Background Subtraction for Pattern Recognition in High Frequency Financial Data

Alex Papanicolaou[1] and Patrick Barkhordarian[1]

Original: January 21, 2016
Revised: April 18, 2016

Abstract

Financial markets produce massive amounts of complex data from multiple agents, and analyzing these data is important for building an understanding of markets, their formation, and the influence of different trading strategies. We introduce a signal processing approach to deal with these complexities by applying background subtraction methods to high frequency financial data to extract significant market making behavior. In foreign exchange, for prices in a single currency pair from many sources, we model the market as a low-rank structure with an additive sparse component representing transient market making behavior. We consider case studies with real market data, showing both in-sample and online results, for how the model reveals pricing reactions that deviate from prevailing patterns. We place this study in context with alternative low-rank models used in econometrics as well as in high frequency financial models and discuss the broader implications of the melding of background subtraction, pattern recognition, and financial markets as it relates to algorithmic trading and risk. To our knowledge this is the first use of high-dimensional signal processing methods for pattern recognition in complex automated electronic markets.

1 Introduction

Modern electronic markets are fast, complex, and involve risky behavior. A conventional market making strategy operated via a human dealer might set a targeted fixed spread from

[1]Integral Development Corporation, 3400 Hillview Ave., Building 4 Palo Alto, CA 94304
Corresponding author email: alex.papanicolaou@integral.com

a midprice determined in relation to order flow to be able to generate profit from the order flow. In automated markets with algorithmic trading, this fixed spread approach can be augmented with faster and richer strategies that only computers can provide. This work is about utilizing pattern recognition, in particular background subtraction, for uncovering complex market behaviors.

The electronic spot Foreign Exchange market is a fragmented and decentralized but highly connected ecosystem consisting of an extensive array of liquidity providers. These providers form a diverse group of organizations such as large money-center banks, smaller regional banks, and non-bank financial institutions. Despite the fragmentation, price discrimination in the form of hundreds of different pricing sources across many currency pairs and lack of "one price," quoted prices between streams in a particular currency pair are highly cointegrated in that the gap between each streams' bids and asks are in some sense stationary. This cointegrated nature forms the basis of our expectation that a background price formation process exists. An important observed aspect in the market is that besides the background price formation process there are transient pricing deviations from the general market.

1.1 Main Result

The main result of this paper is an approach to decomposing and extracting pricing deviatons through a new application of a low-rank plus sparse model. We extract the background price formation process and reveal behavioral responses to trading that would otherwise go unseen. Our goal is to answer the questions, can we detect price deviations from the typical group behavior of the background price formation process and what do these transient deviations say about market making behavior? We demonstrate results for detecting deviatons using both an ex-post analysis of quote data as well as a fast online detection as the price discovery process develops. This work represents a discovery that the behavior of diverse market pricing would adhere to a low-rank structure and that one can extract behaviorial responses

by exploiting a low-rank plus sparse model.

In an intuitive sense, we think of a transient pricing deviation as a change in pricing strategy by a market maker that deviates from their typical strategy as well as in relation to the overall observed patterns in the market. We leverage a broad view of the market from a massive volume of diverse order and quote data to create a notion of a background price formation process from which the transient pricing deviations will represent the detectted foreground behaviors.

We model a collection of time series of quoted prices in a single currency pair as a composition of a few fundamental background signals that represent the general market-wide price formation process. Transient pricing deviations from the general market pattern are modeled with an additive sparse foreground component. We applied the model to quote data obtained from the Integral FXGrid trading platform[1] in two case studies where we show how the model allows one to observe, interpret, and label the transient phenomena in the market. Furthermore, we compare an online approach to making detections compared to the in-sample results and show how results are highly consistent.

While a few case studies is clearly not enough to make broader judgements, the signal extraction methodology makes clear the algorithmic potential for extraction, characterization, and labeling phenomena by manual inspection and with automated inspection over many orders. Establishing these patterns as consistent behaviors longitudinally helps build an understanding of market participants and their market making behaviors.

In the case studies we present, we indentify various types and interpretations for the detected pricing deviations and what the implications are for markets. Since our case studies focus on market reactions to large orders, a notable example of a detection is the risk-averse reaction of some market makers. This suggests a fast, intelligent response to detecting a larger order than these market makers are otherwise privy to. When the detections are then compared

[1]Integral FXGrid is a multi-sided trading platform for foreign exhange that features these diverse streams of quoted prices.

against trading activity, we see a potential for exploitation of the Last Look option[2] where a market maker rejects a trade on an advertised rate only to appear to respond immediately with a much higher price.

We find instances of other risk-aversion strategies that some market makers employ to modify their quoting patterns in reaction to trading. We find a case of a market maker engaging in inventory control following a filling a substantial amount of an order by skewing or shading its prices. This behavior represents an effort by the market maker to charge a premium for more trading on the same side and attract trading on the opposite side to clear the inventory. And while not an intended behavior, we show an instance of time delayed prices from a market-maker that get clearly detected by the background subtraction algorithm.

Detecting behaviors like these help provide a general understanding of the quality of a bank's market making, especially as it relates to cost and resiliency, and to drive execution strategies. To that end, the motivation for extracting a background price formation process varies from different view points. From a trader perspective, it is about mitigating asymmetric information risk that arises from a lack of information or knowledge about the nature of market. From an exchange or facilitator perspective, especially in over-the-counter FX with its hybrid microstructures, the organization, structure, and transparency of the market are key for maintaining integrity and efficiency. From a researcher or regulator perspective, pattern recognition methods that leverage large volumes of data to detect irregularites can extend more broadly to all sorts of analyses, especially for potentially catastrophic events like the 2010 Flash Crash Easley, Lopez de Prado, and O'Hara 2011.

1.2 Comparative Review

The work we present in this paper is related to a diverse range of other topics. We try to provide a general context with respect to markets, low-rank modeling, and high frequency

[2]Last Look is an option for market makers to reject or deny requested trades. The option provides a backstop to market makers so they may quote liquidity far more than broadly than they actually possess.

trading.

There has been work related to addressing issues regarding the fragmentation of markets but to our knowledge none have approached the problem from the perspective of using pattern recognition methods to detect and interpret quoting deviations. Hasbrouck 1995 and Gonzalo and Granger 1995 have addressed price discovery across multiple markets to develop measures of contribution to price formation. With more focus on trading, Laruelle, Lehalle, and Pages 2011 devise a reinforcement learning algorithm for splitting an order over multiple venues. Our work deviates considerably due to the nature of the problem we are trying to address: a method for extracting transient pricing behaviors in quote data.

Factor Analysis and Principle Component Analysis (PCA) are two methods frequently used in econometrics and finance for dimension reduction and contstructing latent variables, often with trading and hedging strategies in mind. The core principle of Factor Analysis and PCA is to find a few latent factors that capture the common variation in the data and leave random noise as the residual. Grimm and Yarnold 1995 and Child 2006 provide a general overview but specifically to finance, a standard example would be factors that capture the fluctuation in asset returns as in the recent works of Pelger 2015 and At-Sahalia and Xiu 2015.

Principle Component Pursuit (PCP), commonly referred to as Robust PCA, is also concerned with capturing a low-rank signal in the data, namely the state of a background signal, but the key differences arise from how PCP uses a structured residual. The extracted structured sparse residual becomes the target of the model in background subtraction applications since it represents the transient signals. PCA has been proposed for background subtraction in Oliver, Rosario, and Pentland 1998 but the unstructured residual poses a challenge in extracting meaningful results. Moreover, while Factor Analysis and PCA can be modeled with small noise, in general financial applications where data is collected across diverse assets, noise can be fairly large such that factors are more challenging to find. Thus for

the application we are proposing that relies on leveraging multiple pricing observations to directly capture information about the background price formation process, an approach like PCP is more applicable since we are not in a high noise environment and inferences on the nature of market makers are made through observations in the sparse component.

Asymmetric bid and ask models have been studied in Hasbrouck 1999 and Zhang, Russell, and Tsay 2008 and the general approach is to try to infer a latent efficient price to give information about asymmetric costs in the market. The asymmetric bid and ask models utilize a bottom up approach to infer time series characteristics for a single observed time series. Hasbrouck 1999 introduces a model for bids and asks as asymmetric costs around the latent efficient price, which itself is modeled as EGARCH process. Zhang, Russell, and Tsay 2008 extend the asymmetric bid and ask model to study the influence of order flow characteristics on costs and how these charactistics drive the asymmetries.

The interesting contrast to our proposed approach utilizing Robust PCA methods comes from how we were able to detect evidence of asymmetric pricing (see the example of Bank 4 in the second case study in Section 3.2.2). The detected pricing deviation found that bid prices out of ordinary with the usual strategy, and likely tied to a liquidity replenishment/inventory control strategy, did not try to leverage any notion of an efficient price. If the background price formation price itself was asymmetrically pricing the asset, this would fail to be detected by a background subtraction methodology. As stated in the previous section, the background subtraction approach is model agnostic and would only recognize certain market maker prices when they are out of the ordinary, not if the background prices are out of ordinary with a latent efficient price.

More broadly, a desirable goal should be to always try to link market microstructure phenomena with results of a background subtraction methodology and see how market microstructure can influence improvements by studying the background price formation process and the transient deviations of market makers.

Extensive interest and research has been conducted recently studying high frequency trading (HFT). A few examples of recent work include: Baron, Brogaard, and Kirilenko 2014, which gives a characterization of HFT and analyzes HFT performance in the E-mini S&P 500 futures market; Brogaard, Hendershott, and Riordan 2014, who studied the role of HFT in price discovery and price efficiency; Carrion 2013, who carries out a study of HFT performance, trading costs, and market efficiency in NASDAQ; Yang et al. 2012, who utilize markov decision process and inverse reinforcement learning model to identify HFT behavior patterns; Hendershott, Jones, and Menkveld 2011, who study the impact of automated quote dissemination and algorithmic trading on NYSE; Clark-Joseph 2013, who perform an analysis of how HFT use small exploratory orders to probe liquidity information in an order book.

The predominant focus in these and similar works is on characterizations of HFT as well as on aggressive HFT orders and their effect on markets, efficiency, and price discovery. The structure of the OTC FX markets, in particular in many cases the transparency of counterparites due to direct credit trading relationships as well as the nature of the available data from many market makers, lends itself to an alternative approach to broadening the understanding of fast, automated electronic markets. Here the focus is on the market making side and how market making differs among (and within) agents and furthermore, how it differs in response to aggressive orders. Our approach is about detecting and extracting very individualistic responses at some of the finest detail levels in the market as opposed to a higher level aggregate market statistics. It should be stated that we envision aggregating the collected results to make broader aggregate conclusions, but we seem to be starting at a much finer grain level.

1.3 Outline

The structure of the paper is as follows. In the next section, we provide a comprehensive setup of the data and the problem formulation. We then present the main results as case studies showing how the market decomposes with a low-rank structure and the various kinds of behavior we extract in the sparse component in response to trading activity. We finish with a discussion on the challenges of background subtraction for high frequency electronic markets.

2 Background Subtraction and Robust PCA

Background subtraction is a well studied problem in video surveillance where the goal is to extract a background scene to reveal important foreground features or moving people and objects. One approach to background subtraction starts with a data matrix, $M \in \mathbb{R}^{m \times n}$, where the rows represent pixel intensity in the video frames and the columns represent the video frames. Because the background signal is typically quite strong and foreground components are not present in all pixels or all frames, one expects there to be a reasonably low-rank representation of M with a few dominant singular values that captures the vast majority of the information of the background signal.

A simple approach to extract this background signal and its dominant singular values would be a low-rank approximation based on Principle Components Analysis (PCA). PCA looks for a singular value decomposition (SVD) approximation of the form $M \approx L = U\Sigma V^T$, where $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$, and $\text{rank}(\Sigma) = r \ll \min(m, n)$. As an optimization problem, this is represented by,

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \|X - Z\|_F \\ \text{subject to} \quad & \text{rank}(X) \leq k, \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm. With the low-rank matrix capturing the background,

one is left with a residual matrix that is likely dense and tricky to interpret. While it may capture the foreground components, it fails to provide a key feature often desired of extracted foreground components: sparsity. In video surveillance, foreground objects do not impact every part of the video frame but rather only localized parts, hence one expects the dominant singular values to compose in the background and the foreground to be spatially localized in the matrix and not contributing much to the singular value decomposition. Thus a good separation of the foreground signal should leave a sparse residual. To improve on standard PCA, one could threshold entries in the residual $M - L$ to get the separated foreground signal. An approach like this is taken in Oliver, Rosario, and Pentland 1998.

The low-rank plus sparse decomposition where $M = L + S$, $L = U\Sigma V^T$ is low rank, and S is sparse, is a generalization of a low-rank model and directly accounts for this sparse residual feature. The low-rank component L will capture the background signal as in PCA while the sparse residual component S will capture a structured representation of the foreground. By obtaining a structured sparse residual S, the foreground is a more desirable form for interpretation and analysis. In our application for high frequency financial data, the structured nature of the extracted sparse component will prove to be useful for revealing the nature of markets and pricing behavior.

Finding an arbitrary low-rank plus sparse decomposition by minimizing the rank of L and the support of S is an NP-Hard problem. Instead, one can utilize the convex relaxation Principal Component Pursuit (PCP) of Candès et al. 2011. The convex optimization problem for Principal Component Pursuit is defined as,

$$\begin{aligned} \underset{L,S}{\text{minimize}} \quad & \|L\|_* + \lambda \|S\|_1 \\ \text{subject to} \quad & M = L + S, \end{aligned} \tag{2}$$

where $\|\cdot\|_*$ is the nuclear norm[3], $\|\cdot\|_1$ is the vectorized ℓ_1-norm, and λ is a regularization

[3]The nuclear norm (also known as the *trace norm*) is defined as $\|A\|_* = \text{trace}(\sqrt{A^*A}) = \sum_{i=1}^{\min(m,n)} \sigma_i$ for $A \in \mathbb{R}^{m \times n}$, where σ_i are the singular values.

parameter. The name *Robust PCA* is often applied to this type of decomposition since it attempts to render PCA robust and therefore we adopt the term Robust PCA throughout.

In the convex optimization problem (2), the nuclear norm essentially minimizes the rank of L and thus induces a low-rank matrix by applying a threshold to singular values of a matrix. The vectorized ℓ_1-norm essentially minimizes the support of S by soft thresholding and therefore induces sparsity. The regularization parameter λ quantifies the trade-off between rank and sparsity. Candès et al. 2011 prove under certain conditions that setting λ equal to $\frac{1}{\sqrt{\max(m,n)}}$ yields exact recovery. Lin, Chen, and Ma 2010 provide an algorithm for fast, exact recovery using an augmented Lagrangian multiplier method, which we present in the appendix.

For reference, see Tibshirani 1996 for the foundations on ℓ_1 penalization in LASSO regression. See Recht, Fazel, and Parrilo 2010 for the use of the nuclear norm to form low-rank reconstructions and how it is similar to ℓ_1 norms inducing sparsity. For a further comprehensive survey on the low-rank plus sparse decomposition and its application in video surveillance, see Bouwmans and Zahzah 2014. And for a general source on convex optimization which is invaluable for problems like these, see Boyd and Vandenberghe 2004.

2.1 Comparison to Factor Analysis and PCA

As has been mentioned, Factor Analysis is frequently used in econometrics and finance for dimension reduction and contstructing latent variables, often with trading and hedging strategies in mind. The core principle of a factor model is to find a few latent factors that capture the common variation in the data and leave random noise as the residual. A standard example in finance would be factors that capture the fluctuation in S&P 500 asset returns as in a recent example like At-Sahalia and Xiu 2015. From Child 2006 or Grimm and Yarnold

1995, a typical time series factor model is structured with a latent variable equation like,

$$Z_t = \Lambda F_t + \varepsilon_t \tag{3}$$

where Z_t are the observed time series processes, Λ are the factor loadings, F_t are the independent latent factor processes, and ε_t is the residual noise with diagonal covariance matrix Ψ and is assumed to be uncorrelated with F_t. In terms of covariance structure, Factor Analysis is described as,

$$\text{Cov}(Z) = \Lambda\Lambda^T + \Psi. \tag{4}$$

Since time series data naturally comes in discrete observations, the factor model is in practice computed through solving a linear algebra or optimization problem. Factor analysis is closely related to Principle Component Analysis (PCA),

$$T = XW \tag{5}$$

where the data matrix X is transformed with an orthogonal matrix W to a new low-dimenstional space. Just like Factor Analysis, PCA can also be expressed via an eigen-decomposition of the covariance matrix Σ,

$$\text{Cov}(X) = \Sigma = WDW^T, \tag{6}$$

where D is the diagonal eigenvalue matrix. In relation to financial applications, the recent work in At-Sahalia and Xiu 2015 gives an approach to performing PCA for Levy process models for high frequency financial data.

Factor Analysis and PCA both lead to a data transformation giving new, unobserved variables and while there are fundamental differences between the two[4], these low rank models

[4]In particular, as related to assumptions on the residual as seen in the covariance structure. For a further discussion on this, see Grimm and Yarnold 1995.

focus much more heavily on the low-rank structure as it relates to explained variance. The residual in either model is considered unstructured and is often as uninteresting since it does not contain important common or significant variation.

As one can see from (2), Principle Component Pursuit is very similar in spirit in how it generates a low-rank decomposition of the data. One key difference to Factor Analysis and PCA is that PCP puts a lot more focus on a structured residual. Moreover, while Factor Analysis or PCA can be applied in relatively high noise environments like financial econometrics as in At-Sahalia and Xiu 2015, Zhou et al. 2010 shows that PCP is best suited for relatively low noise problems.

Ultimately a background subtraction approach using PCP is just more natural because the goal is to leverage repeated stacked observations of prices coming from many different sources to be able to capture information about the background price formation process. The sparse component robustifies estimation of the background signal by extracting parts that do not fit well with the low-rank component. Robust approaches to Factor Analysis have been proposed, for example in Pison et al. 2003 and Zhang, Li, and Liu 2014, but the premise is more based on factor and loading estimates that are robust to outliers corrupting covariances and does not try to identify the outliers.

3 Application of Robust PCA to FX Data

The main result of this work is a showcase with two case studies how a low-rank plus sparse decomposition reveals the nature of pricing in OTC foreign exchange markets and explore the behaviors of market makers and the interpretations of these behaviors from a financial perspective.

In electronic FX spot markets, quoted prices from a single organization market making in a particular currency pair exist as streams of published messages containing a quote book

of bids and asks as well as volume amounts. A stream of quotes can then be interpreted as a stand-alone source of liquidity or venue. Naturally as technology has improved, there is a movement away from single-bank/source trading to aggregation of multiple streams from several organizations. Moreover, many liquidity providing organizations use multiple streams with varying characteristics to target and segment the population of liquidity takers. Aggregating quotes to show best prices is fundamental and common as foreign exchange trading has become fragmented and highly credit-driven. A further common practice in the industry is to split larger orders and route the parts to these various streams to provide price improvement.

In the following, we first introduce the data obtained from price streams as well as key preprocessing steps. We then present observations from two case studies showing what background subtraction can reveal about markets.

3.1 Quote Data

A typical message from a stream is an entire quote book, see the example in Table 1, and is good until a new message is published. Sequences of quote books coming over the same channel form a *stream*. Liquidity providing organizations use multiple streams with various pricing characteristics to target and segment the population of liquidity takers.

As is common in high frequency price data, the quotes form a right-continuous process since they present prevailing prices until replaced. Due to the irregular quote arrivals, different streams end up asynchronous when observed.

To apply the methodology of the low-rank plus sparse matrix decomposition, we need to manipulate the quote data in a few ways. In the interest of simplicity, we restrict to a fixed time period $[0, T]$ and focus on top tier (top of the book) bids and ask prices while ignoring quoted volume amounts as well. By focusing on the top tier prices, we reduce any stream into a bivariate time series (b_t^i, a_t^i), where b_t^i and a_t^i are the i-th stream's bid and ask at time

Table 1: A single tick quote-book showing available spot AUD/USD orders at tiered prices in the book.

Timestamp	Tier	Bid	Bid Amount	Ask	Ask Amount
2015-04-15 01:59:00.033	1	0.76053	1,000,000	0.76104	1,000,000
2015-04-15 01:59:00.033	2	0.76050	2,000,000	0.76106	2,000,000
2015-04-15 01:59:00.033	3	0.76049	2,000,000	0.76110	2,000,000
2015-04-15 01:59:00.033	4	0.76043	2,000,000	0.76113	2,000,000

t, respectively, and $t \in [0, T]$.

Finally, to be able to apply the necessary algorithms, synchronizing the data into a matrix is key. To do this, we resample the quote data to a fixed frequency with time points defined by $t_j = j\frac{T}{n}$, where n is the number of subintervals for the fixed time interval $[0, T]$. This gives the bid/ask time series pairs,

$$b^i = \begin{pmatrix} b^i_{t_1} & \cdots & b^i_{t_n} \end{pmatrix}, \quad a^i = \begin{pmatrix} a^i_{t_1} & \cdots & a^i_{t_n} \end{pmatrix}, \quad i = 1, \ldots, k.$$

3.2 In-Sample/Forensic Model

We begin with demonstrating an in-sample or forensic approach to extracting the background price formation process. This method will leverage information in an optimal sense over a full time interval. We will compare the results of this approach to an online approach in Section 3.3.

We define a time series low-rank plus sparse model for prices snapshots,

$$m_t = \begin{bmatrix} b^1_t & \cdots & b^k_t & a^1_t & \cdots & a^k_t \end{bmatrix}^T \tag{7}$$

$$m_t = \ell_t + s_t = Uw_t + s_t, \quad t \in [0, T] \tag{8}$$

where k is the number of observed pricing streams, $U \in \mathbb{R}^{2k \times r}$ is an orthogonal subspace matrix of rank $r \ll \min(2k, n)$ and contains the factors of the background price formation proces, $w_t \in \mathbb{R}^r$ are weights to form each bid & ask, and s_t contains the sparse transient deviations at time t.

In matrix form, the time series model becomes

$$M = \begin{bmatrix} m_{t_1} & \cdots & m_{t_n} \end{bmatrix} = UW + S = L + S \tag{9}$$

where W and S contain the stacked columns w_t and s_t, respectively. The matrix $M \in \mathbb{R}^{2k \times n}$ contains regularly-sampled bids and asks from k streams over the time period $[0, T]$. Typically, we resample the data to 10ms and consider relatively short time-windows on the order of 1 or 2 minutes to keep the matrix computations reasonable. Moreover, we consider about 250-300 streams for the more liquid currency pairs. We typically use about $k \sim 300$ pricing streams and $n \sim 6000$ observations which results in a matrix with dimensions about 600×6000.

U, W, and S are computed in-sample over the whole time interval by solving (2). Candès et al. 2011 show that $\lambda = \frac{1}{\sqrt{\max(2k,n)}}$ provides exact recovery of the matrix decomposition. In practice, we take λ to be $\frac{\tilde{\lambda}}{\sqrt{\max(2k,n)}}$ for $\tilde{\lambda} \in (0, 2)$.

3.2.1 Case Study 1: Market Response to a Large AUD/USD Order

The first example we consider is a period around a large buy order AUD/USD. The order was for 50 million AUD and it was filled in about 1 second. We took bid and ask quote data from almost 300 available streams over a 2 minute period surrounding the order and resampled the streams to 10ms.

The structure of the market at this time is rather simple: prices are mostly flat for a period of time at, a particular level, and then undergo short bursts of activity that least to the

establishment of a new price level. Figure 1 shows the bid and ask prices from several market making streams and shows how the prices generally track quite tightly. Figure 2 shows a zoomed in view around the order showing how prices ramp as the order fills due to the large order's price impact.

In Figures 3-5, we extract m_t^i, an individual bank's bid/ask time series from M as well as the low-rank and sparse components, ℓ_t^i and s_t^i, respectively. We show the decompositions of the quoted prices from three streams from three distinct large banks. Each of these streams is chosen because they happen to be streams through which parts of the large order was routed. Moreover, we chose these streams because they help to demonstrate some important phenomena extracted by the low-rank plus sparse algorithm.

For each figure, the panels in the plot follow the same format. In the upper two panels, we look at the prices over the course of the whole two minute period. The left panel features the observed prices and the right panel features the low-rank plus sparse prices, ℓ_t^i and s_t^i. To show where the decomposition has picked up a pricing deviation from the rest of the market, the line is colored black when $|s_t^i| > 0$. The lower four panels show the decomposition for a zoomed in two second period for varying $\tilde{\lambda}$: 1.00 (upper left), 0.75 (upper right), 0.50 (lower left), and 0.25 (lower right). We omit the larger values of $\tilde{\lambda}$ since the sparse component generally disappears and the decompositions and their implications appear more interesting for smaller $\tilde{\lambda}$.

Bank 1: Little to No Sparse Component

In Figure 3, we show the low-rank plus sparse decomposition for a stream from a large bank we will call Bank 1. The main takeaway of the decomposition result is that there is very little deviation from the background price formation process of the market during this order. Even as $\tilde{\lambda}$ decreases, the detected deviation remains relatively small since the observed prices appear to closely fit the background process. If anything, we see that for the smallest $\tilde{\lambda}$, the detected deviation only on the ask side (also the order side) emerges as a relatively consistent

deviation from the background process. Generally, the pricing structure of this stream from Bank 1 tracks the overall market and does not deviate much.

Bank 2: Intelligent Detection of the Large Order

Moving to Figure 4, we show the decomposition for a stream from another large bank, which we call Bank 2. Trades are shown as well, with red triangles indicating trade requests rejected using the Last Look option. The total volume of the requests was about 6mm.

This time, the decomposition shows something strikingly different wherein during the order, the pricing deviation shows a large spike in ask prices (the order side). The zoomed-in view of the prices shows this in more detail and how the spike appears as a near contemporaneous reaction on the ask side to trade requests. We note that for each $\tilde{\lambda}$, the general feature of the spike is captured well but as $\tilde{\lambda}$ decreases, more of the spike is captured while aspects of the pricing on the opposite side are detected suggesting a different kind of pricing deviation.

While for most of this period the prices track the general market, Bank 2 appears to have detected the large order and adjusted its ask prices accordingly indicating the likelihood that order information leaked to the bank. This feat of detecting is interesting in its own right, but one potential problem is that right at the cusp of the large spike is a set of attempted executions on this stream that were rejected by this bank. In a credit relationship exchange model like OTC FX where participants are not usually anonymous, this event represents a style of adversarial market-making or front-running (using the knowledge gained against the counterparty) that, while not regulated under FINRA's definition of front-running, has potential to degrade market quality.

An important caveat here is that cause and effect is difficult to discern exactly since data was resampled and system latencies do need to be accounted for. Ultimately, more evidence would be needed for definitive conclusions and coupling this categorization and data mining technique to detect potential events with other methods would prove very effective.

Two other aspects of the market making related to this order: the behavior from this bank

exists as a form of price discrimination since evidence from the bank's other streams shows that not all contain this spiked pattern at this time. Moreover, the change in pricing behavior during execution creates a fundamental need to determine true cost of exucution as distinct from advertised costs.

Bank 3: More Intelligent Detection of the Large Order

Finally, in Figure 5 we show the decomposition for a stream from yet another large bank, which we call Bank 3. The decomposition shows similar behavior to Bank 2 where during the large order, the a pricing deviation is detected as a large spike in ask prices. Moreover, as for Bank 2, at the cusp of the large spike is a set of attempted executions on this stream that were rejected by this bank. For each $\tilde{\lambda}$ the general feature of the spike is captured and as $\tilde{\lambda}$ decreases, more of the spike is captured. Moreover, there is strong detection of a change in behavior in prices on the opposite side from the order.

As with Bank 2, Bank 3 appears also able to detect the large order and adjust its prices accordingly suggesting that more than likely order information leaked to the bank. The issues discussed for Bank 2 regarding cause and effect related to broader evidence of a widespread behavior as well as adversarial market making, price discrimination, and true costs all carry-over here. A further point to raise that is more evident here than with Bank 2 is the nature of a change in pricing behavior on the opposite side of the order. The way in which prices are altered on the bid side during this buy order suggest a kind of risk averse strategy in which there is a hedging against a potential future sell order arriving and the bank not wishing to be caught buying at a suboptimal price. In the second case study, we see a more clear cut example of this risk averse behavior.

Summary

We finish this case study with a quick summary. We note that background modeling ideas have created a tool to highlight some fundamental market-making behaviors related to order information leakage and detection of a large order, risk averse market making, price

discrimination, and the contrast between true costs and advertised costs.

3.2.2 Case Study 2: Market Response to a Large EUR/USD Order

In this example, we consider a period around another large order, this time a sell order for EUR/USD. The order was for 50 million EUR and it was filled in about 1 second. Again we took quote data from almost 300 available streams but only over a 1 minute period surrounding the order and resampled the streams to 10ms.

The market at this time is now a bit more active as one would expect from a more liquid currency pair. Figure 6 shows the bid and ask prices from several market making streams and as in the previous case study shows how the prices generally track quite tightly. As before, prices still experience the see-sawing effect of being flat and steady for periods of time and then undergoing short bursts of activity that lead to the establishment of new price levels. Figure 7 shows a zoomed in view of the market during the order and prices fall as the order fills and then rebound soon after.

As a note on the visualizations in Figure 6 and 7, the vertical streaks that appear in the images correspond to zero prices, indicating that there are very short periods of times when a market maker is unable to provide both a bid and an ask price. To provide at least one side, the other side is voided by setting the value to 0.

In Figures 8-12, as in the previous example we extract m_t^i, an individual bank's bid/ask time series from M as well as the low-rank and sparse components, ℓ_t^i and s_t^i, respectively. The plots are structured as in the previous example showing the observed prices and detected behaviors for varying $\tilde{\lambda}$. The bank labels used in these examples are not consistent with the previous example. We present five streams from four distinct large banks and one smaller regional bank. Four of these streams were chosen as before because the part of the order was routed to them. However, two of the large bank streams were chosen because despite not "seeing" the part of the order, they serve as an interesting contrast.

Bank 1: Little to No Sparse Component

In Figure 8, we show the low-rank plus sparse decomposition for a stream from a large bank we will call Bank 1. In general, the pricing structure of from Bank 1 tracks the overall market and does not deviate much. There is very little deviation from the background price formation process of the market during this order and even as $\tilde{\lambda}$ decreases, the detected deviations remain relatively small since the observed prices appear to closely fit the background price formation process. If anything, we see that for the smallest $\tilde{\lambda}$, detections on the ask side (also the order side) emerge as a relatively consistent deviation from the background.

Bank 2: Risk Averse Market Making In Figure 9, we now see how during the large order, a large pricing deviation is detected as a relatively large drop in bid prices (the order side). The general pattern is that the bank detects activity in the market and engages in a distinct regime switching behavior. Prices on the leading or order side run from market activity to capture a better price given awareness of trading and prices on the trailing or non-order side follow the market slowly as a hedging strategy for a potential future buy order.

The regime switching behavior is evident given the distinct break in pricing in pattern near the beginning of the order but the speed at which the switching takes place is revealed in later periods at the end and after the order. As prices rebound following the filling of the order, bids and asks widen and narrow as the market maker switches between competitive and risk averse strategies.

As discussed in the previous case study, the observed behaviors suggest there is order information leakage or propagation, likely associated with the inherent connectedness of the overall spot FX markets and the background price formation process appears to trigger the switching behavior. Moreover, these behaviors further show how price discrimination (as before, the banks other streams do not all contain this pattern at this time) is used and how costs at execution, i.e. true costs, are not reflected in advertised costs preceding the order.

Bank 3: Risk Averse Market Making

In Figure 10, for smaller values of $\tilde{\lambda}$ we again see a detected regime switching behavior where the market maker uses a similar hedging strategy on trailing side prices as in the previous example. The risk aversion only exists on one side of prices, the non-order side, creating an assymetry in the pricing. With respect to prices on the order side, in this case the bid, prices remain competitive and in track with the background price formation process. When market prices reverse direction, the detected behavior reveals a strategy that quickly switches and begins the hedging strategy on the opposite side and even though we highlight the detected behavior around the large order, it appears all throughout the 1 minute period during market movements.

Bank 4: Liquidity Replenishment

In Figure 11, we again encounter a detected drop in the bid prices but this time the drop is subtle and appears to be timed *following* filling about 10 million of the order. The behavior becomes more apparent for lower values of $\tilde{\lambda}$. As opposed to previous examples of order information leakage or risk averse market making, the detected behavior suggests a liquidity replenishment in response to filling the order. After a short period, the bid side is able to rebound and reset to following the market.

We see this potentially as a demonstration of the concept of price resiliency as covered in Large 2007 as well as inventory control, one of the core aspects of algorithmic market making. The mechanism here for inventory control is to manuever bid and ask prices relative to other market makers, making either side more attractive to counterparties and thus encouraging trading in a particular direction. Due to the filling of trade requests, liquidity was used up so the bank put a premium on more selling. Eventually the skewed pricing fades and thus liquidity appears to have replenished.

This is an example of how asymmetric pricing of the bid and ask is used in response to filled requests. The bank shades or skews its bid and ask prices relative to the general

opinion of the market midrate. It is difficult to see but following the order, the bank's asking price is relatively more attractive. Models for asymmetric pricing are developed in Zhang, Russell, and Tsay 2008 and Hasbrouck 1999 but we note here that the low-rank plus sparse decomposition has not perfectly captured this phenonmen.

Bank 5: Lagged Pricing

Finally, in Figure 12 we come across a set of detected behaviors that appear wholly different from anything we have seen so far. Zooming in on the short period around the order gives a clear view of the phenomenon. In this particular case, prices appear lagged against the background signal. As $\tilde{\lambda}$ decreases, there are fewer component signals in the background price formation process to create the observed lag effect so the prices during movements appear in the sparse component. For smaller values of $\tilde{\lambda}$, one can see how a shadow forms behind the background and the observed prices. In sum, it appears as though pricing here is slow and behind the market. Detecting this lag behavior actually has an analogue in imaging applications of background subtraction: misaligned images would show similar traces that would be cleared up with proper alignment.

Summary

This second case study shows more examples of how background modeling ideas have created a tool to highlight fundamental market-making behaviors related to order information leakage and detection of a large order, risk averse market making, price discrimination, and the contrast between true costs and advertised costs. Moreover, we saw two new behaviors, an example of inventory control and lagged market making.

3.3 Online/Sequential Model

The low-rank plus sparse decomposition we have applied so far has assumed a fixed subspace U in (8) and (9). That approach is in-sample in nature, in that the estimated weights

and sparse componenent depend on data over the full time period and information is used "optimally". One thing we have seen is that transient deviations manifest themselves as complex high dimensional change-points of time series. A natural question is what if we want to be able to monitor and react fast, at or near the speed of the market? Can we instead pose this problem with sequential observations in time and detect change-points on the fly to the same degree as was done using an in-sample approach? This is what we address here.

While online approaches to background subtraction are common, we will limit to one here that represents a natural extension from the previous section's static model. Recall the static time series model of (8) for in-sample analysis,

$$m_t = Uw_t + s_t, \quad t \in [0, T]$$

where U is the low-rank subspace for the background, w_t are the weights for composing the background prices, and s_t are the transient pricing deviations. We can instead try to frame the problem in a dynamic sense with an evolving subspace. This is given by,

$$m_t = U_t w_t + s_t \tag{10}$$

where U_t is modeled as an evolving background *dependent only on past information*. We can relate allowing the subspace U to evolve with the video analogy of tracking a background scene that undergoes a change like lights turning on or off. The goal is to make sequential observations of m_t and perform the update, $(U_{t-1}, w_{t-1}, s_{t-1}) \to (U_t, w_t, s_t)$.

This type of dynamic subspace approach is known as *subspace tracking* for tracking and updating an evolving subspace from sequential data observations. Early forms of the subspace tracking problem can be found in Comon and Golub 1990, Delmas and Cardoso 1998, and Mathew, Reddy, and Dasgupta 1995, where the problem is generally framed as trying to

update eigenvalue/eigenspace estimates for a covariance matrix for the data. To address robust subspace tracking where there are sparse deviations that one wants detected, He, Balzano, and Lui 2011, Mateos and Giannakis 2010, and Qiu and Vaswani 2011 all propose approaches that instead take a form as in (10).

We applied the subspace tracking method GRASTA of He, Balzano, and Lui 2011 to the previous case studies. GRASTA is a non-SVD based approach to separating the background and foreground scenes. The algorithm sequentially solves the Augmented Lagrangian given a new observation m_t,

$$\mathcal{L}(w, s, y, U; m_t) = \|s\|_1 + y^T(Uw + s - m_t) + \frac{\rho}{2}\|Uw + s - m_t\|_2^2,$$

using an alternating minimization routine. Given m_t and a previous subspace estimate U_{t-1}, minimize $\mathcal{L}$ with respect to (w, s, y) to obtain the updated (w_t, s_t, y_t). Then given m_t and (w_t, s_t, y_t), perform a gradient step for U to obtain the updated U_t. In contrast to the Robust PCA approach that utilized rank minimization, here the rank of U_t is fixed. The exact details of the algorithm can be found in Section 6.2.

3.3.1 Comparing Online and In-Sample Models

Figures 13 to 15 show a comparison of the results of the in-sample computation of Robust PCA from Section 2. As in Section 2, we use a 1 minute period around a large EUR/USD order and increase sensitivity a bit by taking $\tilde{\lambda} = 0.25$ since visually we find the results are more compelling in how idiosyncratic pricing behaviors are captured.

We applied GRASTA to a full 2 minute period including around the order. Since the method is online, we first train the subspace estimate on the beginning 5 second period, and then run GRASTA for another 55 seconds before running and comparing it for the 1 minute period. We applied the standard proposed algorithm from He, Balzano, and Lui 2011. Since

Robust PCA generally obtained low-rank matrices of rank 3 to 10, we applied GRASTA with subspace ranks of $2 \sim 5$. Results end up quite similar to each other for different ranks of U_t

Figures 13 shows the example of the Bank 2 in the first case study that appears to have detected the large order. The large spike in ask prices is correctly identified showing the online approach is quite capable of replicating the in-sample result. However, as the order execution finishes, the estimated subspace adapts to the change in pricing strategy and the results between the online and the in-sample models begin to differ.

Figure 14 shows the example of Bank 2 in the second case study that showed an example of risk averse market making. The widened, risk averse prices are again detected by the online method. As in the previous examination, detection of the behavior change begins to falter as the order execution finishes: the updating subspace absorbs part of the detected pattern and obscures the clear view of the behavior. One can see further evidence of this as the behavior quickly switches between behaviors, the online model fails to properly capture the behavior due to the subspace adaptation.

Lastly, Figure 15 shows the example of Bank 4 in the second case study that exhibited a liquidity replenishment/inventory control behavior. The online model appears to correctly identify the liquidity replenishment/inventory control behavior compared to the in-sample estimate. But a recurring theme appears, naturally due to the nature of the adapting subspace approach: results from the online model differ from in-sample computation as time goes on due to subspace adaptation.

Overall, this leads to a natural consideration of the positives and the challenges of trying to detect behaviors in an online manner. As an example of an online approach, GRASTA is very fast and efficient. Moreover, it shows remarkable potential to detect on the fly when compared to the in-sample computation. However, there appears to be sub-optimal calibration of the adapting subspace routine, at least with respect to the problem space of

background price formation and pricing deviations.

While much could be done to tune the calibration of the subspace adaptation, this really gets to a larger point: what in fact constitutes a change in the background price formation process such that there should be an adaptation in the model? How would one distinguish a background change from a transient deviation in this problem context? In relation to the video surveillance problem, this directly corresponds to the challenge of how one would handle a foreground object that stops moving.

Further drawing on this connection to video surveillance, what we find is we are raising questions that have directly analogous questions in the video surveillance problem. But one key difference that needs to be addressed is that these types of questions have answers in the canonical video surveillance problem.

4 Discussion: Challenges in Background Subtraction

Many challenges in background subtraction for video surveillance are well known with some prominent examples noted in Bouwmans and Zahzah 2014, among them illumination changes, inserted background objects, dynamic backgrounds, and a sleeping foreground object. High frequency financial markets will present their own set of challenges and we expect some of these challenges in fact translate over from background subtraction for video surveillance. As seen with applying an online approach, this new application for background subtraction already shows how the challenge of a sleeping foreground object naturally translates over due to the temporal nature of the detection and how a naive subspace adaptation will fail to account for this.

This notion of how a detected pricing deviation relates to a known challenge in video surveillance suggests the need to develop a taxonomy for background subtraction in financial markets. A taxonomy would allow for classification and interpretation of idiosyncratic pricing

behaviors, as in the case studies. Moreover, it would also classify the nature of the background price formation process and any phenomena that might exist as well as the persistent challenges present in the data that must be overcome to provide reliable results. A taxonomy would further allow for the automated classification of phenomena in an analytics pipeline.

Another fundamental challenge we see stems from our deliberate choice to simplify the nature of the data produced in pricing streams. First, we grouped all pricing streams from different banks as well as bids and asks together. Moreover, while each quote is a full book of orders with varying degrees of depth and shape, we merely extracted top of book prices, ignoring size and resting orders deeper in the book. Tackling the deeper data issues of the more complex data structures that exist in reality is fundamentally necessary for giving a complete perspective of markets.

There are broader questions beyond the specifics of modeling and algorithms for background subtraction as well as the complex data structures. How does background subtraction apply in other markets? One should expect given a proper dataset, this methology should translate over and yield similar results. Still, this remains to be addressed. Furthermore, as demonstrated with the online model, how do we extend these ideas to the broader problem of realtime anomaly detection of instabilities due to fast algorithmic trading.

5 Summary

This work represents a new approach to broadening the understanding of electronic financial markets and how pricing formation occurs. We demonstrated how applying a low-rank plus sparse background subtraction method to high frequency foreign exchange quote data can learn a composition of background signals from a collection of currency prices and reveals pricing deviations in the market. We presented case studies that showed how these pricing deviations appear and interpretations for these phenomena and we further demonstrated

how online approaches can be successful as well.

6 Appendix

6.1 Principal Components Pursuit

To solve (2), Candès et al. 2011 first set up the augmented Lagrangian,

$$\ell(L, S, Y) = \|L\|_* + \lambda\|S\|_1 + \langle Y, M - L - S\rangle + \frac{\mu}{2}\|M - L - S\|_F.$$

The augmented Lagrangian can be solved via Alternating Direction Method of Multipliers (ADMM, see Boyd et al. 2011). This is defined by iterated steps of the updating formulae,

$$\begin{aligned} L_{k+1} &= \mathcal{D}_{1/\mu}(M - S_k + \mu^{-1}Y_k) \\ S_{k+1} &= \mathcal{S}_{\lambda/\mu}(M - L_{k+1} + \mu^{-1}Y_k) \\ Y_{k+1} &= Y_k + \mu rho(M - L_{k+1} + S_{k+1}) \end{aligned}$$

where $\mathcal{D}_\mu$ is the singular value thresholding operator and $\mathcal{S}_\mu$ is the element-wise thresholding operator.

6.2 GRASTA Algorithm

This is a brief summary of the algorithm proposed in He, Balzano, and Lui 2011, GRASTA. In (10), U_t, w_t, and s_t all need to be tracked and updated. In order to satisfy the sparsity requirement for s_t, He, Balzano, and Lui 2011 set up the augmented Lagrangian for the problem as,

$$\mathcal{L}(w, s, y, U; m_t) = \|s\|_1 + y^T(Uw + s - m_t) + \frac{\rho}{2}\|Uw + s - m_t\|_2^2, \tag{11}$$

where m_t is the vector of new price observations. In general, a joint optimization on w, s, y, and U is not realistic. Instead, He, Balzano, and Lui 2011 propose a two-step approach: first solve w, s, y holding U fixed to the previous subspace estimate U; then solve for U holding w, s, y fixed to the newly updated estimates. More specifically, given a new set of price observations m_t, then in the first step a solution can be produced quickly using ADMM,

$$\begin{aligned} w_t &= \frac{1}{\rho} {U_{t-1}}^T(\rho(m_t - s_{t-1}) - y_{t-1}) \\ s_t &= \mathcal{S}_{\frac{1}{1+\rho}}(m_t - U_{t-1}w_t - y_{t-1}) \\ y_t &= y_{t-1} + \rho(U_{t-1}w_t + s_t - m_t) \end{aligned}$$

where $S_\mu(x) = \text{sign}(x)\max\{|x| - \mu, 0\}$ is the elementwise soft thresholding operator.

In the second step, He, Balzano, and Lui 2011 propose an incremental gradient descent for updating the subspace using the augmented Lagrangian in (11). The gradient is defined as,

$$\begin{aligned} \Gamma &= (I - UU^T)\left[y_t + \rho(Uw_t + s_t - m_t)\right] \\ \nabla\mathcal{L}(U) &= \Gamma {w_t}^T \end{aligned}$$

and the updating formula is given as,

$$U_t = U_{t-1} + \left((\cos(\eta\sigma) - 1)\frac{U_{t-1}w_t}{\|w_t\|} - \sin(\eta\sigma)\frac{\Gamma}{\|\Gamma\|}\right)\frac{w_t}{\|w_t\|},$$

where η is a step size and σ is the singular value for the rank 1 gradient, $\nabla\mathcal{L}(U)$. He, Balzano, and Lui 2011 propose an adaptive step size mechanism but we feel the details of this are beyond the spirit of capturing the essence of the method. See Section 3.4 of He, Balzano, and Lui 2011 for details on choosing the step size.

References

At-Sahalia, Yacine and Dacheng Xiu (2015). *Principal component analysis of high frequency data*. Tech. rep. National Bureau of Economic Research.

Baron, Matthew, Jonathan Brogaard, and Andrei A. Kirilenko (2014). "Risk and return in high frequency trading". In: *Available at SSRN 2433118*.

Bouwmans, Thierry and El Hadi Zahzah (2014). "Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance". In: *Computer Vision and Image Understanding* 122, pp. 22–34.

Boyd, Stephen et al. (2011). "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends in Machine Learning* 3.1, pp. 1–122.

Boyd, Stephen P. and Lieven Vandenberghe (2004). *Convex optimization*. Cambridge, UK ; New York: Cambridge University Press. ISBN: 0521833787.

Brogaard, J., T. Hendershott, and R. Riordan (2014). "High-Frequency Trading and Price Discovery". en. In: *Review of Financial Studies* 27.8, pp. 2267–2306. ISSN: 0893-9454, 1465-7368.

Candès, Emmanuel J. et al. (2011). "Robust principal component analysis?" In: *Journal of the ACM (JACM)* 58.3, p. 11.

Carrion, Allen (2013). "Very fast money: High-frequency trading on the NASDAQ". In: *Journal of Financial Markets* 16.4, pp. 680–711.

Child, Dennis (2006). *The essentials of factor analysis*. A&C Black.

Clark-Joseph, Adam Daniel (2013). "Three Essays on Trading Behavior". PhD thesis.

Comon, Pierre and Gene H. Golub (1990). "Tracking a few extreme singular values and vectors in signal processing". In: *Proceedings of the IEEE* 78.8, pp. 1327–1343.

Delmas, Jean Pierre and Jean Franois Cardoso (1998). "Performance analysis of an adaptive algorithm for tracking dominant subspaces". In: *Signal Processing, IEEE Transactions on* 46.11, pp. 3045–3057.

Easley, David, Marcos Lopez de Prado, and Maureen O'Hara (2011). "The microstructure of the Flash Crash: Flow toxicity, liquidity crashes and the probability of informed trading". In: *The Journal of Portfolio Management* 37.2, pp. 118–128.

Gonzalo, Jesus and Clive Granger (1995). "Estimation of common long-memory components in cointegrated systems". In: *Journal of Business & Economic Statistics* 13.1, pp. 27–35.

Grimm, Laurence G. and Paul R. Yarnold (1995). *Reading and understanding multivariate statistics.* American Psychological Association.

Hasbrouck, Joel (1995). "One security, many markets: Determining the contributions to price discovery". In: *The journal of Finance* 50.4, pp. 1175–1199.

Hasbrouck, Joel (1999). "The dynamics of discrete bid and ask quotes". In: *The Journal of finance* 54.6, pp. 2109–2142.

He, Jun, Laura Balzano, and John Lui (2011). "Online robust subspace tracking from partial information". In: *arXiv preprint arXiv:1109.3827*.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld (2011). "Does algorithmic trading improve liquidity?" In: *The Journal of Finance* 66.1, pp. 1–33.

Large, Jeremy (2007). "Measuring the resiliency of an electronic limit order book". In: *Journal of Financial Markets* 10.1, pp. 1–25.

Laruelle, Sophie, Charles-Albert Lehalle, and Gilles Pages (2011). "Optimal split of orders across liquidity pools: a stochastic algorithm approach". In: *SIAM Journal on Financial Mathematics* 2.1, pp. 1042–1076.

Lin, Zhouchen, Minming Chen, and Yi Ma (2010). "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices". In: *arXiv preprint arXiv:1009.5055*.

Mateos, Gonzalo and Georgios B. Giannakis (2010). "Sparsity control for robust principal component analysis". In: *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*. IEEE, pp. 1925–1929.

Mathew, George, Vellenki U. Reddy, and Soura Dasgupta (1995). "Adaptive estimation of eigensubspace". In: *Signal Processing, IEEE Transactions on* 43.2, pp. 401–411.

Oliver, Nuria, Barbara Rosario, and Alex Pentland (1998). "A Bayesian Computer Vision System for Modeling Human Interactions". In: *CVPR, The Interpretation of Visual Motion Workshop*. Citeseer, pp. 39–46.

Pelger, Markus (2015). "Understanding systematic risk: A high-frequency approach". In: *Available at SSRN*.

Pison, Greet et al. (2003). "Robust factor analysis". en. In: *Journal of Multivariate Analysis* 84.1, pp. 145–172. ISSN: 0047259X.

Qiu, Chenlu and Namrata Vaswani (2011). "Reprocs: A missing link between recursive robust pca and recursive sparse recovery in large but correlated noise". In: *arXiv preprint arXiv:1106.3286*.

Recht, Benjamin, Maryam Fazel, and Pablo A. Parrilo (2010). "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". In: *SIAM review* 52.3, pp. 471–501.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Yang, Steve et al. (2012). "Behavior based learning in identifying high frequency trading strategies". In: *Computational Intelligence for Financial Engineering & Economics (CIFEr), 2012 IEEE Conference on*. IEEE, pp. 1–8.

Zhang, Jianchun, Jia Li, and Chuanhai Liu (2014). "Robust factor analysis using the multivariate t-distribution". In: *Statistica Sinica* 24, pp. 291–312.

Zhang, Michael Yuanjie, Jeffrey R. Russell, and Ruey S. Tsay (2008). "Determinants of bid and ask quotes and implications for the cost of trading". In: *Journal of Empirical Finance* 15.4, pp. 656–678.

Zhou, Zihan et al. (2010). "Stable principal component pursuit". In: *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE, pp. 1518–1522.

7 Figures

7.1 Case Study 1: Market Response to a Large AUD/USD Order

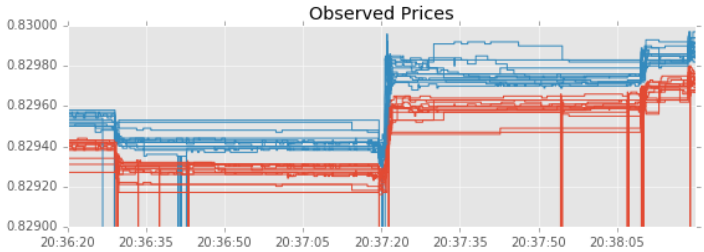

Figure 1: Snapshot of the market for AUD/USD in a 2 minute period around a large buy order. The prices shown represent those through which part of the order was at least attempted to be routed. Bid prices are in red, ask prices are in blue. The market is relatively stable for periods at a time before shifting levels. The downward streaks represent zero values that indicate the market maker stopped pricing on either the bid, the ask, or both.

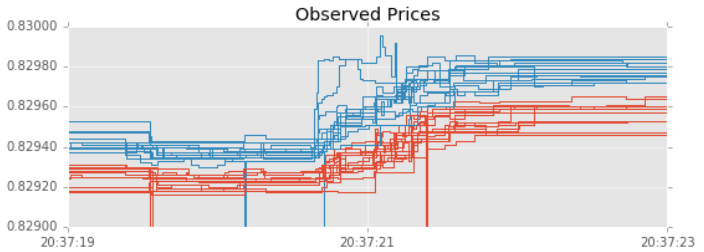

Figure 2: A zoomed in snapshot of the market 4 seconds around the large buy order. The overall pattern is a ramp up in prices as the buy order fills over the course of a second. One can see though that two market makers price out of the norm with respect to the rest of the market. We look at the results of those two market makers.

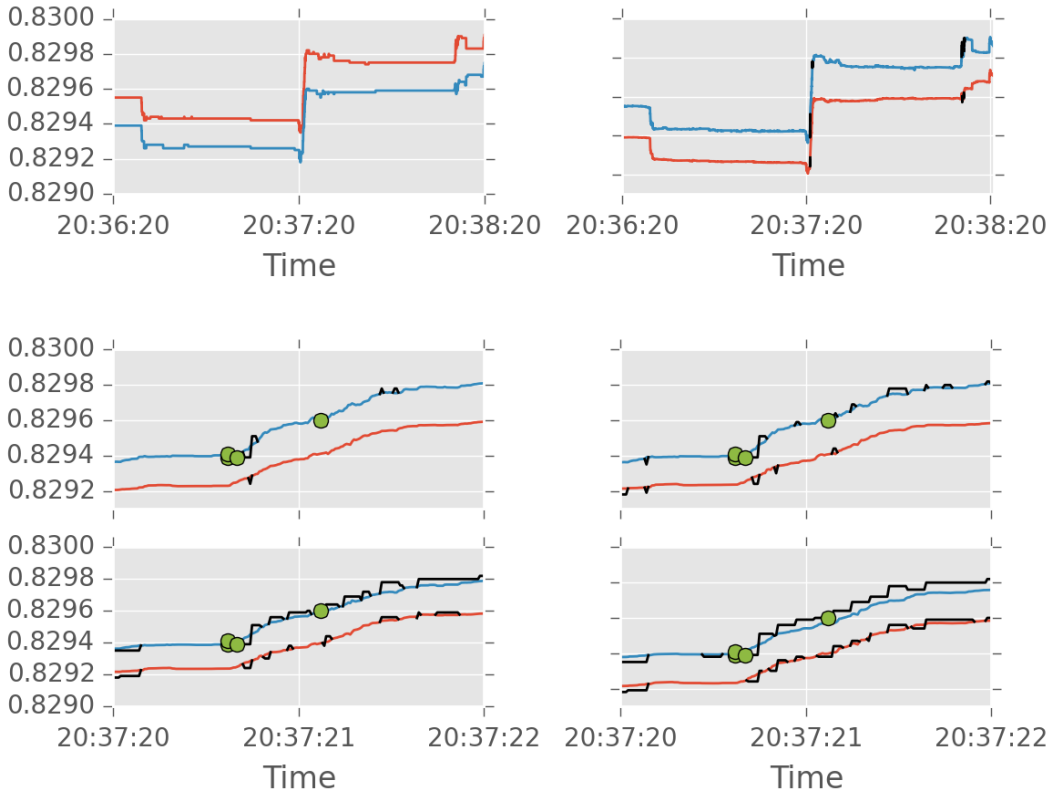

Figure 3: Low-rank plus sparse results for Bank 1. Upper two panels: observed bid and ask prices (left) over the two minute period show the general market behavior; low-rank plus sparse prices (right) for $\tilde{\lambda} = 1$ show the added sparse component (black) is minimal. Lower four panels: low-rank plus sparse prices for zoomed in two second period for $\tilde{\lambda} = 1$ (upper left), 0.75 (upper right), 0.50 (lower left), and 0.25 (lower right). Even as $\tilde{\lambda}$ decreases, the sparse component stays relatively small since the observed prices appear to closely fit the low-rank component. If anything, we see that for the smallest $\tilde{\lambda}$, the sparse component only on the ask side (also the order side) emerges as a relatively consistent deviation its low-rank component.

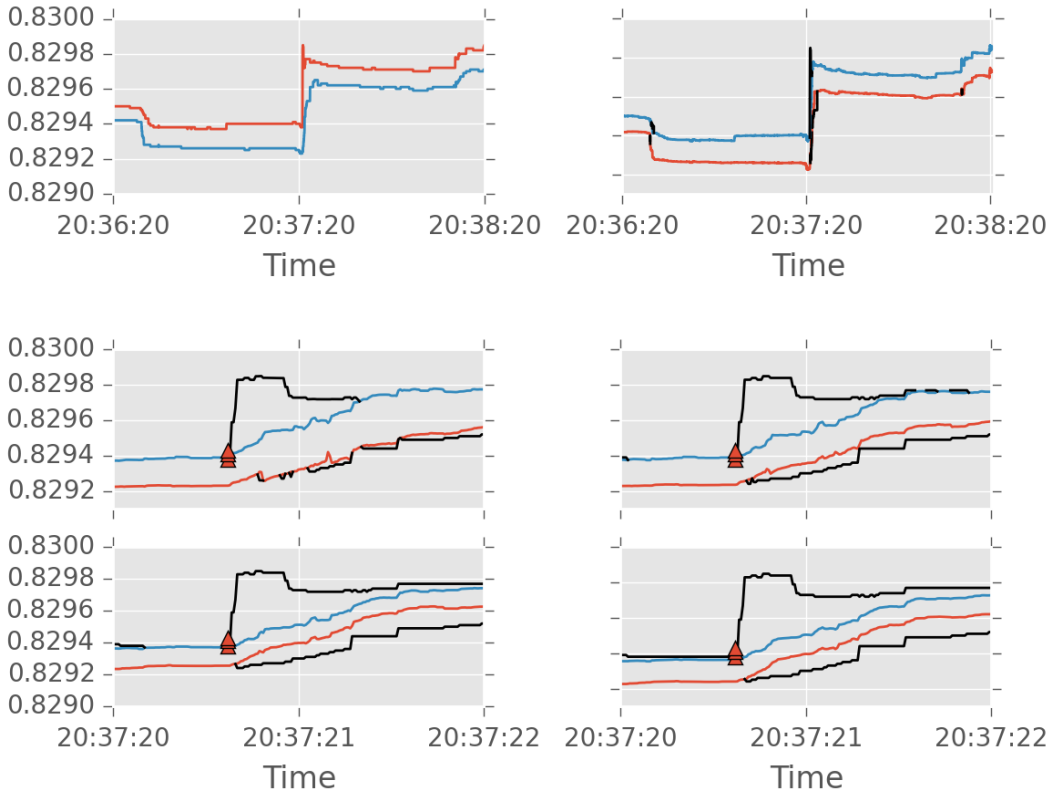

Figure 4: Low-rank plus sparse results for Bank 2. Upper two panels: observed bid and ask prices (left) over the two minute period show the general market behavior with a spike at the order in the middle; low-rank plus sparse prices (right) for $\tilde{\lambda} = 1$ show the spike labeled in the sparse component (black). Lower four panels: low-rank plus sparse prices for zoomed in two second period for $\tilde{\lambda} = 1$ (upper left), 0.75 (upper right), 0.50 (lower left), and 0.25 (lower right). Zoomed in, the spike appears as an immediate reaction on the ask side to three rejected trade requests (red triangle). For each $\tilde{\lambda}$, the general feature of the spike is captured. As $\tilde{\lambda}$ decreases, more of the spike is captured but also prices on the opposite side are picked up in the sparse component suggesting deviation from pricing there as well.

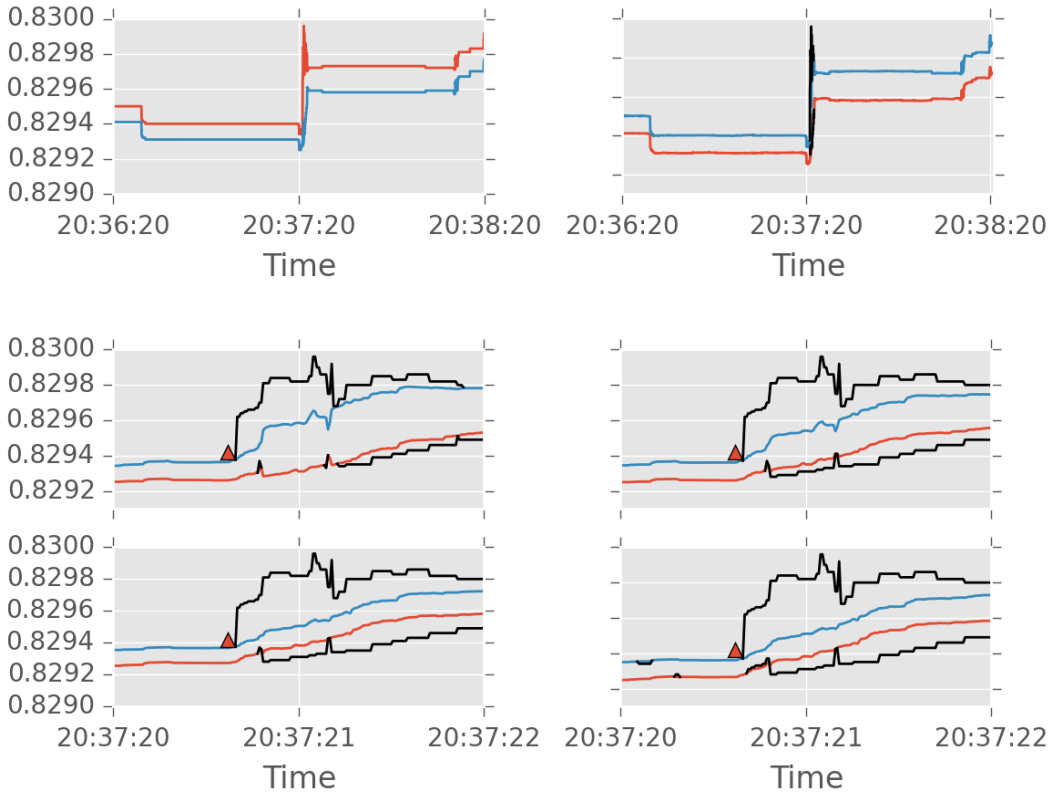

Figure 5: Low-rank plus sparse results for Bank 3. Upper two panels: observed bid and ask prices (left) over the two minute period show the general market behavior with a spike at the order in the middle; low-rank plus sparse prices (right) for $\tilde{\lambda} = 1$ show the spike labeled in the sparse component (black). Lower four panels: low-rank plus sparse prices for zoomed in two second period for $\tilde{\lambda} = 1$ (upper left), 0.75 (upper right), 0.50 (lower left), and 0.25 (lower right). Similar to Figure 4, when zoomed in the spike appears as an immediate reaction on the ask side to a rejected trade request (red triangle). Again, for each $\tilde{\lambda}$ the general feature of the spike is captured and as $\tilde{\lambda}$ decreases, more of the spike is captured. Also again prices on the opposite side are picked up in the sparse component suggesting deviation from pricing there as well.

7.2 Case Study 2: Market Response to a Large EUR/USD Order

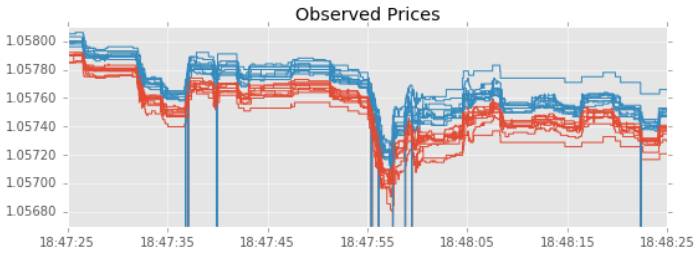

Figure 6: Snapshot of the market for EUR/USD in a 1 minute period around a large sell order. Bid prices are in red, ask prices are in blue. The prices shown represent those through which part of the order was at least attempted to be routed. The market is relatively active during this period, with stable periods being relatively short compared to Case Study 1. As before, the downward streaks represent zero values that indicate the market maker stopped pricing on either the bid, the ask, or both.

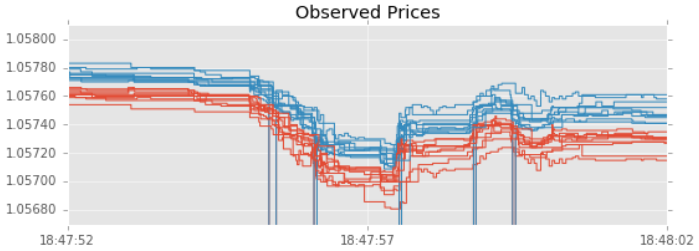

Figure 7: A zoomed in snapshot of the market 10 seconds around the large sell order. To the naked eye, nothing seems particularly out of the ordinary as the prices drop during the sell order. Looking at individual banks, we will see what the detections can reveal.

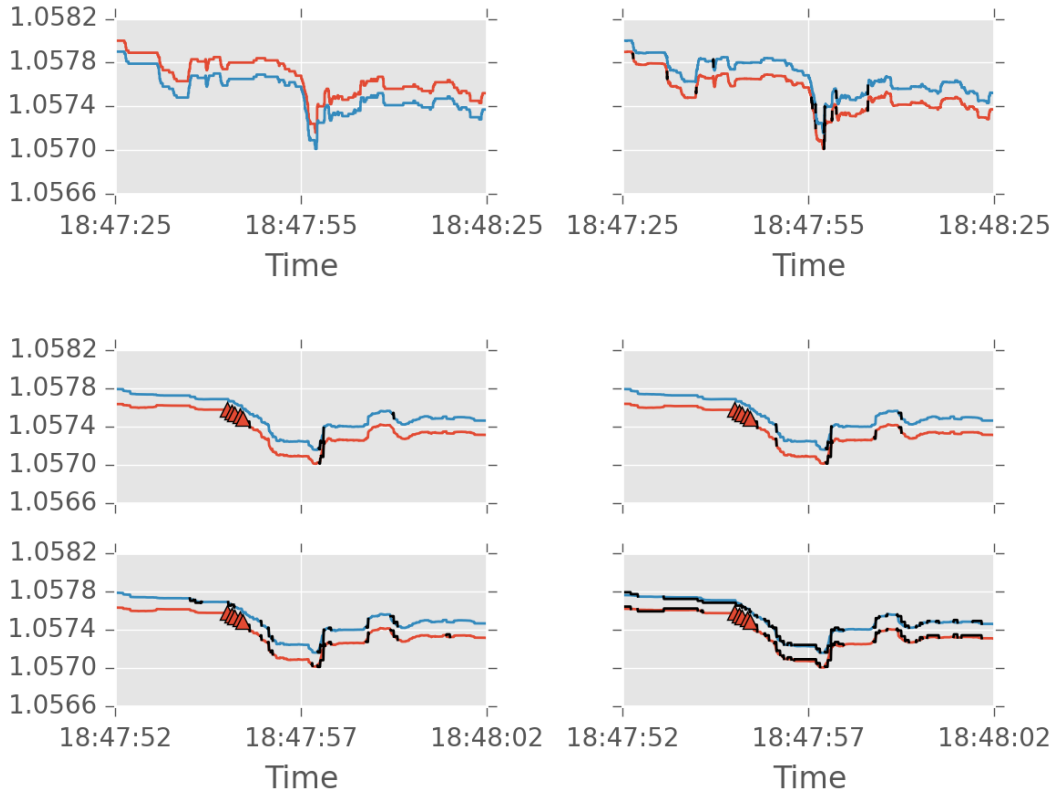

Figure 8: Low-rank plus sparse results for Bank 1. Upper two panels: observed bid and ask prices (left) over the two minute period show the general market behavior; low-rank plus sparse prices (right) for $\tilde{\lambda} = 1$ show the added sparse component (black) is minimal. Lower four panels: low-rank plus sparse prices for zoomed in two second period for $\tilde{\lambda} = 1$ (upper left), 0.75 (upper right), 0.50 (lower left), and 0.25 (lower right). Even as $\tilde{\lambda}$ decreases, the sparse component stays relatively small since the observed prices appear to closely fit the low-rank component. If anything, we see that for the smallest $\tilde{\lambda}$, the sparse component only on the ask side (also the order side) emerges as a relatively consistent deviation its low-rank component.

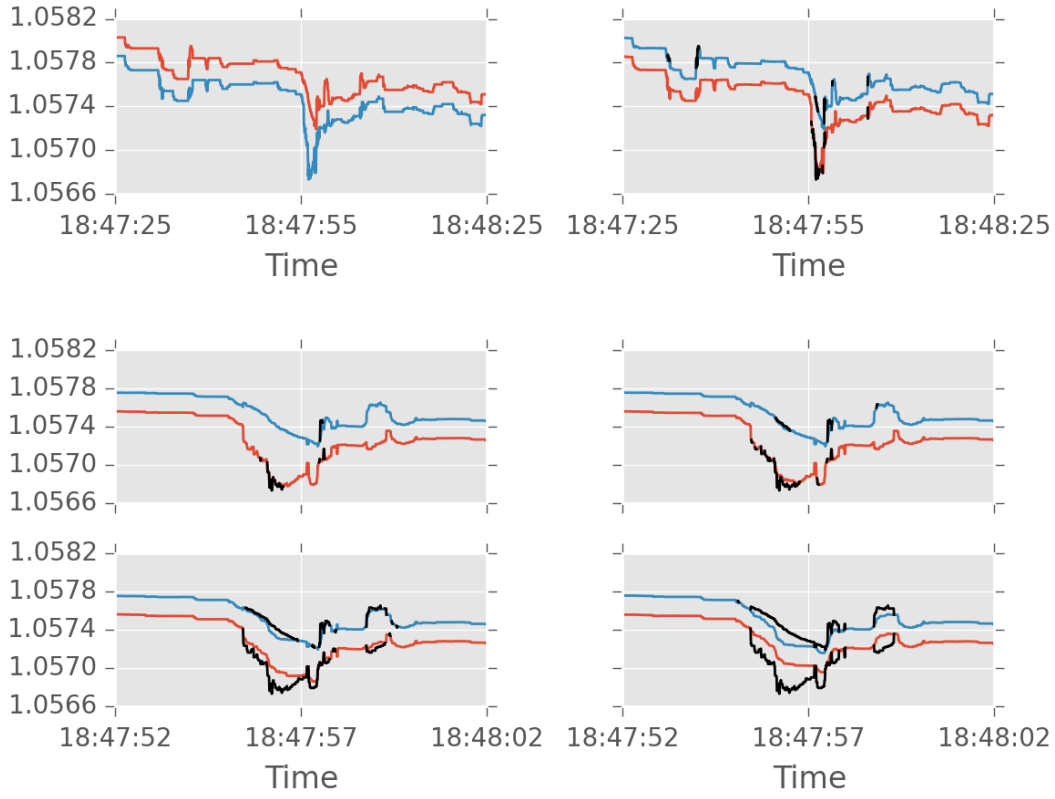

Figure 9: Low-rank plus sparse results for Bank 2. Upper two panels: observed bid and ask prices (left) over the two minute period show the general market behavior with a spike at the order in the middle; low-rank plus sparse prices (right) for $\tilde{\lambda} = 1$ show the spike labeled in the sparse component (black). Lower four panels: low-rank plus sparse prices for zoomed in two second period for $\tilde{\lambda} = 1$ (upper left), 0.75 (upper right), 0.50 (lower left), and 0.25 (lower right). Zoomed in, the ask falls away as an immediate reaction to the large order but note that no trades were routed to this bank. As $\tilde{\lambda}$ increases, the general shape of the drop is captured. Moreover, a slow trailing ask price is picked up that indicates a bit of risk hedging.

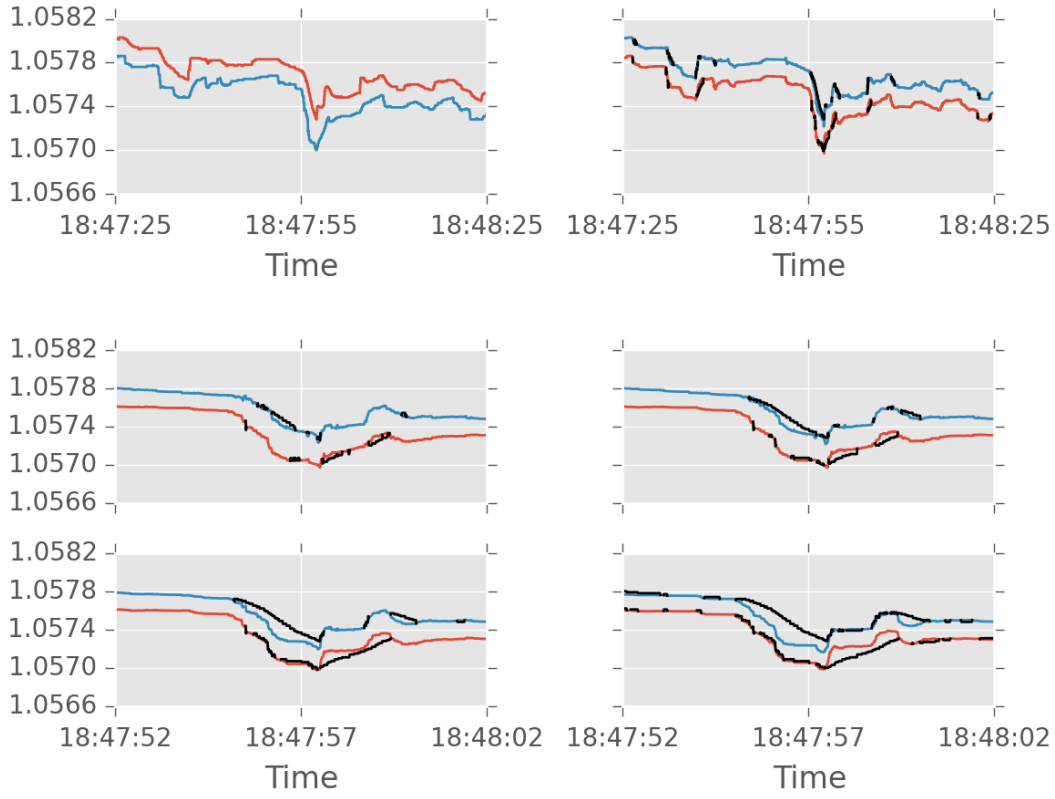

Figure 10: Low-rank plus sparse results for Bank 3. Upper two panels: observed bid and ask prices (left) over the two minute period show the general market behavior with a spike at the order in the middle; low-rank plus sparse prices (right) for $\tilde{\lambda} = 1$ show the spike labeled in the sparse component (black). Lower four panels: low-rank plus sparse prices for zoomed in two second period for $\tilde{\lambda} = 1$ (upper left), 0.75 (upper right), 0.50 (lower left), and 0.25 (lower right). A slow trailing ask price, as in Figure 9, is picked up quite strongly. By slowly following on the opposite side of trading, the bank is able to hedge against other clients buying.

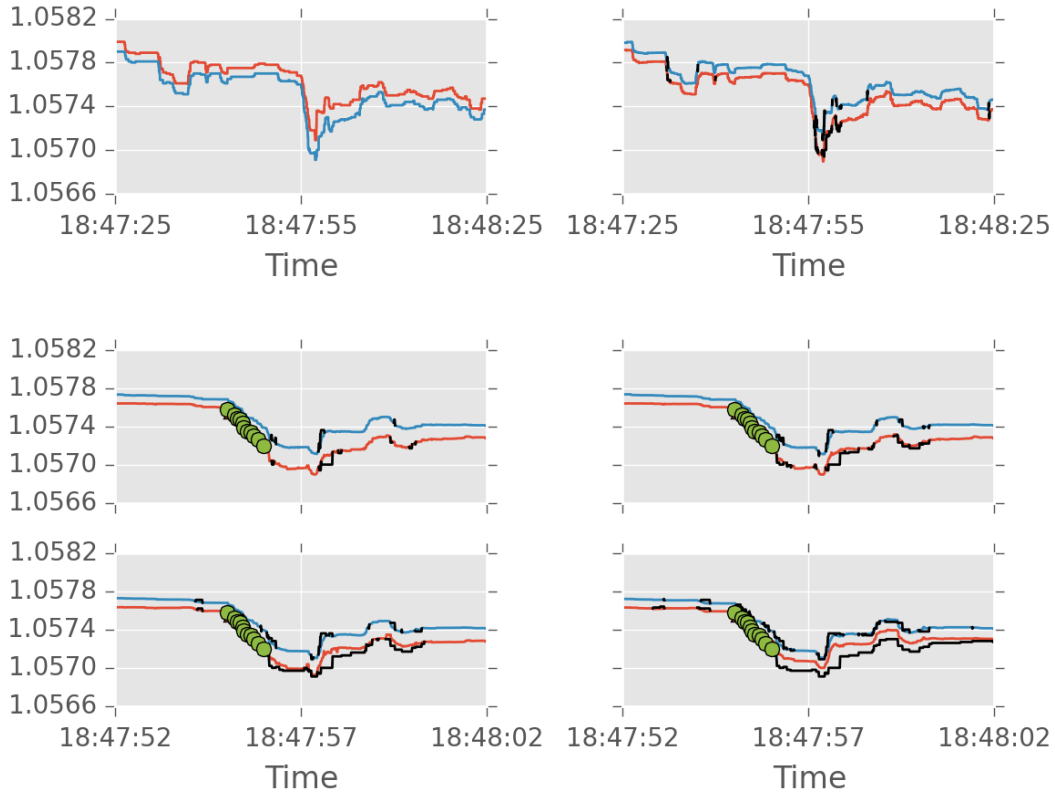

Figure 11: Low-rank plus sparse results for Bank 3. Upper two panels: observed bid and ask prices (left) over the two minute period show the general market behavior with a spike at the order in the middle; low-rank plus sparse prices (right) for $\tilde{\lambda} = 1$ show the spike labeled in the sparse component (black). Lower four panels: low-rank plus sparse prices for zoomed in two second period for $\tilde{\lambda} = 1$ (upper left), 0.75 (upper right), 0.50 (lower left), and 0.25 (lower right). Following a fill of about 10mil, prices on the bid side drop and eventually recover back to tracking the rest of the market. There is liquidity replenishment action that follows the trading and prices become skewed in order for the bank to cover the volume. It cannot be seen in these images and it is difficult to pick up with the low-rank plus sparse decomposition but even to the end of the period, there is price skewing.

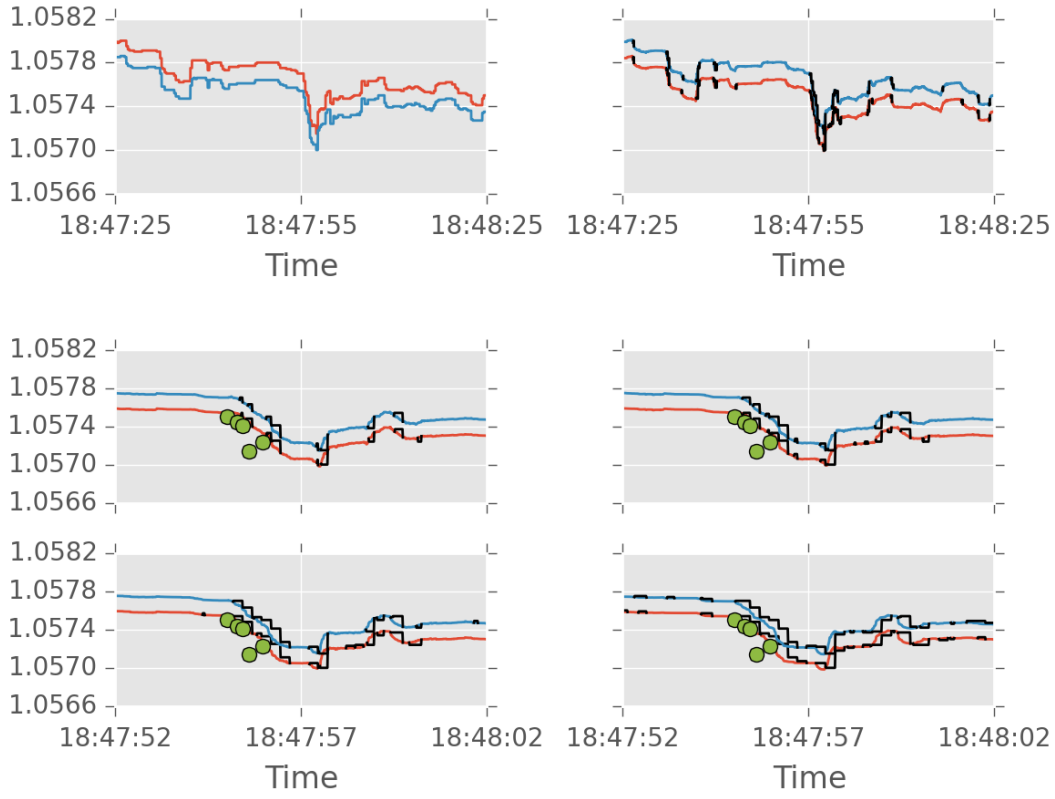

Figure 12: Low-rank plus sparse results for Bank 3. Upper two panels: observed bid and ask prices (left) over the two minute period show the general market behavior with a spike at the order in the middle; low-rank plus sparse prices (right) for $\tilde{\lambda} = 1$ show the spike labeled in the sparse component (black). Lower four panels: low-rank plus sparse prices for zoomed in two second period for $\tilde{\lambda} = 1$ (upper left), 0.75 (upper right), 0.50 (lower left), and 0.25 (lower right). This time, there is no risk averse response to trading. Instead, prices are lagged from the market by a small amount of time. As $\tilde{\lambda}$ decreases, one sees this sharper and sharper. The detected sparse component begins to form a shadow from the low-rank component indicating that if one were to just shift prices back in time, there would be no sparse component.

7.3 Comparison Between Online and In-Sample Models

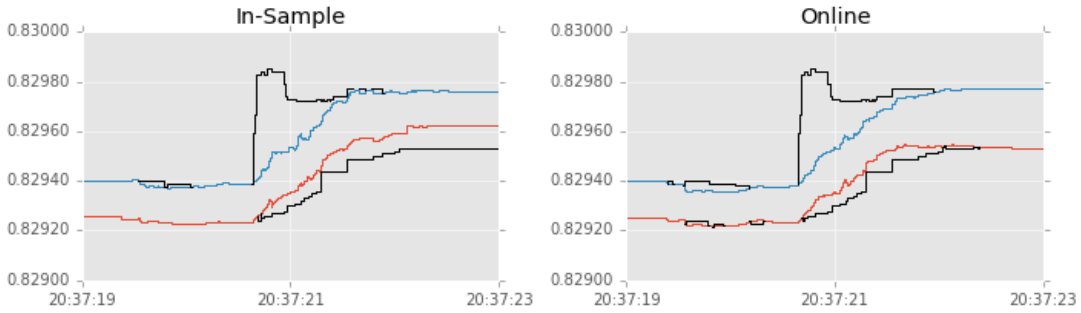

Figure 13: In-sample and online results for Bank 2 from the first case study. Estimated low-rank bids and asks are in red and blue, respectively, and detected deviations are in black. The large spike in ask prices is correctly identified by the online approach. As the order execution finishes, the estimated subspace adapts to the change in pricing strategy and the results between the online and the in-sample models begin to differ.

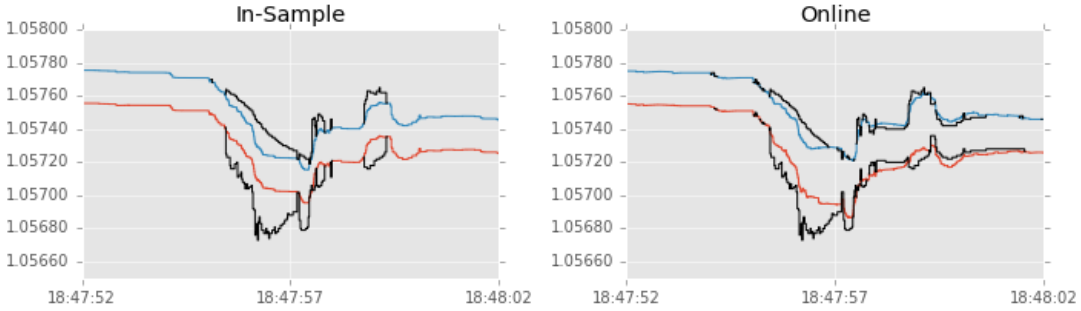

Figure 14: In-sample and online results for Bank 2 from the second case study. Estimated low-rank bids and asks are in red and blue, respectively, and detected deviations are in black. The risk averse prices are again detected by the online method. Detection of the behavior change begins to falter as the order execution finishes: the updating subspace absorbs part of the detected pattern and obscures the clear view of the behavior. As the behavior quickly switches between behaviors, the online model fails to properly capture the behavior due to the subspace adaptation.

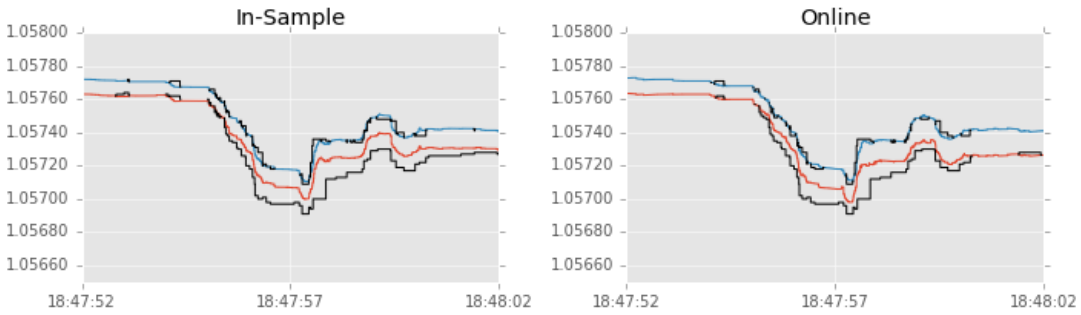

Figure 15: In-sample and online results for Bank 4 from the second case study. Estimated low-rank bids and asks are in red and blue, respectively, and detected deviations are in black. The online model correctly identifies the liquidity replenishment/inventory control behavior compared to the in-sample estimate. Results from the online model differ from in-sample computation as time goes on due to subspace adaptation.