MEASURING TIMELINESS OF ANNUAL REPORTS
FILING BY JUMP ADDITIVE MODELS

BY YICHENG KANG

Bentley University

Foreign public issuers (FPIs) are required by the Securities and
Exchanges Commission (SEC) to file Form 20-F as comprehensive
annual reports. In an effort to increase the usefulness of 20-Fs, the
SEC recently enacted a regulation to accelerate the deadline of 20-
F filing from six months to four months after the fiscal year-end.
The rationale is that the shortened reporting lag would improve the
informational relevance of 20-Fs. In this work, we propose a jump
additive model to evaluate the SEC’s rationale by investigating the
relationship between the timeliness of 20-F filing and its decision
usefulness using the market data. The proposed model extends the
conventional additive models to allow possible discontinuities in the
regression functions. We suggest a two-step jump-preserving estimation
procedure and show that it is statistically consistent. By applying
the procedure to the 20-F study, we find a moderate positive association
between the magnitude of the market reaction and the filing
timeliness when the acceleration is less than 17 days. We also find
that the market considers the filings significantly more informative
when the acceleration is more than 18 days and such reaction tapers
off when the acceleration exceeds 40 days.

1. Introduction. Foreign public issuers (FPIs) are required by the Se-
curities and Exchanges Commission (SEC) to file Form 20-F as compre-
prehensive annual reports. Similar to Form 10-K of U.S. domestic firms, 20-Fs are
regarded as the most valuable source of financial information available to eq-
uity investors of foreign firms. Both the 10-K and the 20-F give a compre-
hensive summary of a company’s financial performance. Detailed information
such as company cashflow, organizational structure, executive compensation
and corporate governance are included. In the accounting literature,
there has been extensive evidence that 10-Ks contain useful information for
investors’ investment decision making (e.g., Griffin 2003; Asthana, Balsam
and Sankaraguruswamy 2004; Callen, Livnat and Segal 2006; You and Zhang
2009; De Franco, Wong and Zhou 2011; Christensen, Heninger and Stice
2013). However, the literature on the decision usefulness of 20-Fs provides
rather mixed results (e.g., Meek 1983; Etter, Rees and Lukawitz 1999; Olibe

Keywords and phrases: Backfitting, decision usefulness, jump detection, jump-
preserving estimation, market reaction, shortened reporting lag.
Y. KANG

2001; Chen and Sami 2008; Chen and Sami 2013; Kim, Li and Li 2012), possibly because 20-Fs differ from 10-Ks in several ways. For instance, the deadline for filing a 20-F is six months after the fiscal year-end, much longer than the two-month or three-month deadline for filing a 10-K. The delay may cause investors to search private information before the release of 20-Fs, rendering 20-Fs less informative. In an effort to improve the information environment for FPIs in the U.S., the SEC recently enacted a regulation to accelerate the deadline of 20-F filing from six months to four months after the fiscal year-end. Effective for the fiscal year ending on or after December 15, 2011, the SEC requires all the FPIs to comply with the new four-month deadline. The aim is to increase the usefulness of 20-Fs by providing investors access to financial information prepared by FPIs in a timelier fashion.

Motivated by the SEC regulation change, this study examines empirically whether timelier filing increases market responses and more importantly, how much acceleration leads to the most significantly improved relevance of 20-F filing. To appropriately address this question, all the other control variables that might cause a difference in market reaction need to be properly taken into account by the empirical analysis. Particularly in our 20-F application, firm-specific variables (e.g., asset and liability) ought to be carefully controlled. It can be challenging because the functional relationships between these variables and the market response are hardly linear. In fact, it is difficult to even specify a parametric form for these functional relationships. Thus, a flexible nonparametric method would be desirable. In the statistics literature, non-additive nonparametric models such as local regression are commonly used. Despite being flexible enough, these models become less useful in dimensions higher than two or three. This problem of dimensionality is often referred to as the curse of dimensionality (Hastie, Tibshirani and Friedman 2009). Additive models (Breiman and Friedman, 1985) mitigate this issue by imposing an additive structure and are still flexible to use without making restrictive parametric assumptions. Thus, we adopt additive models as the framework for our empirical analysis. As for the relationship between the market response and the filing timeliness (i.e., the relationship of interest), in cases when a firm accelerates its 20-F filing significantly (e.g., a month), it might cause an abrupt shift in both the magnitude and the direction of the market response because investors may view such improvement of filing timeliness as a material change in terms of the information relevance of the 20-F. In cases when the acceleration is less significant (e.g., only a few days), the increase in market reaction would be less obvious. By this consideration, there may exist jumps in the functional relationship (Qiu, 2005) between the filing timeliness and the market
response. This is especially true when there is a shift in the direction of the market response. The jump locations indicate the amount of acceleration with which significant shifts in the market reaction are associated, and the jump sizes quantify the magnitudes of such shifts. Thus, it is important for our additive modeling framework to accommodate the existence of possible jumps in the regression functions.

There has been much discussion in the literature about estimation of jump regression curves (e.g., Hall and Titterington 1992; Ma and Yang 2011; Müller 2002; Qiu 2005; Wang 1995; Xia and Qiu 2015). More recent research consider a similar problem of detecting structural changes in the context of time series analysis (e.g., Casini 2018, 2019; Casini and Perron 2019; Wu and Zhao 2007). These existing methods consider the problem of jump detection in cases where there is only one or two predictors. It is unclear whether these methods can be directly extended to cases where there are multiple predictors, such as in our 20-F application. There also has been some recent research on change-point detection in higher dimensions (e.g., Cho and Fryzlewicz 2015; Avanesov and Buzun 2018). However, jump regression is different from change-point detection (Xia and Qiu 2015), in that the mean response could be an arbitrary continuous curve between two consecutive jump points in the former, but is constant between two consecutive change-points in the latter. Thus, change-point detection methods are generally not applicable to our 20-F problem. In this article, we propose a novel jump additive model (JAM) that detects jump locations and preserves jump sizes. We do not impose any parametric assumptions on the functional relationships besides the additive structure, and allow multiple predictors in the model. We show that the estimated number of jumps, the estimated jump locations, the estimated jump sizes, and the estimated regression functions converge to the truths as the sample size grows. Applying JAM to our 20-F timeliness study, we find a moderate positive association between the magnitude of the market reaction and the filing timeliness when the acceleration is less than 17 days. We also find that the market considers the filings significantly more informative when the acceleration is a little more than 18 days and such reaction tapers off when the acceleration exceeds 40 days. JAM is implemented in our R-package jam, which is included in the supplementary material.

The remainder of this article is organized as follows. The research design for our 20-F study is set up in Section 2. The proposed estimation procedure of jump additive models is described in detail in Section 3. Asymptotic properties of the proposed procedure are discussed in Section 4. Some simulated examples are given in Section 5. The empirical findings of our 20-F
study are presented in Section 6. Several concluding remarks are provided in Section 7. Proofs of the theorems and some technical arguments are given in the supplementary material.

2. Research Design. Since stock prices can react differently to the release of information (Beaver 1968; Kim and Verrecchia 1991), a commonly used measure of market response is cumulative abnormal return (e.g., Doyle and Magilke 2013). This measure is also adopted here. More specifically, let

\[ rt_{j,t} = \alpha_j + \beta_j \times rt_{m,t} + \epsilon_{j,t}, \]

where \( rt_{j,t} \) is the daily stock return of the \( j^{th} \) firm on day \( t \), and \( rt_{m,t} \) is the return on the U.S. market index S&P 500 on day \( t \). Let \( t = 0 \) denote the 20-F filing date. Estimate \( \alpha_j \) and \( \beta_j \) using standard linear regression during the period \([-260, -1]\). Denote the estimates for \( \alpha_j \) and \( \beta_j \) by \( \hat{\alpha}_j \) and \( \hat{\beta}_j \), respectively. Since we are concerned with both the direction and the magnitude of the market reaction, we calculate cumulative abnormal return for each 20-F filing date over event window \([T_1, T_2]\) as follows.

\[ y_j = \sum_{t \in [T_1, T_2]} (rt_{j,t} - \hat{rt}_{j,t}), \]

where \( \hat{rt}_{j,t} = \hat{\alpha}_j + \hat{\beta}_j rt_{m,t} \). \([T_1, T_2]\) is chosen to be \([-1, 4]\), because it takes into account potential information leakage and possible delays of 20-F releases. \( y_j \) is called the cumulative abnormal return around the \( j^{th} \) firm’s 20-F filing.

To investigate the relationship between the market response and the filing timeliness after controlling other factors, we propose the following empirical model:

\[ y_j = \phi_0 + \varphi_1(Timeliness) + \phi_2(Assets) + \phi_3(Leverage) + \phi_4(Accruals) + \phi_5(ROA) + \varepsilon_j, \]

where \( \phi_0 \) is the intercept, \( \varepsilon_j \) represents a random error, and \( \{\varphi_1, \phi_2, \ldots, \phi_5\} \) are the unknown regression functions that describe the marginal relationships between the response and the corresponding explanatory variables (\( \varphi_1 \) is denoted differently because it might contain jumps while other \( \phi_j \)'s are assumed to be continuous). \( Timeliness \) is calculated as the day difference between the 20-F filing date of current year and that of the previous year. For instance, suppose that a company filed its 20-F on April 30 this year and its last filing was on May 30 the year before, then the value for \( Timeliness \) is \(-30\). \( Assets \), calculated as the logarithm of the total assets at the fiscal
year-end, is a proxy for firm size. There has been evidence in the literature (e.g., Atiase 1985; Freeman 1987) that firm size is often associated with the magnitude of the market reaction. Leverage, the ratio of the total liability to the total assets, controls for the firm risk and the future growth opportunity. Accruals, calculated as the income minus the operating cash flows scaled by the total assets, is an indicator of financial performance. ROA, the ratio of the net income to the total assets, is a proxy for profitability. As discussed in Section 1, a substantial improvement of filing timeliness may cause a significant shift in the market response. There could be jumps in $\varphi_1$. Therefore, properly estimating the jump locations and the unknown regression functions $\{\varphi_1, \varphi_2, \ldots, \varphi_5\}$ is the primary task of our empirical analysis.

3. Jump Additive Models. Motivated by our study of timeliness of 20-F filing, we propose a jump additive model and a two-step estimation methodology. They are described in the subsequent four subsections. Jump additive models in their own right are of interest in methodological research. Thus, we use general mathematical terms throughout this section.

3.1. Review of Additive Models and ACE. First we give a brief review of additive models (AM) and alternating conditional expectation (ACE) algorithm proposed in Breiman and Friedman (1985). Let $Y$, $X_1$, $\ldots$, $X_p$ be random variables with $Y$ the response and $X_1$, $\ldots$, $X_p$ the predictors. Assume that

$$Y = \phi_0 + \phi_1(X_1) + \cdots + \phi_p(X_p) + \varepsilon,$$

where $\phi_1(\cdot), \ldots, \phi_p(\cdot)$ are arbitrary measurable functions of the corresponding random variables, and $\varepsilon$ is a mean-zero random error and independent of $(X_1, \ldots, X_p)$. Without further restrictions on (2), the solution is not unique, since we can add any constant to one $\phi_j$ and subtract the same constant from some other $\phi_j$. The standard convention is to further assume that $E[\phi_1(X_1)] = \cdots = E[\phi_p(X_p)] = 0$. Without loss of generality, we can also assume that $\phi_0 = 0$. In practice, this can be achieved by centering the response variable. The AM solves (2) for $\phi_1, \ldots, \phi_p$ by minimizing

$$e^2(\phi_1, \phi_2, \ldots, \phi_p) = E \left[ Y - \sum_{j=1}^{p} \phi_j(X_j) \right]^2$$

holding $E(\phi_1) = \cdots = E(\phi_p) = 0$. The solutions are called optimal transformations. The ACE algorithm solves the above minimization problem through a series of bivariate conditional expectations.
ACE Algorithm

(i) Set $\phi_1(X_1), \ldots, \phi_p(X_p) = 0$.
(ii) Iterate until $e^2(\phi_1, \ldots, \phi_p)$ fails to decrease;
   For $j = 1, \ldots, p$, do:
   \[
   \phi_j \leftarrow E \left[ Y - \sum_{k \neq j} \phi_k(X_k) \mid X_j \right];
   \]

The following theorem gives a convergence result for the ACE algorithm.

**Theorem 1.** Define, for $j = 1, \ldots, p$, $\hat{\phi}_j^{(m)}$ after $m$ stages of iteration in the ACE algorithm and $\bar{\phi}_m = \sum_{j=1}^p \hat{\phi}_j^{(m)}$. If (A1) – (A2) in the supplementary material hold, then $\bar{\phi}_m \overset{L^2}{\to} P_X Y$ as $m \to \infty$.

A similar result was obtained in Breiman and Friedman (1985), which also involves estimating the optimal transformation of the response. Since transformation of the response is not needed in our method, we give in the supplementary material a proof tailored for our purpose. Theorem 1 states that the ACE algorithm converges to the optimal transformations.

The ACE algorithm is developed in the context of known distributions, i.e., it assumes that the joint distribution of $(Y, X_1, \ldots, X_p)$ is known. In practice, population distributions are rarely known. Instead, one has a data set \{(\text{\textit{y}}_i, x_{i1}, \ldots, x_{ip}), 1 \leq i \leq n\} that are realizations from $Y, X_1, \ldots, X_p$.

The conditional expectations in the ACE algorithm need to be replaced by suitable univariate nonparametric estimates. Estimates that are commonly used with the ACE algorithm (e.g., Buja, Hastie and Tibshirani 1989; Hastie and Tibshirani 1990; Hastie, Tibshirani and Friedman 2009) include kernel smoothing (Cheng and Lin 1981; Gasser and Müller 1979; Nadaraya 1964; Priestley and Chao 1972; Watson 1964), local polynomial regression (Fan and Gijbels 1996; Hastie and Loader 1993), and splines (Anselone and Laurent 1968; Demmler and Reinsch 1975; Kimeldorf and Wahba 1970; Wahba 1990). These estimation methods assume that the regression functions $\{\phi_j, 1 \leq j \leq p\}$ are smooth and thus can not preserve the jumps well in cases when certain $\phi_j$ is discontinuous. In the next two subsections, we propose a two-step procedure that is jump-preserving.

3.2. Jump Detection. Assume that \{(\text{\textit{y}}_i, x_{i1}, \ldots, x_{ip}), 1 \leq i \leq n\} is a random sample from (2):

\[
\text{\textit{y}}_i = \varphi_1(x_{i1}) + \varphi_2(x_{i2}) + \cdots + \varphi_p(x_{ip}) + \varepsilon_i,
\]
where \( \varphi_1 \) has jumps at \( s_1 < \cdots < s_T, \phi_2, \ldots, \phi_p \) are continuous, and \( x_{ij} \in [a_j, b_j] \) for \( i = 1, \ldots, n, j = 1, \ldots, p \). Both the number of jumps \( T \) and the jump locations \( \{s_t, 1 \leq t \leq T\} \) are unknown. Here \( \varphi_1 \) is denoted differently to indicate the different smoothness assumption.

For a given \( x_1^* \in [a_1 + h_1, b_1 - h_1] \), consider the following locally weighted least square problems:

\[
\min_{c_0, c_1} \sum_{i=1}^{n} \left[ y_i - c_0 - c_1 (x_{i1} - x_1^*) \right]^2 K_+ \left( \frac{x_{i1} - x_1^*}{h_1} \right),
\]

\[
\min_{c_0, c_1} \sum_{i=1}^{n} \left[ y_i - c_0 - c_1 (x_{i1} - x_1^*) \right]^2 K_- \left( \frac{x_{i1} - x_1^*}{h_1} \right),
\]

where \( h_1 \in (0, (b_1 - a_1)/2 \) is a bandwidth parameter, and \( K_\pm \) are one-sided kernel functions with the support included in \([0, 1]\) and \([-1, 0]\), respectively. The bandwidth controls the size of the local neighborhood. The kernel functions control the weights and are often chosen such that observations closer to \( x_1^* \) receive more weights. The solutions to \( c_0 \) in (4) and (5) are one-sided estimates of the following conditional expectation:

\[
P(x_1^*) = E [\varphi_1(X_1) + \phi_2(X_2) + \cdots + \phi_p(X_p) | X_1 = x_1^*]
\]

\[
= \varphi_1(x_1^*) + \sum_{j=2}^{p} E [\phi_j(X_j) | X_1 = x_1^*].
\]

By algebraic manipulations, the solutions to (4) and (5) are:

\[
\hat{P}_-(x_1^*) = \sum_{i=1}^{n} y_i W_{-,i} \quad \text{and} \quad \hat{P}_+(x_1^*) = \sum_{i=1}^{n} y_i W_{+,i},
\]

where

\[
W_{\pm,i} = K_\pm \left( \frac{x_{i1} - x_1^*}{h_1} \right) \frac{w_{\pm,2} - w_{\pm,1}(x_{i1} - x_1^*)}{w_{\pm,0} w_{\pm,2} - w_{\pm,1}^2},
\]

\[
w_{\pm,u} = \sum_{i=1}^{n} (x_{i1} - x_1^*) u K_\pm \left( \frac{x_{i1} - x_1^*}{h_1} \right), \quad u = 0, 1, 2.
\]

Both \( \hat{P}_-(x_1^*) \) and \( \hat{P}_+(x_1^*) \) should estimate \( P(x_1^*) \) well if \( [x_1^* - h_1, x_1^* + h_1] \) does not contain any true jump position. On the other hand, if \( x_1^* \) is a true jump position, then \( \hat{P}_-(x_1^*) \) is a good estimator of \( P_-(x_1^*) \) and \( \hat{P}_+(x_1^*) \) is a good estimator of \( P_+(x_1^*) \), where \( P_\pm(x_1^*) \) denote the one-sided limits of \( P \) at
point \( x_t^* \). Thus, \( \hat{P}_+(x_t^*) - \hat{P}_-(x_t^*) \) should be close to \( d = P_+(x_t^*) - P_-(x_t^*) \). Suppose that the joint distribution function of \((X_1, \ldots, X_n)\) is smooth. Then \( P_+(x^*_1) - P_-(x^*_1) = \varphi_{1,+}(x^*_1) - \varphi_{1,-}(x^*_1) \). As a result, \( \hat{P}_+(x^*_1) - \hat{P}_-(x^*_1) \) is useful for detecting jump locations and estimating jump sizes.

Next, we identify the jump locations. Define

\[
\xi_k = a_1 + (b_1 - a_1) \frac{k}{m}, \quad k = 1, \ldots, m - 1,
\]

where \( m = [(b_1 - a_1)/(2h_1)] + 1 \), and \([x]\) denotes the integer part of \( x \). Then \( \{\xi_k, 1 \leq k \leq m - 1\} \) is sequence of points equally spaced in \([a_1, b_1]\), and the distance between any two neighboring points in \( \{\xi_k, 1 \leq k \leq m - 1\} \) is \((b_1 - a_1)/m\), which is slightly smaller than \( 2h_1 \). For a given point \( x_t^* \), there must be at least one and no more than two points in \( \{\xi_k, 1 \leq k \leq m - 1\} \) located in \((x_t^* - h_1, x_t^* + h_1)\). So we suggest detecting jumps at the points \( \{\xi_k, 1 \leq k \leq m - 1\} \) only. It saves much computation without missing any jumps. Suppose that \( \xi_{k_1}, \ldots, \xi_{k_r} \) are the points in \( \{\xi_k, 1 \leq k \leq m - 1\} \) satisfying

\[
\left| \frac{\hat{P}_+(\xi_{k_l}) - \hat{P}_-(\xi_{k_l})}{\sqrt{\sum_{i=1}^n W_{i,+}^2 + \sum_{i=1}^n W_{i,-}^2}} \right| > q, \quad l = 1, \ldots, r,
\]

where \( q \) is a threshold parameter and \( \sqrt{\sum_{i=1}^n W_{i,+}^2 + \sum_{i=1}^n W_{i,-}^2} \) is the normalizing constant. Then the points \( \{\xi_{k_l}, 1 \leq l \leq r\} \) are flagged as jump candidates. But if \( \xi_{k_l} \) is flagged, some of its neighboring points in \( \{\xi_k, 1 \leq k \leq m - 1\} \) are likely to be flagged as well, even if they are actually continuity points. Thus we suggest the following modification procedure to cancel some of the jump candidates. If there are two integers \( r_1 < r_2 \) such that

\[
\begin{align*}
\xi_{k_{l+1}} - \xi_{k_l} &= \frac{1}{m}, \quad l = r_1, r_1 + 1, \ldots, r_2 - 1, \\
\xi_{k_{r_1}} - \xi_{k_{r_1-1}} &> \frac{1}{m}, \\
\xi_{k_{r_2+1}} - \xi_{k_{r_2}} &> \frac{1}{m},
\end{align*}
\]

then we say that all candidates \( \{\xi_{k_l}, r_1 \leq l \leq r_2\} \) form a tie. For the jump candidates in a tie, we replace all of them by a new candidate defined as \((\xi_{r_1} + \xi_{r_2})/2\). After this modification, the current candidates consist of two types of points: those in \( \{\xi_{k_l}, 1 \leq l \leq r\} \) which do not belong to any tie and the middle points of all ties. Denote the current jump candidates by \( \eta_1 < \eta_2 < \cdots < \eta_r \). We define \( \tau, \{\eta_t, 1 \leq t \leq \tau\} \), and \( \hat{d}_t = \hat{P}_+(\eta_t) - \hat{P}_-(\eta_t), 1 \leq t \leq \tau \) as the estimator of the number of jumps \( T \), the jump locations \( \{s_t, 1 \leq t \leq T\} \), and the jump sizes \( \{d_t, 1 \leq t \leq T\} \), respectively.
3.3. Jump-Preserving Backfitting. Write

\[ \varphi_1(x_1) = \phi_1(x_1) + \sum_{t=1}^{T} d_t \mathbb{1}(x_1 > s_t), \quad x_1 \in [a_1, b_1], \]

where \( \phi_1 \) is a continuous function in \([a_1, b_1]\), and \( \mathbb{1}(\cdot) \) is the indicator function, which takes value of 1 if the argument in the parenthesis is true and takes value of 0 otherwise. The function \( \phi_1 \) is called the \textit{continuity part} of \( \varphi_1 \), and the summation \( J(x_1) = \sum_{t=1}^{T} d_t \mathbb{1}(x_1 > s_t) \) is called the \textit{jump part} of \( \varphi_1 \). The jump part has been estimated in Subsection 3.2. We now estimate the continuous functions \( \{\phi_1, \phi_2, \ldots, \phi_p\} \) from the new data \( \{(y_i - T \sum_{t=1}^{T} \hat{d}_t \mathbb{1}(x_{i1} > \eta_t), x_{i1}, \ldots, x_{ip})\}, 1 \leq i \leq n \).

Since the discontinuities are largely removed in the new data, the backfitting algorithm (Friedman and Stuetzle, 1981) for fitting the conventional additive models can be applied here. It consists of estimating each continuous function holding all the others fixed, then cycling through this process. More specifically, if the current estimates are \( \hat{\phi}_1, \ldots, \hat{\phi}_p \), then each is updated by smoothing the \textit{partial residuals} \( \nu_{ij} = \hat{y}_i^* - \sum_{j' \neq j} \hat{\phi}_{j'}(x_{ij'}) \) against \( x_{ij} \), where \( \hat{y}_i^* = y_i - \hat{J}(x_{i1}) \) with \( \hat{J}(x_{i1}) = \sum_{t=1}^{T} \hat{d}_t \mathbb{1}(x_{i1} > \eta_t) \). The backfitting procedure is summarized below.

**Backfitting Algorithm**

(i) Initialize: \( \hat{\phi}_1 \equiv -\frac{1}{n} \sum_{i=1}^{n} \hat{J}(x_{i1}), \hat{\phi}_j \equiv 0 \) for \( j = 2, \ldots, p \).

(ii) Iterate until the functions \( \hat{\phi}_j \) change less than a pre-specified threshold. For \( j = 1, \ldots, p \) do:

\[
\hat{\phi}_j \leftarrow S_j \left[ \left\{ \hat{y}_i^* - \sum_{j' \neq j} \hat{\phi}_{j'}(x_{ij'}) \right\} \right]_1^n,
\]

\[
\hat{\phi}_j \leftarrow \begin{cases} \hat{\phi}_j - \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\phi}_j(x_{ij}) + \hat{J}(x_{i1}) \right) & \text{if } j = 1, \\ \hat{\phi}_j - \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_j(x_{ij}) & \text{otherwise}; \end{cases}
\]

End For Loop;
End Iteration Loop.

The second step of the above For Loop in (ii) of the backfitting algorithm is to ensure that \( 1/n \sum_{i=1}^{n} \hat{\phi}_1(x_{i1}) = 1/n \sum_{i=1}^{n} \hat{\phi}_2(x_{i2}) = \cdots = \)
\[ \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_p(x_{ip}) = 0, \] since it is assumed in (3) that \( E[\varphi_1(X_1)] = E[\phi_2(X_2)] = \cdots = E[\phi_p(X_p)] = 0. \] Above, \( S_j \) in the backfitting algorithm can be any appropriate univariate smoother; here we consider local polynomial regression (Fan and Gijbels 1996). Specifically, for a given \( j \) and any given \( x^*_j \in [a_j + h_2, b_j - h_2] \), solve the following minimization problems:

\[
\min_{g_0, g_1, \ldots, g_p} \sum_{i=1}^{n} \left[ y_{ij} - g_0 - g_1(x_{ij} - x^*_j) - \cdots - g_p(x_{ij} - x^*_j)^p \right]^2 K \left( \frac{x_{ij} - x^*_j}{h_2} \right),
\]

where \( \kappa \) is the order of the local polynomial regression, \( h_2 \in (0, \min_j (b_j - a_j)/2) \) is the bandwidth parameter, and \( K \) is the kernel function with the support in \([-1, 1]\). The solution to \( g_0 \) in (7) is defined to be the \( \kappa^{th} \) order local polynomial estimate at \( x^*_j \). In particular, procedure (7) is often called local constant kernel (LCK) smoothing and local linear kernel (LLK) smoothing, when \( \kappa = 0 \) and \( \kappa = 1 \), respectively.

After the estimates \( \hat{\phi}_1, \ldots, \hat{\phi}_p \) are obtained by the backfitting algorithm, the estimated regression function \( \hat{\varphi}_1 \) is defined by

\[
\hat{\varphi}_1(x_1) = \hat{\phi}_1(x_1) + \sum_{t=1}^{\tau} \hat{d}_t \mathbb{1}(x_1 > \eta_t), \quad x_1 \in (a_1, b_1).
\]

3.4. Data-Driven Parameter Selection. There are three parameters to choose in our proposed procedure: \( h_1, q \) and \( h_2 \). We select them sequentially. Let us first consider the choice for \( h_2 \), assuming that the jump locations and jump sizes have been properly estimated. The purpose of \( h_2 \) is for estimating the continuous functions \( \{\phi_1, \ldots, \phi_p\} \) from data where discontinuities have been largely removed. Hence, \( h_2 \) plays the same role as a bandwidth parameter does in fitting continuous additive models. In the literature on bandwidth selection in nonparametric regression, plug-in procedures (e.g., Gasser, Kneip and Köhler 1991; Loader 1999; Ruppert, Sheather and Wand 1995) are easy to compute and often enjoy nice theoretical properties. Particularly for fitting conventional additive models using local polynomial smoothing, Opsomer and Ruppert (1998) proposed a direct plug-in (DPI) bandwidth selection method with the following formula.

\[
h_{2,\text{DPI}} = \left( \frac{(\kappa + 1)(\kappa)!^2 R(K(\kappa)) \sigma^2 \sum_{j=1}^{p} (b_j - a_j)}{2n \mu_{\kappa+1}(K(\kappa))^2 \sum_{j=1}^{p} \theta_{jj} (\kappa + 1, \kappa + 1)} \right)^{1/(2\kappa+3)},
\]

where

\[
\theta_{jj}(\kappa_1, \kappa_2) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\phi_j^{(\kappa_1)}(x_{ij})}{\phi_j^{(\kappa_1)}} - \frac{\phi_j^{(\kappa_2)}(x_{ij})}{\phi_j^{(\kappa_2)}} \right) \times \left( \frac{\phi_j^{(\kappa_1)}(x_{ij})}{\phi_j^{(\kappa_1)}} - \frac{\phi_j^{(\kappa_2)}(x_{ij})}{\phi_j^{(\kappa_2)}} \right).
\]
with \( \hat{\phi}_{j}^{(\kappa_{r})} = (1/n) \sum_{i=1}^{n} \phi_{j}^{(\kappa_{r})}(x_{ij}) \) for any positive integer \( \kappa_{r} \), \( r = 1, 2 \) and \( j = 1, \ldots, p \), and \( R(K(\kappa)) \) and \( \mu_{\kappa+1}(K(\kappa)) \) are quantities involving the kernel function \( K \) only. Their exact expressions are provided in the supplementary material. Opsomer and Ruppert (1998) showed that \( h_{2,\text{DPI}} \) achieves the optimal rate of convergence in the mean average squared error (MASE), defined by

\[
\text{MASE} = \frac{1}{n} \sum_{i=1}^{n} E \left[ \left( \sum_{j=1}^{p} \hat{\phi}_{j}(x_{ij}) - \sum_{j=1}^{p} \phi_{j}(x_{ij}) \right)^{2} \right].
\]

Because of this optimality, we adopt the DPI method for selecting \( h_{2} \). For simplicity, we focus on the special case of (9) when LLK smoothing is used (i.e., \( \kappa = 1 \)). The expression for \( h_{2,\text{DPI}} \) simplifies to

\[
h_{2,\text{DPI}} = \left( \frac{\sigma^{2} R(K) \sum_{j=1}^{p}(b_{j} - a_{j})}{n \mu_{2}(K)^{2} \sum_{j=1}^{p} \theta_{jj}(2, 2)} \right)^{1/5},
\]

where \( R(K) = \int K(u)^{2} \, du \) and \( \mu_{2}(K) = \int u^{2} K(u) \, du \). There are several unknown quantities (i.e., \( \theta_{jj}(2, 2) \) and \( \sigma^{2} \)) in (10) to be estimated before it can be used directly in practice. \( \theta_{jj}(2, 2) \) involves the second order derivatives of \( \phi_{j} \). Following Opsomer and Ruppert (1998), we obtain pilot estimates \( \{ \hat{\phi}_{j}^{(2)}(x_{ij}) \} \) by applying the backfitting algorithm with local cubic smoothing to data \( \{ (\hat{y}_{i}^{*}, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n \} \), in which an initial bandwidth \( h_{2,0} \) is used. The performance of the DPI method is relatively insensitive to the choice of \( h_{2,0} \). Our numerical experience suggests that \( h_{2,0} = 0.75 \max_{j}(b_{j} - a_{j}) \) works reasonably well. As a byproduct of fitting the backfitting algorithm to \( \{ (\hat{y}_{i}^{*}, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n \} \) with local cubic smoothing, we can estimate \( \sigma^{2} \) by

\[
\hat{\sigma}^{2} = \frac{1}{n} \sum_{i=1}^{n} \left[ \hat{y}_{i}^{*} - \sum_{j=1}^{p} \hat{\phi}_{j}(x_{ij}) \right]^{2}.
\]

Next, we consider the choice for \( q \). Let \( \phi_{j|1}(x_{i1}) = E[\phi_{j}(X_{j})|X_{1} = x_{i1}] \) and \( \delta_{ij} = \phi_{j}(x_{ij}) - \phi_{j|1}(x_{i1}) \) for \( j = 2, \ldots, p \). Consequently, \( \delta_{ij} \) and \( \phi_{j}(X_{j}) - E[\phi_{j}(X_{j})|X_{1}] \) have the same distribution and \( \{ \delta_{ij}, i = 1, \ldots, n \} \) are i.i.d.
with mean 0. Denote the variance of $\delta_{ij}$ by $\tau^2_j$. Now write

\[
\hat{P}_+(x^*_1) = \sum_{i=1}^{n} \varphi_1(x_{i1})W_{+,i} + \sum_{j=2}^{p} \sum_{i=1}^{n} \phi_{j|1}(x_{i1})W_{+,i} + \sum_{j=2}^{p} \sum_{i=1}^{n} \delta_{ij}W_{+,i} + \sum_{i=1}^{n} \varepsilon_iW_{+,i},
\]

\[= P_+(x^*_1) + O(h^2_1) + 1' \sum_{i=1}^{n} W_{+,i} \left( \varepsilon_i \right), \tag{11}\]

where $\mathbf{1}$ is a vector of length $p$ with all elements equal to 1 and $\mathbf{d}_i = (\delta_{i2}, \ldots, \delta_{ip})'$. Under the null hypothesis that $x^*_1$ is not a jump point, we have

\[
\frac{\hat{P}_+(x^*_1) - \hat{P}_-(x^*_1)}{\sqrt{\sum_{i=1}^{n} (W^2_{+,i} + W^2_{-,i})}} \xrightarrow{d} N \left( 0, \mathbf{1}'\Sigma\mathbf{1} \right), \tag{12}\]

where $\xrightarrow{d}$ denotes convergence in distribution, and $\Sigma$ is the covariance matrix of $(\varepsilon_i, \mathbf{d}_i)'$. The technical arguments for deriving (11) – (12) are provided in the supplementary material. Since $\varepsilon_i$ is independent with $(X_{i1}, \ldots, X_{ip})$, $\Sigma$ has the following form

\[
\Sigma = \begin{pmatrix}
\sigma^2 & 0 & 0 & \cdots & 0 \\
0 & \tau^2_2 & \tau_{23} & \cdots & \tau_{2p} \\
0 & \vdots & \vdots & \ddots & \vdots \\
0 & \tau_{2p} & \tau_{3p} & \cdots & \tau^2_p
\end{pmatrix},
\]

where $\tau_{j_1j_2} = E[\delta_{ij_1}\delta_{ij_2}]$ for $j_1, j_2 = 2, \ldots, p$. Based on (12), $q$ can be chosen to be $z_{1-\alpha/2} \sqrt{\hat{\Sigma} \mathbf{1}}$, where $\hat{\Sigma}$ is an estimate for $\Sigma$, $\alpha$ is the significance level, and $z_{1-\alpha/2}$ denotes the 100(1 - $\alpha/2$) percentile of $N(0, 1)$. Commonly used $\alpha$ values include 0.05, 0.01, 0.001, etc. In our theoretical justifications in the supplementary material, the asymptotic condition (B5) is stated in terms of $q$ for convenience. It can be equivalently stated in terms of $\alpha$. That is, (B5) quantifies the rate of the type I error (i.e., $\alpha$) approaching to zero as the sample size grows.

To obtain an estimate $\hat{\Sigma}$, we suggest the following 2-step procedure. First, obtain pilot estimates $\{\hat{\varphi}_1, \hat{\delta}_j, j = 2, \ldots, p\}$ by applying the backfitting algorithm with local cubic smoothing to the original data $\{(y_i, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n\}$, where an initial bandwidth $h_{1,0}$ is used. Similar to $h_{2,0}$, the choice
of $h_{1,0}$ is not critical and we find that an initial bandwidth that covers about 75% data points works well. To mitigate the effect of possible jumps in $\varphi_1$, we suggest estimating $\sigma^2$ using a robust statistic, such as median absolute deviation (MAD). We adopt the MAD estimator $\hat{\sigma} = 1.4826 \cdot \text{med}_i \{|\hat{\varepsilon}_i - \text{med}_k \{\hat{\varepsilon}_k\}\}$ (Rousseeuw and Croux, 1993), where $\text{med}$ denotes the sample median and $\{\hat{\varepsilon}_i\}$ are the residuals from the pilot estimates, i.e., $\hat{\varepsilon}_i = y_i - \hat{\varphi}_1(x_{i1}) - \sum_{j=2}^{p} \hat{\phi}_j(x_{ij})$. Second, obtain residuals $\{\hat{\delta}_{ij}\}$ by applying LLK smoothing to $\{(x_{i1}, \hat{\phi}_j(x_{ij})), i = 1, \ldots, n\}$ for each $j = 2, \ldots, p$ and then estimate
\[
\hat{\tau}_{j1j2} = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\delta}_{ij1} - \bar{\delta}_{j1} \right) \left( \hat{\delta}_{ij2} - \bar{\delta}_{j2} \right),
\]
where $\bar{\delta}_{jr} = (1/n) \sum_{i=1}^{n} \hat{\delta}_{ijr}$ for $r = 1, 2$.

Next, we consider the choice of $h_1$, which is used for jump detection. Let $S = \{s_t, t = 1, \ldots, T\}$ denote the set of the true jump locations. Recall that, for a given $h_1$, we only examine (6) at locations $\{\xi_k = a_1 + (b_1 - a_1)k/m : k = 1, \ldots, m - 1\}$. Let $\hat{S} = \{\hat{\xi}_k : \xi_k = \text{argmin}_{l=1,\ldots,m-1} |s_t - \xi_l|, t = 1, \ldots, T\}$. Using the method described in Section 3.2, we obtain an estimate $\hat{S}$. The Hausdorff distance between $\bar{S}$ and $\hat{S}$, $d_H(\bar{S}, \hat{S})$ (see Section 4 for the formal definition), takes discrete values in $\{(b_1 - a_1)k/m : k = 0, \pm 1, \pm 2, \ldots\}$. Ideally, $d_H(\bar{S}, \hat{S}) = 0$ with high probability. In the literature, Gijbels and Goderniaux (2004) proposes a bootstrap procedure to estimate $P(d_H(\bar{S}, \hat{S}) = 0)$ and to select the bandwidth for which this estimated probability is maximal. Following Gijbels and Goderniaux (2004), we select $h_1$ as follows.

- **Step 1: Computation of Residuals.** For a given $h_1$, with $q$ and $h_2$ calculated by the aforementioned procedures, construct the estimates $\hat{S}$ and $\{\hat{\varphi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p\}$. Define $\hat{\varepsilon}_i = y_i - \hat{\varphi}_1(x_{i1}) - \sum_{j=2}^{p} \hat{\phi}_j(x_{ij})$.
- **Step 2: Bootstrap Simulation.** Let $\varepsilon_{1}^{(1)}, \ldots, \varepsilon_{n}^{(1)}$ be a resample randomly drawn with replacement from the set $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$. Define
\[
y_{i}^{(1)} = \hat{\varphi}_1(x_{i1}) + \sum_{j=2}^{p} \hat{\phi}_j(x_{ij}) + \varepsilon_{i}^{(1)}.\]
Then $\{(y_{i}^{(1)}, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n\}$ is the first bootstrap sample.
- **Step 3: Determination of the Bootstrap Probability.** Compute the analog $\hat{S}^{(1)}$ for the first bootstrap sample. From $B$ bootstrap replications, obtain $B$ values of $\{\hat{S}^{(b)} : 1 \leq b \leq B\}$, and estimate the discrete
probability $P(d_H(\tilde{S}, \hat{S}) = 0)$ via

$$\hat{P}(d_H(\tilde{S}, \hat{S}) = 0) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{d_H(\tilde{S}, \hat{S}^{(b)}) = 0\}.$$ 

- Step 4: Determination of $h_1$. For a grid of potential values for $h_1$, choose the value for which $\hat{P}(d_H(\tilde{S}, \hat{S}) = 0)$ is maximum.

The theoretical underpinning for the bootstrap procedure is established in Gijbels, Hall and Kneip (2004). More specifically, by the same arguments in Theorem 2 in Gijbels, Hall and Kneip (2004), for all the discrete probabilities $\hat{P}(d_H(\tilde{S}, \hat{S}) = (b_1 - a_1)k/m), k = 0, \pm 1, \pm 2, \ldots$, we have

$$\sup_{k=0, \pm 1, \ldots} \left| \hat{P}(d_H(\tilde{S}, \hat{S}) = \frac{(b_1 - a_1)k}{m}) - P(d_H(\tilde{S}, \hat{S}) = \frac{(b_1 - a_1)k}{m}) \right| \to 0$$

in probability.

The bootstrap procedure is computationally intensive since it requires repeating our proposed procedure many times. To reduce running time, we employ parallel computing in the implementation of the companion R package jam, which is included in the supplementary material.

4. Asymptotic Properties. We first establish the almost sure consistency of the jump detection procedure. Define $D$ to be a set $\{x_1, \ldots, x_n\}$ of $n$ points in $p$–dimensional space, i.e., $x_k = (x_{k1}, \ldots, x_{kp})$. Let $\mathcal{D}_n$ be the collection of all such $D$. We have the following theorem.

**Theorem 2.** Assume that (B1) – (B6) in the supplementary material hold. Then, for any given $D \in \mathcal{D}_n$,

1. $d_H(\tilde{S}, \hat{S}) = O(h_1)$ a.s., where $d_H(A, B)$ denotes the Hausdorff distance between two point sets $A$ and $B$ as defined by

$$d_H(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} d_E(a, b), \sup_{b \in B} \inf_{a \in A} d_E(b, a) \right\}$$

with $d_E(\cdot, \cdot)$ being the Euclidean distance. Furthermore, when $n$ is sufficiently large, there is exactly one detected jump in each $(s_t-h_1, s_t+h_1)$, $t = 1, \ldots, T$.

2. $|\hat{\tau} - \tau| = o(1)$ a.s. for $t = 1, \ldots, T$.

Theorem 2 states that $\tau$ is an almost surely consistent estimator of $T$. Additionally, the estimators of jump locations and jump magnitudes are all almost surely consistent. Next, we give the consistency result for the backfitting algorithm.
Theorem 3. Let \(\hat{\phi}_j(\cdot; m), j = 1, \ldots, p\) denote the functions after repeating the backfitting loops \(m\) times with the LCK smoother. Repeat the ACE loop \(m\) times, getting functions denoted by \(\phi_j(\cdot; m), j = 1, \ldots, p\). Under the same conditions as those in Theorem 2, for any small number \(\epsilon > 0\), as \(n \to \infty\),

\[
E \left[ \left| \frac{\hat{\phi}_j(\cdot; m) - \phi_j(\cdot; m)}{a_j, b_j, n} \right|^2 \left| X_1 \notin S_\epsilon, D \right. \right] \to 0, \quad j = 1, \ldots, p,
\]

where \(S_\epsilon = \bigcup_{t=1}^T [s_t - \epsilon, s_t + \epsilon]\), and \(\left| \hat{\phi}_j(x; m) - \phi_j(x; m) \right|^2_{I,n} = 1/n \sum_{i,j \in I} |\hat{\phi}_j(x_{i,j}; m) - \phi_j(x_{i,j}; m)|^2\) for \(j = 1, \ldots, p\) with \(I\) being an interval.

The proofs of Theorem 2 and Theorem 3 are given in the supplementary material.

5. Simulation Study. We present some simulation results regarding the numerical performance of the jump detection procedure, the backfitting procedure, and the parameter selection procedure. Throughout this section, the kernel function \(K\) in (7) is chosen to be the Epanechnikov kernel \(K(x) = 0.75(1 - x^2)\mathbb{1}(|x| < 1)\). The one-sided kernels \(K_{\pm}\) in (4) – (5) are chosen to be \(K_{-}(x) = 2K(x)\mathbb{1}(x < 0)\) and \(K_{+}(x) = 2K(x)\mathbb{1}(x > 0)\). In the local kernel smoothing literature, the Epanechnikov kernel is often used because of its good theoretical properties (Fan and Gijbels 1996). In our simulation, the data is generated from the following model

\[
Y = 0.2 + \varphi_1(X_1) + \phi_2(X_2) + \phi_3(X_3) + \epsilon,
\]

\[
\varphi_1(x_1) = e^{-x_1} - 0.2 \times \mathbb{1}(x_1 > 0.5) - c_1,
\]

\[
\phi_2(x_2) = 0.5x_2^3 - c_2, \quad \phi_3(x_3) = 0.5x_3^3 - c_3,
\]

\[
X_1 \overset{d}{=} U, \quad X_2 \overset{d}{=} \rho U + (1 - \rho)V_2, \quad X_3 \overset{d}{=} \rho U + (1 - \rho)V_3,
\]

where \(d\) denotes equality in distribution, \(U, V_2\) and \(V_3\) are independent random variable with identical distribution uniform\([0, 1]\), \(c_1, c_2\) and \(c_3\) are constants such that \(E[\varphi_1(X_1)] = E[\phi_2(X_2)] = E[\phi_2(X_3)] = 0, \rho = 0.3\) (it implies that \(\text{corr}(X_1, X_2) = \text{corr}(X_1, X_3) \approx 0.4\), and the random noise \(\epsilon\) are generated from normal distribution \(N(0, \sigma^2)\). We consider the cases when the sample size \(n = 200, 500\) and the noise level \(\sigma = 0.1, 0.15, 0.2\). Throughout this section, the grid of values for \(h_1\) is specified to be \(\{0.1 + 0.01 \cdot k : k = 1, \ldots, 49\}\). We find the grid with step size of 0.01 fine enough for our example, and the largest permitted should not exceed 0.5, which is half the
length of the design interval for $X_1$. In addition, the number of bootstrap samples $B = 100$, and the significance level $\alpha = 0.05$. Let

$$h_{1,\text{opt}} = \arg\max P \left( d_H \left( \hat{S}, \tilde{S} \right) = 0 \right),$$

where the probability is evaluated based on 100 replicated simulations. Similarly, let $h_{2,\text{opt}}$ denote the optimal $h_2$ for which the MASE is minimized. Furthermore, let $h_{1,\text{br}}$ and $h_{2,\text{DPI}}$ denote the bandwidths selected by the bootstrap procedure and the plug-in procedure, respectively. The realizations of $\{(Y, X_1, X_2, X_3)\}$ when $n = 200$ and $\sigma = 0.15$ are shown in Figure 1(a) – (c). Plots of $\varphi_1$, $\varphi_2$ and $\varphi_3$ along with their estimates are shown respectively in Figure 1(c) – (e), where $h_{1,\text{opt}}$ and $h_{2,\text{DPI}}$ were used. It can be seen that (i) the data is quite noisy that the jump structure can hardly been visually detected in the scatterplots, and (ii) the proposed procedure is able to preserve the jump well and largely eliminate the noise. It should also be pointed out that the nominal noise level $\sigma$ is not the effective noise level for our jump detector (6). As shown in (12), the effective noise level is $\sqrt{Y \Sigma Y}$, which is greater than $\sigma$. It explains why the jump structure can hardly be seen in Figure 1(a), where the jump size is 1.33 times the magnitude of the nominal noise level.

![Scatterplot and Plots](image)

**Fig 1.** (a)-(c): Scatterplot of $Y$ versus $X_1, X_2$ and $X_3$, respectively. (c) - (e): Plot of $\varphi_1$ and $\hat{\varphi}_1$, $\varphi_2$ and $\hat{\varphi}_2$, and $\varphi_3$ and $\hat{\varphi}_3$, respectively.

Next, we evaluate the numerical performance of the proposed procedure...
quantitatively. Table 1 summarizes the simulation results of the estimation procedure and the parameter selection. The columns denoted by optimal are associated with the cases when the optimal parameters are used (i.e., $h_{1,\text{opt}}$ and $h_{2,\text{opt}}$). The columns denoted by data-driven are associated with the cases when the parameters are chosen by the data-driven procedure (i.e., $h_{1,\text{bt}}$ and $h_{2,\text{DPI}}$). $\Pr$ denotes $P(d_H(\hat{S}, \tilde{S}) = 0)$ and $\text{sePr}$ denotes the standard error of $P(d_H(\hat{S}, \tilde{S}) = 0)$. $\text{seMASE}$ denotes the standard error of the MASE. The numbers in rows of MASE and $\text{seMASE}$ are in the unit of $10^{-4}$. The results are based on 100 replicated simulations. It can be seen from Table 1 that both MASE and $\text{seMASE}$ decrease as $n$ increases and as $\sigma$ decreases. Similarly, $P(d_H(\hat{S}, \tilde{S}) = 0)$, which measures the accuracy of our jump detection procedure, increases as $n$ increases and as $\sigma$ decreases. This is intuitively reasonable and consistent with our theoretical analysis in Section 4. In addition, across various cases of sample size and noise level, our data-driven parameter selection works well in the sense that both $P(d_H(\hat{S}, \tilde{S}) = 0)$ and MASE calculated using $h_{1,\text{bt}}$ and $h_{2,\text{DPI}}$ are not significantly different from those calculated using the optimal parameters. However, it is worth noting that, in cases when the sample size is moderate (i.e., $n = 200$), our procedure performs reasonably well when the nominal noise level $\sigma$ is half

<table>
<thead>
<tr>
<th>n</th>
<th>$\sigma = 0.1$</th>
<th>$\sigma = 0.15$</th>
<th>$\sigma = 0.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>optimal data-driven</td>
<td>optimal data-driven</td>
<td>optimal data-driven</td>
</tr>
<tr>
<td>200</td>
<td>$h_1$</td>
<td>.48</td>
<td>.43</td>
</tr>
<tr>
<td></td>
<td>$\Pr$</td>
<td>.93</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td>$\text{sePr}$</td>
<td>.026</td>
<td>.030</td>
</tr>
<tr>
<td></td>
<td>$h_2$</td>
<td>.40</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>MASE</td>
<td>14.43</td>
<td>15.44</td>
</tr>
<tr>
<td></td>
<td>$\text{seMASE}$</td>
<td>4.23</td>
<td>4.49</td>
</tr>
<tr>
<td>500</td>
<td>$h_1$</td>
<td>.29</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>$\Pr$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>$\text{sePr}$</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>$h_2$</td>
<td>.30</td>
<td>.22</td>
</tr>
<tr>
<td></td>
<td>MASE</td>
<td>3.70</td>
<td>3.92</td>
</tr>
<tr>
<td></td>
<td>$\text{seMASE}$</td>
<td>.43</td>
<td>.46</td>
</tr>
</tbody>
</table>

Table 1: Numerical performance of the proposed procedure and the parameter selection. The columns denoted by optimal are associated with the cases when the optimal parameters are used. The columns denoted by data-driven are associated with the cases when the parameters are chosen by the data-driven procedure. $\Pr$ denotes $P(d_H(\hat{S}, \tilde{S}) = 0)$ and $\text{sePr}$ denotes the standard error of $P(d_H(\hat{S}, \tilde{S}) = 0)$. $\text{seMASE}$ denotes the standard error of the MASE. The numbers in rows of MASE and $\text{seMASE}$ are in the unit of $10^{-4}$. The results are based on 100 replicated simulations.
the jump size, but its performance starts to deteriorate as $\sigma$ is of the same magnitude as the jump size. Thus, we consider our procedure better at detecting large jumps.

6. **Empirical Analysis.** In this section, we apply our proposed procedure to analyzing timeliness of the 20-F filings.

6.1. **Data.** Our sample period covers 2008 - 2011 because the SEC required all the FPIs to follow the new four-month filing deadline starting from the fiscal year ending on or after December 15, 2008. The SEC also provided a transition period of three years, which allowed FPIs to comply with this new rule effective for the fiscal year ending on or after December 15, 2011. Following the data selection method in Doyle and Magilke (2013) and Liu (2016), we obtained 20-Fs, and 20-F filing dates from the SEC EDGAR system; daily stock returns and index data from the CRSP U.S. stock database; firm financial data from the COMPUSTAT database; and firm major operation country, filer status, and accounting standard from the Audit Analytics database. Sample firms were initially obtained from the SEC’s official annual summary of “International Registered and Reporting Companies”. Then we deleted the firms not listed on either the NASDAQ or the NYSE, the Canadian firms (because of the special filing requirements for Canadian firms), over-the-counter firms, debt issuing firms, and the firms that lack complete data during the test period. We identified 272 filings in the final sample.

6.2. **Parameter Values.** Since the predictors are of different units (e.g., Assets are denominated in U.S. dollars and Leverage ratio is a fraction), each predictor is centered and standardized before our procedure is applied. Using the method described in Subsection 3.4, the following grid of values is specified for selecting $h_1$: \{$k/\text{SE}_1 : k = 2, 3, \ldots, 12$\}, where $\text{SE}_1$ denotes the sample standard deviation of Timeliness. $h_1$ is chosen to be $6/\text{SE}_1 = 0.22$ by our bootstrap procedure, where $B = 1000$. The asymptotic variance in (12) is estimated to be $1^{\prime}\hat{\Sigma}1 = .0035$ and the plug-in $\hat{h}_{2,\text{plug}}$ is calculated to be 0.81.

6.3. **Results.** The jump additive model (1) estimated by the proposed procedure is shown in Figure 2. Jump detection criterion (6) is plotted in Figure 2(a), which suggests a possible jump location around where Timeliness is $-18$. From Figure 2(b), it can be seen that the magnitude of market reaction (i.e., the absolute value of cumulative abnormal return) increases moderately as firms improve their filing timeliness by less than 17 days (the
more negative the *Timeliness* value, the timelier the filing). When firms accelerate their filing by 18 days, there is an abrupt shift in the direction of market response, indicating that the market changes its directional view. This is possibly because market investors consider such filings significantly more relevant and the financial information disclosed in these 20-Fs is then quickly priced in by the market, triggering the change of view. Nonetheless, when firms further accelerate their filings by more than 40 days, we find little market reaction. One explanation is that timelier filing reduces the preparation time and results in more errors in financial information. The market may view filings with too much acceleration not as reliable. Hence, there is a trade-off between timeliness and reliability. Taken together, the market reaction is the most significant when the filing is accelerated by a little more than 18 days. As for the control variables, it can be seen from Figure 2 (c) – (f) that firms with more assets (larger firms), healthy leverage ratio (good growth potential but not deeply in debt), and higher ROA (more profitable) tend to associate with more positive market reaction. This is consistent with our intuition. Confidence bands have also been added to Figure 2(b) – (f). There are several ways to obtain confidence bands after the jump part has been removed. One way is to follow the F-distribution approximation approach suggested by Hastie and Tibshirani (1990). It needs to assume that the errors follow a normal distribution. An alternative approach is to create bootstrap residuals as in Subsection 3.4 and compute the backfitting estimates of the bootstrap response on the predictors. The bootstrap approach is adopted here for producing the confidence bands in Figure 2, where the number of bootstrap samples is chosen to be 1000.

7. Conclusions. In this study, we investigate the relationship between the filing timeliness of 20-F forms and the market reaction. We propose a novel jump additive model that accounts for possible discontinuities in the relationship, which includes jumps in the magnitude and shifts in the direction of market reaction. We also propose a two-step jump-preserving estimation procedure and show that it consistently estimates the number of jump positions, the jump locations, the jump sizes and the regression functions. A major feature of our proposed procedure is that it does not impose restrictive parametric assumptions so it is flexible to use. A data-driven parameter selection procedure is suggested as well. Therefore, the proposed procedure can be fully automated. By applying the jump additive model to our 20-F study, we find that the magnitude of the market reaction increases moderately when the filing timeliness is improved by less than 17 days. Market investors consider the 20-Fs significantly more relevant when
Fig 2. The jump additive model (1) estimated by the proposed procedure. (a): Plot of the jump detection criterion. (b) – (f): The estimated regression function (dashed line) and 99% confidence bands (dotted line) of $\phi_1$, $\phi_2$, $\ldots$, $\phi_5$, respectively.
the filings are accelerated by a little more than 18 days. However, as firms further accelerate their filings by more than 40 days, the market reaction tapers off possibly because of concerns that filings with too much acceleration are prone to errors and thus less reliable.

There is still room for further improvement. Methodologically, our estimation procedure assumes that there is only one regression function in the model with possible discontinuities. Removing such restriction is certainly an interesting direction to pursue. In addition, we have assumed that random errors $\varepsilon_i$ are i.i.d. Modifying our method to accommodate heteroscedasticity in errors is a useful extension. To further investigate the costs and benefits of shortened 20-F filing, it requires to examine the relationship between the filing timeliness and the reporting reliability and to determine whether such relationship plays a role in diminishing market reactions to filings with large accelerations.

Acknowledgment. The author would like to thank the Editor, an Associate Editor, and two referees for many constructive comments and suggestions which substantially improved the quality of the paper.

SUPPLEMENTARY MATERIAL

Supplement 1: Proofs of the Theoretical Results

Supplement 2: The Companion R Package jam

References.


