y_t, r_t) &= p(y_t, r_t|y_{1:t-1})p(y_{1:t-1}) \\ &\propto p(y_t, r_t|y_{1:t-1}) \\ &= \sum_{r_{t-1}} p(y_t, r_t|y_{1:t-1}, r_{t-1})p(r_{t-1}) \\ &= \sum_{r_{t-1}} p(y_t|y_{1:t-1}, r_{t-1})p(r_t|y_{1:t-1}, r_{t-1}) \\ &= \sum_{r_{t-1}} p(y_t|y_{t-r_{t-1}:t-1})p(r_t|y_{1:t-1}, r_{t-1}) \end{aligned}$$

Here, $p(y_t|y_{t-r_{t-1}:t-1})$ is an Underlying Predictive Models (UPM) which here, is the Gaussian Process, as used in [19]. $p(r_t|y_{1:t-1}, r_{t-1})$ is a Hazard function, which is choosen as a constant function in this study.

Intuitively, the $run\ length$ represents the length of a time segment with similar statistical behavior. At each time t, GP is used to compute conditional probabilities $p(y_t|y_{(t-r):(t-1)})$ for all the possible values of the run length $r_{t-1} \in [1, t-1]$. Such probabilities are then used to determine the run length based on the recursive formula above.

BOCPD invokes GP on each possible combination of subsequence of data records. Give an input time series of n data records, there are $n(n+1)/2$ subsequences. Given that GP has a computational complexity of $O(n^3)$, the overall computational complexity of BOCPD is $O(n^5)$. This cost is justified by its power in

procedure BOCPD($y_{1:n}$) ▷ Input data
  $p(r_0{=}1) \leftarrow 1; p(r_0 \neq 1) \leftarrow 0$
  $t \leftarrow 1$
  for $t < n$ **do**
    $t \leftarrow t + 1$
    $f_H(r_t{=}1) \leftarrow c$ ▷ f_H: Hazard Function
    for all $j > 1$ **do**
      $f_H(r_t{=}j) \leftarrow p(r_{t-1}{=}j-1)$
    end for
    $rlen \leftarrow 1; tot = 0$
    while $t - rlen > 0$ **do**
      $f(y_t, rlen) \leftarrow p_{GP}(y_t | y_{t-rlen:t-1})$ ▷ p_{GP}: Gaussian Process
      $f(r_t{=}rlen) \leftarrow f(y_t, rlen) \cdot f_H(r_t{=}rlen)$
      $tot \leftarrow tot + f_(r_t{=}ren)$
      $rlen \leftarrow rlen + 1$
    end while
    for all $rlen$ **do**
      $p(r_t{=}rlen) \leftarrow f(r_t{=}rlen)/tot$ ▷ Normalize
    end for
  end for
  return $(p(r_1), \cdots, p(r_t)))$ ▷ The dist. of run lengths
end procedure

Figure 4: BOCPD algorithm

detecting subtle changes in important applications [2, 3, 18, 19]. However, in most published reports, CPD algorithms typically handle only a few hundred data points. In the next section, we describe a strategy that could significantly reduce the computational cost and make BOCPD suitable for large data sets.

4 Semi-Separable Matrices

In this section, we describe a technique that takes advantage of the algebraic structure in the matrix K in Equation 3 to effectivly reduce the cost of solving n Gaussian Processes in $O(n)$ time.

When using a GP on a 1-dimensional time series, where x_t for each time t is a real number the covariance matrix K in (3), which, here, is based on the Matérn function (2), has a special matrix structure. To illustrate, we assume that $x_1 < x_2 < \cdots < x_n$ have been arranged in ascending order. We can rewrite K as

$$K = D + \mathbf{triu}\left(PQ^T\right) + \left(\mathbf{triu}\left(PQ^T\right)\right)^T, \tag{5}$$

where we have used the Matlab notation $\mathbf{triu}(\cdot)$ to denote the strictly upper triangular part of a given matrix;

and

$$\begin{aligned} D &= \mathbf{diag}\left(\sigma^2+\sigma_n^2, \sigma^2+\sigma_n^2, \cdots, \sigma^2+\sigma_n^2\right), \\ P &= \sigma \begin{pmatrix} \mathbf{e}^{\left(\frac{\sqrt{3}x_1}{\ell}\right)} & x_1\mathbf{e}^{\left(\frac{\sqrt{3}x_1}{\ell}\right)} \\ \vdots & \vdots \\ \mathbf{e}^{\left(\frac{\sqrt{3}x_n}{\ell}\right)} & x_n\mathbf{e}^{\left(\frac{\sqrt{3}x_n}{\ell}\right)} \end{pmatrix}, \\ Q &= \sigma \begin{pmatrix} \left(1+\frac{\sqrt{3}x_1}{\ell}\right)\mathbf{e}^{\left(-\frac{\sqrt{3}x_1}{\ell}\right)} & -\frac{\sqrt{3}}{\ell}\mathbf{e}^{\left(-\frac{\sqrt{3}x_1}{\ell}\right)} \\ \vdots & \vdots \\ \left(1+\frac{\sqrt{3}x_n}{\ell}\right)\mathbf{e}^{\left(-\frac{\sqrt{3}x_n}{\ell}\right)} & -\frac{\sqrt{3}}{\ell}\mathbf{e}^{\left(-\frac{\sqrt{3}x_n}{\ell}\right)} \end{pmatrix}. \end{aligned}$$

Equation (5) also holds for the more general Matérn function in (1), with a diagonal matrix D and $n\times(s+1)$ matrices P and Q.

Matrices of the form (5) are known as *Semi-Separable* matrices, with a large literature on their fast factorization and inversion [8, 26]. Below we describe a recursive procedure to compute $u^T K^{-1} v$ for any vectors $u = (u_1, \cdots, u_n)^T$ and $v = (v_1, \cdots, v_n)^T$ in $O(n)$ time. For this purpose, write

$$D = \mathbf{diag}\left(d_1, \cdots, d_n\right),$$

$$P = \begin{pmatrix} p_1^T \\ \vdots \\ p_n^T \end{pmatrix},$$

and

$$Q = \begin{pmatrix} q_1^T \\ \vdots \\ q_n^T \end{pmatrix}.$$

To begin, define

$$\begin{aligned} \widehat{A}_k &= \begin{pmatrix} q_k & \cdots q_n \end{pmatrix} K^{-1}_{k:n,k:n} \begin{pmatrix} q_k^T \\ \vdots \\ q_n^T \end{pmatrix}, \\ \delta_k &= \begin{pmatrix} u_k & \cdots u_n \end{pmatrix} K^{-1}_{k:n,k:n} \begin{pmatrix} v_k \\ \vdots \\ v_n \end{pmatrix}, \\ \widehat{U}_k &= \begin{pmatrix} q_k & \cdots q_n \end{pmatrix} K^{-1}_{k:n,k:n} \begin{pmatrix} u_k \\ \vdots \\ u_n \end{pmatrix}, \\ \widehat{V}_k &= \begin{pmatrix} q_k & \cdots q_n \end{pmatrix} K^{-1}_{k:n,k:n} \begin{pmatrix} v_k \\ \vdots \\ v_n \end{pmatrix}, \end{aligned}$$

where we have used the Matlab notation $K_{k:n,k:n}$ to denote the tailing $(n-k+1)\times(n-k+1)$ submatrix of K. The following recursion allows for the computation of all of $\widehat{A}_k, \widehat{U}_k, \widehat{V}_k$ and δ_k in $O(n)$ time without explicitly inverting any $K_{k:n,k:n}$.

Let

$$\begin{aligned}\widehat{A}_{n+1} &= 0,\\ \widehat{U}_{n+1} &= 0,\\ \widehat{V}_{n+1} &= 0,\\ \delta_{n+1} &= 0.\end{aligned}$$

Do for $k = n, n-1, \cdots, 1$:

$$\begin{aligned}\widehat{d}_k &= d_k - p_k^T \widehat{A}_{k+1} p_k,\\ \tau_k &= q_k - \widehat{A}_{k+1} p_k,\\ \widehat{u}_k &= u_k - p_k^T \widehat{U}_{k+1},\\ \widehat{v}_k &= v_k - p_k^T \widehat{V}_{k+1},\\ \delta_k &= \delta_{k+1} + \frac{\widehat{u}_k \widehat{v}_k}{\widehat{d}_k},\\ \widehat{A}_k &= \widehat{A}_{k+1} + \frac{\tau_k \tau_k^T}{\widehat{d}_k},\\ \widehat{U}_k &= \widehat{U}_{k+1} + \frac{\widehat{u}_k \tau_k}{\widehat{d}_k},\\ \widehat{V}_k &= \widehat{V}_{k+1} + \frac{\widehat{v}_k \tau_k}{\widehat{d}_k}.\end{aligned}$$

Now let $v = (y_1 \quad y_2, \cdots, y_n)^T$. To compute y_* in equation (4), all that is needed is to apply the above recursion with $u = K_*$.

However, there is a rather remarkable feature about the above recursion. By their definitions, δ_k is in fact the prediction y_* based on the points $(x_k, y_k), \cdots, (x_n, y_n)$ for every $1 \le k \le n$. In other words, we have computed all n predictions for y_* in $O(n)$ time. The computation of the variances in equation (4) follows a similar pattern.

The dominant cost of the probabilities $p(y_t | y_{(t-r):(t-1)})$ is in the GP predictions of y_t and their variances given $y_{(t-r):(t-1)}$ for all $r < t$. The above recursion can thus be used to compute all these predictions and variances, and therefore all the probabilities $p(y_t | y_{(t-r):(t-1)})$, in $O(t)$ time, leading to amortized $O(1)$ time for each GP and each probability, and quadratic time for BOCPD.

For the optimal use of GP and BOCPD, the hyperparameters can be selected through an optimization procedure, such as maximum likelihood. The semi-separable matrix structure of the covariance matrix K can also be exploited to perform hyperparameter training in linear time.

The squared-exponential, $\kappa(x, x') = \sigma^2 exp(\frac{\|x-x'\|_2^2}{2\ell^2})$, is another popular covariance function. It is known to be well-approximated by the Matérn functions. This approximation allows us to utilize the recursion above to perform rapid GP regression and BOCPD for the squared-exponential as well.

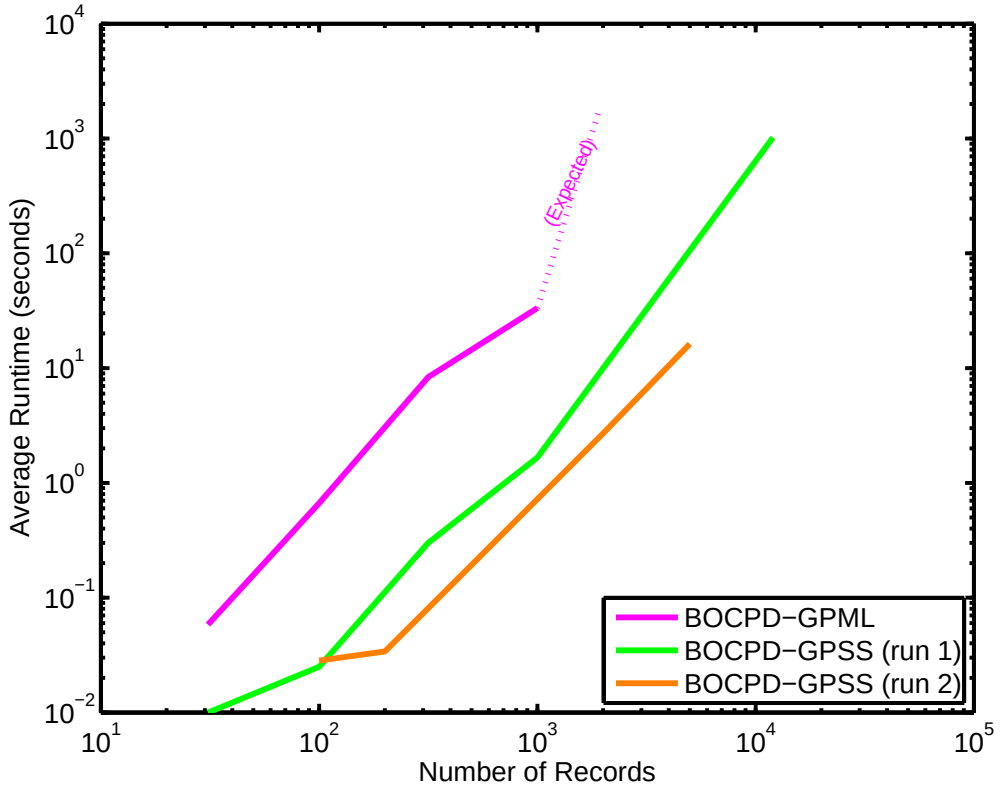

Figure 5: The time (seconds) used by the two versions of the BOCPD algorithm using GPML and our GPSS.

This vast improvement in scalability allows us to run BOCPD in Matlab with more than 10,000 data points on a laptop overnight, a previously huge task that could be realistically attempted with only the fastest supercomputers.

5 Experimental Results

We have implemented a version of BOCPD in Matlab following the description by the original authors [19]. The initial version of the code uses GPML to solve the Gaussian Processes [17][2]. A faster version is also implemented using the algorithm described in the previous section to solve the Gaussian Processes. In this section, we will present some timing results to compare the two versions of BOCPD and discuss the changes points detected.

The 1-dimensional time series used in our study consists of prices at different time period. The raw time series is expected to have one value per hour, however, there are some hours with missing values. In addition, we use smaller samples in many timing tests. To accomodate these variations, we explicitly record the time values, which are the x_i in the earlier discussions. The y_i values are the corresponding prices.

[2]Information about GPML is available at `http://www.gaussianprocess.org/gpml/`.

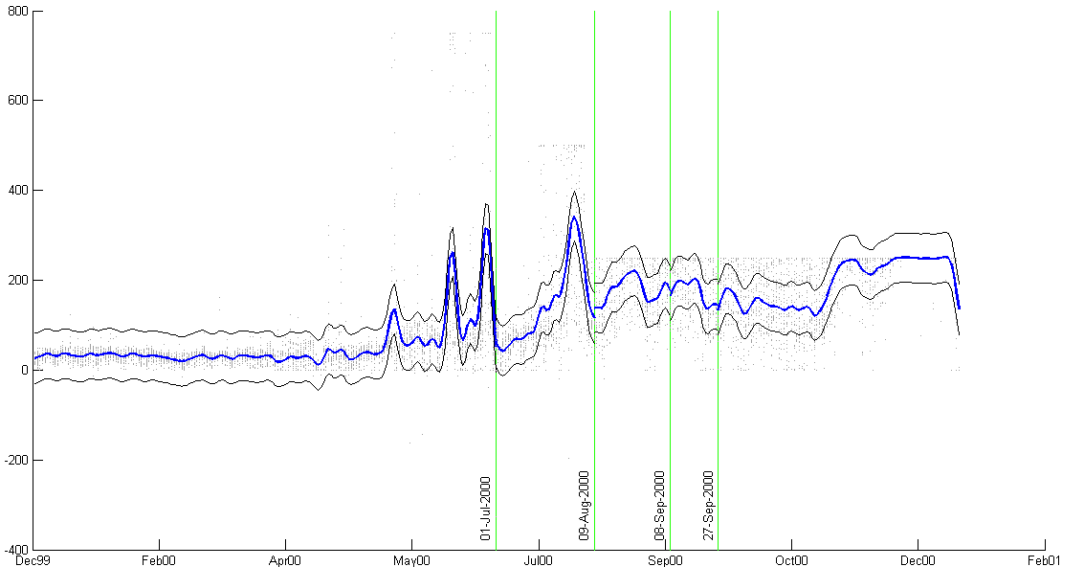

Figure 6: The ISO prices during 2000 (blue dots) with Gaussian processes from different runs separated by change points (green lines).

Figure 5 shows the time used by the two versions of the BOCPD algorithm. The test runs with different sized samples of the hourly ISO prices from 2008 to 2001. Because the GPML code uses the efficient Matlab built-in functions to solve the linear systems, the BOCPD with GPML actually can handle up to 1000 data points in a reasonable amount of time. Event in this case, the new method, marked BOCPD-GPSS, is at least 10 times faster than BOCPD with GPML. Furthermore, we see that BOCPD with GPSS can easily handle 10,000 data points even though our algorithm uses interpreted Matlab statements, while the BOCPD with GPML did not finish within 24 hours.

An execution of our algorithm on the market data produces Figures 6 and 7, which display several runs during 2000 and 2001, the years of the California Electricity Crisis. These runs are separated by change points, represented by green lines, and sometimes coincide with the dates in Table 1, a chronology of important events relating to the Crisis.

Among the change points detected in year 2000, see Figure 6, the first change point was July 1. This is the date when the price cap was reduced from \$750/MWh to \$500/MWh. Prior to this date, there were signification volatility in the ISO prices; and there are also evidence of price manipulations from sources including the Enron email archive [20, 23, 24]. The price cap reduced again to \$250/MWh in early August, 2000. However, by this time, even the ISO prices during the off-peak hours are quite high. The two change points in September appear to be related to two instances where the minimum prices in each day have reached over \$100/MWh. Of course, these are only anecdotal observations, further work is needed to systematically test the validity of all change points.

Table 1: List of change points detected and their possible associated events. List of events extracted from various published sources [9, 10, 25]. The terms "Fat Boy", "Death Star", "counter-flow", and "ricochet" refer to market manipulation schemes identified by investigators [15, 12, 27]. Email messages are from top Enron managers [14, 24].

date YYYY/MM/DD	**Related events**
1998/08/06	5-day long heat wave ($> 100°$F)
1998/09/14	Unseasonably warm ($> 95°$F); reaching price cap \$250
1998/12/31	Cold winter
1999/08/13	CA ISO authorized price cap increase to \$750; Enron emails mentioned "Fat Boy" for the 1st time on 06/25, and "Death Star" on 08/23
1999/10/01	Price ceiling raised to \$750
2000/07/01	CA ISO reduced price cap to \$500; near 100,000 Pacific Gas & Electric (PG&E) customers suffer black outs on 06/14; Enron emails mentioned "counter-flow" for the 1st time on 04/13, and "ricochet" on 03/14;
2000/08/02	Price ceiling lowered to \$250.
2000/08/09	CA ISO reduced price cap to \$250 on 08/07; fossil fuel price rises increased cost of peak-electricity producers
2000/09/08	FERC launched investigation of Enron on 08/23; San Diego Gas & Electric (SDG&E) Company failed bankruptcy; Enron emails mentioned "ricochet" 221 times on 09/12
2000/09/27	FERC met in San Diego on 09/12; FERC was to allow "flexible" price cap in December
2000/12/15	FERC rejected firm cap requested by California, but approved "flexible" cap
2001/02/11	Blackouts affected 100,000s on 01/17-18; State of emergency declared on 01/17; Enron emails mentioned "ricochet" 324 times on 02/22 and "Death Star" 95 times on 02/28
2001/03/03	Blackouts affected millions on 03/19-20; Enron emails mentioned "ricochet" 380 times on 03/08 and "counter-flow" 78 times on 04/19
2001/03/14	FERC orders increase in natural gas and reduction of energy demand
2001/03/28	FERC discovers El Paso Natural Gas Company of market manipulation and orders cessation of its illicit practices.
2001/04/21	PG&E filed for bankruptcy;
2001/05/20	California authorized bonds to buy electricity using long-term contracts
2001/06/09	FERC announces a price-mitigation plan; lower demands reduced spot prices blow long-term contract prices; nearly two weeks of high temperatures in early August
2001/12/02	Enron, the main company behind the market manipulation, files for bankruptcy
2001/08/16	Prices capped at \$100
2002/03/08	FERC opened investigation of market manipulation
2002/04/11	Long-term contract sales begin to replace spot sales.
2002/05/23	Enron email mentioned "Death Star" one last time on 05/24; CPUC re-impose regulations
2003/03/10	FERC released investigation report
2003/11/17	Anti-Manipulation laws finalized and put into effect.

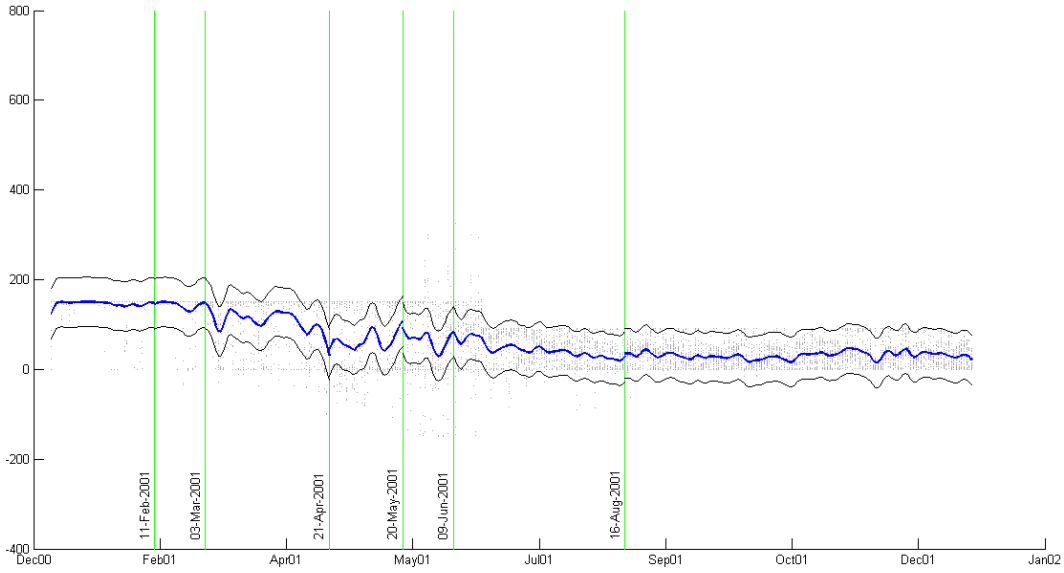

Figure 7: The ISO prices during 2001 (blue dots) with Gaussian processes from different runs separated by change points (green lines).

6 Conclusion

In this paper, we present a strategy to significantly accelerate the Gaussian Process on 1-dimensional time series by taking advantage of the structure of the matrices. The technique represents the covariance matrix in semi-separable form and then applies a recursive solution procedure. The overall effect is that we are able to reduce the computational complexity of the Bayesian Online Change Point Detection (BOCPD) from $O(n^5)$ to $O(n^2)$ on a time series of n records. Since GP is at the core of many machine learning techniques, reducing the solution time for GP could benefit many applications.

To demonstrate the efficiency of the new GP algorithm, we apply it to a change point detection procedure that makes extensive use of GP. In our timing measurements, we see that that the new GP algorithm significantly reduces overall execution time (by more than a factor of 10). We further demonstrate that BOCPD can effectively identify important events around the California Electricity Crisis from the price information alone. The changes detected include seasonal and policy changes, as well as market manipulations. Therefore, we believe the Change Points detected are useful in monitoring market activities.

Additional work is needed to further develop the GP and establish the effectiveness of BOCPD. Our solution strategy currently only apply to 1-Dimensional time series, we are working on extending this to 2- and 3-Dimensional cases. In the discussion of the Gaussian Process, we mentioned that different kernels that could be used in GP. One direction of future work would be explore ways to accelerate Gaussian Processes using a variety of different kernels. The current implementation of the fast Gaussian Process is in Matlab scripts. We plan to rewrite the software in C or C++. This has the potential to speed up the software considerably.

7 Acknowledgements

We would like to thank Dr. Emily Fisher for providing a concise description of ISO pricing and for reviewing the drafts of this paper.

This work is supported in part by the Director, Office of Laboratory Policy and Infrastructure Management of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

[1] Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

[2] Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes - Theory and Application*. Information and System Sciences Series. Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

[3] István Berkes, Edit Gombay, Lajos Horváth, and Piotr Kokoszka. Sequential change-point detection in GARCH(p, q) models. *Econometric Theory*, 20:1140–1167, 2004.

[4] E. Wes Bethel, David Leinweber, Oliver Rübel, and Kesheng Wu. Federal market information technology in the post-flash crash era: Roles for supercomputing. *The Journal of Trading*, 7(2):9–25, 2012. `http://dx.doi.org/10.3905/jot.2012.7.2.009`.

[5] Carl Blumstein, Lee S Friedman, and Richard Green. The history of electricity restructuring in california. *Journal of Industry, Competition and Trade*, 2(1-2):9–38, 2002.

[6] Severin Borenstein. The trouble with electricity markets: understanding california's restructuring disaster. *The Journal of Economic Perspectives*, 16(1):191–211, 2002.

[7] James Bushnell. California's electricity crisis: a market apart? *Energy Policy*, 32(9):1045–1052, 2004.

[8] Srinivasan Chandrasekaran, Patrick Dewilde, Ming Gu, T Pals, and Alle-Jan van der Veen. *Fast stable solver for sequentially semi-separable linear systems of equations*. Springer, 2002.

[9] Charles J Cicchetti, Jeffrey A Dubin, and Colin M Long. *The California electricity crisis: What, why, and what's next*. Springer, 2004.

[10] Richard D Farmer, Dennis Zimmerman, and Gail Cohen. *Causes and lessons of the California electricity crisis*. US Congressional Budget Office, Sept. 2001. A CBO paper.

[11] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft, October 2009.

[12] Paul L Joskow. California's electricity crisis. *Oxford Review of Economic Policy*, 17(3):365–388, 2001.

[13] David Xianglin Li. On default correlation: a copula function approach. *Journal of Fixed Income*, 9(4):43–54, 2000.

[14] Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Intell. Res.(JAIR)*, 30:249–272, 2007.

[15] Robert McCullough. Fat boy report, 2003. `http://www.mresearch.com/pdfs/84.pdf`.

[16] Christopher J Paciorek and Mark J Schervish. Nonstationary covariance functions for gaussian process regression. *Advances in neural information processing systems*, 16:273–280, 2004.

[17] Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT Press, 2006.

[18] Bonnie K Ray and Ruey S Tsay. Bayesian methods for change-point detection in long-range dependent processes. *Journal of Time Series Analysis*, 23(6):687–705, 2002.

[19] Yunus Saatçi, Ryan D Turner, and Carl E Rasmussen. Gaussian process change point models. In *ICML-10*, pages 927–934, 2010.

[20] Jitesh Shetty and Jafar Adibi. The Enron email dataset database schema and brief statistical report. Technical Report 4, Information sciences institute technical report, University of Southern California, 2004.

[21] Jian Qing Shi and Taeryon Choi. *Gaussian process regression analysis for functional data*. CRC Press, 2011.

[22] Arie Shoshani and Doron Rotem. *Scientific Data Management: Challenges, Technology, and Deployment*, volume 3. Chapman and Hall/CRC, 2010.

[23] K. Stockinger, D. Rotem, A. Shoshani, and K. Wu. Analyzing enron data: Bitmap indexing outperforms MySQL queries by several orders of magnitude. Technical Report LBNL-61083, Lawrence Berkeley National Laboratory, 2006.

[24] Kurt Stockinger, John Cieslewicz, Kesheng Wu, Doron Rotem, and Arie Shoshani. *Using Bitmap Indexing Technology for Combined Numerical and Text Queries*, volume 3 of *Annals of Information Systems*, pages 1–23. Springer, 2008. Preprint appeared as LBNL Tech Report LBNL-61768.

[25] James L Sweeney. *The California electricity crisis*. Hoover Press, 2008.

[26] Raf Vandebril, Marc Van Barel, and Nicola Mastronardi. *Matrix Computations and Semiseparable Matrices: Linear Systems (Volume 1)*. The Johns Hopkins University Press, 2007.

[27] Christopher Weare. *The California electricity crisis: causes and policy options*. Public Policy Instit. of CA, 2003.