# Accurate Change Point Detection for Electricity Market Analysis

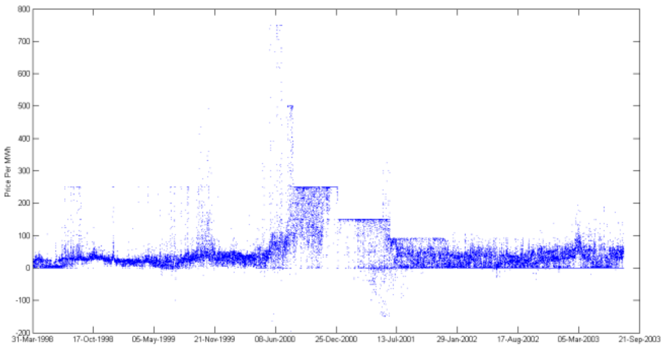

Will Gu

Berkeley Lab & UC Berkeley

wcgu@lbl.gov


## I: Introduction

In what is now known as the California Electricity Crisis, the state experienced severe shortages in electric power from May 2000 to December 2001. It was caused by several contributing factors, including weather, state deregulation policies, as well as illicit market manipulation by energy companies, mainly Enron. Electricity prices, skyrocketed by up to a factor of 800%, as depicted by the figure below.



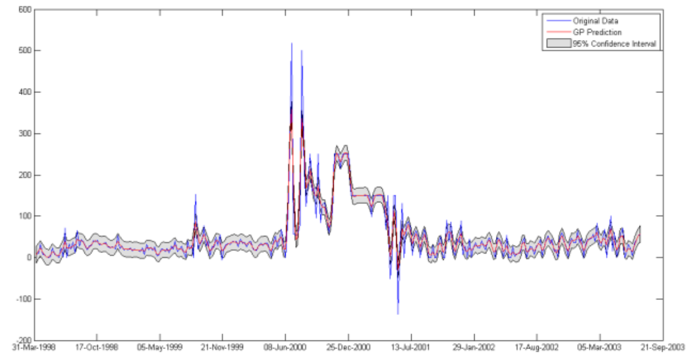

In this paper, our goal is to detect points of unnatural price changes within our time frame and possibly associate them to instances of external price-changing events.

## II: Gaussian Process

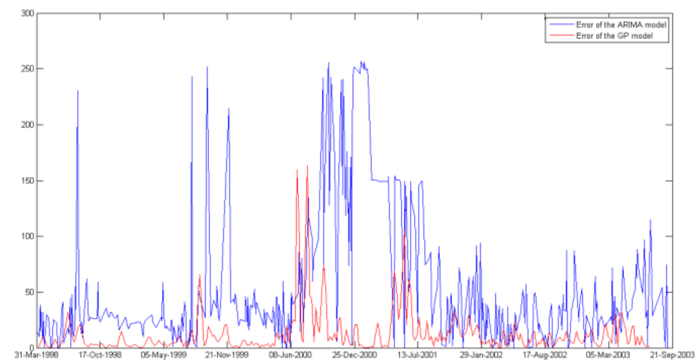
The Gaussian Process (GP) is a popular regression tool. At its core, GP is a stochastic process that assigns its input points to a Gaussian distribution and uses it to make predictions at new values. A non-parametric model, GP makes no underlying assumptions about its inputs and is instead specified by a mean function (m), which is set to 0 (m(x) = 0,) and the covariance function (κ), which is set to be one of the Matern covariance functions.

$$(1) \qquad \kappa(x, x') = \sigma_f \left[1 + \sqrt{3}\frac{\|x-x'\|}{\rho}\right] exp\left(-\sqrt{3}\frac{\|x-x'\|}{\rho}\right).$$

Although GP is computationally expensive, taking cubic time ($O(n^3)$), its non-parametric nature and its ability to provide a confidence interval allows it to better adapt to the changes of the data than a typical parametric model could, thus yielding superior predictions. The next figure illustrates the GP approximating the data and a 95% confidence interval.

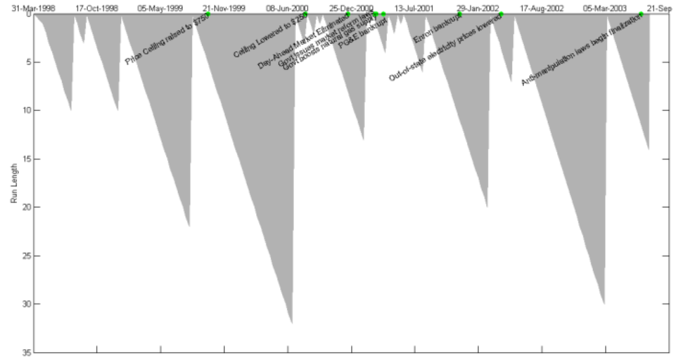



The figure below compares GP with the less expensive, parametric ARIMA model; GP' smaller errors testify to its higher accuracy.




## III: BOCPD

The Bayesian Online Change Point Detection (BOCPD) is a Bayesian change detection algorithm that utilizes GP as an Underlying Predictive Model (UPM). The key concept is the *run length*, the length of time segment with similar statistical behavior. For each t, GP is used to compute conditional probabilities $p(y_t|y_{(t-r):(t-1)})$ for all $r \in [1, t-1]$. Such probabilities are then used to determine the run length based on a recursive formula. The main computational cost of this algorithm is in GP for all time segments, as BOCPD attempts to fit a GP to every possible time segment for the purpose of finding the best series of GPs to fit the data. In utilizing the BOCPD, we can identify the points in between runs as change points, or points during which the market experienced unusual price anomalies, including market manipulations.




## IV: Semi-Separable Matrices

Change detection takes quadratic time ($O(n^2)$) for each time segment, so incorporating GP models to BOCPD results in an algorithm that runs $O(n^5)$. Since our data contains 48000 values, scalability proves heavily problematic, as running just the first 200 points would consume several hours.

It turns out that GP can be rapidly computed for most financial time series data, which are typically 1-dimensional (i.e., $x_t$ for each t is a real number.) In particular, for the Matern function (1), the covariance matrix in GP is a semi-separable matrix with off-diagonal rank 2. In other words, there exist a diagonal matrix D, and matrices $P, Q \in R^{n\times 2}$ such that

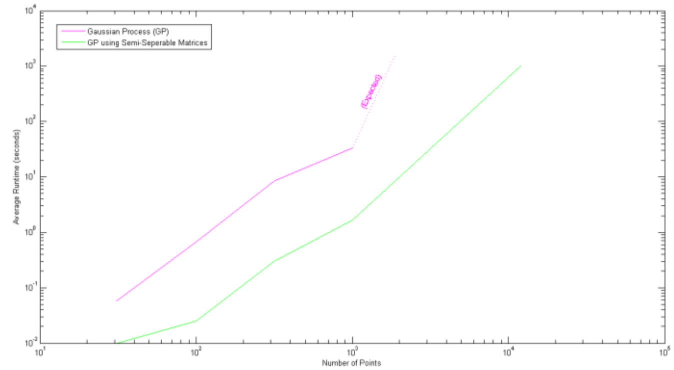
$$K = \begin{bmatrix} \kappa(x_1,x_1) & \kappa(x_1,x_2) & \dots & \kappa(x_1,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(x_n,x_1) & \kappa(x_n,x_2) & \dots & \kappa(x_n,x_n) \end{bmatrix} + \sigma_n^2 I = D + triu(P\,Q^T) + (triu(P\,Q^T))^T$$

where we have used the Matlab notation $triu(\cdot)$ to denote the upper triangular part of a given matrix. With this algebraic matrix structure, the cost of GP can be reduced to O(n) (Chandrasekaran, *et al*, 2002). Based on a fast semi-separable Cholesky update formula, the probabilities $p(y_t|y_{(t-r):(t-1)})$ for all r < t can be computed in O(t) time, leading to amortized O(1) time for each GP.

The semi-separable matrix structure exists for all Matern covariance functions. The more popular covariance function, the squared-exponential,

$$\kappa(x, x') = \sigma_f \, exp\left[-\frac{\| x - x' \|_2^2}{2\ell^2}\right],$$

is known to be well-approximated by the Matern functions, allowing rapid GP regression as well. This algorithmic advance results in a quadratic time method for change point detection.




This vast improvement in scalability allows us to run BOCPD with 48000 point data set on a laptop overnight, a task previously possible with the fastest supercomputers.

## References:


Adams, P.A., and MacKay, D.J.C. 2007 *Bayesian Online Changepoint Detection*. Technical report. University of Cambridge, Cambridge, UK.

Borenstein, Severin. "The trouble with electricity markets: understanding California's restructuring disaster." The Journal of Economic Perspectives 16.1 (2002): 191-211.

Chandrasekaran, S. and Dewilde, P. and Gu, M. and Pals, T. and van der Veen, A.J. , *Fast stable solver for sequentially semi-separable linear systems of equations*, Lecture Notes in Computer Science, vol. 2552, Springer-Verlag, Heidelberg, 2002.

Cunningham, J.P., Zoubin G., and Rasmussen, C.E. *Gaussian Processes for time-marked time-series data*. University of Cambridge, Cambridge, UK.

Rasmussen, C. and Williams C. 2006. Gaussian Processes for Machine Learning. MIT Press. Cambridge.

Saatci, Y., Turner, R. and Rasmussen, C.E. 2010. Gaussian Process Change Point Models, International Conference on Machine Learning (ICML).