Analyzing Financial Time Series With Persistent Homology

Elias Boughosn and Quentin Le Roux

Université Côte d'Azur, Sophia-Antipolis, France

Abstract. We aim to analyze the evolution of daily log-returns of four key US stock markets indices – DowJones, Nasdaq, Russell 2000, S&P500 – over the period from 1989 to 2016 using persistent homology, and determine whether increasing volatility risk on financial markets can be predicted. To do so, we reproduce the approach proposed by Marian Gidea and Yuri Katz [1] and discuss the replication and extensible aspect of their process.

Keywords: Persistent Homology · Persistence Landscapes · Time Series Analysis · Econometrics · Financial Early Warning Signs

1 Paper Overview

1.1 Context

Sharp changes in the behavior of financial markets have led to deep social, political and economic upheavals. The destabilization of economies, countries, and people's lives at a systemic level during the past two decades – with crises such as the 2000-2001 Dotcom Bubble, 2008 Financial Crisis, and 2012 Euro Debt Crisis – have highlighted the regulatory need for proactive and predictive policies.

This has mainly transpired in the field of econometrics with, for instance, the development of the conditional-value-at-risk, or ΔCoVaR by Tobias Adrian and Markus Brunnermeier [2]. Though econometrics exert a strong influence on policy-making worldwide, predicting rises in volatility and potentially subsequent crashes remain an arduous, yet unsolved task.

In this context, Marian Gidea and Yuri Katz have proposed using persistence homology to evidence sharp changes in financial markets' behavior [1]. With financial markets being complex systems, Topological Data Analysis (TDA) may help highlight evolving market features and underlying behavior without having to model the intrinsically stochastic processes that are market movements.

1.2 Methodology and Workflow

The paper relies on analyzing time series built from the daily log-returns (preprocessed from daily adjusted closing prices extracted from Yahoo Finance) of 4 key US stock market indexes: S&P500, Dow Jones, NASDAQ, Russell 2000.

Starting with daily adjusted closing prices, the paper uses a multi-step process to transform adjusted closing prices into a topological embedding (persistence landscapes) and a resulting L^p norm time series on which statistical treatment and analysis can be performed.

From price to log-return – The adjusted closing prices from the selected US stock market indexes (a $n \times 4$ time series, with n the number of days covered) is processed into a a $(n-1) \times 4$ matrix of log-returns (See Fig. 1 in Annex for the absolute price and respective log-returns for the four financial indexes).

The logarithm of returns is an ubiquitous quantitative finance transformation procedure of price data. Given a price time series P of length $n \in N_+$ with components $p_i \ \forall i \in \{0, \ldots, n\}$, P can be transformed into a times series R of returns $r_i \ \forall i \in \{1, \ldots, n\}$ such that:

$$\forall i \in \{1, \ldots, n\}, \ r_i = \frac{p_i - p_{i-1}}{p_{i-1}} \tag{1}$$

$$r_i^{log} = log(p_i) - log(p_{i-1}) \tag{2}$$

The benefits of using relative metrics like returns compared to absolute metrics are log-normalization (prices are usually assumed to be log-normally distributed in Finance), approximate raw-log equality ($\forall i \in \{1, \ldots, n\}$, $r_i \approx log(1 + r_i)$), time additivity (Given an ordered sequence of n prices p, the corresponding compounding return r is $1 + r_n = \prod_{i=1}^{n}(1 + r_i)$), and numerical stability (addition is preferred to multiplication to avoid floating point issues).

TDA parameters selection – A few key parameters are pre-selected and will drive the whole TDA process, the sliding/scaling window w and p parameter(s) of the end L^p norm time series. While the paper considers two window sizes $w \in \{50, 100\}$, we will consider window sizes $w \in \{40, 80, 120\}$. We will keep the same parameter p with two L^p norms ($p = 1$ and $p = 2$).

Extraction of time-dependent point cloud data sets – The paper defines a point cloud data set X_i as a $w \times d$ matrix. Each of the d columns of a point cloud corresponds to a w-window-sized slice of one of the 4 stock market log-return time series. When applied to the data, the 4D log-return time-series matrix is split into subsets (point cloud data sets) of size $w \times d$ with a time step of 1, resulting in $n - w$ matrices. As such, the resulting set of $n - w$ point cloud data sets is called the "time-ordered sequence of point clouds."

Measuring topological persistence with persistence homology – Given a range of radii $\epsilon > 0$, a Vietoris-Rips complex $R(X, \epsilon)$ is constructed at each ϵ-resolution for each point cloud $\forall i \in \{1, n-w\}, X_i$ that composes the previously computed time-ordered sequence of point clouds.

Of note, $\forall k \in \{0, 1, 2, \ldots\}$, a k-simplex of vertices $\{X_{i1}, \ldots, X_{ik}\}$ is part of $R(X, \epsilon)$ if and only if the mutual distance between any pair of vertices is less than ϵ such that for some distance d: $d(x_{ij}, x_{il}) < \epsilon$.

These complexes $R(X, \epsilon)$ form a filtration (See Fig. 2 in Annex for a diagram of the process). From each complex, one can compute k-dimensional homology classes, part of the homology group $H_k(R(X, \epsilon))$, such that:

$$\forall \epsilon > 0 \tag{3}$$
$$\text{Given } \epsilon_1 < \epsilon_2 \tag{4}$$
$$R(X, \epsilon_1) \subseteq R(X, \epsilon_2) \tag{5}$$
$$H_k(R(X, \epsilon_1)) \subseteq H_k(R(X, \epsilon_2)) \tag{6}$$
$$\tag{7}$$

For instance, a class in the 1-dimensional homology group $H_1(R(X, \epsilon))$ corresponds to an independent loop in $R(X, \epsilon)$. Each homology class α is characterized by a point $z_\alpha = (b_\alpha, d_\alpha)$ and multiplicity $\mu_\alpha(b_\alpha, d_\alpha)$ where b_α and d_α are respectively called "birth" and "death" values such that:

$$b_\alpha = \epsilon_1 \tag{8}$$
$$d_\alpha = \epsilon_2 \tag{9}$$
$$\epsilon_1 < \epsilon_2 \tag{10}$$
$$\text{s.t. } \alpha \in H_k(R(X, \epsilon_1)) \qquad \text{s.t. } \forall \delta > 0,\ \alpha \notin H_k(R(X, \epsilon_1 - \delta)) \tag{11}$$
$$\alpha \notin H_k(R(X, \epsilon_2)) \qquad \text{s.t. } \forall \text{ small enough } \delta,\ \alpha \in H_k(R(X, \epsilon_2 - \delta)) \tag{12}$$
$$\mu_\alpha(b_\alpha, d_\alpha) \qquad \text{amount of classes } \alpha \text{ born at } b_\alpha \text{ dead at } d_\alpha \tag{13}$$

As such, the paper computes a Vietoris-Rips complex filtration for each of the $n - w$ point clouds.

Encoding in a persistence diagram – For each k-dimensional homology class α, one can construct a persistence diagram P_k by projecting the points z_α of the homology group $H_k(R(X, \epsilon))$ on a 2D plot. Additionally, the points $\{(x, y)|x = y\}$ (i.e. the diagonal) are also projected on the same 2D plot. They represent the classes instantly born and dying at each level ϵ.

As such, the paper computes a persistence diagram for each Vietoris-Rips complex filtration of the starting $n - w$ point clouds. Of note, the paper only focuses on 1-dimensional homology from then on ("persistence homology of loops," p3; "In the sequel, we will use only the 1-dimensional homology," p5). The paper's goal is to study the evolution of loops in the topological representation of the data and see whether market behavior can be evidenced from the corresponding persistence diagram/landscape.

Encoding in a persistence landscape – The metric space of persistence diagrams $(\mathcal{P}, W_P)$ (with the Wasserstein distance) does not lend itself as-is to time-series analysis – the goal of the paper. To do so, the paper proposes to embed the space of persistent diagrams constructed from the underlying financial data into a space of persistence landscapes.

The persistence landscape of a persistence diagram P_k is the sequence of functions $\lambda_l(x)$ constructed such that $\forall k \in \{0, 1, 2, \ldots\}$ and $\forall (b_\alpha, d_\alpha) \in P_k$:

$$f_{(b_\alpha, d_\alpha)} = \begin{cases} x - b_\alpha & \text{if } x \in (b_\alpha, \frac{b_\alpha + d_\alpha}{2}) \\ -x + d_\alpha & \text{if } x \in (\frac{b_\alpha + d_\alpha}{2}, d_\alpha) \\ 0 \quad \text{otherwise} \end{cases} \tag{14}$$

$$\forall l \in N_+, \ \lambda_l(x) \colon R \to [0, 1] \tag{15}$$

$$= \begin{cases} l\text{-max}\{f_{(b_\alpha, d_\alpha)}(x) | (b_\alpha, d_\alpha) \in P_k\} \\ 0 \quad \text{otherwise} \end{cases} \tag{16}$$

As such, a persistence landscape is computed for each of the $n-w$ persistence diagram (of the 1-dimensional loops) previously computed (See Fig. 3 in Annex for a diagram of a persistence landscape).

Computing of L^p norm times series – Given the properties of persistence landscapes (e.g. Banach space, see the more detailed section in the available Jupyter Notebook), a L^p norm can be computed for each obtained persistence landscape (one per window), producing a time series of L^p norms such that:

$$L^p\text{-norm of } \eta = ||\eta||_p = \big(\sum_{i=1}^{\infty} ||\eta_k||_p^p\big)^{\frac{1}{p}} \tag{17}$$

The paper only computes the L^1 and L^2 norms from each corresponding persistence landscape.

Visualization of potential trends with statistics – From then on, the paper states that the L^1 and L^2 norms can be used to highlight market trends.

The paper proposes to also compute further statistics: rolling indicators with a z-day time window and a 1-day shift step. Such indicators are the variance, with $z = 500$, the spectral density at low frequencies, with $z = 500$, and the first lag of the autocorrelation function (ACF), with $z = 250$.

The Mann-Kendall test is also proposed to be used to assert monotonic upward or downward movement trends in the L^1 and L^2 time series. The MK test tests whether to reject the null hypothesis $H0$ (No monotonic trend) and accept the alternative hypothesis H_a (Monotonic trend is present).

2 Persistence Landscape Implementation and Synthetic Data Testing

We decide to implement a custom persistence landscape function (See the related Jupyter Notebook) and replicate the synthetic data processes covered in the paper in order to check for errors (See Fig. 4 in Annex for example displays of computed persistence diagrams and landscapes with synthetic data).

2.1 Noisy Hénon Maps

Hénon maps are used to model economic cycles. They are defined by two recursive operations x_{n+1} and y_{n+1}, to which further parameters can be added. The papers relied on adding a timestep update of the parameter a (growing from 0 to 1.4 by preset small timesteps), a noise intensity σ, and a Gaussian noise W to emulate a change of regime in a stochastic process. Further details on the generation methodology are available in the Jupyter Notebook.

By reproducing the paper's use of Hénon maps (See Fig. 5 in Annex for an example Hénon map process and its resulting L^1 and L^2 norm time series), we are able to verify that persistence landscapes can capture changes in regime in a stochastic process.

2.2 White noise with growing variance or with Gamma-distributed inverse variance

The paper relies on synthetic white noise (with either growing variance or with Gamma-distributed inverse variance) to check whether the use of persistence landscapes is able to handle increasing levels of noises or changes in the noise regime. Further details on the generation methodology of both types of noise are available in the Jupyter Notebook.

By reproducing the paper's use of white noise (See Fig. 6 and 7 in Annex for an example white noise processes and their resulting L^1 and L^2 norm time series), we are able to verify that persistence landscapes can capture changes in regime in a stochastic process while also being robust to increasing amount of noises.

2.3 Observations with Synthetic Data

Using the synthetic examples covered in the paper, we replicate, and thus demonstrate the validity of, the paper's workflow and results on that synthetic data.

As such, we can be confident in asserting, as the paper did, that the use of persistence landscapes could be a useful econometric indicator in order to assess market behavior – knowing the chaotic nature of market prices. This sensitivity to state transitions (quote "from a regular to a 'heated' regim") and robustness is a very sought-after basis for any financial early warning sign (EWS) metrics.

3 Applying the Methodology to Financial Data

3.1 Replication and Observations

Using window sizes w of 40, 80, and 120 (the authors used a window size of 50 or 100 as shown in the paper's Figure 9, page 20, and Figure 10, page 21), we find that we obtain very similar results to that of the original paper (See Fig. 8 in Annex for the resulting L^1 and L^2 norm time series).

Though scales are not identical, we can visually identify the same trends on the 1000-trading-day snapshots the article highlighted (See Fig. 9 in Annex for the resulting L^1 and L^2 norm time series with a focus on the Dotcom Bubble Crash period), providing proof of the paper's reproducibility and of the soundness of its methodology.

Furthermore, we can further highlight the identified trends at or around market volatility events such as the Dotcom Bubble or the 2008 Financial Crisis by computing several statistics: 500-day rolling variance, low-pass filtering, first lag of autocorrelation, and the state identified by a Mann-Kendall test (See Fig. ref10, ref11, ref12 in Annex for the resulting statistics for each window).

3.2 A Point of Contention

Some of the paper's assertions and conclusions should be dampened, however. Reproducing the paper's analysis on the whole range of datapoints (i.e. all data points from 1989 to 2016), we found that the paper presented their data only at specific intervals, exemplified with the paper's Figure 9 (See Fig. 13 in Annex for the reproduction). Limiting the data visualization of the L^p norms to a date range prior and shortly after the America One-Time Warner merger may overstate the actually observed upward trend purported to be shown. Indeed, if we increase the display range (See Fig. 8 and 9 in Annex) such a representation eschews larger spikes that happen right after the merger's date.

This does not disprove the paper's method. However it undercuts some of the paper's conclusion. Though the America One-Time Warner merger is considered among many to be one of the warning signs of the Dotcom bubble period, it cannot be construed as the start of a crisis. As such, taking a slice of time prior to the crisis itself may paint only a partial picture of what the persistence landscapes' L^1 and L^2 norms can offer as early warning sign indicators.

3.3 Further Comments

We also want to provide some reserve with regards to the paper's conclusion: that the empirical analysis shows that "the time series of the L^p norms exhibit strong growth around the primary peak emerging during a crisis." This is, in verification, true but to a lesser extent that seems to be defended in the paper – especially through the paper's Figure 9 and 10. It happens that as we previously alluded, if we can clearly see in advance rises in the L^1 and L^2 norms, those

increases cannot be construed for crisis predictors – but rather rises in volatility only.

We also find that the increases in L^1 and L^2 norm values are not proportional to the depth of the crisis they precede. Indeed, given the 2008 Financial Crisis was the worst economic crisis since 1929 (besides the Second World War), we find that the height of the two norms during the Dotcom Bubble was at least three times higher than for the 2008 crisis.

Since market behavior, financial innovation, regulatory oversight, and other market conditions evolve over time, it is also rather important to note that using whole data ranges as the paper did (1989 to 2016) can be useful to note change in underlying market conditions. However, these changes in market conditions, often marked by volatility, are not synonymous with crisis.

That "TDA offers a novel econometric method and is yielding the new category of early warning signals of an imminent market crash" is true but should be further examined and tested – Looking at curve graphs for insight on market behaviors recalls the practice of technical analysis, criticized among many finance circles.

4 Further Explorations

4.1 Norm of the Difference Between Landscapes

We can compute the norm of the difference between consecutive landscapes, yielding a 1-dimensional time series that evidences large movements in terms of landscapes day-by-day (See Fig. 14 in Annex for the norm of the difference between consecutive persistence landscapes computed with windows $w \in \{40, 80, 120\}$).

Based on the hypothesis that such a norm displays volatility behavior on financial markets, we obtain results similar to that of the previously introduced L^1 and L^2 norms of the persistence landscapes. We also find that the norm of the difference seems to highlight increases in volatility. However, the upward trends prior to an eventual financial crash, if these trends ever existed, are milder or absent compared to the paper's method.

4.2 Bottleneck Distance Between Diagrams

We can compute the bottleneck distance between consecutive diagrams, yielding a 1-dimensional time series that evidences large movements in terms of diagrams day-by-day. Based on the hypothesis that such a distance displays volatility behavior on financial markets, we obtain the following with the previously computed diagrams:

Persistence diagrams, as unions of points and of a diagonal, can be formed into a metric space such that the bottleneck distance between diagrams $diag_1$ and $diag_2$ is defined as:

$$\mathrm{d}_b(diag_1, diag_2) = \inf_{\text{matching } m} \max_{(p,q)\in m} \|p - q\|_\infty$$

We find here that the bottleneck distance, though catching switches in behaviors at the time of a crisis does not seem to display much, if any, predictive capability as most spikes seemingly happen after a financial trigger event.

4.3 Replacing the S&P500 with its Derivative Volatility Index: the VIX

Because of its ubiquity in Finance, the volatility index VIX [3] is interesting to look at as part of this process of crisis prediction. Because it is based on the S&P500, we decided to keep the DowJones, Nasdaq, Russell 2000 times series but swap the S&P 500 for the VIX data.

It appears that switching in the VIX, we obtain a much noisier L^1 and L^2 norm time series (See Fig. 16 in Annex for the resulting L^1 and L^2 norm time series). Here again, we have a hard time actually delineating market predictive properties except maybe for the ones created with a window size of 120. This should not be construed as crisis prediction however, but rather the ability to catch some underlying market volatility, strongly influenced by the inclusion of the VIX, rather than the S&P500 – which might bias the results.

4.4 Testing the Methodology on Data Up to 2022

Finally, given the available data (for the VIX, Russell 2000 and Nasdaq as the S&P500 and DowJones have rescinded their authorization for Yahoo to share their daily history), we perform the same process but with data up to February 3rd, 2022.

Given the same results for the period 1989 to 2016, we are interested in seeing what trends we can identify from 2017 onward. As such, we see some changes in market behavior around the most recent events like the last American elections and of course the Covid-19 Pandemic (See Fig. 17 in Annex for the resulting L^1 and L^2 norm time series).

5 Conclusion

Overall, the paper offers a novel approach to the field of econometrics and the development of metrics enabling the detection of upcoming adverse events, i.e. early warning signs. Using a multi-step process relying on Topological Data Analysis allows the extraction of valuable information, such as volatility, from intrinsically noisy, multi-dimensional time series data. Indeed, we have proven that persistence landscapes can identify underlying structures in financial data data, from which we can expect to yield useful information on market volatility. In as such, the method is promising.

Nonetheless, the proposed methodology, though it can accurately describe change in market conditions, may be less convincing in terms of whether or not it can work in any predictive fashion – a grail in econometrics. Indeed, we have seen that, though the method can evidence volatility, such evidence is hardly

forward-looking when compared to the timeline of a given crisis (in our case the Dotcom Bubble and the 2008 Crisis). Rises in volatility metrics should not be construed for early warning signs of an upcoming crisis.

Finally, as an area of exploration, were we able to have access to intra-day data (of stock indexes or companies for instance), we could envision the use of the proposed persistence landscape method to study and highlight change in log-return tail behavior, a key target in econometrics in terms of regulatory metrics such as with the often mandatory reporting on Value-at-Risk or Expected Shortfall in North America and the European Union.

References

1. Gidea, M., Katz, Y.: "Topological Data Analysis of Financial Time Series: Landscapes of Crashes" (2017)
2. Tobias, A., Brunnermeier, M. K.: "CoVaR" American Economic Review, 106 (7): 1705-41. (2016)
3. Chicago Board Options Exchange. "The CBOE Volatility Index -VIX" (2009)

Annex

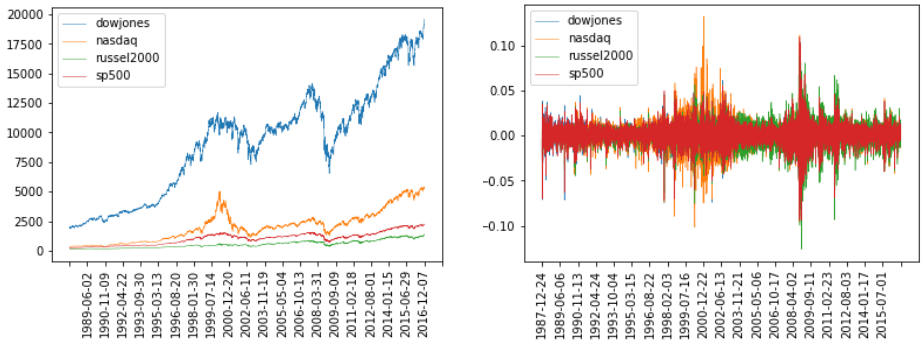

Fig. 1. Adjusted Closing prices and corresponding log-returns of the 4 key US stock indexes from 1989 to 2016

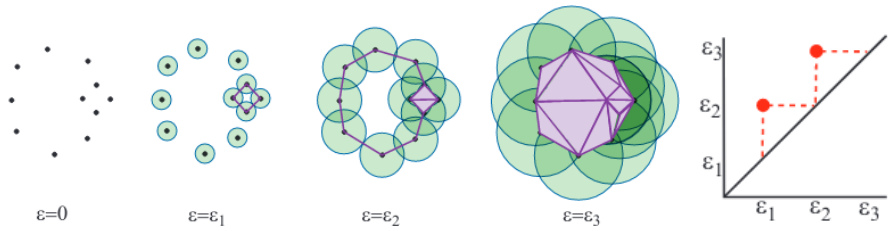

Fig. 2. Rips filtration of simplicial complexes illustrating the birth and death of loops along with the corresponding persistence diagram[1]

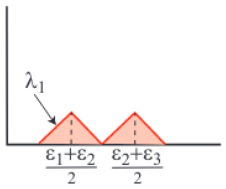

Fig. 3. Persistence landscape of the previously displayed persistence diagram [1]

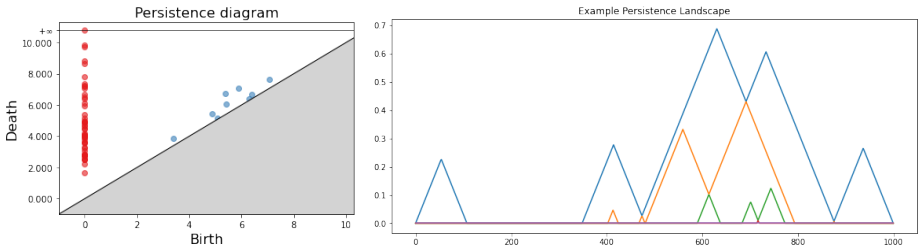

Fig. 4. Example persistence diagram and landscape computed on a $w - 50$ cloud data points from a white noise with growing variance stochastic process as described in [1].

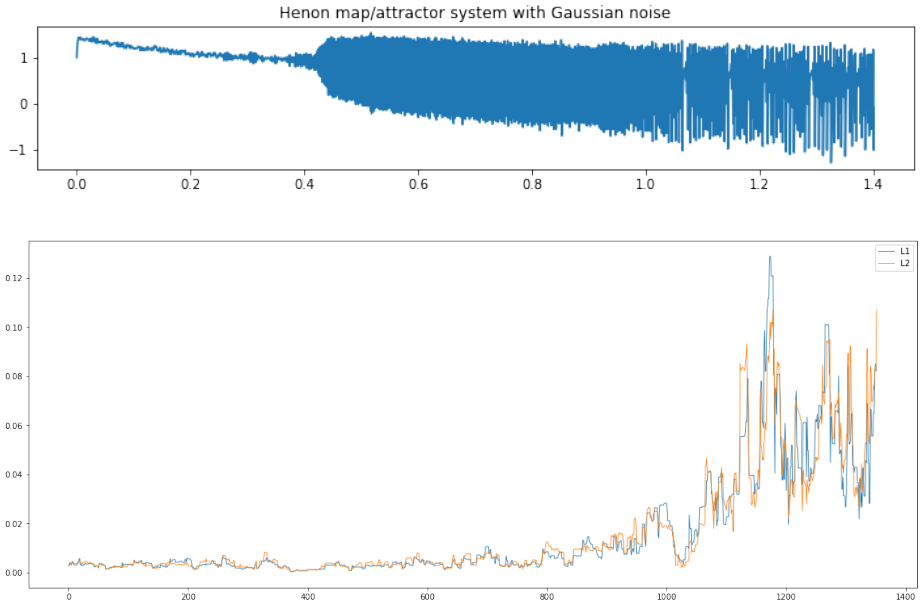

Fig. 5. Replication of the use of noisy Hénon maps as displayed in [1].

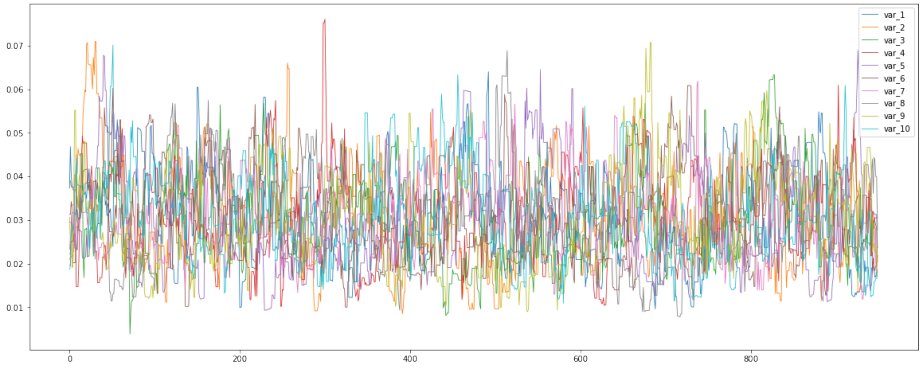

Fig. 6. Replication of the use of noisy white noise with growing variance – the increasing variance does not affect the computation of the L^p norm.

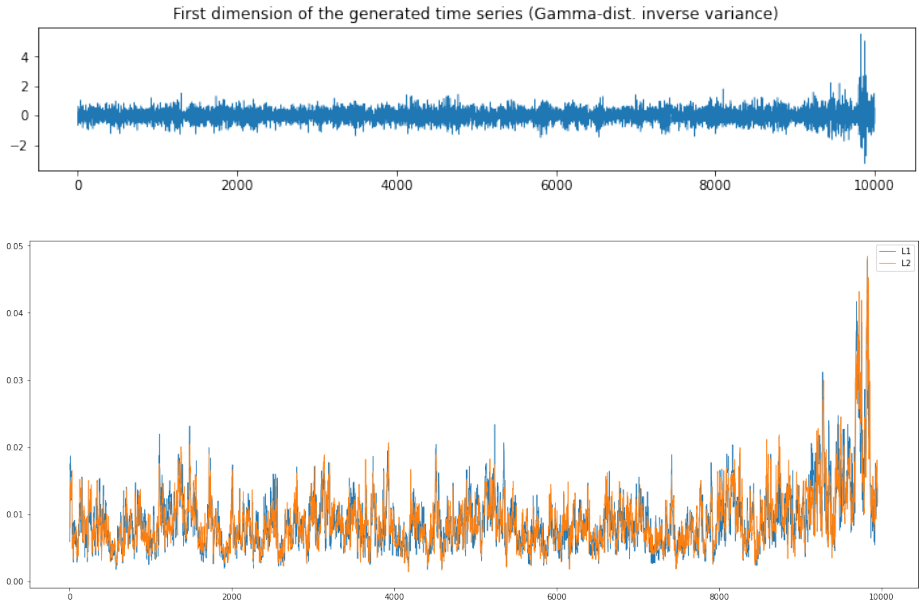

Fig. 7. Replication of the use of noisy white noise with Gamma-distributed inverse variance as displayed in [1] – the change in regime happens during the last 25%.

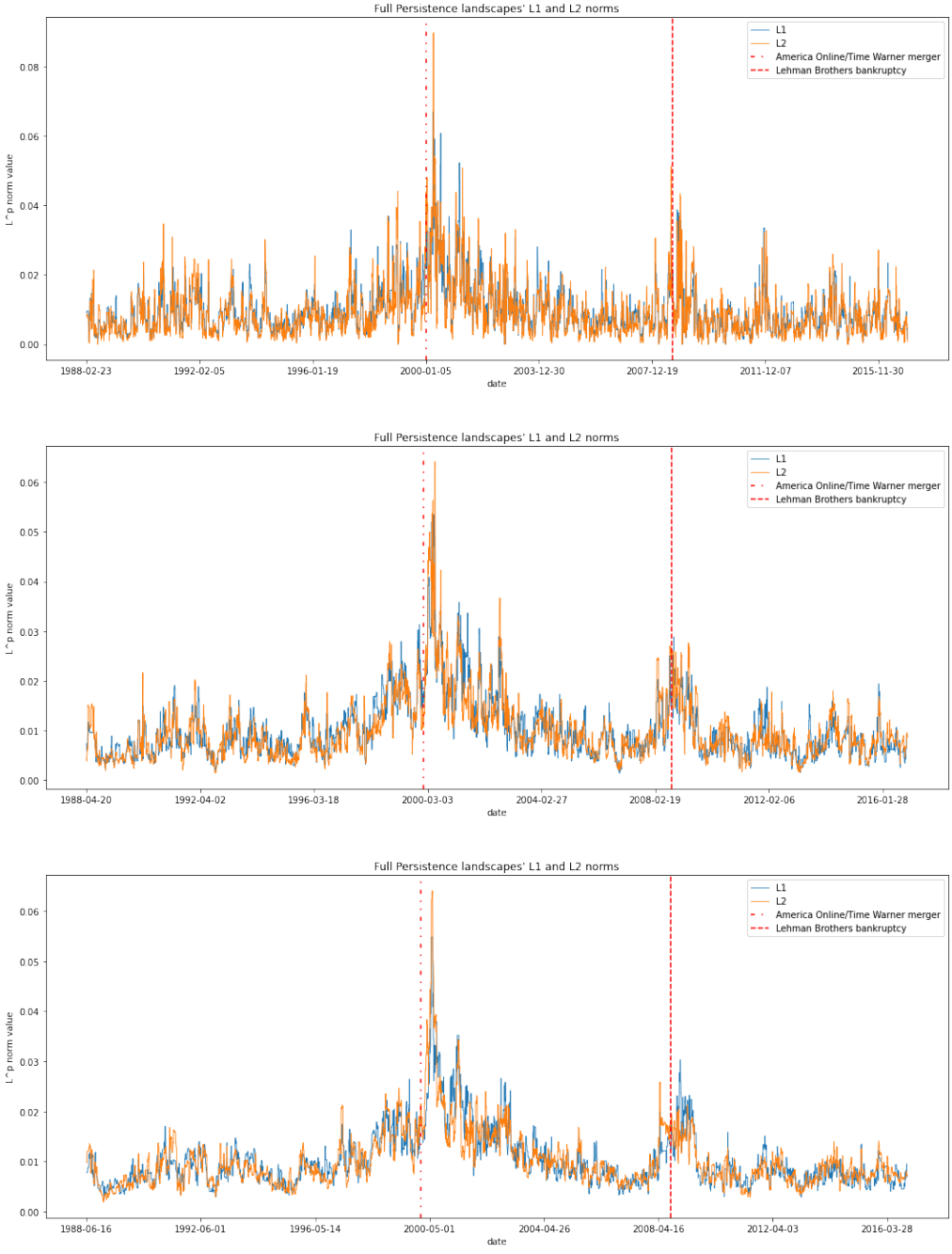

Fig. 8. Resulting L^1 and L^2 norms from performing the paper's persistence landscape workflow[1] with window sizes, in order, of 40, 80, and 120

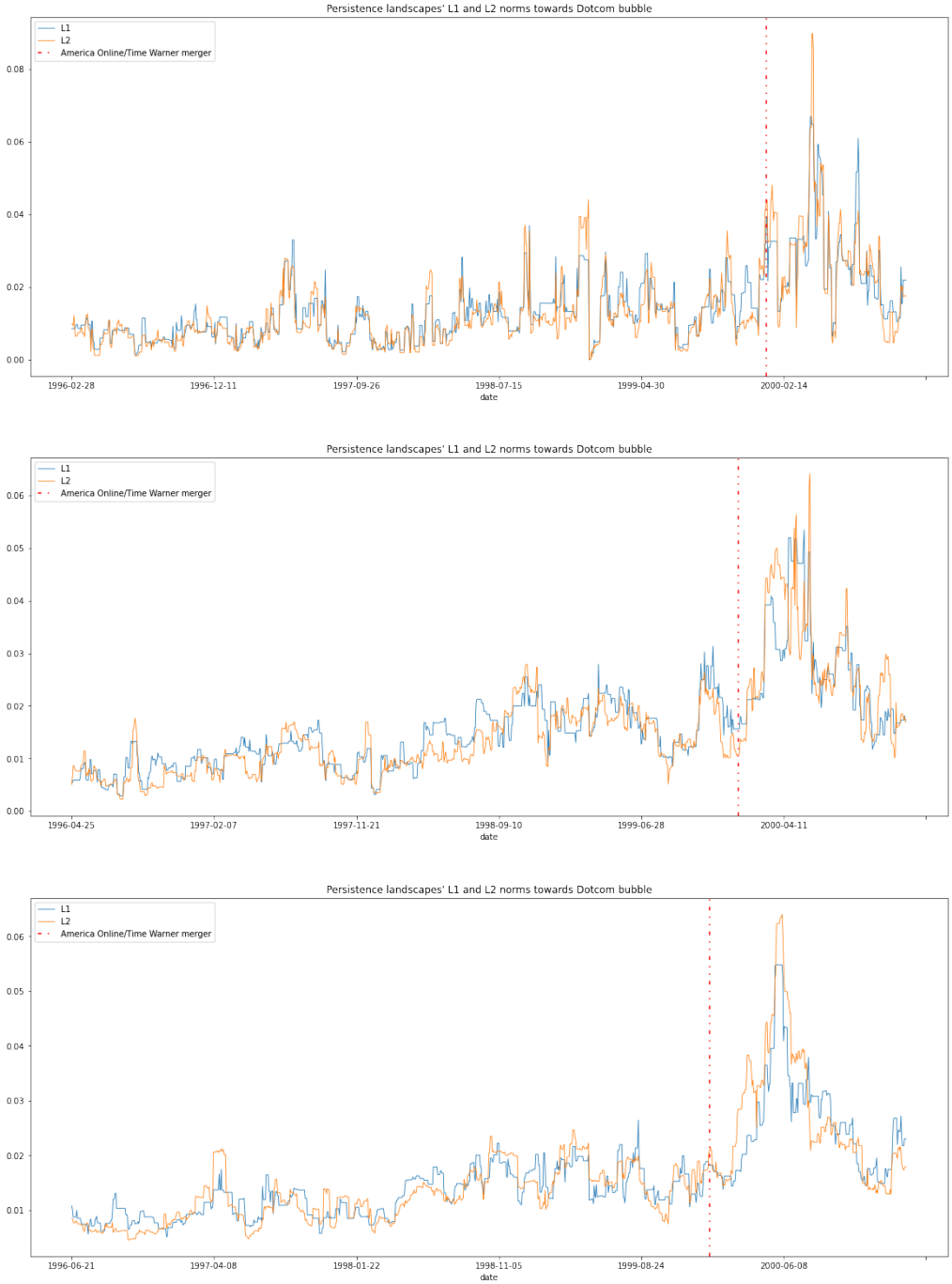

Fig. 9. Zoom on the period prior and during the Dotcom Bubble Crsis on the resulting L^1 and L^2 norms from performing the paper's persistence landscape workflow[1] with window sizes, in order, of 40, 80, and 120

Fig. 10. 500-day rolling variance, low-pass filtering, first autocorrelation lag, Mann-Kendall-identified state of the L^1 and L^2 norms resulting from the paper's workflow[1] with a window size of 40.

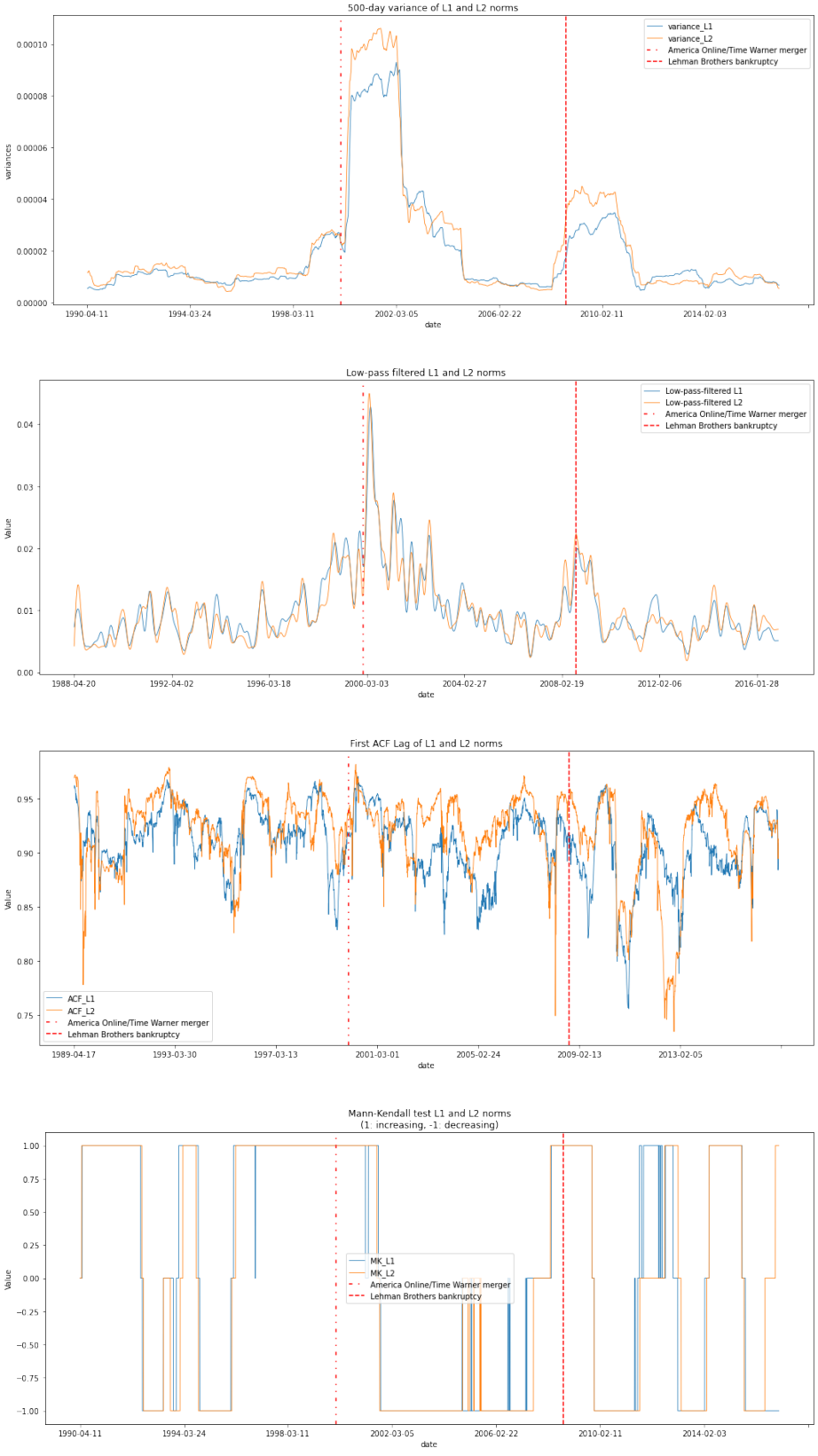

Fig. 11. 500-day rolling variance, low-pass filtering, first autocorrelation lag, Mann-Kendall-identified state of the L^1 and L^2 norms resulting from the paper's workflow[1] with a window size of 80.

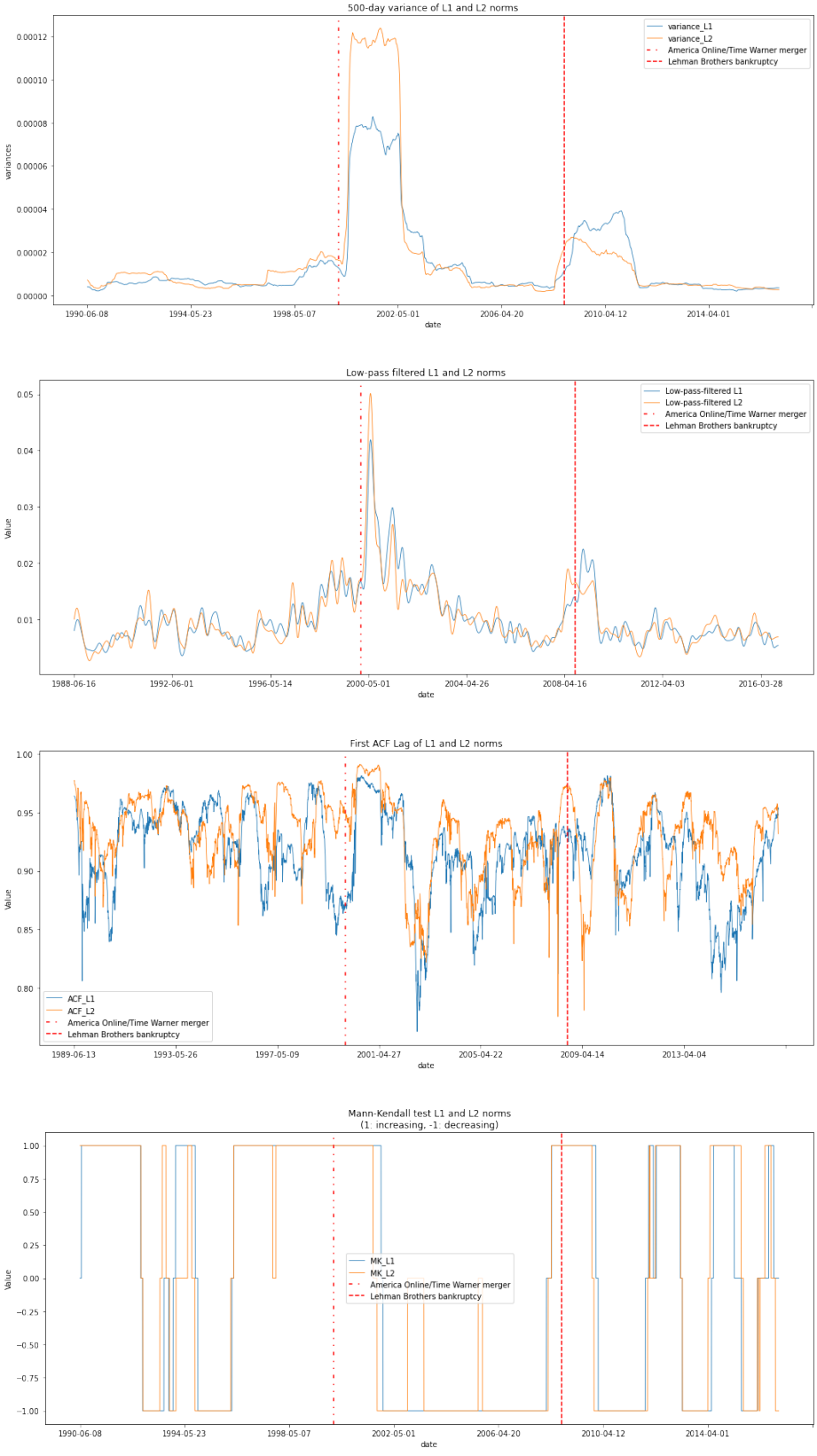

Fig. 12. 500-day rolling variance, low-pass filtering, first autocorrelation lag, Mann-Kendall-identified state of the L^1 and L^2 norms resulting from the paper's workflow[1] with a window size of 120.

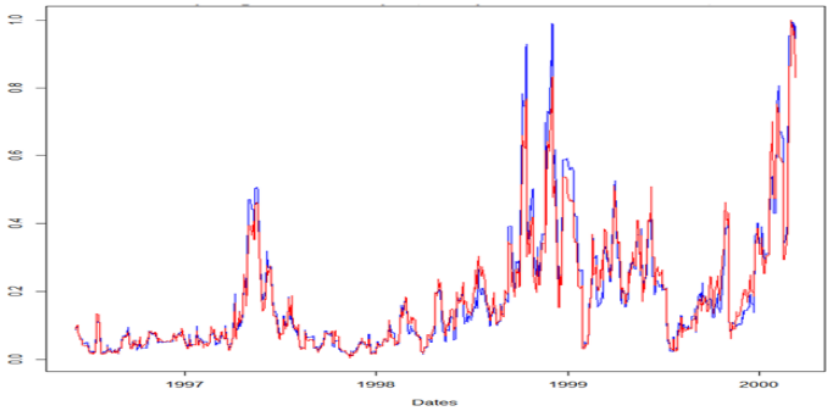

Figure 9: The time series of normalized L^1 (blue line) and L^2 (red line) norms of persistence landscapes calculated with the sliding window of 50 days. Color online.

Fig. 13. Reproduction of Figure 9 from [1].

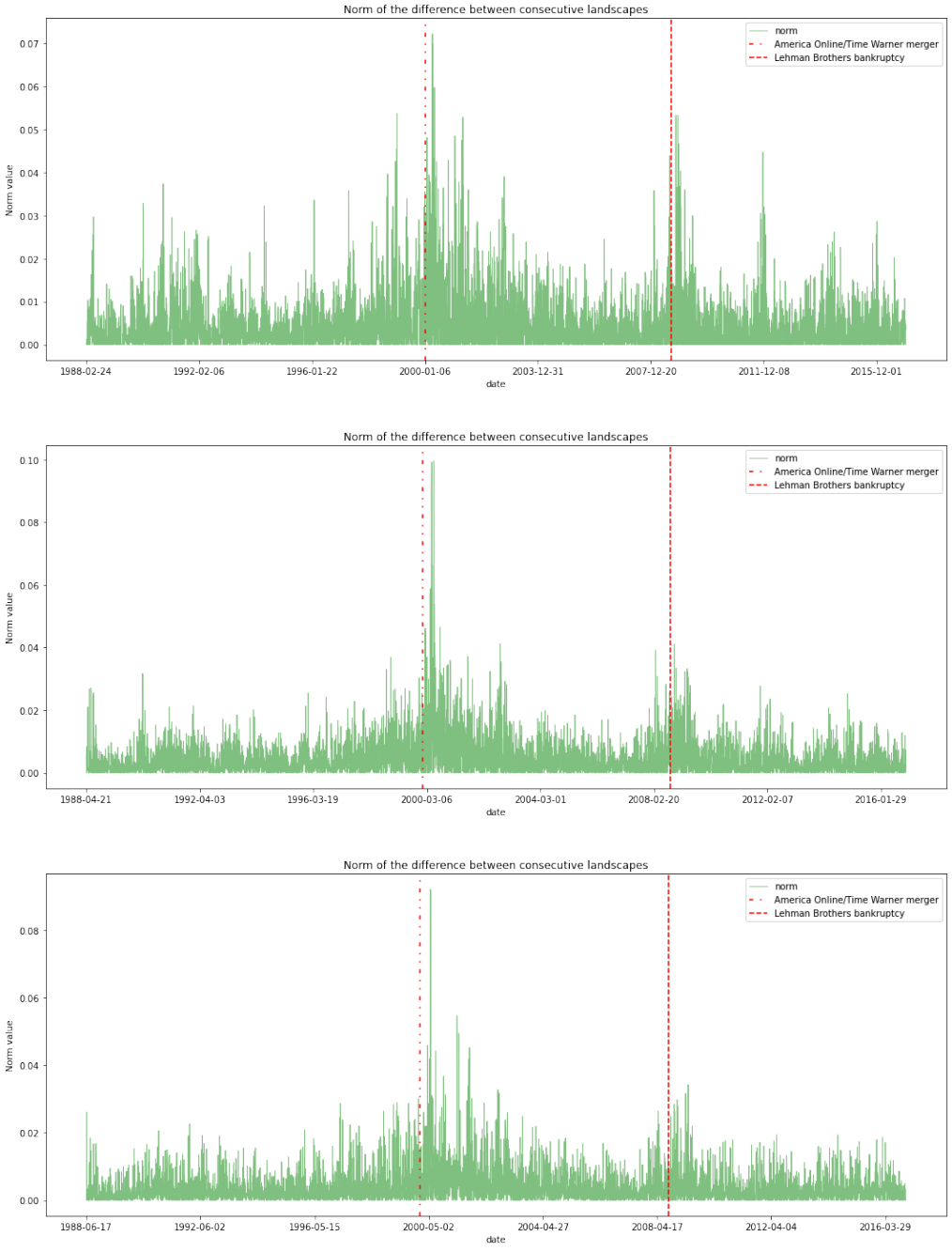

Fig. 14. Resulting norm of the difference between persistence landscapes computed with window sizes, in order, of 40, 80, and 120

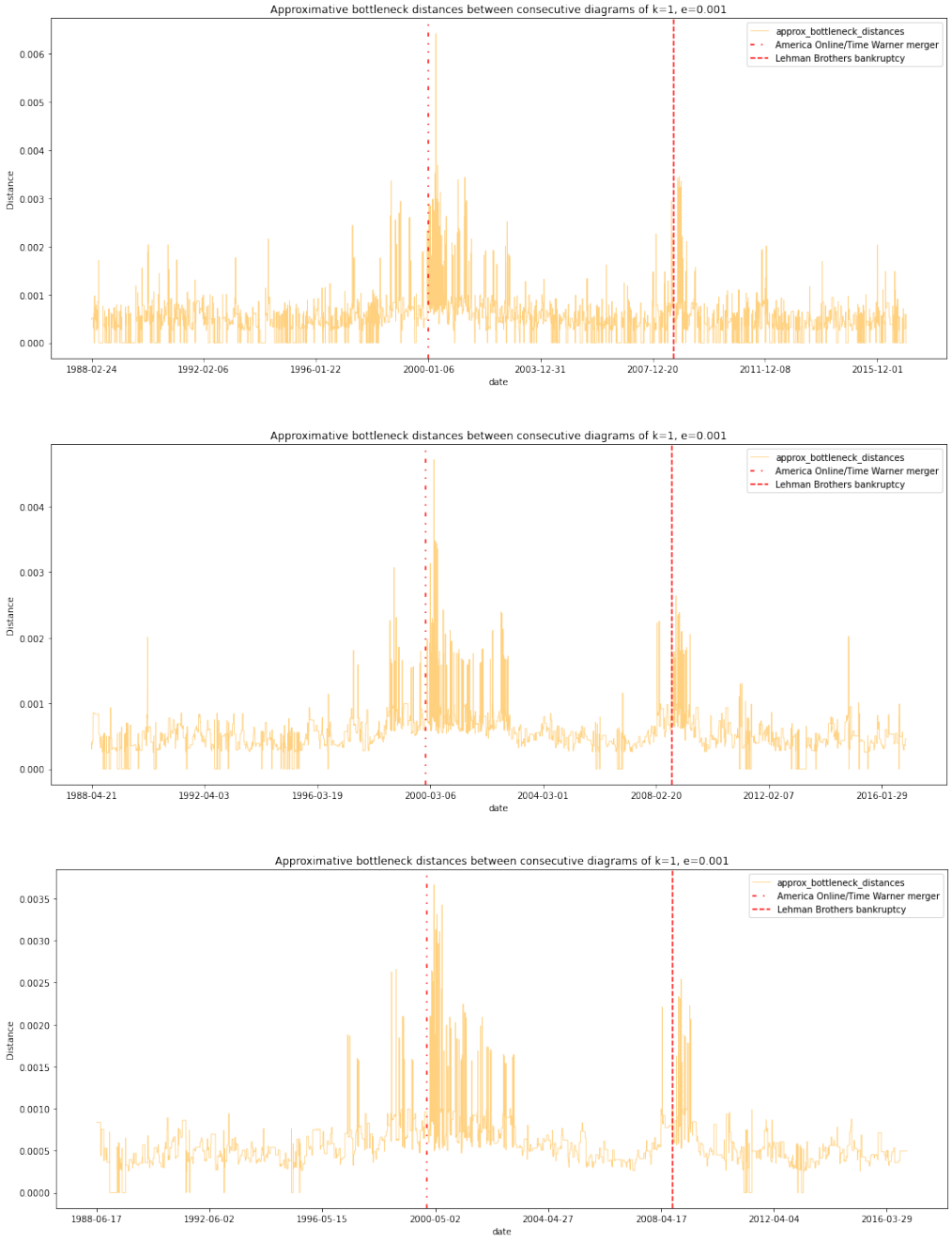

Fig. 15. Resulting bottleneck distance between persistence diagrams computed with window sizes, in order, of 40, 80, and 120

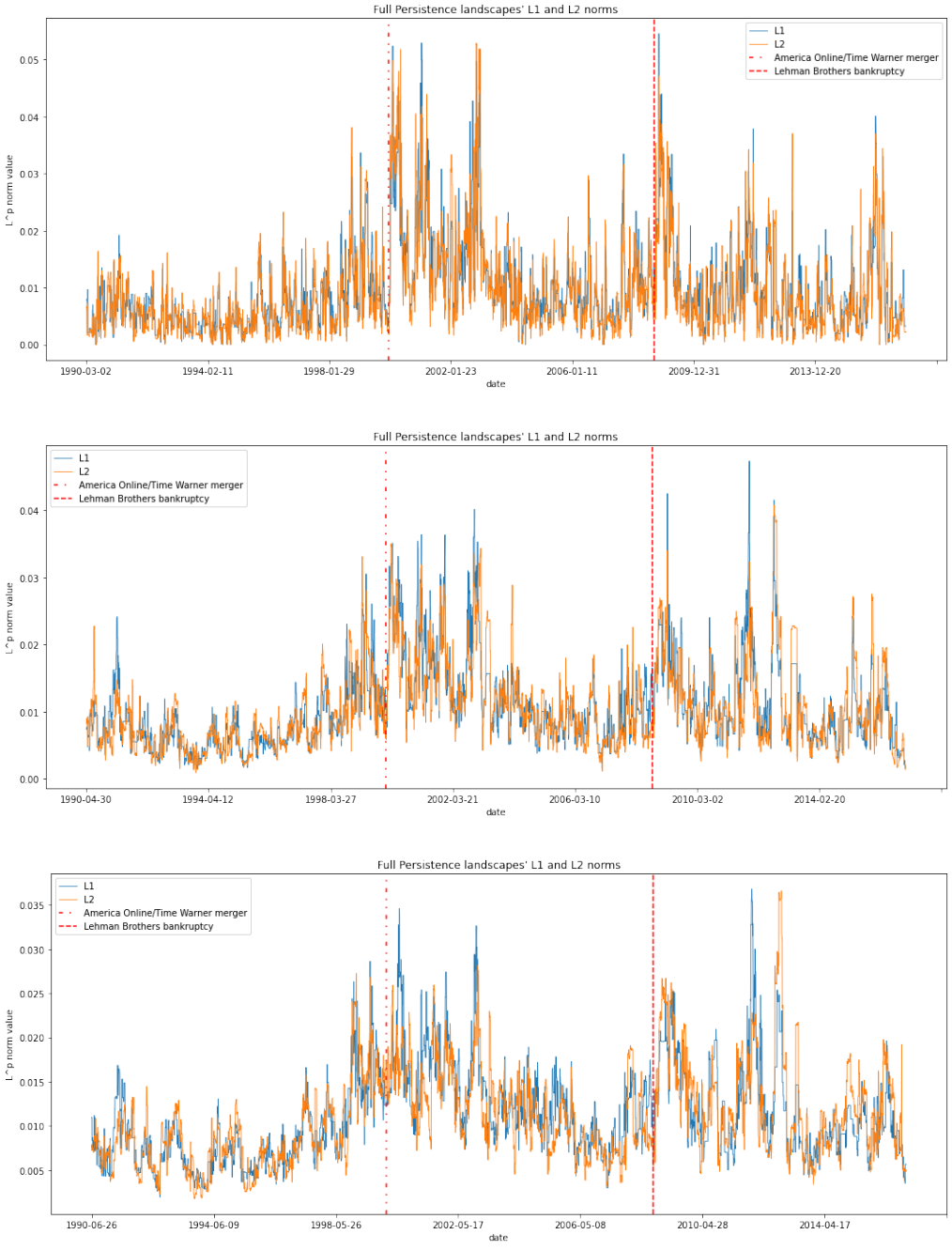

Fig. 16. Resulting L^1 and L^2 norms from performing the paper's persistence landscape workflow[1] with window sizes, in order, of 40, 80, and 120 – with the VIX index data instead of the S&P500

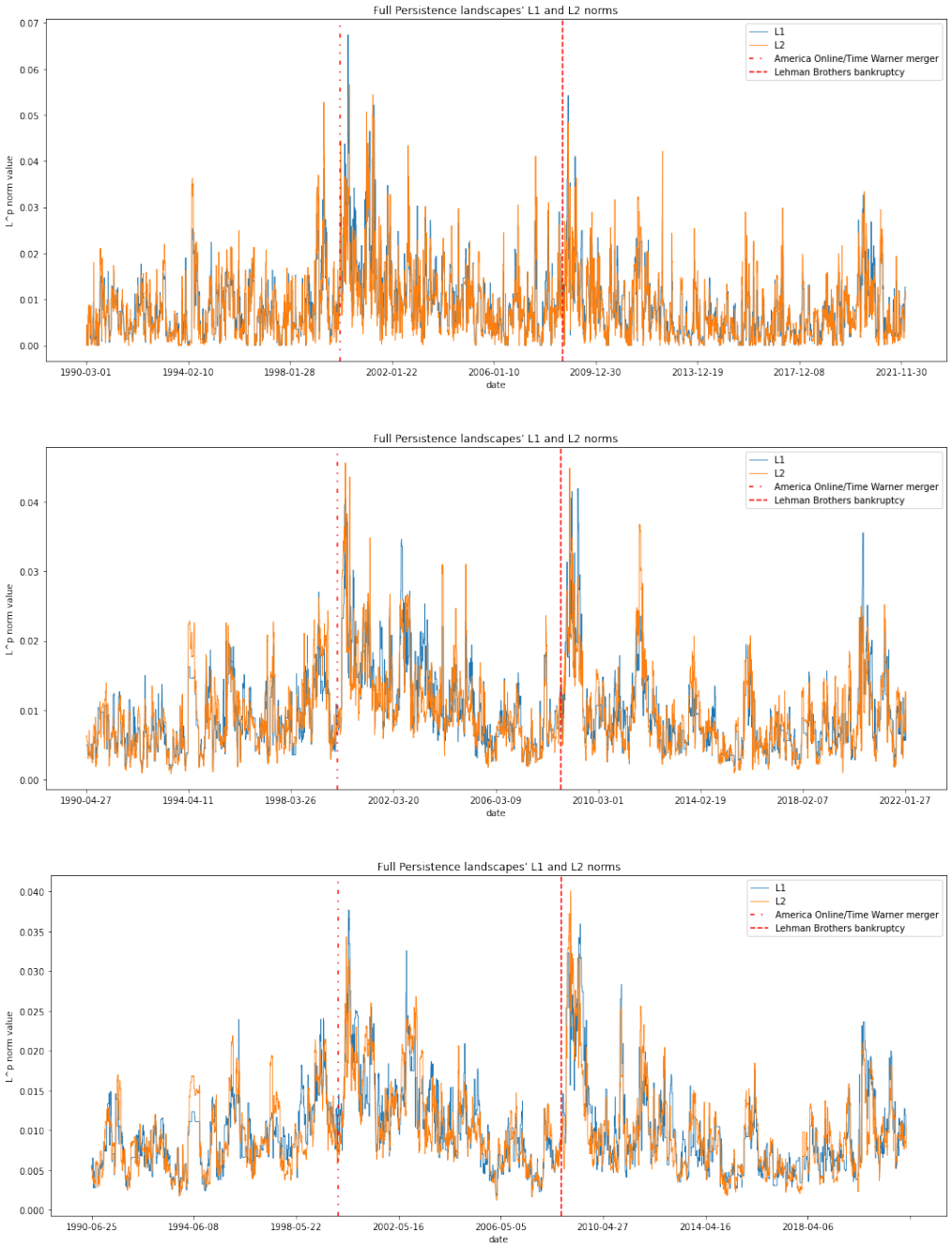

Fig. 17. Resulting L^1 and L^2 norms from performing the paper's persistence landscape workflow[1] with window sizes, in order, of 40, 80, and 120 – with the VIX index, Russell 2000 and Nasdaq data up to February 3rd, 2022