Efficient Large-Scale Join Point Estimation Under the Influence of Heavy-Tailed Residual Distributions

Myrl G. Marmarelis

Abstract

This study details the construction of a technique that estimates an optimal number of linear segments connected via join points for noisy time series. The method starts by saturating the model with join points and proceeds to greedily remove the least important ones until a Bayesian Information Criterion reaches its minimum. The criterion is developed based on assumptions of the underlying residual distributions; first a Gaussian setting is explored, then robustness is increased by switching to Laplace-distributed and finally Lomax-distributed residuals. The Gaussian model performs well against a macroeconomic index sampled weekly, whereas the robust versions are needed when fitting higher-frequency foreign exchange and Bitcoin data.

1 Introduction

Time-series analysis involves the study of a process that evolves over time. Usually this realizes into the search for patterns in the series itself or interactions with other variables. In fields like econometrics, it is of immense value to be able to identify features such as trends in otherwise noisy data. The extraction of this information helps to make predictions about the future. In the traditional setting, one treats the time series as a sequence of random variables and then looks for non-stationarities, which are changes in the variables' distributions over subsequent periods of time.

Previous work on segmentation of stochastic time series has mostly focused on locating change-points (also known in econometric literature as structural breaks) in mean and variance. In these models, the mean and variance of the variables shift at discrete points in time. The task of estimating both the number and the locations of these change points is vastly nontrivial, and a multitude of techniques have been developed to tackle it [1, 2], [3, 4, 5, 6]. This study works on a different construction: instead of looking for sudden changes in statistical characteristics, decompose the signal into a sequence of connected linear segments. These segments can be viewed as the integral of a shifting mean function. So from within the mean-shift framework, the present task is to minimize (using the Euclidean norm)

$$\sum_{t\in T}\left[X(t)-\int_0^t \mu(\tau)d\tau\right]^2, \text{ as opposed to } \sum_{t\in T}[x(t)-\mu(t)]^2.$$

Clearly, while there is room for cross-pollination in both directions (e.g. [7]), this is an entirely different proposition. Clinical medicine has widely adopted a "joinpoint regression" that shares this study's goal and is largely spearheaded by Kim et al. [8, 9, 10]. In its applied form, the technique can lay heavy on computation. As a result most studies are severely limited to a mere handful of join points. A new technique is presented in this study with the aim to robustly and efficiently estimate a more flexible number of join points. It starts with a very large amount of hypothetical join points, then proceeds to greedily merge the least effectual ones until the fit has been reduced to the most significant trends found in the time series. Whereas most join-point regression techniques so far (with the exception of [11] and a few others) tend to start with one join point and recursively add more, here the opposite procedure is followed.

1.1 Background & Motivation

In 2000, Kim et al. published a seminal paper [8] that set the standard for join-point regression in cancer research. Since then, it has been subject to incremental improvements and employed in various medical studies, e.g. [12]. Its procedure is roughly as follows: perform a statistical test between H_0: there are k_0 join points and H_1: there are k_1 join points, with $0 \leq k_0 < k_1$. If H_0 is rejected, test against new hypotheses with $(k_1 + 1)$ join points. Within each test, a grid search through the join points' time components finds the optimal least-squares fit for a fixed k_0. Then the residuals of this model are taken and shuffled, and k_1 join points are fit in the same way on the new data set with permuted residuals. The two fits are compared via the F-statistic, and this whole procedure is repeated in a Monte Carlo fashion until the desired significance is achieved. In recent years, Schwarz' Bayesian Information Criterion [13] and its variants were introduced as an alternative mechanism for selecting between models with different values for k [9]. When residuals are assumed to be Gaussian, the plain BIC has been shown to overestimate the number of join points [14].

Residuals of other—particularly heavier-tailed—distributions have not been studied extensively. Mandelbrot was the first to observe that price fluctuations in financial markets are more wild than would be expected within Gaussian models [15]. Rather, they appear to be distributed by a power law. This study analyzes residuals under three separate distributional assumptions: first the Gaussian, then the Laplace, and finally the Lomax power-law distribution.

To expand the scale of applicability for join-point estimation, no exhaustive search is used when looking for optimal join-point locations with the proposed methodology. Instead, a saturating number of join points are placed on the time series and then greedily pruned, one by one, until only the ones that are perceived to be the most important are left.

2 Methods

2.1 The Generalized Model

Consider a sequence of points $\boldsymbol{v} : T \to \mathbb{R}^n$, where $T \subset \mathbb{R}$ is finite. In the present setup, assume that $\boldsymbol{v}(t)$ consists of piecewise-connected linear segments plus noise. Denote the smaller set of join points (connecting the segments) located at $\widetilde{T} \subset T$ by $\boldsymbol{\nu}\colon \widetilde{T} \to \mathbb{R}^n$. These points are connected contiguously through a linear-interpolation function $t \mapsto \boldsymbol{l}_{\widetilde{P}}(t)$ that expands the domain of $\boldsymbol{\nu}$ onto T and takes as parameter the generated set $\widetilde{P} = \left\{(\tau, \boldsymbol{\nu}(\tau))\middle| \tau \in \widetilde{T}\right\}$. Hence $\boldsymbol{v}(t)$ is expressed as the sum of $\boldsymbol{l}_{\widetilde{P}}(t)$ and some noise process $\boldsymbol{e}(t)$. A possible definition for $\boldsymbol{l}_{\widetilde{P}}$ is as follows:

$$\boldsymbol{l}_{\widetilde{P}}(t) = \frac{(t - x_l)\boldsymbol{y}_r + (x_r - t)\boldsymbol{y}_l}{x_r - x_l}, \qquad \begin{aligned} (x_l, \boldsymbol{y}_l) &= \operatorname*{arg\,max}_{(x,\boldsymbol{y})\in\widetilde{P}} \{x \leq t\} \\ (x_r, \boldsymbol{y}_r) &= \operatorname*{arg\,min}_{(x,\boldsymbol{y})\in\widetilde{P}} \{x > t\} \end{aligned} \tag{1}$$

The problem of simultaneously recovering both $\widetilde{T}$ and $\boldsymbol{\nu}(\tau)$ is nontrivial. Even estimating $k = |\widetilde{T}|$ has been approached via a number of statistical methods, each with their own tradeoffs and compromises. This study will attempt a bottom-up formulation of an algorithm to efficiently and robustly estimate $\left(\widetilde{T}, \boldsymbol{\nu}\right)$. First, $\boldsymbol{\nu}(\tau)$ will be calculated on the assumption of a fixed known estimate of $\widetilde{T}$. Then, a greedy reduction algorithm will be introduced to find $\widetilde{T}$ given k; finally, methods for choosing k will be discussed. An iterative composition of the solutions to these three optimization problems will create a technique for estimating all the unknown parameters in a single shot. From now on, estimates for the join points are denoted with the hat symbol (e.g. $\widehat{T}$). The overall objective is to study $\widehat{\boldsymbol{v}}(t) = \boldsymbol{l}_{\widehat{P}}(t) + \widehat{\boldsymbol{e}}(t)$.

2.2 Estimating $\widehat{\boldsymbol{\nu}} = f(\boldsymbol{v}(t), T; \widehat{T})$

Ideally, one would wish to maximize the whiteness of the noise $\widehat{\boldsymbol{e}}(t)$. But to make the problem more tractable, one often resorts to minimizing the error's Euclidean norm over all the points t: $\hat{e}^2 = \sum_{t\in T} \|\widehat{\boldsymbol{e}}(t)\|^2$. Without loss of generality, each dimension of $\widehat{\boldsymbol{e}}(t)$ can be separated into its own independent subproblem, and thus the optimization task reduces to finding the set of scalars

$\widehat{\nu}_i(\tau)$ for $\tau \in \widehat{T}$, within each dimension i such that $1 \leq i \leq n$.

The following reduced-dimension version of the error is to be minimized:

$$\hat{e}^2 = \sum_{t \in T} \left[v(t) - l_{\widehat{P}}(t) \right]^2 . \tag{2}$$

Recall that T and its corresponding $v(t)$ give the complete data set. By setting the derivative of $\hat{e}^2$ with respect to each $\widehat{\nu}(\tau \in \widehat{T}\,)$ to zero, utilizing Equations (1), one obtains a local solution where each point is defined in terms of the fixed variables as well as its neighbors, i.e. $[\hat{x}_l, \hat{y}_l, \hat{x}_r, \hat{y}_r]$, borrowing the notation employed to define $\boldsymbol{l}_{\widehat{P}}$ above.

Let $\omega(\tau; P)$ denote the local solution for $\widehat{\nu}(\tau)$ given a set of prior approximation points P, and recall that it depends only on the two immediate neighbors of τ in P. Thus a global approximation is obtained via repeated local optimizations, each time using the previous iteration's results as priors. The first iteration is a guess—the better the guess, the faster the convergence to an optimal solution. To illustrate the algorithm, successive iterations are represented by ${}^{(j)}\widehat{\nu}(\tau)$:

$$\begin{aligned}
{}^{(1)}\widehat{\nu}(\tau) &= v(\tau), \\
{}^{(2)}\widehat{\nu}(\tau) &= \omega(\tau; \widehat{P}_1), \qquad & \widehat{P}_1 &= \left\{ \left(x, {}^{(1)}\widehat{\nu}(x)\right) \middle| x \in \widehat{T} \right\}, \\
{}^{(3)}\widehat{\nu}(\tau) &= \omega(\tau; \widehat{P}_2), \qquad & \widehat{P}_2 &= \left\{ \left(x, {}^{(2)}\widehat{\nu}(x)\right) \middle| x \in \widehat{T} \right\}, \\
{}^{(4)}\widehat{\nu}(\tau) &= \omega(\tau; \widehat{P}_3), \qquad & \widehat{P}_3 &= \left\{ \left(x, {}^{(3)}\widehat{\nu}(x)\right) \middle| x \in \widehat{T} \right\}, \\
&\vdots
\end{aligned} \tag{3}$$

Empirical data on the rate of convergence demonstrate that only a few iterations are needed to reach a plateau when the initial guess is $v(\tau)$. If the guess is vastly off, then the first iteration of the approximator tends to overshoot in the other direction (to compensate for each point's erroneous neighbors); hence, the second iteration overshoots again in the opposite direction. This resembles a damped oscillator.

2.3 Estimating $\widehat{T} = g_f(\boldsymbol{v}(t), T; \widehat{k}\,)$

Equipped with an estimator for $\boldsymbol{\nu}(\tau)$, the investigator's next step is to optimize $\widehat{T} \ni \tau$. Would it be possible to follow an approach similar to the one employed above in finding $\widehat{\nu}(\tau)$? Such a technique would involve minimizing an analogue to e^2, but with respect to τ within T. An alteration of a τ is an alteration on the domains of interpolation; thus, it is not possible to arrive at a closed-form solution for minimizing the error as defined in Equation (2). It could potentially be done iteratively, but an exhaustive search (deterministic or randomized) has too big of a state space to be done efficiently.

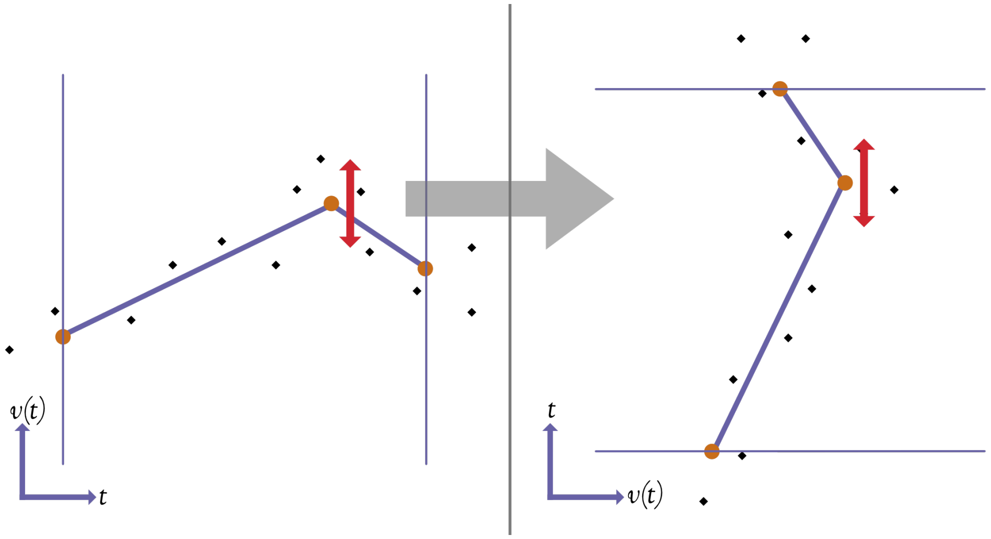

Figure 1: Rotating the local setup by 90° to treat τ as $\widehat{\nu}(\tau)$ and vice versa. Orange circles are $\widehat{\nu}$ and black diamonds are v, with the connecting lines showing the interpolation $l_{\widehat{P}}$.

2.3.1 The prospect of reusing the $\widehat{\nu}(\tau)$ optimizer

Perhaps one could restate the problem of optimizing $\widehat{T}$ by flipping the axes on a graph of $\widehat{\nu}(\tau)$ versus $\tau \in \widehat{T}$, as depicted in Figure 1, to reuse the vertical optimizer ω. One issue is that since $l_{\widehat{P}}(t)$ is not bijective, its inverse is not well-defined. Hence we must augment it with a label signaling the relevant line segment; for instance, ${}^{\text{up}}l_{\widehat{P}}^{-1}(\nu, \tau)$ would indicate the t where the segment above τ takes the value of ν. The variable τ here is used as an index for join points—note that $\forall(\tau \in \widehat{T})\exists y[(\tau, y) \in \widehat{P}]$, so the difference between $<$ and $\leq$ matters when selecting the boundaries in Equation 4.

$$
{}^{\text{up/down}}l_{\widehat{P}}^{-1}(\nu, \tau) = \frac{(\nu - y_d)x_u + (y_u - \nu)x_d}{y_u - y_d}, \quad \begin{array}{ll} \text{up} & \begin{cases} (x_d, y_d) = \arg\max_{(x,y)\in\widehat{P}} \{x \leq \tau\} \\ (x_u, y_u) = \arg\min_{(x,y)\in\widehat{P}} \{x > \tau\} \end{cases} \\ \text{down} & \begin{cases} (x_d, y_d) = \arg\max_{(x,y)\in\widehat{P}} \{x < \tau\} \\ (x_u, y_u) = \arg\min_{(x,y)\in\widehat{P}} \{x \geq \tau\} \end{cases} \end{array} \tag{4}
$$

For implementation purposes, it is worth remarking that swapping the pairs (x_d, y_d) and (x_u, y_u) in Equation 4 (as well as (x_l, y_l) and (x_r, y_r) in Equation 1) ends up negating both the numerator and the denominator and so the results are the same as before the swap—thus proper ordering can be safely neglected.

These modifications entail a less elegant definition for the error to be minimized: namely,

$$\xi^2 = \sum_{\tau\in\widehat{T}} \sum_{t\in\widehat{U}(\tau)} \left[t - {}^{\mathrm{up}}l_{\widehat{P}}^{-1}(v(t),\tau)\right]^2, \quad \widehat{U}(\tau) = \left\{ t\in T \,\middle|\, \tau \le t < \min_{\tau'\in\widehat{T}}[\tau' > \tau] \right\}. \tag{5}$$

In essence, Equation 5 loops over $\tau \in \widehat{T}$ and sums up the t-errors within each T-interval ranging from the current τ to the next τ. The present task is to find each τ that solves $\frac{\partial\xi^2}{\partial\tau} = 0$, which simplifies to resemble the solution to Equation 2. This method is similar in spirit but not in implementation to [16]; as a consequence, it suffers from the same dependence on the choice of initial values.

The next section proceeds to explore a technique that renders the τ-optimizer optional. It is still valuable to keep the above analysis in mind, since it may prove useful in the future. Currently, it is not; though it could be used for post-refinement, this τ-optimizer is a bit unwieldy and with its practical necessity reduced to marginal, it will not be examined in the Results.

2.4 Estimating $\widehat{k} = h_{f,g}(\boldsymbol{v}(t), T)$ via Greedy Reduction

A possible estimation method for $\widehat{k} = |\widehat{T}|$ would be to start with an initial guess that vastly overestimates the number of join points, then to gradually reduce them by discarding the least significant/effectual point $(\tau, \widehat{\boldsymbol{\nu}}(\tau))$ every iteration. After each removal, the algorithm would have to readjust the relevant $\widehat{\boldsymbol{\nu}}(\tau)$-values. If $\mu(\tau; P)$ is allowed to denote the significance metric, then the algorithm would roughly proceed as follows:

$$\begin{aligned}
&\widehat{T}_1 = \left\{ t_{\mathrm{min}} + \frac{i-1}{\widehat{k}_{\mathrm{guess}} - 1}(t_{\mathrm{max}} - t_{\mathrm{min}}) \,\middle|\, i = 1, \ldots, \widehat{k}_{\mathrm{guess}} \right\}, && \\
&\widehat{T}_2 = \widehat{T}_1 - \left\{ \operatorname*{arg\,min}_{\tau\in\widehat{T}_1} \mu(\tau; \widehat{P}_1) \right\}, && \widehat{P}_1 = \left\{ \left(\tau, {}^{(1)}\widehat{\nu}(\tau)\right) \,\middle|\, \tau \in \widehat{T}_1 \right\}, \\
& && \widehat{P}_2 = \left\{ \left(\tau, {}^{(2)}\widehat{\nu}(\tau)\right) \,\middle|\, \tau \in \widehat{T}_2 \right\}, \\
&\widehat{T}_3 = \widehat{T}_2 - \left\{ \operatorname*{arg\,min}_{\tau\in\widehat{T}_2} \mu(\tau; \widehat{P}_2) \right\}, && \widehat{P}_3 = \left\{ \left(\tau, {}^{(3)}\widehat{\nu}(\tau)\right) \,\middle|\, \tau \in \widehat{T}_3 \right\}, \\
&\quad\vdots &&
\end{aligned} \tag{6}$$

Equations 6 borrow the generation of $\left\{ {}^{(j)}\widehat{\nu}(\tau) \,\middle|\, j = 1, 2, \ldots \right\}$ via $\omega(\tau; P)$ as it is used in Equations 3. Therefore it is clear that $|\widehat{T}_l| \le \widehat{k}_{\mathrm{guess}} - l + 1$. Now two things are left; first, one must define an adequate $\mu(\tau; P)$. Second, one needs an effective stopping criterion for $|\widehat{T}_l|$. The issue of knowing when to stop is more important here than it was in Equations 3 because unlike before, it is not safe to over-iterate when $\widehat{T}_l$ shrinks monotonically.

2.4.1 The stopping criterion for $|\widehat{T}_l|$

The Bayesian information criterion (BIC) has been lauded for striking a balance between minimizing the chance of overfitting and maximizing the expressiveness of a model [13]. Its use involves the minimization of a metric that considers the goodness of a fit and penalizes against the number of free parameters in that fit—the BIC is commonly written as follows:

$$B = \ln(n_B)k_B - 2\ln(L), \tag{7}$$

where L is the likelihood of the given model with (near-)optimal parameters, n_B is the sample size, and k_B is the number of free parameters. The subscript "$_B$" is used to differentiate between the notation used in the previous sections and that introduced in Equation 7. Hence $k_B = 2|\widehat{T}_l|$ and $n_B = |T|$. To estimate $\ln(L)$, termed the maximum log-likelihood, it is necessary to make a few assumptions about the distribution of the residuals. First, one assumes that they are independently and identically distributed (i.i.d). Since the derived optimizers operate under the assumption that it is best to minimize the sum of squared errors, one could go even further and declare that the residuals are Gaussian (which is a logical consequence to the assumption that likelihood is maximized via a least-squares norm). Now the log-likelihood is transformed as

$$2\ln(L) + C = -n_B \cdot \ln\sum_{i=1}^{n_B}(x_i - \widehat{x}_i)^2 = -n_B \ln e^2_{\widehat{\theta}} = -|T|\ln\sum_{t\in T}\Big[v(t) - l_{\widehat{P}}(t)\Big]^2. \tag{8}$$

Here $e^2_{\widehat{\theta}}$ denotes the minimized square error for a given parameter space $\Theta \ni \widehat{\theta}$, and the "$+C$" notifies of the existence of some additive constant that can be ignored, since it does not depend on the model. The full equation for a Gaussian BIC, B^G, can be written like so:

$$B^G_l = 2|\widehat{T}_l|\ln|T| + |T|\ln\sum_{t\in T}\Big[v(t) - l_{\widehat{P}_l}(t)\Big]^2, \tag{9}$$

and one simply needs to find $l_{\text{opt}} = \arg\min_{l=1,2,\ldots}(B_l)$, where B_l is the appropriate choice of BIC for the given situation.

2.4.2 The choice of significance metric $\mu(\tau; P)$

In Section 2.4.1, one is forced to impose a hypothetical distribution on the residuals. Locally, the join-point optimizer minimizes the sum of squared errors for each $\widehat{\nu}(\tau)$ calculation; globally, however, the investigator is free to choose any norm through picking the residual distribution to be assumed for the log-likelihood estimation. Intuition calls for a join-point significance metric consistent with the chosen BIC, akin to the change in global log-likelihood that would result from discarding the join point under question. The assumed residual distribution is parametrized globally. Denote this approximated change by

$$\widehat{\Delta \ln L}(\tau; P) = \widehat{\ln L}(P') - \ln L(P), \text{ where } P' = P - \{(\tau, y) | y \in \mathbb{R}\}\,. \tag{10}$$

Note that $\ln L(P)$ does not depend on τ, so in effect it can be ignored when comparing the approximate changes $\widehat{\Delta \ln L}(\tau; \widehat{P})$ across $\tau \in \widehat{T}$. In the Gaussian case, monotonicity of the logarithm can be exploited to simplify the computation of $\widehat{\Delta \ln L}$, sacrificing commensurability across models for efficiency:

$$\widehat{\Delta \ln L}^{G}(\tau; P) = -\sum_{t \in \Lambda} \left[v(t) - l_{P - \{(\tau, y)|y \in \mathbb{R}\}}(t) \right]^2 , \tag{11}$$

where $\Lambda(\tau; P)$ is the “local T”: it gives the values of t embedded in P that are within the neighboring join points of τ. This measure only bears meaning when comparing different join points within a specific model, and it is only approximate because the join points are not adjusted after the pruning of τ. Nonetheless, it is adequate for an efficiently computable significance metric

$$\mu(\tau; P) = -\widehat{\Delta \ln L}(\tau; P). \tag{12}$$

Remark. In most cases, it is expected that even $\min_{\tau \in \widehat{T}} \mu(\tau, \widehat{P})$ would be non-negative, since the removal of free parameters tends to increase the error norm. This intuition cannot be certain because the discrepancy between the local and global norms in use may drive the minimum μ below zero.

2.5 Towards Robustness

One could re-derive the BIC with respect to a cost function that penalizes outliers less; namely, the L_1-norm that is precisely used in situations that demand more robustness. Minimizing the absolute value instead of the square of the errors corresponds to fitting a model with residuals that belong to a Laplace distribution. Dropping the Gaussian assumption and replacing it with the Laplacian, the new log likelihood becomes

$$\ln L^R = -|T| \ln \sum_{t \in T} \left| v(t) - l_{\widehat{P}}(t) \right| + C_{|T|}, \tag{13}$$

where $C_{|T|}$ is the constant that depends only on $|T|$ and not the model. Hence the new BIC, termed the robust BIC, is

$$B_l^R = |\widehat{T}_l| \ln |T| + |T| \ln \sum_{t \in T} \left| v(t) - l_{\widehat{P}_l}(t) \right|. \tag{14}$$

The optimizer still operates on the Euclidean norm, however. And since a closed-form local solution for $\widehat{\nu}(\tau)$ is only known in the context of minimizing the sum of squared errors, that part will have to be kept. But the significance metric $\mu(\tau; P)$, which utilizes the global log-likelihood estimate, can be modified from Equation 11 for coherence with the Laplace assumption:

$$\widehat{\Delta \ln L}^{R}(\tau; P) = -\sum_{t \in \Lambda} \left| v(t) - l_{P-\{(\tau,y)|y\in\mathbb{R}\}}(t) \right|. \tag{15}$$

Now the trends are locally fit via minimizing the L_2-norm, but the reduction procedure seeks to minimize the global L_1-norm.

Remark. Even if it were feasible to solve analytically for a local optimum with respect to the L_1-norm, such a result would be undesirable. One can demonstrate visually that this norm is invariant to certain translations of the line of best fit; the solution is rarely unique.

2.6 The Power Law

In economics there may arise situations in which the L_1-norm is not robust enough. In these cases, one has to resort to residual distributions with even heavier tails, e.g. some materialization of a power law. The Pareto family of distributions contains simple and viable candidates. The simplest one that does not exhibit a singularity at $x = 0$ is a special case of the Pareto Type II called the Lomax distribution [17]. Its probability density function (p.d.f) is commonly written as follows.

$$f(x; \alpha, \lambda) = \frac{\alpha}{\lambda \cdot (1 + x/\lambda)^{\alpha+1}}, \tag{16}$$

taking a shape parameter $\alpha > 0$ and a scale parameter $\lambda > 0$. It is necessary to feed in the absolute value of the residuals, since the domain of the p.d.f is $[0, \infty)$. The value of α that maximizes likelihood given a set of data $\boldsymbol{x} = \{x_1, \dots, x_n\}$ and an estimate for λ is given in closed form as

$$\hat{\alpha}(\boldsymbol{x}; \hat{\lambda}) = \frac{n}{\sum_{i=1}^{n} \ln\left(1 + x_i/\hat{\lambda}\right)}; \tag{17}$$

unfortunately, it is impossible to find an analogous closed-form solution for the optimal $\hat{\lambda}$. Henceforth an iterative solution is employed. For the hill-climbing (i.e. gradient ascent), one must compute

$$\frac{\partial \ln L}{\partial \lambda}(\boldsymbol{x}; \hat{\alpha}) = n \cdot \frac{\hat{\alpha}}{\lambda} - \sum_{i=1}^{n} \frac{\hat{\alpha} + 1}{x_i + \lambda} \tag{18}$$

for an initial guess $\hat{\lambda}_0$ with $\hat{\alpha}_0 = \hat{\alpha}(\boldsymbol{x}; \hat{\lambda}_0)$, and use the update rule

$$\hat{\lambda}_{i+1} \leftarrow \hat{\lambda}_i + \gamma \cdot \frac{\partial \ln L}{\partial \lambda}(\boldsymbol{x}; \hat{\alpha}_i)\bigg|_{\lambda=\hat{\lambda}_i} \quad \text{until} \quad \frac{\left|\hat{\lambda}_{i+1} - \hat{\lambda}_i\right|}{\gamma \cdot \hat{\lambda}_{i+1}} \leq \varepsilon. \tag{19}$$

The new BIC is thusly

$$B_l^P = |\widehat{T}_l| \ln |T| - (\hat{\alpha} + 1) \sum_{t \in T} \ln \left(\left| v(t) - l_{\widehat{P}_l}(t) \right| + \hat{\lambda} \right). \tag{20}$$

This criterion, with its associated logarithmic norm, is vastly more lenient towards outliers. To minimize the number of gradient-ascent iterations performed before each BIC evaluation, the final $\hat{\lambda}$ utilized in a B_l^P can be used as the initial guess for the next round, B_{l+1}^P. The distribution parameters $(\hat{\alpha}, \hat{\lambda})$ should only change gradually between successive values $l, (l+1), (l+2) \ldots$ because the pruning algorithm is greedy.

3 Results

3.1 Comparison to Other Methodologies

The existing state of the art largely employs a grid search in locating optimal join-point locations; therefore those methods are incapable of handling such a plethora of join points as are present in the data sets under study. It is infeasible to compare results from previously established methods *except* for Muggeo's [16], which escapes the need for grid search by opting for an iterative procedure (this was briefly discussed in the Methods). Reliance on the choice of initial locations is diminished through bootstrap restarting [18], and during the fitting procedure, justification for one additional join point is affirmed via a statistical test [19]. The method developed by Muggeo will be used as a benchmark upon which to validate the present technique—refer to Figure 4 for the outcome.

3.2 Experimental Setup

Say we have a range $T = \{t \in \mathbb{Z} | t_{\text{min}} \leq t \leq t_{\text{max}}\}$. To generate a simulation, one has to pick $\tau \in \widetilde{T} \subset T$. The simplest way to do it is through uniform subsampling $P[\tau \in \widetilde{T} | \tau \in T] = \frac{k}{|T|}$, while enforcing $|\widetilde{T}| = k$. Once $\widetilde{T}$ is created, it is adequate enough to independently sample $\nu(\tau \in \widetilde{T}) \sim \mathcal{N}(0, \sigma^2)$. Then $v(t \in T)$ can be filled out using the definition given at the beginning of the Methods: $v(t) = l_{\widetilde{P}}(t) + e(t)$. The noise process can be defined in a number of different ways; in this case it was deemed beneficial to make it a "leaky" Gaussian that introduces fake trends (as random walks) that persist for an adjustable amount of time:

$$e(t) = \alpha e(t-1) + g(t; \beta\sigma^2), \tag{21}$$

where $0 \leq \alpha < 1$ and $g(t; s)$ is a realization of purely Gaussian noise centered at 0 with a variance of s, i.e. $\mathcal{N}(0, s)$. This definition of $e(t)$ is analogous to passing pseudo-white noise through a filter characterized by an exponentially decaying impulse response function. The parameters α and β directly control the signal-to-noise ratio of the simulated experiment. Due to the Central Limit Theorem, $e(t)$ is also Gaussian. It can be shown that the variance of the noise asymptotically and monotonically approaches a fixed quantity by expressing the recursive relation $e(t)$ as an infinite sum:

$$\begin{aligned}\sigma_e^2 &= \sum_{i=0}^{N} \mathrm{Var}(\alpha^i G) = \sum_{i=0}^{N} \alpha^{2i} \mathbb{E}[G^2] = \beta\sigma^2 \sum_{i=0}^{N} \alpha^{2i} \\ &= \left(\frac{\beta - \alpha^{2(N+1)}}{1-\alpha^2}\right)\sigma^2 \\ &\xrightarrow[N\to\infty]{} \left(\frac{\beta}{1-\alpha^2}\right)\sigma^2, \qquad G \sim \mathcal{N}(0, \beta\sigma^2), \quad 0 \leq \alpha < 1.\end{aligned} \tag{22}$$

3.3 Choice of Parameters

The most important parameters to select are the density/frequency of join points $d = k/|T|$ and the signal-to-noise ratio $r = \sigma_e/\sigma = \sqrt{\beta/(1-\alpha^2)}$ from Equation 22. Of course, for a fixed r, there are infinite combinations of values for (α, β) that in a sense control the degree of whiteness in $e(t)$. The special case of $(0, r^2)$ turns the noise completely pseudo-white by eliminating its memory. On the other hand, an assignment of $(\sqrt{1-\varepsilon}, \varepsilon r^2)$ such that $\alpha \to 1$ and $\beta \to 0$ makes the noise closer to Brownian.

For a large enough $|T|$, the exact values of the numerator and denominator in d should not wield a noticeable influence on the results. Therefore, all the simulations use a large T, and for the purposes of this experiment the sample points are evenly spaced. So $T = \{1 \dots 500\}$ with an adjustable $k = |\widetilde{T}|$. Likewise reasoning applies for r and σ^2, so σ^2 is set to 1 and (r, α) are kept flexible. β is then computed by rearranging Equation 22.

3.4 Model Evaluation

Each simulation may have a slightly different manifestation of noise, and to correct for that, it is useful to calculate the ratio of the estimation error $\hat{e}^2$ to the inherent error e^2. Also for a given join-point density d, the fraction $\widehat{k}/k$ produces a more meaningful number than the plain $\widehat{k}$ because it is commensurable across varying setups.

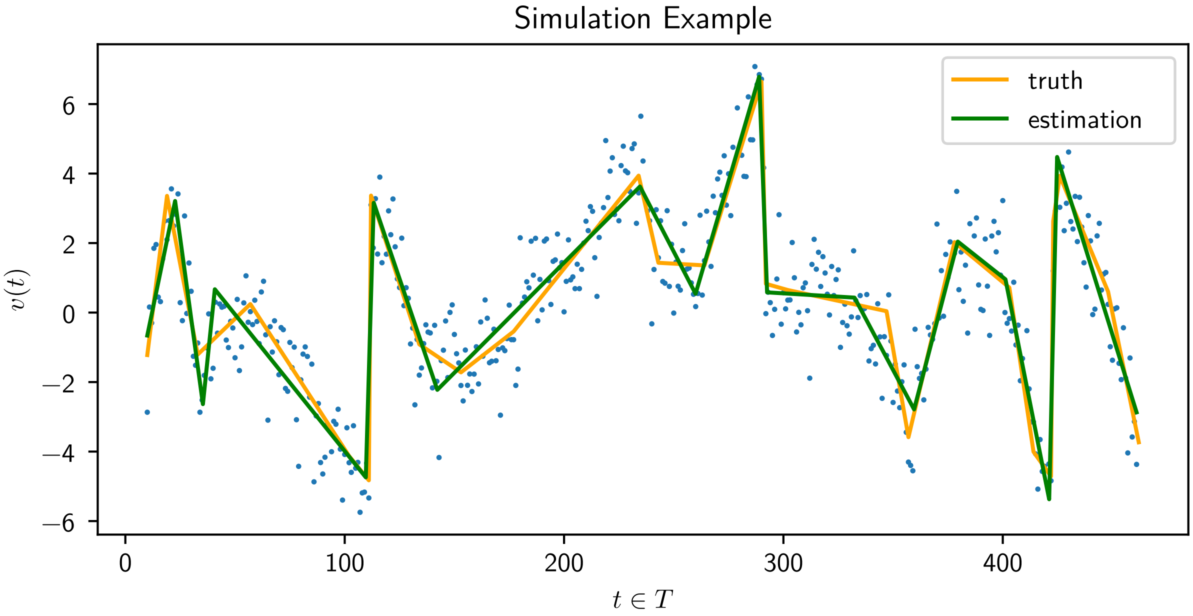

Figure 2: Example of a single trial simulation run.

A third metric would be to measure the average distance in T from each estimated join point to the nearest true join point, and vice versa. In other words, this distance function $\gamma^2(\widehat{P}, \widetilde{P})$ would operate as follows:

$$\begin{aligned} \gamma^2(P_1, P_2) = \frac{1}{2l^2}\left[\frac{1}{|X_1|}\sum_{x_1\in X_1}\min_{x_2\in X_2}\left[x_1 - x_2\right]^2 + \frac{1}{|X_2|}\sum_{x_2\in X_2}\min_{x_1\in X_1}\left[x_1 - x_2\right]^2\right], \\ X_1 = \{x|(x,y)\in P_1\},\ X_2 = \{x|(x,y)\in P_2\}, \\ l = (t_{\max} - t_{\min})/(|X_1| + |X_2|), \end{aligned} \tag{23}$$

where the l^2 factor normalizes against a soft worst-case length between join points from P_1 and P_2 (which occurs when they are evenly spaced). Note that this serves as a scaling parameter, and $\gamma^2(\widehat{P}, \widetilde{P})$ can still be greater than 1 under particularly noisy circumstances. All three of the aforementioned evaluation metrics will illuminate on the effectiveness of the study's approach in different scenarios.

3.5 Simulated Results

Experiments on key parameter combinations were performed and summarized results displayed in Figure 3. For each set of parameters, 200 trials were run and statistics on the evaluation metrics were

Inputs			Outputs		
$1/d$	r	α	$\hat{e}^2/e^2$	$\gamma^2(\widehat{P}, \widetilde{P})$	$\widehat{k}/k$
100	$1/3$	0	1.01 ± 0.04	0.25 ± 0.32	0.89 ± 0.20
50	$1/3$	0	1.04 ± 0.05	0.39 ± 0.40	0.81 ± 0.17
10	$1/3$	0	1.19 ± 0.12	0.75 ± 0.38	0.57 ± 0.09
$[10, 100]$	$2/3$	0	1.11 ± 0.12	1.53 ± 2.18	0.48 ± 0.18
$[10, 100]$	1	0	1.13 ± 0.12	4.04 ± 7.70	0.35 ± 0.20
$[10, 100]$	2	0	1.09 ± 0.06	12.40 ± 10.48	0.17 ± 0.14
$[10, 100]$	$1/3$	$1/4$	1.02 ± 0.09	0.57 ± 0.40	0.74 ± 0.19
$[10, 100]$	$1/3$	$2/4$	0.84 ± 0.12	0.68 ± 0.52	0.88 ± 0.28
$[10, 100]$	$1/3$	$3/4$	0.44 ± 0.12	1.79 ± 1.51	1.51 ± 0.98

Figure 3: Results of experiments with different parameter arrangements. Each row is averaged over 200 trials. A parameter supplied as a range means that it was uniformly sampled from that interval.

recorded. Each trial generated a fit like the one shown in Figure 2. For performance reasons, $\widehat{k}_{\text{guess}}$ is set to 250, equivalent to a starting $1/\widehat{d}$ of 2. When an entry in the table contains a $1/d$ value of $[10, 100]$, it means that it was independently and uniformly sampled from that interval in each trial.

The first column in the Outputs of Figure 3 is the error ratio $\hat{e}^2/e^2$. Normally, this ratio tends to be greater than 1, with values closer to 1 signifying a better estimation. It is intriguing that under high levels of noise, which were tested in the bottom rows of the table, the error ratio sometimes goes below 1. A high α especially decreases the ratio significantly. This discovery can be rationalized with the hypothesis that "trendy" (i.e. semi-Brownian) noise makes it possible to fit a set of lines with lower overall error than the original construction, by incorporating those phantom trends in the fit.

The second column is the distance measure $\gamma^2(\widehat{P}, \widetilde{P})$. Lower is always better with this metric. It is clear that as the noise is amplified, γ^2 increases sharply. Also the standard deviation becomes very large—this suggests that a minority of fits have extremely large γ^2 values, since it is unlikely to encounter a γ^2 close to 0 under so much noise. Since the ultimate goal is to recover the join-point locations, this metric more directly measures the quality of the estimation.

The third column, $\widehat{k}/k$ or the "length ratio", is also an indicator of estimation quality. In this case, the closer to 1 the better, regardless of which side the metric falls on. The majority of the results in Figure 3 do show a length ratio less than 1. Perhaps it is beneficial that the BIC tends to err on the side of underestimation for the number of lines? It appears that a quality model naturally underestimates k, and Brownian noise (with high α) introduces more false trends. The takeaway here is that one must be careful of noise that is not identically and *independently* distributed.

Figure 4 exhibits results from identical fabricated data sets being fed into both the proposed technique (termed Marmarelis) and Muggeo's. Like in Figure 3, each row displays the successful outcomes of 200 attempted trials. Muggeo's method as provided in his public R package `segmented`

Inputs		Method	Outputs		
r	α		$\hat{e}^2/e^2$	$\gamma^2(\widehat{P},\widetilde{P})$	Success Rate
1/3	0	Marmarelis	1.01 ± 0.05	0.47 ± 0.42	
		Muggeo	1.01 ± 0.13	0.30 ± 0.44	60.5%
2/3	0	Marmarelis	1.00 ± 0.04	0.69 ± 0.46	
		Muggeo	0.98 ± 0.03	0.46 ± 0.46	51.0%
1	0	Marmarelis	0.99 ± 0.03	0.72 ± 0.53	
		Muggeo	0.97 ± 0.02	0.54 ± 0.63	45.5%
1/3	1/4	Marmarelis	1.01 ± 0.06	0.50 ± 0.41	
		Muggeo	1.00 ± 0.15	0.38 ± 0.47	66.0%
1/3	2/4	Marmarelis	0.98 ± 0.08	0.46 ± 0.35	
		Muggeo	0.94 ± 0.09	0.41 ± 0.46	63.0%
1/3	3/4	Marmarelis	0.86 ± 0.12	0.58 ± 0.49	
		Muggeo	0.84 ± 0.24	0.63 ± 0.76	74.0%

Figure 4: Results of experiments comparing the proposed method to Muggeo's.

throws an exception if the join points in the data are particularly elusive. In practice this is avoided by first testing for the strong existence of said join points before fitting them, but in this case that would severely underestimate k. In order to "level the playing field," each method is given the true k (ranging from 5 to 10 with $|T| = 500$) in Figure 4. *Marmarelis* saturates the model with join points and then reduces to exactly k, bypassing the need for an information criterion; *Muggeo* evenly lays out k join points and then refines their positions incrementally. Higher k-values significantly reduce the success rate of Muggeo's method. Notwithstanding, the respective fits produced by "Marmarelis" are similar but of slightly lesser quality than "Muggeo" in each experimental instance. Evidently the value in pruning lies not in its stellar accuracy for a targeted k but in its effectiveness when iterating through possible k-values without starting over each time.

4 Discussion

4.1 Experiments on Real Data

It is time to test the method on real macroeconomic data. There is only one free parameter now: the starting number of join points $\widehat{k}_{\text{guess}}$. This has to be as high as possible, computational resources permitting. The implemented algorithm is fast enough to run on the below time series with $\widehat{k}_{\text{guess}} = |T|$, i.e. practically the maximum number of starting points possible. In fact, the entire process of estimating a fit for each possible $\widehat{k}$ and selecting the best one runs in $\mathcal{O}(|T|^2)$ time.

A time series of the United States Dollar trade-weighted index from 1980 to 2017 sampled weekly was retrieved from FRED, which is operated by the Federal Reserve Bank of St. Louis. A

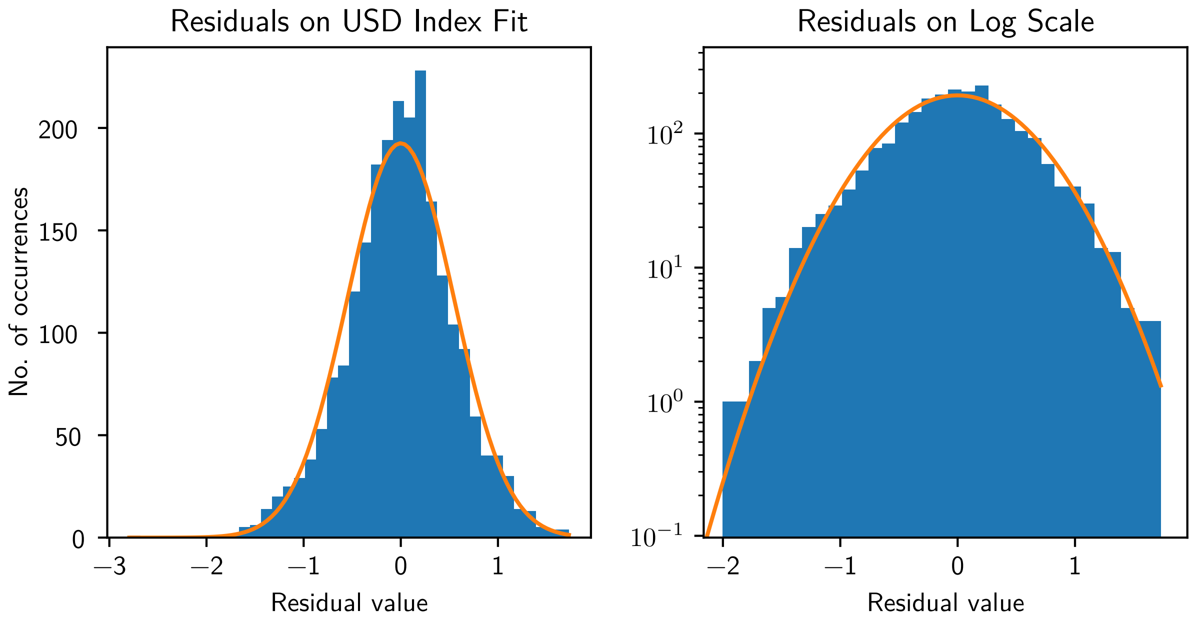

Figure 5: A Gaussian fit on the trade-weighted United States Dollar index, with a resulting join-point reduction from 2000 to 218. Mean log-likelihood is -0.82.

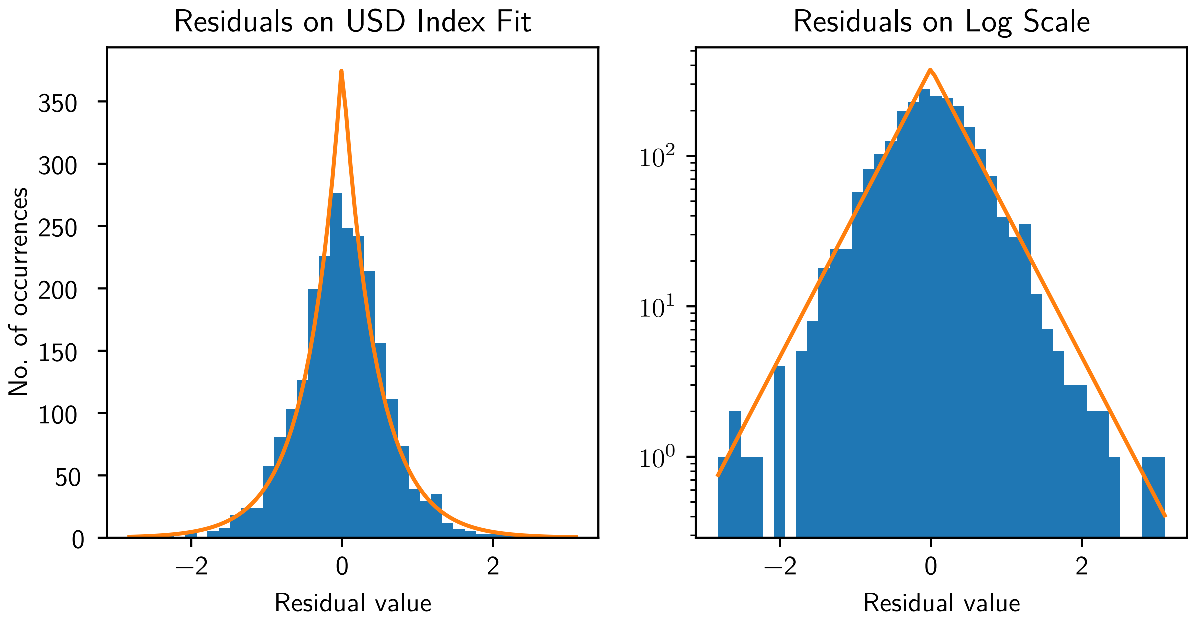

Figure 6: A Laplace fit on the trade-weighted United States Dollar index, with a resulting join-point reduction from 2000 to 214. Mean log-likelihood is -0.90—worse than Figure 5.

model with Gaussian residuals was fit and the histogram of the residuals plotted in Figure 5. Due to

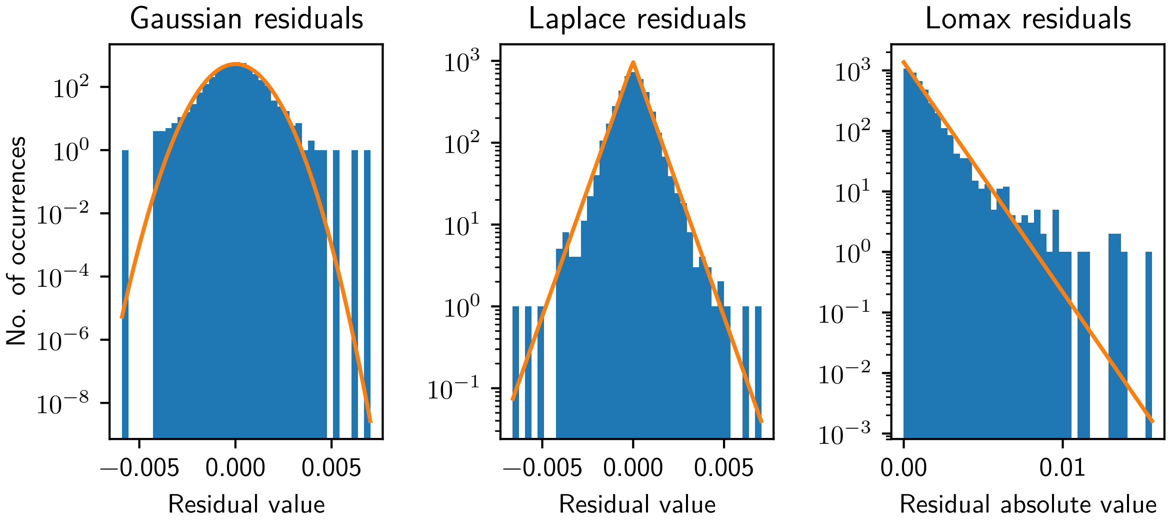

Figure 7: EUR/USD mean log-likelihoods from left to right: $5.52, 5.58, 5.77$. The final Lomax parameters are $\hat{\alpha} = 63137, \hat{\lambda} = 72.2$.

the Gaussian assumption inherent in the implemented BIC, a normal distribution with the sample mean and variance is overlaid. It is hoped that the histogram does not deviate too much from this hypothetical distribution; in particular, one should look out for inflated tails (which flag the presence of unexpected outliers). To easily examine the tails, the same chart is put on a logarithmic scale in the right panel of Figure 5.

The USD trade-weighted index is exemplar of a macroeconomic time series with Gaussian fluctuations. The mean log-likelihood, a goodness-of-fit measure that only carries meaning within a specific data set and that is defined as

$$\overline{\ln L} = \ln\left[\prod_{i=1}^{n} f(x_i)\right]^{\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^{n} \ln f(x_i), \tag{24}$$

is -0.82 for Gaussian residuals (see Figure 5) and -0.90 for Laplace-distributed (see Figure 6). Therefore the Gaussian assumption leads to a fit of higher quality. As a side note, Mandelbrot's self-similarity principle is violated in this time series because it lacks the necessary power-law fluctuations. It is thus reasonable to conclude that a different sampling frequency could produce a different type of residual distribution.

A higher-resolution financial time series such as the hourly EUR/USD exchange rate (4,000 points from May 2016 to January 2017 obtained from http://fxhistoricaldata.com/) could turn non-Gaussian. The noisiness of such data precludes the need to consider distributions on the other end of the spectrum, those that are lighter-tailed than normal. So a Laplace distribution is first evaluated as an alternative to Gaussian, and then a Lomax distribution is tried. Figure 7 shows that the Lomax version produces a mean log-likelihood of 5.77 versus the 5.58 and 5.52 for Laplace and Gauss, respectively. However the final estimated distributional parameters are not satisfactory. They are too large for numerical stability and come to approximate a Laplace distribution by producing a straight line in the log scale. Even though it has the best likelihood, the Lomax version is less desirable than the Laplace version in this particular instance.

The above process was repeated for an hourly sampling (4,181 points acquired from Bitfinex) of recent Bitcoin (BTC/USD) prices, selected because this exchange is believed to be a more speculative market than the EUR/USD foreign exchange. This time the results support a Lomax model, as witnessed in Figure 8. Maybe the cause of power-law price fluctuations is speculation. See Figures 9, 10, and 11 for supplementary information on the Bitcoin fits.

5 Conclusion

5.1 Interpretation of Findings

In the context of economics, the statistics of both residuals and linear trends are a way to characterize the volatility of a market. The patterns may reveal true market behavior or they may not; one must be careful before arriving at drastic conclusions because even a martingale develops trends that could be mistaken for temporal dynamics. Nevertheless, it is intriguing that the three time series

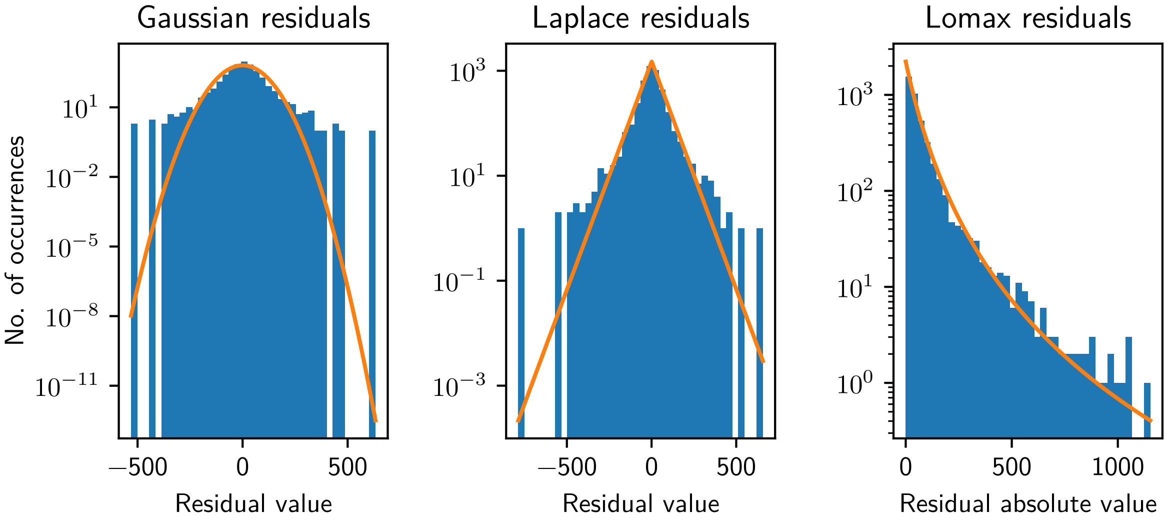

Figure 8: Bitcoin mean log-likelihoods from left to right: $-5.74, -5.60, -5.30$. The final Lomax parameters are $\hat{\alpha} = 3.27, \hat{\lambda} = 177$.

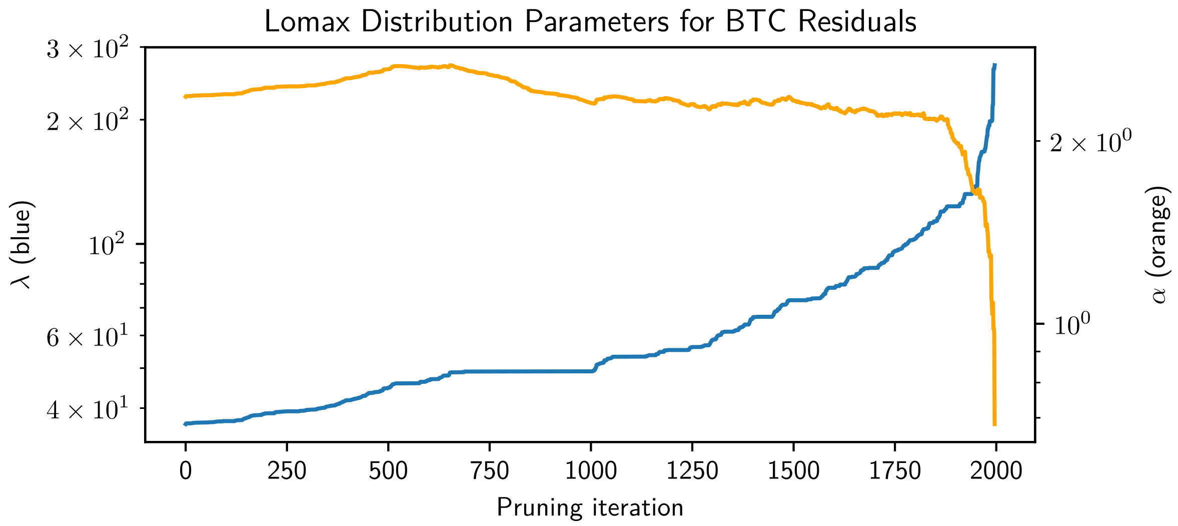

Figure 9: Estimated distributional parameters for the Lomax model as they evolve through the reduction procedure. The final values are at the 1,803rd iteration, corresponding to $2000 - 1803 = 197$ join points.

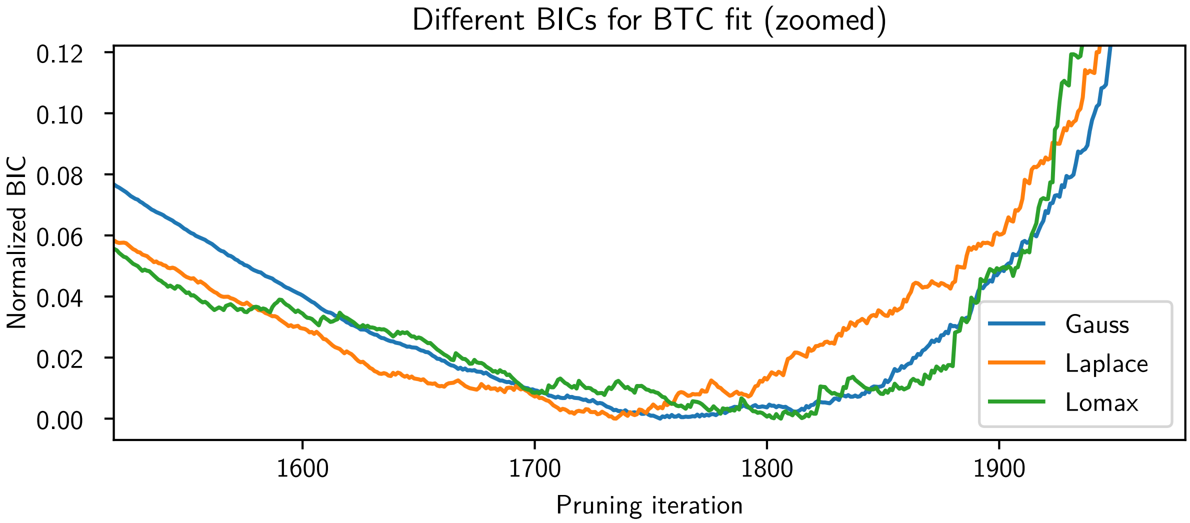

Figure 10: BICs, scaled to $[0, 1]$, for the three different models under study. Their minima stipulate how many join points will be left in the final model.

investigated in this study had residuals best approximated as Gaussian, Laplace, then Lomax, in the order as they were presented.

The extent to which certain markets fluctuate by the power law and others do not, and how the latter change by the selected time scale, is a topic that deserves further study. Before these studies can be allowed to come to fruition, a universally acceptable model must be adopted to separate the

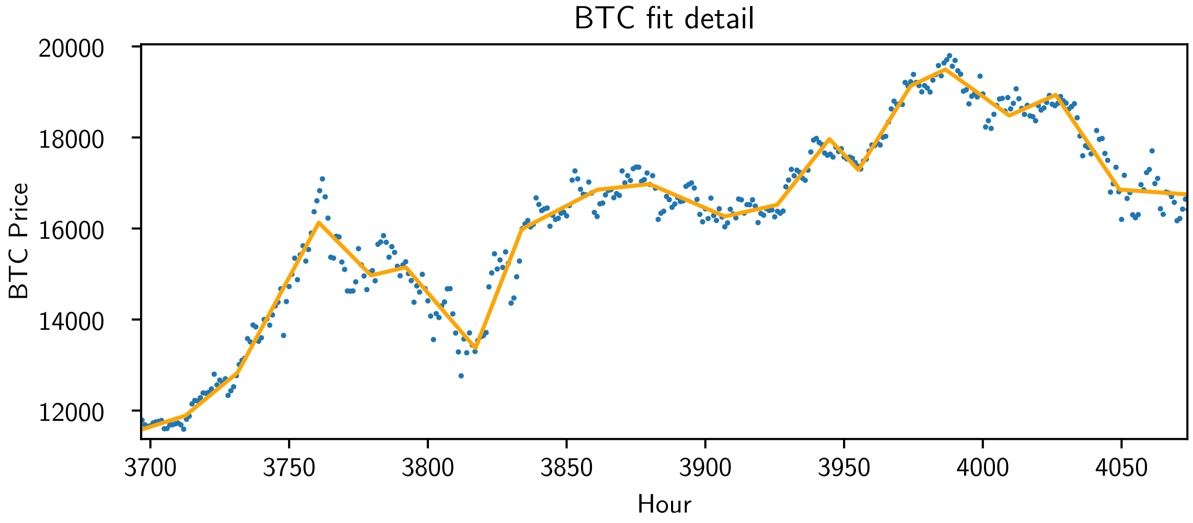

Figure 11: A snippet of the Bitcoin fit with Lomax-distributed residuals.

long-term trends from movements that are short-lived. In the real world these two phenomena behave differently and they should be treated as such.

5.2 From Retrospection to Prospection

Many would argue that the ultimate purpose of time-series analysis is to develop a capacity for prediction. This paper has explored a technique for extracting sequences of linear trends from time-series data. The next intuitive step would be to attempt to predict the nearest future trend following the supplied observations. One could look at a fit of the present model as a sequence of movements $\Delta X = \{\Delta x_1 \ldots \Delta x_m\}$ and $\Delta Y = \{\Delta y_1 \ldots \Delta y_m\}$ and feed those into some other generative model. The steps are defined as

$$\Delta x_i = \widehat{T}_{i+1} - \widehat{T}_i, \quad \Delta y_i = \widehat{\nu}(\widehat{T}_{i+1}) - \widehat{\nu}(\widehat{T}_i). \tag{25}$$

In many fields, the anticipated change in y is more important than the duration Δx. Thus one could frame the goal as the search for an estimate

$$\widehat{\Delta y}_{m+1} = p(\Delta X, \Delta Y) \tag{26}$$

where $p(\ldots)$ is the predictive function under investigation. This problem is left as the subject of future study. Below is a list of possible directions that have received anecdotal support from this study.

- Take advantage of any significant autocorrelation in the residuals. While this would formally violate the i.i.d assumption, it will provide valuable insight if combined with the next item.
- Look for negative serial correlations in ΔY.
- The first and—more importantly—the last estimated trends have different characteristics from the rest of the fit. Since the time series cuts short before a new trend may be justified by the model, these "boundary trends" probably contain remnants of neighboring trends that would exist if the data set were longer.

References

[1] V. Jandhyala, S. Fotopoulos, I. MacNeill, and P. Liu, "Inference for single and multiple change-points in time series," *Journal of Time Series Analysis*, vol. 34, 2013.

[2] A. Aue and L. Horváth, "Structural breaks in time series," *Journal of Time Series Analysis*, vol. 34, 2013.

[3] J. A. Aston and C. Kirch, "Detecting and estimating changes in dependent functional data," *Journal of Multivariate Analysis*, vol. 109, 2012.

[4] P. Fryzlewicz, "Wild binary segmentation for multiple change-point detection," *The Annals of Statistics*, vol. 42, no. 6, 2014.

[5] H. Dette and D. Wied, "Detecting relecant changes in time series models," *Journal of the Royal Statistical Society B*, vol. 78, 2016.

[6] M. Vogt and H. Dette, "Detecting gradual changes in locally stationary processes," *The Annals of Statistics*, vol. 43, no. 2, 2015.

[7] V. M. R. Muggeo and G. Adelfio, "Efficient change point detection for genomic sequences of continuous measurements," *Bioinformatics*, vol. 27, no. 2, 2011.

[8] H.-J. Kim, M. P. Fay, E. J. Feuer, and D. N. Midthune, "Permutation tests for joinpoint regression with applications to cancer rates," *Statistics in Medicine*, vol. 19, 2000.

[9] H.-J. Kim, B. Yu, and E. J. Feuer, "Selecting the number of change-points in segmented line regression," *Statistica Sinica*, vol. 19, 2009.

[10] J. Kim and H.-J. Kim, "Consistent model selection in segmented line regression," *Journal of Statistical Planning and Inference*, vol. 170, 2016.

[11] M. A. Martinez-Beneito, G. García-Donato, and D. Salmerón, “A bayesian joinpoint regression model with an unknown number of break-points,” *The Annals of Applied Statistics*, vol. 5, no. 3, 2011.

[12] M. A. Prince and S. A. Maisto, “The clinical course of alcohol use disorders: Using joinpoint analysis to aid in interpretation of growth mixture models,” *Drug and Alcohol Dependence*, vol. 133, 2013.

[13] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[14] Y.-C. Yao, “Estimating the number of change-points via shwarz’ criterion,” *Statistics & Probability Letters*, vol. 6, 1988.

[15] B. B. Mandelbrot, “Stochastic volatility, power laws and long memory,” *Quantitative Finance*, 2001.

[16] V. M. R. Muggeo, “Estimating regression models with unknown break-points,” *Statistics in Medicine*, vol. 22, 2003.

[17] K. S. Lomax, “Business failures: Another example of the analysis of failure data,” *Journal of the American Statistical Association*, vol. 49, no. 268, 1954.

[18] S. N. Wood, “Minimizing model fitting objectives that contain spurious local minima by bootstrap restarting,” *Biometrics*, vol. 57, no. 1, 2001.

[19] V. M. R. Muggeo, “Testing with a nuisance parameter present only under the alternative: a score-based approach with application to segmented modelling.,” *Journal of Statistical Computation and Simulation*, vol. 86, 2016.