IARIW 2022

Monday 22 - Friday 26 August

Nowcasting with Signature Methods

Samuel Cohen
(University of Oxford, Alan Turing Institute)

Silvia Lui
(Office for National Statistics, Economic Statistics Centre of Excellence)

Will Malpass
(Office for National Statistics)

Giulia Mantoan
(Office for National Statistics)

Lars Nesheim
(University College London, Alan Turing Institute)

Aureo de Paula
(University College London, Alan Turing Institute)

Craig Scott
(Office for National Statistics)

Emma Small
(Office for National Statistics)

Lingyi Yang
(University of Oxford, Alan Turing Institute, Office for National Statistics)

Nowcasting with signature methods

Samuel N. Cohen[1,3], Silvia Lui[2,5], Will Malpass[2], Giulia
Mantoan[2,3], Lars Nesheim[3,4], Aureo de Paula[3,4], Craig Scott[2],
Emma Small[2], and Lingyi Yang[1,2,3]

[1]*University of Oxford*
[2]*Office for National Statistics*
[3]*The Alan Turing Institute*
[4]*University College London*
[5]*Economic Statistics Centre of Excellence*

July 2022

Abstract

Nowcasting refers to the "forecast" of the current ("now") state of the economy. This is necessary as key economic variables are often published with a significant delay of over a month. The nowcasting literature has arisen to address the need to have fast, reliable estimates of delayed economic indicators. The path signature is a mathematical object which captures geometric properties of sequential data; it naturally handles missing data from mixed frequency and/or irregular sampling – issues often encountered when merging multiple data sources – by embedding the observed data in continuous time. Calculating path signatures and using them as features in models have achieved state-of-the-art results in other fields such as finance, medicine, and cyber security. We look at the nowcasting problem by applying regression on signatures, a simple linear model on these nonlinear objects that we show subsumes the popular Kalman filter. We quantify the performance via a simulation exercise and application to US GDP growth, where in the latter we demonstrate no loss of performance compared with the dynamic factor model. By embedding discrete information in continuous time, this approach allows greater flexibility for future applications on data with complex sampling patterns.

Disclaimer: The views expressed are those of the authors and may not reflect the views of the Office for National Statistics or the wider UK Government.

1 Introduction

Households, businesses, and policymakers need up-to-date information to make decisions. However, information is often incomplete or delayed because collection and compilation of the data underlying key economic indicators takes time. For example, UK monthly GDP is published with approximately a 6-week lag. Incomplete or delayed information hinders the ability for rapid responses and decision making. Decision-makers want to be able to query the UK GDP now, they do not want to wait 6 weeks. As a result, there has been an increasing demand for reliable early predictions using any and all available information. These early predictions or estimates are called nowcasts. They provide a prediction/inference of the current state of the economy.

The economic literature on nowcasting has grown dramatically in the past 20 years (e.g. Stock and Watson, 2002; Giannone et al., 2006; Bok et al., 2018). This literature has focused on three main issues. First, when incorporating a large number of predictors into a model, how one imposes structure on the model to reduce dimension. Second, how one incorporates data into a model when there are missing observations or when there are mixed or irregular sampling frequencies (Kapetanios et al., 2018; Ghysels and Marcellino, 2018). Third, how one allows for time-varying parameters or for non-linearities.

1.1 Dimension reduction

In most nowcasting exercises, a large number of variables are used to predict a single target. For example, the NY Federal Reserve nowcasting model uses more than 160 variables to nowcast US GDP (FRBNY, 2016). The leading approaches in the macroeconomic nowcasting literature are dynamic factor models (DFM), Bayesian vector auto-regressions (BVAR), and penalised estimation methods like the LASSO.[1]

Dynamic factor models were introduced into economics by Geweke (1977) and are reviewed in Stock and Watson (2017); Bai and Ng (2008). These models are based on the idea that most of the time series variation in a large set of economic variables is driven by the dynamics of a small number of unobserved common factors. These factors can be extracted from the complete set of economic variables using singular value decomposition of the data matrix. Then, the dynamics of the factors are analysed using standard

---

[1]Recent macroeconomic papers have also investigated the use of machine learning methods including random forests and artificial neural networks (Richardson et al., 2021). Furthermore, outside of economics, there is a large literature in nowcasting across a range of disciplines including statistics, meteorology, computer science, engineering, etc.

time series methods such as vector autoregression. There has been rapid development of DFM methods in applied macroeconomic analysis, and they have been used extensively in economic forecasting and nowcasting, see e.g., Giannone et al. (2006), Doz et al. (2006), and Bok et al. (2018).

Bayesian vector auto-regressions, developed in macroeconomics in Litterman (1979); Sims (1980), and reviewed in Karlsson (2013), impose restrictions on high-dimensional models by specifying prior beliefs on the parameter space and then using Bayesian methods to estimate the posterior distribution of the parameters. The posterior distribution then is used to nowcast the economic variables of interest.

Penalised estimation methods impose restrictions on high dimensional models by including a penalty term in the estimation objective function. For example, the LASSO estimator, popularised in statistics by Tibshirani (1996), adds a penalty based on the sum of the absolute values of the parameters (an L1 penalty). This shrinks parameter estimates towards zero and, due to the non-smooth objective, selects a subset of variables with nonzero coefficients. The ridge estimator (Hoerl and Kennard, 1970) instead adds a penalty based on the sum of the squares of the parameters (L2 penalty). This also has the effect of shrinking parameter estimates towards zero. Many other forms of the penalty function are possible. The penalty terms allow researchers to include very large numbers of predictors in nowcasting models and, in many settings, result in nowcasts that have lower mean squared error than unpenalised alternatives. These methods have been used in nowcasting in Babii et al. (2021), for example.

1.2 Missing observations and mixed frequency or irregular sampling

Due to varying publication lags, it is often the case that some variables in a dataset have missing values for recent time periods. This issue is often referred to as the "ragged-edge" problem. In many other cases, one would like to predict low frequency variables (monthly or quarterly GDP) using high frequency ones (daily, weekly, and monthly prices or expenditure).

Several solutions to this problem have been developed in the macroeconometrics literature. One option is to use the classic discrete time Kalman filter (e.g. Kalman, 1960; Stock and Watson, 2002; Bańbura and Modugno, 2014). This method assumes a linear time series structure and imputes missing values using the Kalman filter.

A second option is the bridge method. This approach consists of a two-equation system, linking the high frequency predictor x_t^h to the low frequency

target y_t^l

$$y_t^l = \beta_0 + \lambda y_{t-1} + \beta(L)x_t^l + \varepsilon_t.$$

The lag polynomial $\beta(L)$ of order p is defined as $\beta(L) = \sum_{i=0}^{p} \beta_{i+1}L^i$ with $Ly_t = y_{t-1}$ where p is the number of equidistant high-frequency periods one can divide the low-frequency period; i.e. $p = 3$ in the case of monthly (high-frequency) data are used to nowcast quarterly (low-frequency) target variable.

Given the high-frequency predictors x_t^h, the user has to aggregate the indicator forecast $x_t^l = \omega(L^{1/p})x_t^h$ before proceeding to step estimate y_t^l. Note that as written above, the aggregation can be applied to complete quarters only. Which means that, the practitioner has to wait until the three monthly observations of the indicator corresponding to the calendar quarter are available before applying the aggregator function. This can be resolved by forecasting the predictor for any missing months. In the literature, typically very simple univariate models are chosen to predict the high-frequency indicator. For example, a simple AR forecast equation for x_t^h such as

$$\hat{x}_t^h = \alpha_0 + \alpha(L^{1/p})x_{t-1/p}^h + \varepsilon_t^h,$$

can be be used in aggregation to obtain x_t^l which is then used in Equation (1.2). Although simple to use, the bridge model relies on estimated values derived from the higher frequency variable to nowcast the lower frequency target variable. Hence, forecast accuracy of missing values directly affects nowcast accuracy.

A third approach that directly relates high frequency indicators to a low frequency target is the mixed-data sampling (MIDAS) model (Ghysels et al., 2004). MIDAS accommodates data sampled at different frequencies by incorporating higher frequency variables using distributed lag terms with fractional lags. For example, in a quarterly model in which each time period is composed of three months, a monthly variable can be incorporated as a variable that is lagged by one third or two thirds of a model period. Ghysels and Marcellino (2018) describe the method as a "tightly parameterised, reduced form regression that involve processes sampled at different frequencies". The MIDAS equation for y_t^l is given by:

$$y_t^l = \beta_0 + \lambda y_{t-r} + \beta_1 B(L^{1/p}, \theta)x_t^h + \varepsilon_t.$$

The term $B(L^{1/p}, \theta)$ is a lag polynomial

$$B(L^{1/p}, \theta) = \sum_{k=0}^{K} b(k, \theta)L^{k/p}$$

In the MIDAS literature, typically functional lag polynomials are chosen for $B(L^{1/p}, \theta)$ to avoid parameter proliferation for long high-frequency lags K. A popular functional form of the polynomial is the exponential Almon lag

$$b(k, \theta) = \frac{\exp(\theta_1 k + \theta_2 k^2)}{\sum_{j=1}^{k} \exp(\theta_1 j + \theta_2 j^2)}$$

with parameters $\theta = [\theta_1, \theta_2]$. For a given θ, the lag function provides a parsimonious way to consider a large number of K high-frequency lags of the indicators.

To-date, most empirical studies on nowcasting deals with regular missing data patterns. This is because nowcasting typically utilises official published data which are released on a fairly regular schedule. However, in recent years there has been an increase in the use of alternative data sources e.g. web-scraped data and scanner data. These alternative sources can have complicated missingness patterns as a results of collection methods. Therefore, there is a need for a nowcasting methodology to be able to handle these cases.

1.3   Time varying parameters and nonlinearities

Classic time series methods such as the autoregressive integrated moving average (ARIMA) model assume that the dynamics of a low dimensional variable, after suitable differencing to ensure stationarity, can be modelled with a linear model with autoregressive and moving average components. After differencing, the data are stationary, the model is linear, and the parameters of the model are assumed to be constant. In nowcasting settings, all three of these assumptions can be too restrictive.

State space models offer one approach to relax these restrictions (Durbin and Koopman, 2012). These models are highly flexible and can allow for non-stationarity, non-linearity, and time-varying parameters. Many papers, including Hamilton (1994), Nielsen and Berg (2014), Carter and Kohn (1994), Kim et al. (1999), Kim (1994), also allow for time-varying parameters.

1.4   Summary of contributions

Our contribution is in introducing signature methods to the nowcasting literature. The path signature is a mathematical object to describe time-series-like data, arising from the theory of rough path analysis. They have properties that are desirable in applications. We shall refer to the family of methods that utilises the path signature for prediction tasks as *signature methods*. Signature methods naturally allow for missing data due to mixed frequency

and irregular sampling by modelling time series processes in continuous time. They have been used to great success in a range of applications including Chinese handwriting recognition (Graham, 2013), sepsis detection (Morrill et al., 2020), and malware detection (Cochrane et al., 2021).

When the process is in continuous time but the measurements of the process are discrete or irregular, one can still construct the signature of the underlying process and use the signature in a nowcasting model. In addition, as in state space models, signatures can be constructed and used in non-stationary settings and in settings where the underlying dynamics are nonlinear. Given the success of utilising signatures in other areas, we benchmark the performance against the well known, aforementioned nowcasting methods.

Signatures on their own do not solve the problem of dimension reduction. However, signatures can be combined with dimension reduction methods. As a second contribution, we thus show how to combine the flexible dynamics of signatures with standard dimension reduction techniques like those used in DFMs. We develop a signature DFM model (SDFM) in which low dimensional factors are extracted from high dimensional data using the singular value decomposition and the dynamics of the factors are analysed using signature methods.

The rest of the paper is organised as follows: Section 2 gives some background on continuous time state-space models and path signature; Section 3 proves how the Kalman filter can be equivalently written as a linear regression on the signature space; Section 4 shows how to use regression on signature to nowcast; Section 5 shows how it is possible to replicate the performance of the Kalman filter by using regression on signature in a controlled environment; 6 applies the signature model to nowcast US GDP growth; finally Section 7 concludes.

2 Background

In this section, we outline some of the underlying theory we will rely on to build the signature method that we introduce.

2.1 Discrete time state-space models

We begin by introducing the notation that we use throughout the paper and recapping the classic discrete time Kalman filter (see also Bertsekas (2012)).

Suppose that we have a hidden process $Y \in \mathbb{R}^k$ that we cannot directly measure but would like to infer with the observed data $X \in \mathbb{R}^d$. At each time t, the Kalman filter splits into two stages. The first is the *prediction* stage: given our previous estimate of Y_{t-1}, how can we model its evolution to Y_t? The second is the *correction* stage: as more measurements are made, and we obtain information X_t, how can we use this to update the estimate of Y_t?

To be precise, let us assume that the ground truth is given by

$$Y_t = AY_{t-1} + W_t, \qquad X_t = CY_t + V_t,$$

with the starting distribution $Y_0 \sim N(\mu_{0|0}, P_{0|0})$. Here W, V are white noise processes in $\mathbb{R}^k$ and $\mathbb{R}^d$ respectively, with $W_t \sim N(0, \Gamma)$ and $V_t \sim N(0, \Sigma)$ for all t (all values independent). Note A and Γ are in $\mathbb{R}^{k \times k}$, and $C \in \mathbb{R}^{d \times k}$, $\Sigma \in \mathbb{R}^{d \times d}$. We assume all parameters are known.

For the prediction stage, we know that $Y_t | \mathcal{X}_t \equiv Y_t | (X_1, X_2, ..., X_t)$ is normal (and similarly so is $Y_t | \mathcal{X}_{t-1}$), and let us write $Y_t | \mathcal{X}_s \sim N(\mu_{t|s}, P_{t|s})$. Using the dynamics of Y and X, we can obtain the prediction equations

$$\begin{aligned} \mu_{t|t-1} &\equiv E[Y_t | \mathcal{X}_{t-1}] \\ &= A\mu_{t-1|t-1}, \\ P_{t|t-1} &\equiv \mathrm{var}(Y_t | \mathcal{X}_{t-1}) \\ &= AP_{t-1|t-1}A^\top + \Gamma. \end{aligned}$$

Before computing the Kalman correction step, let us define the "innovation" process η and its variance S

$$\begin{aligned} \eta_t &\equiv X_t - E[X_t | \mathcal{X}_{t-1}] = X_t - C\mu_{t|t-1}, \\ S_t &\equiv \mathrm{var}(\eta_t | \mathcal{X}_{t-1}) = \mathrm{var}(X_t | \mathcal{X}_{t-1}) = CP_{t|t-1}C^\top + \Sigma. \end{aligned}$$

The innovation process η tells us what information we "learn" from X_t. For the correction, we compute the new mean estimate to be

$$\begin{aligned} \mu_{t|t} &\equiv E[Y_t | \mathcal{X}_t] = E[Y_t | \eta_t, \mathcal{X}_{t-1}] \\ &= \mu_{t|t-1} + K_t \eta_t, \end{aligned}$$

where K is the "Kalman gain" process, which allows us to optimally incorporate new information,

$$K_t = P_{t|t-1}C^\top S_t^{-1} = (S_t^{-1}CP_{t|t-1})^\top.$$

Finally the variance correction equation is given by

$$\begin{aligned} P_{t|t} &\equiv \mathrm{var}(Y_t | X_t, \mathcal{X}_{t-1}) \\ &= (I - K_t C)P_{t|t-1}. \end{aligned}$$

This method also generalises to the case where the parameters are time-varying.

2.2 Continuous time state-space models

In order to treat variable timings of observations, it is natural to embed the discrete-time model in a continuous-time framework. The most natural continuous-time extension of the Kalman state-space model is the Kalman–Bucy filter[2]. As in the discrete time setting, we assume a hidden process Y and observed process X. We assume a ground truth of the form

$$dY_t = (F_t Y_t + f_t)dt + \sigma_t dV_t$$
$$dX_t = (H_t Y_t + h_t)dt + dW_t,$$

where $F \in \mathbb{R}^{d \times d}$, $\sigma \in \mathbb{R}^{d \times p}$, $f \in \mathbb{R}^d$, $H \in \mathbb{R}^{m \times d}$, $h \in \mathbb{R}^m$, so Y is d-dimensional, X is m-dimensional. The processes V, W are Brownian motions, and W is independent of Y.

The filter, which estimates the current state of the underlying process Y, is then given by the pair of equations

$$d\hat{Y}_t = (F_t \hat{Y}_t + f_t)dt + R_t H_t^\top (dX_t - (H_t \hat{Y}_t + h_t)dt)$$
$$\frac{dR_t}{dt} = \sigma_t \sigma_t^\top + F_t R_t + R_t F_t^\top - R_t H_t^\top H_t R_t.$$

With these equations, we have the result that, given our observations up to time t, Y_t has conditional distribution $\mathcal{N}(\hat{Y}_t, R_t)$. We can see that R_t is the solution to a Riccati equation, and in particular is deterministic (so does not depend on the observations X).

2.3 Paths and signatures

The path signature (Chen, 1958) is a property of continuous paths. It is motivated from the theory of rough path analysis, and have been shown to be effective in prediction tasks. In this section, we introduce the terminology and definitions related to paths and signatures. For a more detailed introduction see Chevyrev and Kormilitzin (2016) and Lyons et al. (2007).

A path in $\mathbb{R}^d$ is defined as $X : [a, b] \to \mathbb{R}^d$, where each component is a 1-D path $X^{(k)} : [a, b] \to \mathbb{R}$.

---

[2]For a detailed study of these equations, and derivation of the filter, see for example Bain and Crisan (2008) (whose notation we broadly follow) or Cohen and Elliott (2015, Chapter 22).

Let $X : [a, b] \to \mathbb{R}^d$ be a d-dimensional path and let

$$S^1(X^k)_{a,t} \equiv \int_{a<s<t} dX^k_s = X^k_t - X^k_a,$$

noting that the superscript denotes the number of iterated integrals we take, and the subscript indicates the limits of the outermost integral. Next let us define the double iterated integral as

$$S^2(X^k X^l)_{a,t} \equiv \int_a^t S^1(X^k)_{a,t} dX^l_s = \int_a^t \int_a^s dX^k_r dX^l_s.$$

Similarly, we define all the higher order index terms as iterated integrals, then the signature of the path is the ordered infinite collection of all such terms

$$S(X)_{a,b} \equiv \Big(1, S^1(X^1)_{a,b}, \ldots, S^1(X^d)_{a,b}, S^2(X^1 X^1)_{a,b}, S^2(X^1 X^2)_{a,b}, \ldots \Big).$$

The path signature captures geometric information, such as the order of events. The first level signature terms give the increment in each dimension between the beginning and end of the path. The second level terms are linked to the area bound by the path, in particular, cross terms represent the Lévy area. The Lévy area captures which dimensions are changing first.

There is a Stone–Wierstrass theorem which gives universality of this approximating class. Any continuous function on paths can be approximated arbitrarily well through a linear combination of signature terms.

As signature terms are iterated path integrals, they also inherit their invariance properties, that is, the terms are invariant to translation of the path and the reparameterisation in time.

The signatures describe the continuous path rather than the discrete points. As a result, choice of interpolation can matter as different paths are generated. But this also means that if sufficiently similar paths are generated, then we would not expect the signature to be greatly affected by the frequency of observation/sampling.

There are well-developed theories to support applications of signatures. Chen's identity gives us a way to compute the signature of a concatenated path as a algebraic product of individual paths, and this can be utilised for efficient computation of signatures for moving windows. The log-signature has a formal algebraic definition which reduces dimension without losing information. Where there are many random paths, the expected signature gives us a way to explore the space of paths.

2.4 Practical signature computation

In the previous section, we saw how the signature is a natural object/basis to consider for paths and time-series data in particular.

When working with real data, we only observe variables at discrete intervals. As a result, we need to interpret new data releases as coming from some underlying continuous time series and exploit connections to model the target variable. Using signatures as features is a natural way for market participants to make sense of data releases continuously in real-time and sidesteps data issues discussed below.

A common feature/issue for datasets is that different variables are collected at different frequencies. Some things are easy/cheap to measure and we may get daily updates, for example electricity and oil prices. Other data may be more expensive or difficult to measure, for example unemployment rate or personal income. Another issue is that there may be missing data. Survey data may not be collecting the same information on the same subject across time. Finally, the data collection can be very irregular and sporadic. By interpreting the time series as a path object, these problems can be resolved; provided the dimension of the path (number of variables) is fixed, the number of signatures at each level is fixed too.

Signatures can be efficiently computed through with the `iisignature` package in Python. Due to the invariance to re-parameterisation property of signatures mentioned in Section 2.3, it is common to manually add time as one of the input variables (as is typical in prediction problems).

There are two different ways to consider incoming data. We can either consider an *expanding window*, appending the new information to our existing data, or we can take a a *rolling window*: fixing a window size and discarding the older data.

In the case of the expanding window, we note that certain terms in the signature may be increasing. For example, the first level signatures terms are just the differences between the start and the end of the path, and therefore those corresponding to time would increase as more observations become available.

A classic result in rough paths is the factorial decay of the signature terms Lyons et al. (2007). Let $Y : [0, T] \to V$ be a path with finite one-variation. Then, for each $k \geq 1$,

$$\int_{0 < u_1 < \cdots < u_k < T} dY_{u_1} \otimes \cdots \otimes dY_{u_k} \leq \frac{||Y||^k_{1,[0,T]}}{k!}.$$

Therefore, as we increase the depth of signature terms, since the terms are decreasing factorially, we still obtain a reasonable approximation by trun-

cating the signature at a sufficiently high level (certain lower level terms may be increasing in the expanding window case).

However, it is important to note there is also an explosion in the number of terms. For a d dimension object if we truncate at k levels then there are $\sum_k d^k$ number of signature terms. So whilst signatures can universally approximate any path objects, we need to be wary of the exponentially increasing number of terms, which would increase model complexity, and may result in over-parameterisation.

It can be useful to add additional information by providing e.g. time since the last observation or counting the number of measurements observed in certain variables. These have been observed to boost performance in practice (Morrill et al., 2021).

In the last few years, signatures have been utilised in a range of applications with great success, these include Chinese handwriting recognition (Graham, 2013), malware detection (Cochrane et al., 2021), and time series generation (Ni et al., 2021).

3 Signatures generalize (linear) Kalman filters

Our claim is that the filter can be *equivalently written* as a linear regression problem on the signature space. In other words, we can express $\hat{Y}$ as a linear combination of the iterated integrals of the extended observation process (t, X_t).

3.1 Derivation

To do this semi-explicitly, we write

$$A_t \equiv F_t - R_t H_t^\top H_t$$

$$\xi_t \equiv \int_0^t (f_s - R_s H_s^\top h_s) ds + \int_0^t R_s H_s^\top dX_s,$$

which gives us the simplified expression

$$d\hat{Y}_t = A_t \hat{Y}_t dt + d\xi_t. \tag{1}$$

A naive approximation of this equation would be to take

$$\hat{Y}_t \approx \hat{Y}_0 + A_0 \hat{Y}_0 t + \xi_t.$$

11

However, this is only appropriate for small values of t, and assumes that A does not vary much through time. The goal of signature methods is to give a better approximation for this system, which can be used effectively for a wider range of time horizons.

Denote the iterated integrals of A and ξ by

$$\mathbb{A}_t^n = \int_0^t \int_0^{t_1} \cdots \int_0^{t_{n-1}} \left(A_{t_1} A_{t_2} \cdots A_{t_n} \right) dt_n \cdots dt_1$$

$$\Xi_t^n = \int_0^t \int_0^{t_1} \cdots \int_0^{t_{n-1}} \left(A_{t_1} A_{t_2} \cdots A_{t_n} \xi_{t_n} \right) dt_n \cdots dt_1,$$

with the conventions $\mathbb{A}^0 = \mathrm{I}_d$ (the identity matrix) and $\Xi_t^0 = \xi_t$. Note that these are both formally solutions of the recurrence relation $Q_t^n = \int_0^t A_s Q_s^{n-1} ds$ with different values for Q^0 (and in different dimensions). We note that the nth iterated integral is super-polynomially small, and in particular the infinite sums $\sum_{n \geq 0} \mathbb{A}_t^n$ and $\sum_{n \geq 0} \Xi_t^n$ are both well defined.

Writing $\hat{Y}$ in integral form, we have

$$\hat{Y}_t = \hat{Y}_0 + \int_0^t A_s \hat{Y}_s ds + \xi_t.$$

The value of $\hat{Y}$ can then be expressed through the series solution

$$\hat{Y}_t = \sum_{n \geq 0} \left(\mathbb{A}_t^n \hat{Y}_0 + \Xi_t^n \right).$$

Given the recurrence relation mentioned above, this is easily seen to be the (unique) solution to the integral equation. This shows that $\hat{Y}_t$ is a (linear) function of its initial value $\hat{Y}_0$ and the iterated integral processes $\mathbb{A}^n$ and Ξ^n. The expansion given is very closely related to the Picard series approximation of the stochastic differential equation defining $\hat{Y}$.

It remains to show that $\mathbb{A}$ and Ξ can be expressed in terms of the (joint) signature of (t, X_t). As A is a continuous deterministic function of time, we know that (over any finite time horizon) it can be approximated arbitrarily well by a polynomial (by the Stone–Weierstrass theorem). As the signature of time is simply the sequence $\mathbb{S}(t) = (1, t, t^2/2, ..., t^n/n!, ...)$, we see that A can be written as (matrix valued) a linear function of the signature of t.

The process ξ is slightly more delicate, as it depends on both time and the observations X. The first term $\int_0^t (f_s - R_s H_s^\top h_s) ds$ is deterministic, so again has a polynomial expansion in terms of the signature of t. Considering

the second term $\int_0^t R_s H_s^\top dX_s$, we see that if H is continuous[3], $R_s H_s^\top$ has an expansion in terms of the signature of time, so $\int_0^t R_s H_s^\top dX_s$ has an expansion in terms of the signature of (t, X_t), with the special form where the only integral with respect to X is the outermost one.

As both A and ξ have expansions in terms of the signatures of t and (t, X) respectively, it follows that their iterated integrals $\mathbb{A}$ and Ξ also have expansions of this type. For our purposes the explicit values of this expansion are not of particular interest (and do not have a simple algebraic form), but the definition of $\mathbb{A}$ immediately shows that if A can be written as a polynomial in time, then so can $\mathbb{A}^n$ for each n. Similarly for Ξ^n, but now this will involve iterated integrals with a single X integral included. (The absence of iterated integrals with more than one X integral is equivalent to the linear dependence of the filter $\hat{X}$ on the observations X.)

Example 1. *Consider the situation where* $d = p = m = 1$, $f = h = 0$, $F = -1$, $H = 1$ *and* $\sigma = \sqrt{3}$. *Then our filter equations simplify to*

$$d\hat{Y}_t = -\hat{Y}_t dt + R_t(dX_t - \hat{Y}_t dt)$$

$$\frac{dR_t}{dt} = 3 - 2R_t - R_t^2.$$

For simplicity, assume the initial variance is at the steady state $R_0 = R_t = 1$, *so we have*

$$d\hat{Y}_t = -2\hat{Y}_t dt + dX_t$$

which can be solved as

$$\hat{Y}_t = e^{-2t}\hat{Y}_0 + \int_0^t e^{-2(t-s)} dX_s.$$

The signature expansion of $\hat{Y}$ *can also be computed, as was done for* (1), *with the identity* $A_t = -2$ *and* $\xi_t = X_t$, *to give*[4]

$$\mathbb{A}_t^n = (-2)^n \frac{t^n}{n!} = (-2)^n S^n(t \cdots t)$$

$$\Xi_t^n = (-2)^n S^{n+1}(xtt \cdots t),$$

---

[3]If H is not continuous, then we need to use the fact that polynomials in time are dense in the $L^2([0, t])$ space, which is the appropriate space to consider given we have an outer integral with respect to the process X. In this case we will still have a polynomial approximation $p_n(s) \approx R_s H_s^\top$, such that the integral $\int_0^t p_n(s) dX_s \to \int_0^t R_s H_s^\top dX_s$ converges in a mean-square sense as $n \to \infty$.

[4]Here we write $S^n(...)$ for the n-fold iterated integral with respect to the sequence indicated (which must be of length n, with the innermost integral listed first); i.e. $S^3(xtt) = \int_0^t \int_0^{t_1} \int_0^{t_2} 1 \, dX_{t_3} \, dt_2 \, dt_1$. In this example all signature terms are defined for $[0, t]$ so we drop the subscript for convenience.

and hence

$$\hat{Y}_t = \sum_n \left(\mathbb{A}_t^n \hat{Y}_0 + \Xi_t^n \right)$$

$$= \hat{Y}_0 + \sum_{n \geq 1} \left([(-2)^n \hat{Y}_0] S^n(t \cdots t) + [(-2)^{n-1}] S^n(xtt \cdots t) \right)$$

$$= \hat{Y}_0 \left(1 - 2t + 2t^2 - \frac{8}{6}t^3 + \frac{16}{24}t^4 + \dots \right)$$

$$+ Y_t - 2\int_0^t X_s \, ds + 4 \int_0^t \int_0^{t_1} X_s \, ds \, dt_1 - 8 \int_0^t \int_0^{t_1} \int_0^{t_2} X_s \, ds \, dt_1 \, dt_2 + \dots$$

In particular, if t is small, the first few terms of this series provide a good approximation for the value of $\hat{Y}_t$, in terms of a linear function of these signature terms. Note that the naive approximation $\hat{Y}_t \approx \hat{Y}_0 - 2\hat{Y}_0 t + X_t$ appears as the first terms in this expansion.

However, we should remember that the simple structure we obtain is due to the assumptions we have made on our state-space model (in particular, the absence of signature terms of the form $S^n(t...txt...t)$ is due to the assumption the variance is in its steady state).

The key advantage of this approach, even while restricting our attention to the Kalman–Bucy state-space model, is that we now have an expansion which is valid, in theory for all t, and in practice for all t not too large. This simplifies dramatically the problem of working with data at mixed frequencies, as we can evaluate the filter state at any t, in terms of the corresponding signature terms, rather than having to compute (as is done, for example, in a MIDAS model) a version of the filter which depends on the timing of observations.

3.2 Nonlinear examples

A further advantage of the use of signature methods is that it allows us to easily incorporate nonlinearity in our state-space models. This can easily occur if our observations do not have a linear impact on the underlying system. For example, if our observations X were replaced in the above example by $\bar{X} = \log(X)$. Then, by the chain rule (which holds for Stratonovich integrals), we know

$$dX_t = e^{\bar{X}_t} d\bar{X}_t.$$

An immediate consequence of this is that the Kalman–Bucy filter can be written in terms of the modified observation process $\bar{X}$, and by very similar

arguments to before, this would have an expression in terms of the signature, but with terms involving multiple integrals with respect to $\bar{X}$.

This suggests that using the signature expansion is robust to the specification of the observation time series. That is, the traditional issues around the choice of transformations (whether to use a series, its logarithm, differences, etc...) are less significant, as they can be absorbed into the signature expansion.

4 Nowcasting via signature regression

Given the above representation of the Kalman filter mean in terms of the signature, we gain an easy approach to nowcasting, using regression on the signatures of our observations.

Let Y_t^{lf} be the (low-frequency) variable we want to nowcast. We write $X_{k,t}^{\mathrm{hf}}$, $k = 1, \ldots, K$, for the (high-frequency) observed explanatory variables at t. Motivated by the approximation of the Kalman filter in terms of signatures, we assume that

$$Y_t^{\mathrm{lf}} = \sum_{k=1}^{K} (\alpha_k + \beta_k Y_{t-}^{\mathrm{lf}}) \psi_{k,t} + \epsilon_t, \tag{2}$$

where

- Y_{t-}^{lf} is the most recent (at t) prior observation of the low frequency variable,

- ϵ_t is a mean-zero error term,

- $\psi_{k,t}$ is a sequence (for each value of t) of signature terms, including iterated integrals of t and the different components of the observed process X^{hf}, calculated over the interval from from the previous observation of Y to the present,

- α_k, β_k are two sequences of real regression coefficients.

As outlined above, in order to replicate the Kalman filter, we would have $\beta_k = 0$ whenever ψ_k corresponds to a signature term depending on X, and would only include signature terms which depend either purely on t, or on a single component of X appearing once in the iterated integral. Both of these restrictions can be relaxed, leading to a richer class of models than considered by the (linear) Kalman filter.

4.1 Regularisation

As the number of signature terms can still be numerous, the regression framework allows for the well-known regularisation techniques to be applied. In particular, we utilise L1-regularisation (Lasso) which also performs feature selection.

The fitted parameters $\hat{\beta}_k$ are obtained by Lasso regression, i.e. minimising

$$\min_{\beta_0, \beta_k} \sum_{t=0}^{T} \left(Y_t^{lf} - \beta_0 - \sum_{t=1}^{p} Y_{t-i}^{lf} - \sum_{k=1}^{K} \beta_k \psi_{k,t} \right)^2 \quad \text{s.t.} \quad \sum_{k=1}^{K} |\beta_k| \leq c, \quad (3)$$

where c is a free parameter that determines the degree of regularisation.

5 Simulation Exercise

In this section we show that it is possible to almost replicate the performance of the Kalman filter by using regression on signatures for a simulated experiment. For this, we look at the case when the underlying state is a Markov process in continuous time, and suppose that we observe the data at irregular intervals.

Practically, let us take an AR(1) process

$$Y_t = AY_{t-1} + \sigma W_t^{(2)} \tag{4}$$

$$X_t = BY_t + \gamma W_t^{(1)}, \tag{5}$$

where $W_t^{(1)}, W_t^{(2)}$ are independent, standard Gaussian random variables.

After simulating the complete time series, we then downsample (randomly) the data to some approximate proportion and apply our signature method on the reduced series. As the downsampling is random, this means that the gaps between observations will be random too.

5.1 Experiment set-up

Let us look at a 1-dimensional problem, and suppose that we want to nowcast the final value of the hidden state Y_T using only the initial value of the hidden state Y_0 along with the observations made at irregular times $X_0, \ldots X_T$.

Take $A = 0.99$, $B = 1$, $\sigma = 0.01$, and $\gamma = 0.1$. Let us simulate 400 trajectories of the above path for $T = 1000$ timesteps, and downsample randomly to only keep 20% of the data.

Suppose that we know the parameter values are known, that is, we have perfect information and do not need to learn any parameters from data.

Then, as this problem is linear, the Kalman filter should give the optimal solution.

For regression on signatures, we proceed as discussed in Section 4. Here we do not know any parameter values. From the observation values, we compute truncated signature terms, and scale it with the initial value of the hidden variable. We use 80% of the generated trajectories as training data (to calibrate the regression model) and the rest as the test/evaluation set.

5.2   Results

The application of the filter on a particular example, along with the Y_T value predicted by the signature method is given in Figure 1.
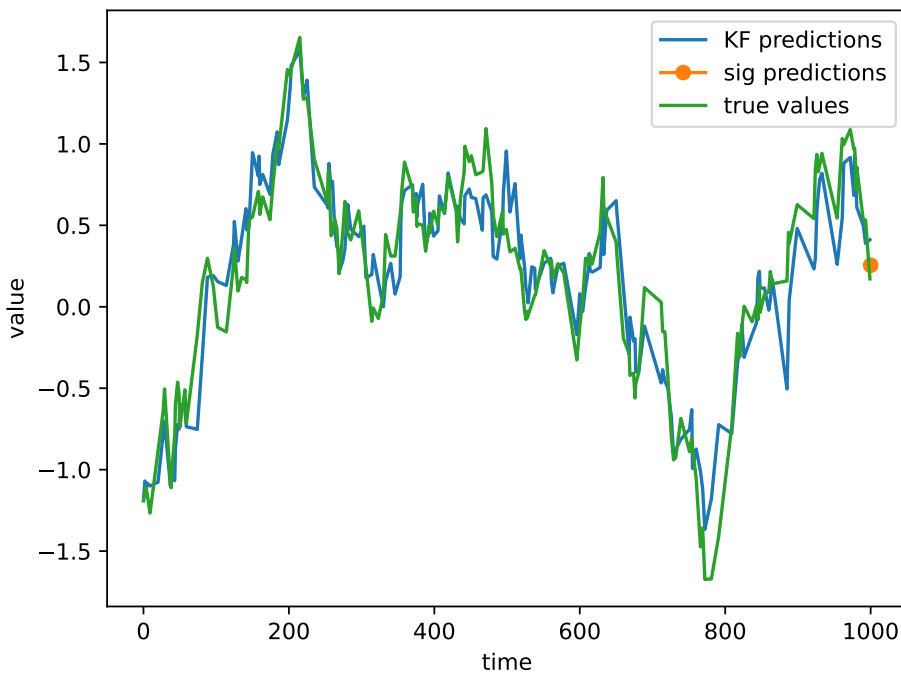

Figure 1: Predicted trajectory of the Kalman filter, along with the predicted final Y_T value of the signature.

For the Kalman filter, the mean of the residual is $\mu_{\mathrm{kf}} = 0.003$ and variance is $\sigma^2_{\mathrm{kf}} = 0.044$. The signature method has a mean of $\mu_{\mathrm{sig}} = -0.014$ and larger variance at $\sigma^2_{\mathrm{sig}} = 0.129$. The plot of the residuals of each method can be seen in Figure 2.

Plotting the residuals of the Kalman filter against that of the signature method, we obtain Figure 3. The line of best fit (minimal mean square error)

(a) Histogram of the residuals of the Kalman filter.

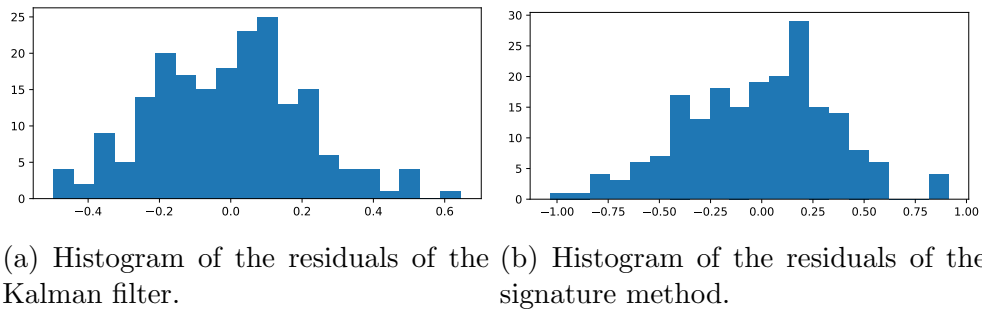(b) Histogram of the residuals of the signature method.

Figure 2: Residuals for simulation data

has gradient 0.96, and intercept -0.01. Therefore we see good alignment of the errors of the signature against the ideal filter given full information. In this plot, we see that the residuals are up to double the size for the signatures method compared with the Kalman filter, but that is not unexpected, given the Kalman filter has the exact parameters provided and is therefore theoretically optimal.

Overall, this experiment demonstrates that regression on signatures provides a competitive method to infer parameters from data.
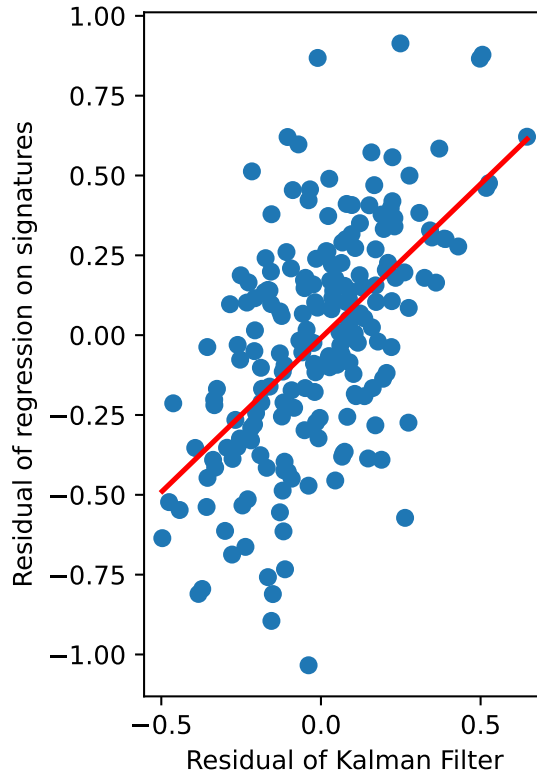
Figure 3: Residuals of the Kalman filter against the signature method.

6   Nowcasting US GDP growth

In this section we apply the signature methods to nowcast the quarter on quarter GDP growth for the United States. The statistical imprecision of early GDP estimates and the delay of its publication by a month since the end of reference quarter, has posed a challenge to economists and practitioners to monitor the state of the economy in real-time. In order to detect economic fluctuations in a more timely manner, economists, market participants, and media have scrutinized more frequent data series and built indicators (such as the composite leading indicator by OECD and the set of real-time indicators released by ONS). Unlike professional forecasts, that combined a number of models and judgement, Giannone et al. (2008) proposed a single formal model to utilise a variety of data to monitor economic conditions in real-time. The model proposed characterises current economic activity by condensing the information into a few factors that summarise economic conditions and it would be referred to here as the "New York Fed Staff Nowcast model".

In this application, we look at the nowcasting problem using a dynamic

factor model with the same variables underlying the New York Fed Staff Nowcast model in Bok et al. (2018), and compare this result with regression on signatures as described in Section 4.

6.1 Data

In order to nowcast the real GDP growth, we use 32 of the 36 variables proposed in Bok et al. (2018), since "ISM nonmanufactory: NMI composite index", "ISM mfg: Prices index", "ISM mfg.: PMI composite index" and "ISM mfg.: Employment index" are not publicly available. The dataset includes a selection of monthly variables that covers housing, income, manufacturing, labor, surveys, trade, and consumption. As in Bok et al. (2018), we apply the same factor loading. All the variables are affected by the "Global" factor, a "soft" block is included to model correlations in survey data; "real" and "labour" to model real and labour market variables respectively. For more details please refer to Table 5.1 of Bok et al. (2018).

6.2 The New York Fed Staff and the Signature Nowcast models

The model proposed by Giannone et al. (2008) and used in Bok et al. (2018) is a dynamic factor model (DFM). The DFM is a natural candidate to tackle the task at hand; the problem is cast into a state-space model form and hence the inference can be done using Kalman filtering techniques. From the discussion in Section 3, we have shown how linear regression on signature space generalises a Kalman filter, making the DFM model a suitable candidate to compare nowcast performance against the methodology introduced in this paper.

A dynamic factor model assumes that the target variable y_t is driven by r unobservable factors $(f_{1,t}, \ldots, f_{r,t})$, while the features that are specific to individual series, such as measurement errors are captured by $(e_{1,t}, \ldots, e_t)$. The state-space model is summarized in the following equations:

$$
\begin{aligned}
y_t &= \lambda_1 f_{1,t} + \cdots + \lambda_r f_{r,t} + e_t \\
f_{j,t} &= a_j f_{j,t-1} + u_{j,t}
\end{aligned}
\tag{6}
$$

which relates the data y_t to the r latent common factors $f_{1,t}, \ldots, f_{r,t}$. The factors and the idiosyncratic components are unobserved states. A model of this kind was first introduced by Stock and Watson (1989) to extract a single common factor ($r = 1$) from a small set of monthly indicators, which were then extended by others (e.g. Mariano and Murasawa, 2003; Aruoba

et al., 2009). However, large datasets like the one we use pose the issue of dealing with "large n (the number of predictors), small T (the length of time-series)". Bok et al. (2018) address this issue in two steps. First, they computed the estimations iteratively using the Kalman smoother and the EM algorithm, which is initialised by computing principal components. Second, given the estimated parameters, they update the common factors using a Kalman Smoother. Maximum likelihood estimation is obtained by computing these two steps until convergence.

We compare this model with the linear regression on the signature space as presented in Section 4. We use the same factors as computed by the DFM model, compute signatures terms up to order 3 and estimate the parameter of a linear regression model by OLS. Theoretically, the two models should be comparable, since we have shown in the previous sections how we can replicate the performance of the Kalman filter by using a regression on the signature. In the next section we replicate the exercise in Bok et al. (2018) and compare it with our nowcast from signature.

6.3 Results

Figure 4a reports the evolution of the nowcasts of annualised QoQ real GDP growth in quarter four of 2016 obtained with DFM and signature model. Annualised growth rates show the value that would be registered if the quarter-on-previous quarter rate of change were maintained for a full year. A new nowcast has been produced every time a new release of one or more of the 32 variables is available. The lines in the figure are the progression of the nowcasts throughout the reference quarter (1$^{\text{st}}$ of October to 31$^{\text{st}}$ of December) and one month after the reference quarter, until the BEA publishes the release. The official estimate for 2016Q4 (red circle) was published on January the 27$^{\text{th}}$. To consider the revision error, we have also included the latest estimate for 2016Q4, which was published the 30$^{\text{th}}$ of July 2020 (green square).

We can produce a nowcast very close to the official release with both models as early as mid-December; this evidence is in favour of our claim that regression in the signature space is comparable with Kalman filter used in DFM. However, comparing the two models' nowcasts, it seems that the DFM is more stable than the signature model since its estimates fluctuate less then the alternative model and display a smaller error in Figure 4b. It must be mentioned that to check the robustness of this result we need to compare the performance of the two models over a evaluation period longer than one quarter (as done here). This exercise will be presented in future versions of

this paper. From this picture we can also judge the marginal gain from using more timely information: it seems that after mid-December, the updates in predictors do not impact the nowcasts much. However it appears that the signature method picks up some signal in the final two sets of data releases as the estimate increases, which aligns more with the later releases of the GDP growth, compared with the DFM.

7 Conclusion

To summarise, the path signature, arising from theory, captures relevant geometric properties of sequential data. Signature methods naturally allow for missing data from mixed frequency and irregular sampling, issues often encountered in nowcasting, by embedding the observed data in continuous time.

In this paper we demonstrated the application of regression on signatures to nowcasting. First, we introduced the theory of the path signatures and illustrated how they are computed. We then showed that regression in the signature space subsumes the linear Kalman filter, which is commonly used in the nowcasting literature. Focused on practicality, we detailed the methodology of how to estimate nowcasts via signature regressions.

Second, we showed how it is possible to almost replicate the performance of the Kalman filter (when the parameters are known) by using regression on signatures (applied only on the data) for a simulated experiment. We then applied the model proposed in this paper to a well-known empirical exercise: nowcast GDP growth for US. We have shown that the results obtained are comparable with the ones published in Bok et al. (2018).

Although the application demonstrated the performance of signature methods to nowcast regularly sampled data, it is a very flexible tool that can address many of the issues encountered when dealing with micro-data or very irregularly sampled dataset; applications that could be shown in future extensions of this work.
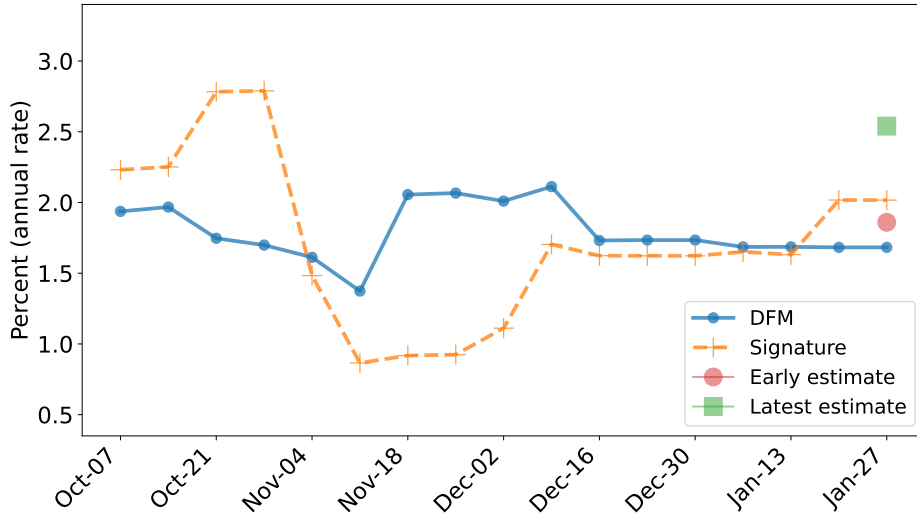
(a) Results for DFM and signatures to nowcast 2016 Q4 GDP growth. Note: the "Early estimate" is released in January 27th, 2016, while the "Latest" refers to the released in July 30th, 2020,
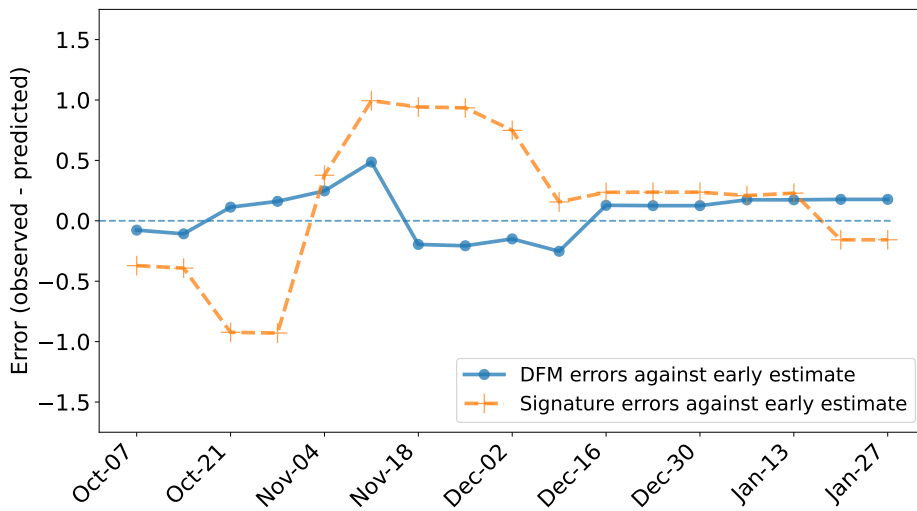
(b) Nowcast Error for DFM and Signature model

Figure 4: Results on the US GDP growth application.

References

Aruoba, S. B., Diebold, F. X., and Scotti, C. (2009). Real-time measurement of business conditions. *Journal of Business & Economic Statistics*, 27(4):417–427.

Babii, A., Ghysels, E., and Striaukas, J. (2021). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, pages 1–23.

Bai, J. and Ng, S. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics*, 3(2):89–163.

Bain, A. and Crisan, D. (2008). *Fundamentals of Stochastic Filtering*. Stochastic Modelling and Applied Probability. Springer New York.

Bańbura, M. and Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, 29(1):133–160.

Bertsekas, D. P. (2012). *Dynamic programming and optimal control*. Athena Scientific, 4th edition.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., and Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Available at SSRN 3102227*.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Chen, K.-T. (1958). Integration of Paths–A Faithful Representation of Paths by Noncommutative Formal Power Series. *Transactions of the American Mathematical Society*, 89(2):395–407.

Chevyrev, I. and Kormilitzin, A. (2016). A Primer on the Signature Method in Machine Learning. *arXiv*, abs/1603.03788.

Cochrane, T., Foster, P., Chhabra, V., Lemercier, M., Salvi, C., and Lyons, T. (2021). SK-Tree: a systematic malware detection algorithm on streaming trees via the signature kernel. *arXiv*, abs/2102.07904.

Cohen, S. and Elliott, R. (2015). *Stochastic Calculus and Applications*. Probability and Its Applications. Springer New York, 2nd edition.

Doz, C., Giannone, D., and Reichlin, L. (2006). A quasi maximum likelihood approach for large approximate dynamic factor models. *European Central Bank Working paper No. 674/September 2006.*

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38. OUP Oxford.

FRBNY (2016). Nowcasting report.

Geweke, J. (1977). *The dynamic factor analysis of economic timeseries models. In: Latent variables in socio-economic models.* Amsterdam: North-Holland Publishing Company.

Ghysels, E. and Marcellino, M. (2018). *Applied economic forecasting using time series methods.* Oxford University Press.

Ghysels, E., Pedro, S.-C., and Rossen, V. (2004). The midas touch: mixed data sampling regressions. *Discussion paper UNC and UCLA.*

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of monetary economics*, 55(4):665–676.

Giannone, D., Reichlin, L., and Small, D. H. (2006). Nowcasting gdp and inflation: the real-time informational content of macroeconomic data releases. *European Central Bank Working paper No. 633/May 2006.*

Graham, B. (2013). Sparse arrays of signatures for online character recognition. *arXiv*, abs/1308.0371.

Hamilton, J. D. (1994). State-space models. *Handbook of econometrics*, 4:3039–3080.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45.

Kapetanios, G., Papailias, F., et al. (2018). Big data & macroeconomic nowcasting: Methodological review. *Economic Statistics Centre of Excellence Discussion Paper 2018-12.*

Karlsson, S. (2013). Forecasting with bayesian vector autoregression. *Handbook of economic forecasting*, 2:791–897.

Kim, C.-J. (1994). Dynamic linear models with markov-switching. *Journal of econometrics*, 60(1-2):1–22.

Kim, C.-J., Nelson, C. R., et al. (1999). State-space models with regime switching: classical and gibbs-sampling approaches with applications. *MIT Press Books*, 1.

Litterman, R. (1979). Techniques of forecasting using vector autoregressions. Technical report, Federal Reserve Bank of Minneapolis.

Lyons, T. J., Caruana, M., and Lévy, T. (2007). *Differential equations driven by rough paths.* Springer.

Mariano, R. S. and Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of applied Econometrics*, 18(4):427–443.

Morrill, J., Kidger, P., Yang, L., and Lyons, T. (2021). Neural Controlled Differential Equations for Online Prediction Tasks. *arXiv*, abs/2106.11028.

Morrill, J. H., Kormilitzin, A., Nevado-Holgado, A. J., Swaminathan, S., Howison, S. D., and Lyons, T. J. (2020). Utilization of the signature method to identify the early onset of sepsis from multivariate physiological time series in critical care monitoring. *Critical Care Medicine*, 48(10):e976–e981.

Ni, H., Szpruch, L., Sabate-Vidales, M., Xiao, B., Wiese, M., and Liao, S. (2021). Sig-Wasserstein GANs for Time Series Generation. *arXiv*, abs/2111.01207.

Nielsen, A. and Berg, C. W. (2014). Estimation of time-varying selectivity in stock assessments using state-space models. *Fisheries Research*, 158:96–101.

Richardson, A., van Florenstein Mulder, T., and Vehbi, T. (2021). Nowcasting gdp using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, 37(2):941–948.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48.

Stock, J. and Watson, M. (2017). *Dynamic Factor Models. In: Clements MP, Henry DF Oxford Handbook of Economic Forecasting.* Oxford University Press.

Stock, J. H. and Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER macroeconomics annual*, 4:351–394.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.