Brexit: Tracking and disentangling the sentiment towards leaving the EU

Miguel de Carvalho[a*], Gabriel Martos[b]

[a] *University of Edinburgh, UK.* [b] *Universidad Torcuato di Tella, Argentina*

Abstract

On 23 June 2016 the UK held a referendum so to decide whether to stay or leave the European Union. The uncertainty surrounding the outcome of this referendum had major consequences in terms of public policy, investment decisions, and currency markets. We discuss some subtleties entailed in smoothing and disentangling poll data at the light of the problem of tracking the dynamics of the intention to Brexit, and propose a multivariate singular spectrum analysis method that produces trendlines on the unit simplex. The trendline yield via multivariate singular spectrum analysis is shown to bear a resemblance with that of local polynomial smoothing, and singular spectrum analysis presents the nice feature of disentangling directly the dynamics into components that can be interpreted as changes in public opinion or sampling error. Merits and disadvantages of some different approaches to obtain smooth trendlines on the unit simplex are contrasted, both in terms of local polynomial smoothing and of multivariate singular spectrum analysis.

Keywords: European Union politics, Local polynomial regression, Smoothing, Tracking public opinion, Multivariate singular spectrum analysis, UK's EU referendum, Unit simplex.

1. Introduction

23 June 2016, one of the most important days on the European Union (EU) and UK public policy agenda. A day marked in history, and a front-and-center referendum that opposed Europhiles against Europsceptics. Whatever the decision made by voters, perhaps the only truly accurate 'forecast' was made by former Prime Minister David Cameron, who remarked in a speech preceding the referendum (Cameron, 2013):

> "If we leave the EU, we cannot of course leave Europe. It will remain for many years our biggest market, and forever our geographical neighborhood."

The aftershock of a possible Brexit was widely analyzed by the media, analysts, and scholars and this has lead to a stream of materials on this matter on the press, working papers (Dhingra and Sampson,

*Corresponding author: School of Mathematics, James Clerk Maxwell Building, The King's Buildings, Peter Guthrie Tait Road, Edinburgh, EH9 3FD. Email: *Miguel.deCarvalho@ed.ac.uk*

2016), and academic research (Qvortrup, 2016b). For instance, in Cressey (2016), the author discuss the possible consequences of this referendum for research:

> "Millions in research funding, collaborations and the employment status of thousands of scientists could be affected by the outcome."

The final outcome of the referendum was 51.9% in favor to leave the EU.

In this article we examine methods for tracking and disentangling the dynamics governing public opinion in favor of Brexit, with a focus on how to obtain smooth trendlines that add up exactly to 100%—or in other words that live on the unit simplex

$$\Delta_D = \{\mathbf{y} = (y_1, \ldots, y_D) : y_1 + \cdots + y_D = 1, \text{where } y_d \geqslant 0 \text{ for } d = 1, \ldots, D\}. \tag{1}$$

Such normalization constraint is natural so that one obtains proportion estimates on the sentiment towards Brexit, Bremain, and undecided that obey the minimal requirement of adding up to one—at each period in time. Even though the applied focus here will be on the EU referendum—and thus $D = 3$ (Brexit, Bremain, undecided)—the approaches proposed in this article to produce trendlines that sum to one, apply more generally to the setting where a poll on $D - 1$ options (say, presidential candidates) are available to voters, and with a Dth option representing a group of undecided voters. We discuss both the situation where all poll data lives on the unit simplex—and thus can be regarded as compositional data—as well as the situation where the raw data are allowed to be $\varepsilon > 0$ apart from being in the unit simplex, thus living on a 'quasi-simplex,'

$$\Delta_{D,\varepsilon} = \{\mathbf{y} = (y_1, \ldots, y_D) : 1 - \varepsilon \leqslant y_1 + \cdots + y_D \leqslant 1 + \varepsilon, \text{where } y_d \geqslant 0 \text{ for } d = 1, \ldots, D\}. \tag{2}$$

The latter situation is actually the one occurring on our Brexit case study, as indeed the raw poll data actually lives in (2).

Since smoothing methods are widely applied by several outlets (e.g. *pollster.com*), we take here the opportunity to underscore that local polynomial methods (Fan & Gijbels, 1996) lead to estimates on the unit simplex as long as the raw data are on the unit simplex, and if the same bandwidth is used for all dimensions; this trivial feature—discussed in detail in Section 3.2—is handy from an applied viewpoint, as in particular it implies that the estimates yield through local polynomial regression can add up to one, if data from each poll individually adds up to one. Another contribution of this article rests on the proposal of a multivariate singular spectrum analysis (MSSA) method that can be used to smooth poll data and that yields trendlines on the unit simplex. Singular spectrum analysis (SSA) and MSSA have been widely applied recently in forecasting and nowcasting in a variety of contexts of

applied interest (e.g. Hassani et al., 2009, 2013; Golyandina & Korobeynikov, 2014; Golyandina et al., 2015; de Carvalho & Rua, 2017), but the potential of MSSA by poll data scientists may not have been yet fully appreciated. The trendline yield via MSSA is shown to bear a resemblance with that of local polynomial regression, while it presents the nice feature that it allows for disentangling the dynamics into components that can be interpreted as changes in public opinion or sampling error. Note that there is an important difference in the paradigm used to obtain the trendlines: Nonparametric regression models the conditional mean, whereas (M)SSA decomposes the trajectory of a stochastic process into so-called elementary reconstructed components. Finally, the paper offers an automatic criterion for selecting what components to aggregate so to obtain trendlines, by resorting to a Kolmogorov–Smirnov test on the cumulative periodogram. The proposed approach for learning about components from data is related to that of de Carvalho & Rua (2017), yet whereas in the latter the focus was on extracting components with hidden periodicities, here the goal is on learning about what components are related with trendlines.

The paper is organized as follows. We start the analysis in Section 2 with a data description and by conducting a preliminary exploratory analysis. In Section 3 we focus on local polynomial methods and particularly on how these can yield trendlines on the unit simplex. In Section 4, we introduce our multivariate singular spectrum analysis method for producing trendlines on the unit simplex. The `R` code to reproduce experiments in Section 4 is available in the supplementary materials.

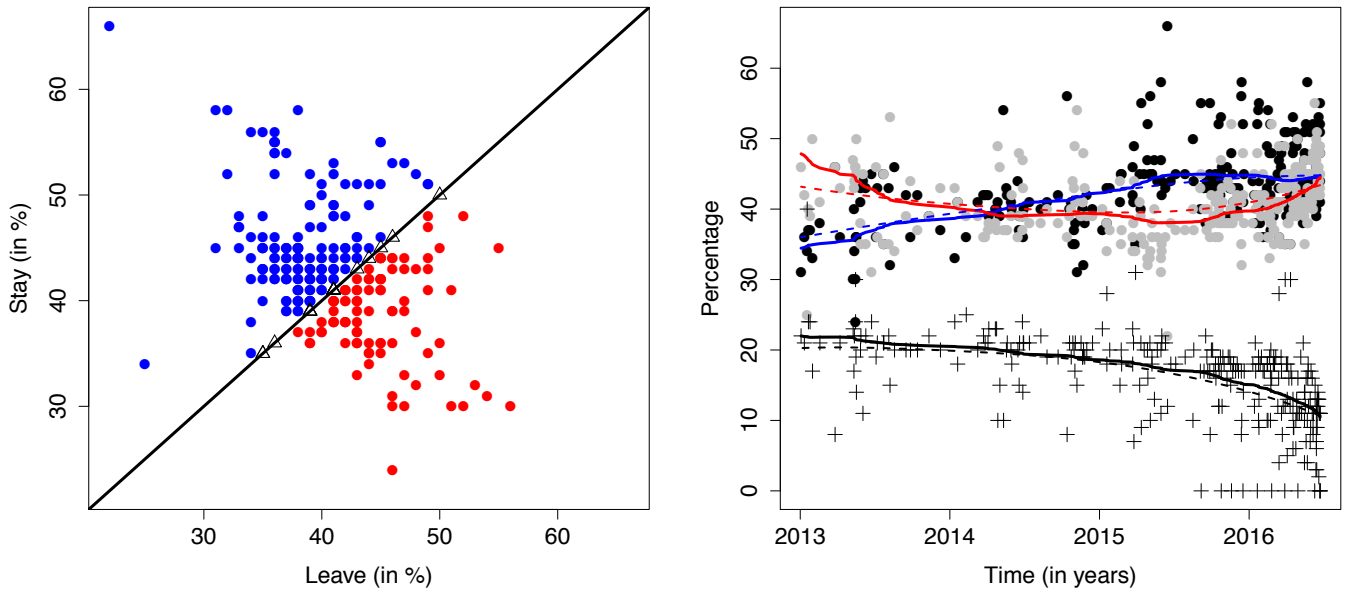
2. Poll tracker data and preliminary exploratory analysis

2.1. Brexit poll tracker data

The data were gathered from the Financial Times (FT) Brexit poll tracker, and are available from:

`https://ig.ft.com/sites/brexit-polling/`

The data consist of 267 polls which have been conducted by a variety of pollsters, including YouGov, ICM, Survation, and so on, by order of number of polls conducted. In Figure 1 we plot the data; the sample sizes range from 500 to 20058. In approximately 58% of the samples, stay was the 'winner', in the sense that more sampled subjects were in favor of staying. The amount of undecided was always fairly large though it has reduced over time. In Table 1 we provide the list of the pollsters entailed in this exercise, along with some summary statistics. The first poll was conducted in September 2010. However, before 2013 only other five polls were conducted. Thus, throughout the paper, we focus on



(a) Leave–Stay plot

(b) Chronological plot

Figure 1: Left: Leave–Stay plot representing the distribution of percentages between leave and stay in each poll. Polls where Brexit would win (•), loose (•), and ties (△). Right: Trendlines of the proportion in favor of leave (•), stay (•), and undecided (+); the dashed and solid lines represent smoothers obtained through local linear regression (Section 3.3) and singular spectrum analysis (Section 4), respectively, for leave (—), stay (—), and undecided (—).

presenting the results from Jan. 1, 2013 onwards; all data was however used for producing the figures reported in the paper.

In Figure 1, we plot what we will refer to as a 'leave–stay plot' (a) and a chronology of the outcome of all FT polls (b). Smoothing in Figure 1(b) was obtained through local linear methods (Fan & Gijbels, 1996); more on this in Section 3.3. Beyond sampling variability, one should expect to observe in these data variability due to the use of different methodologies ('house effect'), as well as due to the dynamics of public opinion over time. In particular, with respect to the latter point, it is interesting to observe how the number of undecided voters seems to diminish over time, as suggested from the local linear fit in Figure 1(b).

2.2. What can we learn from the last polls?

We start the analysis by focusing on the latest polls, which took place on Jun 22, 2016, and which has been conducted by Populus. Throughout, we will focus on the decided voters. In this part of the analysis, we will resort to standard methods for confidence intervals for a proportion, and the goal is on obtaining an (approximate) confidence interval for the population proportion of subjects in favor of stay. A popular confidence interval for a binomial proportion, p, is the Wald confidence interval $\text{CI}_\alpha^{\text{W}} = (\widehat{p} - z_{\alpha/2}\widehat{\sigma}_{\widehat{p}}, \widehat{p} + z_{\alpha/2}\widehat{\sigma}_{\widehat{p}})$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile from a standard normal distribution $N(0,1)$, and $\sigma_{\widehat{p}} = \sqrt{\text{var}(\widehat{p})} = \sqrt{p(1-p)/n}$. Although such approximate confidence intervals are implemented in many statistical packages—and widely used in practice—their performance has been heavily criticized (Brown et al., 2002), thus for precaution it seems sensible supplementing the analysis with another approach. There is however an interval with the same form, but with a different center $\widetilde{p}$, and a modified value for n, which is known to have better coverage properties, although it tends to overcover. Indeed, quoting Brown et al. (2002, p. 186):

> "To summarize, the conclusion is that the Agresti–Coull interval dominates the other intervals in coverage, but is also longer on an average and is quite conservative for p near 0 or 1."

Such modified interval is of the form:

$$\widetilde{\text{CI}}_\alpha^{\text{AC}} = (\widetilde{p} - z_{\alpha/2}\widetilde{\sigma}_{\widetilde{p}}, \widetilde{p} + z_{\alpha/2}\widetilde{\sigma}_{\widetilde{p}}), \quad \widetilde{\sigma}_{\widetilde{p}} = \sqrt{\widetilde{p}(1-\widetilde{p})/\widetilde{n}},$$

where $\widetilde{p} = \widetilde{X}/\widetilde{n}$, $\widetilde{n} = n + z_{\alpha/2}^2$, and $\widetilde{X} = X + z_{\alpha/2}^2/2$. Such intervals are known as Agresti–Coull confidence intervals (Agresti & Coull, 1998). Another popular approach is the Clopper–Pearson interval

$$\text{CI}_\alpha^{\text{CP}} = (B(\alpha/2; X, n-X+1), B(1-\alpha/2, X+1, n-X))$$

where B are the quantiles of a beta distribution. The Clopper–Pearson confidence interval is exact, and thus has a coverage at least equal to $1-\alpha$, for all $p \in (0,1)$, whereas the Wald and the Agresti–Coull are approximate confidence intervals, and thus may have coverage less than $1-\alpha$ for some values of p.

Table 1: Summary statistics for the 16 houses included in the poll tracker.

Polling house	Mean number of respondents	Number of polls
BMG Research	1464	8
ComRes	2258	16
Greenberg Quinlan Rosner Research	2327	2
Harris	2114	1
ICM	1883	47
Ipsos MORI	909	12
Lord Ashcroft Polls	20058	1
Opinium	1970	13
ORB	1294	19
Panelbase	1547	2
Pew Research Center	1006	2
Populus	3368	2
Survation	1879	22
TNS	1414	12
YouGov	1882	113

As advocated by Agresti & Coull (1998), one is often better off by using an approximate confidence interval than an exact, since the latter may be too conservative. However, there are situations where it may be beneficial to apply exact intervals. As posed by Thulin (2014, p. 818):

> "The benefit of using an exact interval is obvious: one does not risk that the actual coverage falls below $1-\alpha$. For this reason, some regulatory authorities require that exact intervals be used."

For the last poll, the Wald, Agresti–Coull, and Clopper–Pearson intervals would lead to a similar conclusion in the sense that there would be no differences by just examining the first three decimal places; the corresponding confidence interval for a population proportion in favor of stay is then $(0.536, 0.564)$. In Figure 2, we present the Agresti–Coull confidence interval for the proportion in favor Bremain for all polls focusing on decided voters only. It goes without saying that: i) there are obvious limitations with the application of these intervals here (e.g., random sample hypothesis may not apply); ii) there is a wealth of other approaches which could have been used for supplementing the results in this section, many of which documented in Fleiss et al. (2003). Still—with this disclaimer in mind—all in all, the evidence from the latest polls seemed to be fairly mixed but perhaps with a little more evidence in favor of a Bremain.

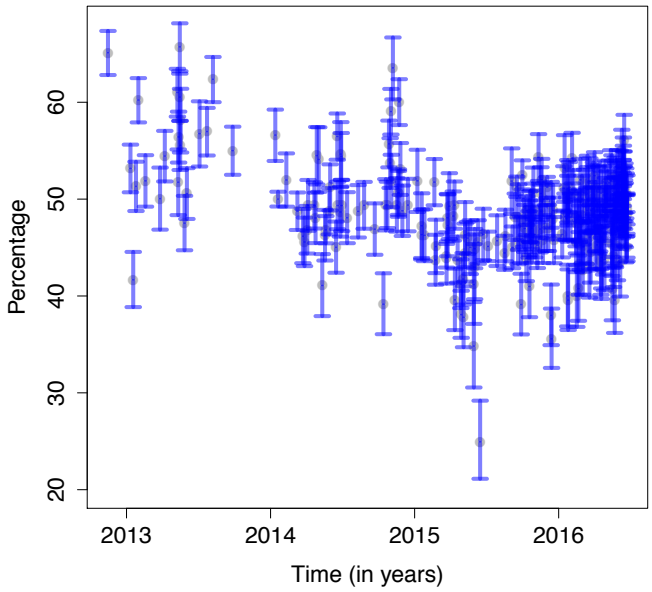


Figure 2: Agresti–Coull confidence intervals for the proportion in favor of Bremain for all polls, focusing on decided voters only.

3. Local polynomial smoothing of poll data

3.1. We are actually smoothing estimates

Kernel smoothing can be readily applied to obtain plots such as Figure 1(b), but what does this kind of plot rigorously means is another matter. There are indeed some key conceptual subtleties entailed in smoothing estimated proportions, which may pass unnoticed at first sight. Let N be the total number of polls, and $\widehat{p}_i$ be the estimated proportion in favor of Brexit, for $i = 1, \ldots, N$. Suppose for now that we only have binary leave–stay data, along with the times at which the polls took place $\{(t_i, \widehat{p}_i)\}_{i=1}^N$. The first subtle point is that one is actually smoothing estimates $(\widehat{p}_i)$—and thus we are smoothing a variable whose distribution changes with the number of observations (say, if $\widehat{p}_i$ is consistent, then it degenerates into a constant in the limit, as $n_i \to \infty$). To make this subtlety more precise, below we focus on the Nadaraya–Watson estimator (Nadaraya, 1964; Watson, 1964), but the ideas apply more generally to local polynomial smoothing Fan & Gijbels (1996). Let

$$\widetilde{p}(t) = \sum_{i=1}^{N} \pi_{i,h}(t) \widehat{p}_i^{(n_i)}, \tag{3}$$

be the smooth estimate of the proportion in favor of Brexit, where $\pi_{i,h}(t) = K_h(t - t_i) / \sum_{j=1}^N K_h(t - t_j)$, $h > 0$ is a smoothing parameter (bandwidth), and $K_h(\cdot) = h^{-1} K(\cdot)$, with K being a non-negative function integrating one

3.2. The estimate is on the unit simplex, if data are on the unit simplex

A second interesting aspect of smoothing estimated proportions is that, when data are of the type leave–stay–undecided, $(\widehat{p}, \widehat{s}, \widehat{u})$, a similar approach as in Equation (3) yields:

$$\widetilde{p}(t) = \sum_{i=1}^{N} \pi_{i,h}(t)\widehat{p}_i, \quad \widetilde{s}(t) = \sum_{i=1}^{N} \pi_{i,h}(t)\widehat{s}_i, \quad \widetilde{u}(t) = \sum_{i=1}^{N} \pi_{i,h}(t)\widehat{u}_i. \tag{4}$$

It thus follows that, if $\widehat{p}_i + \widehat{s}_i + \widehat{u}_i = 1$ for $i = 1, \ldots, N$, then

$$\widetilde{p}(t) + \widetilde{s}(t) + \widetilde{u}(t) = \sum_{i=1}^{N} \pi_{i,h}(t)(\widehat{p}_i + \widehat{s}_i + \widehat{u}_i) = 1,$$

and thus the estimates obtained through local smoothing still add up to one, for all t. A key ingredient for this to hold is of course that we are assuming the bandwidth to be the same along the three dimensions. This principle applies more generally for local polynomial regression, and if we observe proportions, $\widehat{\mathbf{p}} = (\widehat{p}_{i,1}, \ldots, \widehat{p}_{i,D})$, on the unit simplex Δ_D, as defined in (1). Indeed, by fitting a local polynomial regression individually on each dimension—as in (4), i.e.:

$$\widetilde{p}_1(t) = \sum_{i=1}^{N} \pi_{i,h}(t)\widehat{p}_{i,1}, \quad \ldots \quad , \widetilde{p}_D(t) = \sum_{i=1}^{N} \pi_{i,h}(t)\widehat{p}_{i,D}, \tag{5}$$

it follows that $\widetilde{\mathbf{p}}(t) = (\widetilde{p}_1(t), \ldots, \widetilde{p}_D(t))$ is in Δ_D, that is, $\widetilde{p}_1(t) + \cdots + \widetilde{p}_D(t) = \sum_{i=1}^{N} \pi_{i,h}(t)(\widehat{p}_{i,1} + \cdots + \widehat{p}_{i,D}) = 1$ for all t, if $\widehat{\mathbf{p}}_t \in \Delta_D$ for $t = \{t_1, \ldots, t_N\}$.

The Nadaraya–Watson estimator consists of a rolling weighting mean with weights defined by the kernel function K_h. Given the well-known boundary–bias issues of the Nadaraya–Watson method at the end of the sample period, it seems more sensible considering local linear weights (Wand & Jones, 1995), which yield

$$\widetilde{p}(t) = \sum_{i=1}^{N} \omega_{i,h}(t)\widehat{p}_i, \quad \widetilde{s}(t) = \sum_{i=1}^{N} \omega_{i,h}(t)\widehat{s}_i, \quad \widetilde{u}(t) = \sum_{i=1}^{N} \omega_{i,h}(t)\widehat{u}_i, \tag{6}$$

where

$$\omega_{i,h}(t) = \frac{\{\widehat{s}_2(t;h) - \widehat{s}_1(t;h)(t_i - t)\}}{\widehat{s}_2(t;h)\widehat{s}_0(t;h) - \widehat{s}_1(t;h)^2} K_h(t_i - t),$$

with $\widehat{s}_r(t;h) = N^{-1} \sum_{i=1}^{N} (t_i - t)^r K_h(t_i - t)$. Since some of the polls on the poll tracker data do not add up to one—in other words the poll tracker data are in $\Delta_{D,\varepsilon}$ as defined in Equation (2)—it follows that $\widetilde{p}(t) + \widetilde{s}(t) + \widetilde{u}(t)$ do not add up to one. A different approach for smoothing on the simplex, which relies on the geometry of the unit simplex, has been recently proposed by Bergman & Holmquist (2014).

3.3. Comments on smoothing parameters

Despite the fact that $\widehat{\mathbf{p}}_t \in \Delta_D$ implies that $\widetilde{\mathbf{p}}(t) \in \Delta_D$, for all t (cf. Equation (5))—provided a single bandwidth is used—one should bear in mind however that for some analysis—especially for moderate to large D—the cost of using the same bandwidth for all dimensions may not be small. The main inconvenience when using the same bandwidth to estimate each component of $\widetilde{\mathbf{p}}(t)$ is the optimality loss since $h_d^{\text{opt}} = c_d N^{-1/5}$ and the constants c_d need not to be equal, for $d = 1, \ldots, D$. In this sense, a natural way to choose asymptotically optimal bandwidths $\mathbf{h}^* = (h_1^*, \ldots, h_D^*)$ is by an 'aggregated' least squares cross validation (LSCV), so to minimize

$$\text{LSCV}(\mathbf{h}) = \sum_{d=1}^{D} \left\{ \frac{1}{N} \sum_{j=1}^{N} \int \left(\widetilde{p}_d^{(j)}(t, h_d)\right)^2 \mathrm{d}t - \frac{2}{N} \sum_{j=1}^{N} \widetilde{p}_d^{(j)}(t_j, h_d) \right\}, \tag{7}$$

where $\widetilde{p}_d^{(j)}(t, h_d)$ denotes the leave–one–out estimated smooth proportion $\widetilde{p}_d$ when the jth observation, $(\widehat{p}_{j,d}, t_j)$, is excluded from the sample using the bandwidth h_d. The optimal smoothing vector parameter is then $\mathbf{h}^* = \arg\min_{\mathbf{h} \in \mathcal{H}(N, \boldsymbol{\gamma}, \boldsymbol{\delta})} \text{LSCV}(\mathbf{h})$, where $\mathcal{H}(N, \boldsymbol{\gamma}, \boldsymbol{\delta}) = \{\mathbf{h} : N^{-1/5}\boldsymbol{\gamma} \leqslant \mathbf{h} \leqslant N^{-1/5}\boldsymbol{\delta}\}$ with constants $0 < \gamma_d < \delta_d < \infty$ such that $\gamma_d < c_d < \delta_d$, for $d = 1, \ldots, D$; in practice, we do not know (γ_d, δ_d) and simply search the minimum of $\text{LSCV}(\mathbf{h})$ over a suitable grid of values for $\mathbf{h}$. It is precisely the estimate in Equation (7) that it is presented in Figure 1(b).

Notice that if we impose the constraint $h_1 = \cdots = h_D$ when computing $\mathbf{h}^*$ by LSCV, we obtain smooth estimators which may not be 'optimal' (in the sense of (7)) but are 'coherent'—in the sense that they add up to one—if the raw data are on the unit simplex. Therefore a trade–off must be solved, in favor of optimality or coherency depending on the situation. An heuristic way to proceed is to use asymptotically optimal bandwidths if the raw data are not on the unit simplex, and project the obtained estimates onto the simplex in a second stage.[1] When data are on the unit simplex, the trade–off between the two alternatives should be carefully analyzed. When $\widetilde{\mathbf{p}}(t)$ produces 'incoherent estimators,' that is $\widetilde{\mathbf{p}}(t) \notin \Delta_D$, a projection of the estimates onto the unit simplex can be used, as we formally discuss in the next section (cf. Equation (16)), so to obtain estimates in the unit simplex—even if a different bandwidth is used over different dimensions, or even when the raw data are on the quasi-simplex as defined in (2). As remarked earlier, the latter situation is actually the one occurring on our Brexit case study, as indeed the raw poll data actually lives in (2).

[1] See the discussion surrounding Equation (16) for details on how the projection into the unit simplex could be computed.

3.4. Local regression is a poll of polls

The idea underlying smoothing as in Equation (6) is to interpolate the dynamics governing the state of opinion and to *borrow strength across polls* and from this viewpoint, the estimates produced through Equations (3)–(6) are a 'poll of polls;' the bandwidth controls how much the attention is confined simply to polls that take place near time t—by setting a lower bandwidth—or if we want to take on board more information from polls which are further away in time—by setting a higher bandwidth. Roughly speaking, polls of polls are approaches for aggregating poll data, and at the moment there is no consensus on what aggregation strategy is best (Pasek, 2015, p. 5). The FT itself did considered one such aggregation strategy:

> "The FT poll of polls is calculated by taking the last seven polls from unique pollsters up to a given date, removing the two polls with the highest and lowest shares for 'remain', and calculating an adjusted average of the five remaining polls, where the more recent polls are given a higher weight."

An important question remains however unanswered: Given the longitudinal nature of the data, is it fair pooling all the data and applying local linear methods?

3.5. But what about the longitudinal nature of the data?

A more natural way of regarding the poll tracker dataset would be as longitudinal data, $\{(\widehat{p}_{i,j}, t_{i,j})\}$, with $\widehat{p}_{i,j}$ denoting the proportion in favor of Brexit at time $t_{i,j}$, as assessed by the jth poll of the ith pollster, for $i = 1, \ldots, m$, and $j = 1, \ldots, J_i$. A surprising and counterintuitive result by Lin & Carroll (2000), suggests however that when faced with longitudinal data one is actually better off by pooling the data and running a standard nonparametric regression analysis, possibly accounting for variability, than taking into account the correlation structure. The intuition for the result is as follows (Lin & Carroll, 2000, p. 521):

> "As the bandwidth becomes smaller, the chance that correlated observations from the same cluster fall in the same bandwidth vanishes and the observations essentially behave independently."

Thus, one can ignore the within-subject correlation and pretend that observations are independent; for strategies which take advantage of the within-subject correlation see for instance Yao & Li (2013) and references therein.

We now switch gears and discuss how singular spectrum analysis would perform at tracking the dynamics of the intention to Brexit.

4. Tracking public opinion with singular spectrum analysis

In this section we show that singular spectrum analysis yields a trendline that is shown to bear a resemblance with that of local linear regression. On top of this, it can be used for disentangling the dynamics of poll data into components that can be either interpreted as changes in public opinion or due to sampling error. But what is singular spectrum analysis?

4.1. A concise description on the univariate setting

The method entails two phases, namely decomposition and reconstruction, and each of these phases includes two steps; the phase of decomposition includes the steps of embedding and singular value decomposition, which we discuss below. Let $\mathbf{p} = (p_1, \ldots, p_N)$ denote a univariate time series of proportion data from which we intend to extract information on the state of opinion.

Embedding. This is the preliminary step of the method. SSA starts by organizing the original time series of interest, $\mathbf{p}$, into a so-called trajectory matrix $\boldsymbol{X}$, i.e., a matrix whose columns consist of rolling windows of length L, i.e.

$$\boldsymbol{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_K] = \begin{bmatrix} p_1 & p_2 & \cdots & p_K \\ p_2 & p_3 & \cdots & p_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ p_L & p_{L+1} & \cdots & p_N \end{bmatrix}. \tag{8}$$

Here, L denotes the window length, and $K = N - L + 1$; L is a signal–noise separation parameter that must be assigned by the user and which plays a similar role to that of the bandwidth in nonparametric regression. Keeping in mind theoretical considerations in Hassani & Mahmoudvand (2013, Section 4.1), we suggest considering the window length $L = \lceil (N+1)/2 \rceil$, where $\lceil \cdot \rceil$ is the ceiling function.[2] Since all elements over the diagonal $i + j = const$, are equal, the trajectory matrix $\boldsymbol{X}$ is a Hankel matrix.

Singular value decomposition. In the second step we perform a singular value decomposition of the trajectory matrix. Let $\lambda_1, \ldots, \lambda_L$ denote the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, presented in decreasing order, and let $\mathbf{u}_1, \ldots, \mathbf{u}_L$ denote the corresponding eigenvectors. In this step, we decompose the trajectory matrix $\boldsymbol{X}$ as follows

$$\boldsymbol{X} = \sum_{i=1}^{d} \boldsymbol{X}_i, \tag{9}$$

[2]The latter is also the default option in the `Rssa` package (Golyandina & Korobeynikov, 2014).

where $\boldsymbol{X}_i = \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^{\mathrm{T}}$, $\mathbf{v}_i = \boldsymbol{X}^{\mathrm{T}}\mathbf{u}_i/\sqrt{\lambda_i}$, and $d = \max\{i \in \{1,\ldots,L\} : \lambda_i > 0\}$.

Below we discuss the second phase of the method—reconstruction, which entails the steps of grouping cyclical components and diagonal averaging.

Grouping. Not all terms in Equation (9) contain relevant information on the state of opinion, and hence we confine ourselves to a subset I of $\{1,\ldots,d\}$, so to compute $\boldsymbol{X}_I = \sum_{i\in I}\boldsymbol{X}_i$. The objective of the grouping step is on disentangling signal from noise, and in such case we typically write $\boldsymbol{X} = \sum_{i\in I}\boldsymbol{X}_i + \sum_{i\notin I}\boldsymbol{X}_i$, with the components in $I \subset \{1,\ldots,d\}$ representing the signal. Further details on the grouping stage are presented later in this section, where we resort to a cumulative periodogram-based method for learning about the signal [cf Equation (12)].

Diagonal averaging. In this step we average over all the elements of the diagonal $i+j = const$ of $\boldsymbol{X}_I$ so to obtain a Hankel matrix, from where our trendline indicator, $\widetilde{\mathbf{p}} = (\widetilde{p}_1,\ldots,\widetilde{p}_N)$, can be constructed. Essentially, our trendline indicator is constructed by averaging the matrix over the 'antidiagonals' $i+j = k+1$. Let $x^I_{i,j} = [\mathbf{X}_I]_{i,j}$ then for $k=1$, $\widetilde{p}_1 = x^I_{1,1}$; $k=2$, yields $\widetilde{p}_2 = (x^I_{1,2}+x^I_{2,1})/2$; $k=3$, yields $\widetilde{p}_3 = (x^I_{1,3}+x^I_{2,2}+x^I_{3,1})/3$; *etc.* Generalizing this simple idea, we can construct our trendline indicator through the map

$$\widetilde{\mathbf{p}} = \overline{\mathbb{D}}(\boldsymbol{X}_I) = \left(\frac{1}{|\mathcal{A}_1|}\sum_{(i,j)\in\mathcal{A}_1} x^I_{i,j},\ldots,\frac{1}{|\mathcal{A}_N|}\sum_{(i,j)\in\mathcal{A}_N} x^I_{i,j}\right), \tag{10}$$

where $|A|$ denotes cardinality of a set A, and where the sequence of sets

$$\mathcal{A}_k = \{(i,j) : i+j = k+1, i \in \{1,\ldots,L\}, j \in \{1,\ldots,K\}\}, \quad k = 1,\ldots,N, \tag{11}$$

defines the elements of the N 'antidiagonals' of the matrix $\boldsymbol{X}_I$.

Some comments on the grouping and diagonal averaging stages are in order. In practice, criteria such as the ratio $\lambda_i/\sum_{j=1}^d \lambda_j$, the scree plot $\{(j,\lambda_j)\}_{j=1}^d$ or the pair plot of eigenvectors, are often used to guide on the selection of I; see for instance Hassani & Mahmoudvand (2013), Rodrigues & Mahmoudvand (2018), and references therein. For automatically learning about trendlines, we resort to spectral analysis over the grouping and diagonal averaging stages. Recall that the periodogram of a time series $\mathbf{y} = (y_1,\ldots,y_N)$ consists of $\{(\omega_j, \mathbb{I}(\omega_j))\}_{j=1}^J$, where $\omega_j = 2\pi j/N$ are Fourier frequencies, for $j = 1,\ldots,J = \lfloor N/2\rfloor$, with $\lfloor\cdot\rfloor$ denoting the floor function, and

$$\mathbb{I}(\omega_j) = \frac{1}{N}\left|\sum_{t=1}^N y_t \mathrm{e}^{-\mathrm{i}t\omega_j}\right|^2.$$

We found the following cumulative periodogram-based criterion to be a sensible option for learning about components underlying trendlines:

Targeted grouping based on the Kolmogorov–Smirnov statistic

Set $i = 1$ and execute the steps:

Step 1. Compute the residuals $\boldsymbol{\varepsilon}_p = \mathbf{p} - \widetilde{\mathbf{p}}$, yield from the trendline based on $I = \{1, \ldots, i\}$.

Step 2. Compute the cumulative periodogram of $\boldsymbol{\varepsilon}$, and test the null hypothesis of white noise using the Kolmogorov–Smirnov test based on the statistic

$$\sqrt{J} \max\{|C(\omega_j) - j/J|\}_{j=1}^{J}, \quad C(\omega_j) = \frac{\sum_{i=1}^{j} \mathbb{I}(\omega_i)}{\sum_{i=1}^{J} \mathbb{I}(\omega_i)}. \tag{12}$$

If the null is rejected and $\int_0^\pi C(\omega)\, \mathrm{d}\omega > \pi/2$, then increment i and repeat Steps 1 and 2. Otherwise stop.

The proposed approach is related to that of de Carvalho & Rua (2017), who also resort to inference for the spectrum so to automatically select components; our approach differs however from the latter as their target was on learning about components related with hidden periodicities, whereas here the goal is on keeping track of components that actually contribute to the trend. Details on the Kolmogorov–Smirnov test on the cumulative periodogram in (12) can be found in Brockwell & Davis (1991, Section 10.2), and it is known that asymptotically the maximum deviation from the straight line corresponding to white noise has a law given by that of the Kolmogorov–Smirnov statistic, having an approximate 95% limit of $1.358/[\sqrt{J} + 0.11 + 0.12/\sqrt{J}]$ (Venables & Ripley, 2002, Section 14.1). The supplementary materials contain Monte Carlo evidence illustrating the performance of the automatic criterion based on Steps 1 and 2 above. Although our simulation study suggests that the cumulative periodogram-based method tends to perform well, in practice we suggest supplementing the automatic recommendation of the criterion with further diagnostics, and to conduct inferences over a range of values nearby the values suggested by the method.

4.2. Revisiting the dynamics of poll tracker data: Take I

In Figure 3, we plot the first elementary reconstructed components resulting from applying singular spectrum analysis; here we took $L = \lceil (N+1)/2 \rceil = 137$ using the criterion discussed on p. 11. For learning about components which account for the trendline, we resort to the automatic criterion based on the cumulative periodogram discussed in Section 4.1. The criterion suggests as trendline estimators for the proportion in favor of Brexit, the first two components relative to the trajectory matrix, that is $\widetilde{\mathbf{p}} = \mathbb{D}(\boldsymbol{X}_{I_p})$, with $I_p = \{1, 2\}$. A similar analysis was conducted for the estimated trendline regarding the proportion in favor of Bremain to obtain $\widetilde{\mathbf{s}} = (\widetilde{s}_1, \ldots, \widetilde{s}_N)$ and the estimated trendline regarding

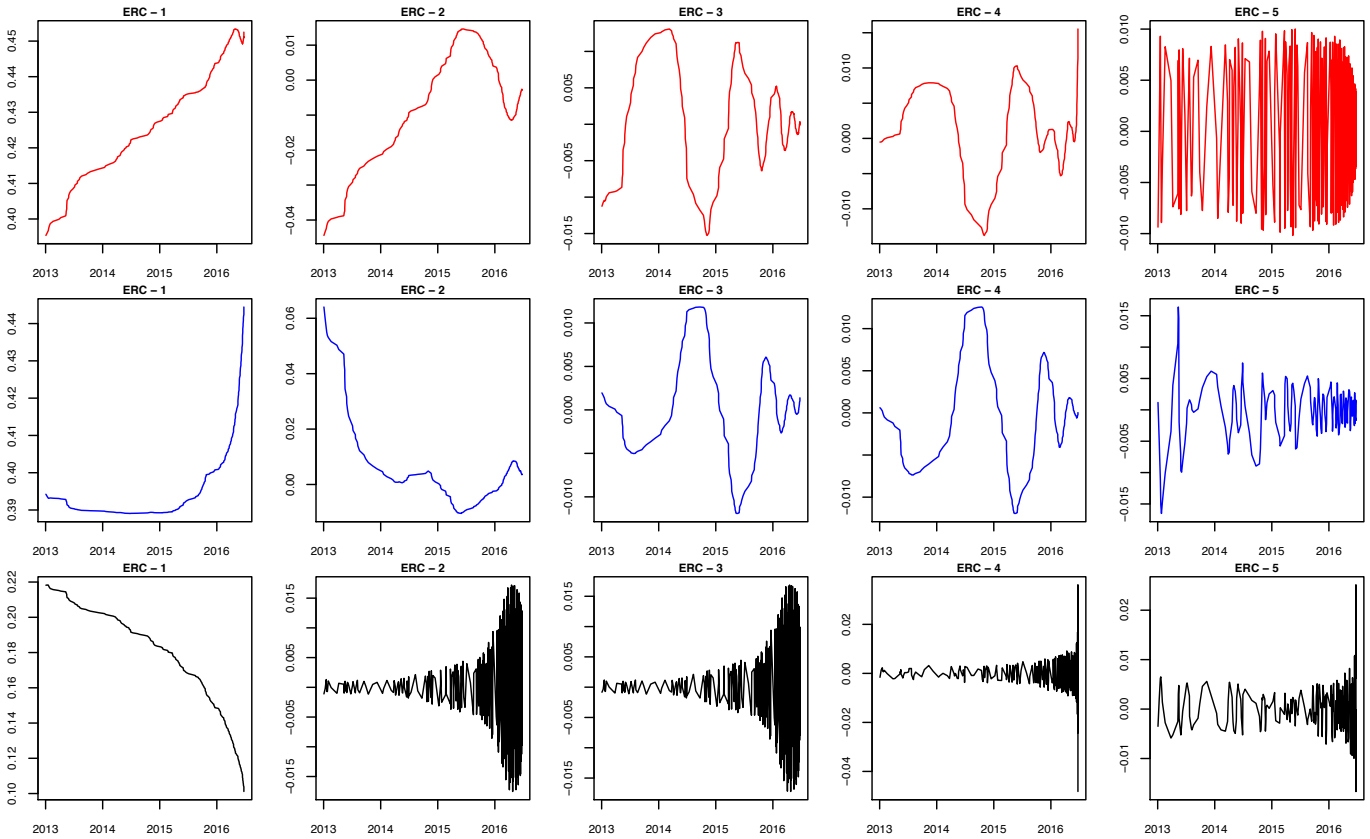


Figure 3: SSA elementary reconstructed components (ERC) for leave (—), stay (—), and undecided (—).

the proportion of undecided voters $\widetilde{\mathbf{u}} = (\widetilde{u}_1, \ldots, \widetilde{u}_N)$, respectively based on $I_s = \{1, 2\}$ and $I_u = \{1\}$. In Figure Figure 1(b) we present the estimated trendline triplet $(\widetilde{\mathbf{p}}, \widetilde{\mathbf{s}}, \widetilde{\mathbf{u}})$, the estimated dynamics in public opinion fits with the tendency of the polls to bounce up the opinion in favor of Stay from 2013 to the end of 2015 and then remain more or less constant until the day of elections. The proportion of voters in favor of Brexit starts to diminish considerably at the end of year 2014 and later start to grow fast from the second semester of year 2015. The proportion of undecided voters diminishes as the election approaches.

The univariate SSA trendline estimator does not take into account the correlation between the trends and does not ensure the estimated proportion triplet belongs to the probability simplex, that is $\mathbf{1} = \widetilde{\mathbf{p}} + \widetilde{\mathbf{s}} + \widetilde{\mathbf{u}}$, where $\mathbf{1}$ is an N vector of ones. Next, we propose a multivariate singular spectrum analysis (MSSA) that take into account these drawbacks.

4.3. MSSA yielding trendlines on the unit simplex

In the multivariate case, we consider D time series. In our case $D = 3$ and the array of time series is $(\mathbf{p}, \mathbf{s}, \mathbf{u}) = ((p_1, \ldots, p_N); (s_1, \ldots, s_N); (u_1, \ldots, u_N))$. Next, we follow the four steps below; the steps can be extended in a straightforward way to the setting where D is any positive integer.

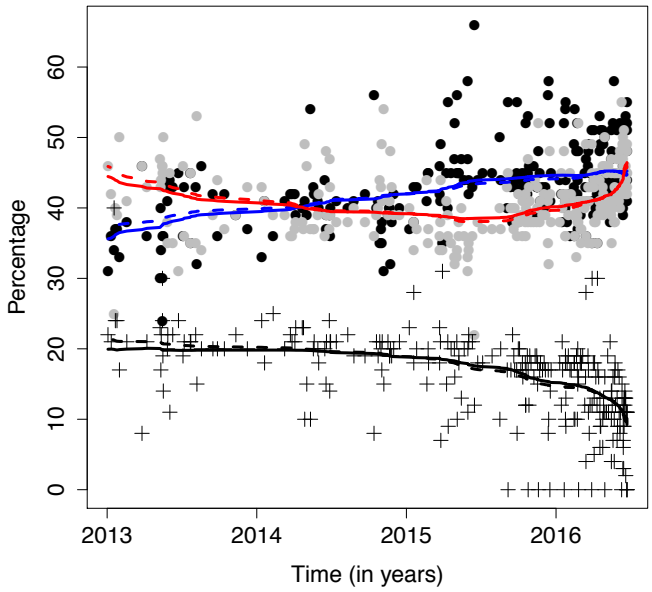


Figure 4: Trendlines of the proportion in favor of leave (•), stay (•), and undecided (+); the dashed and solid lines represent smoothers obtained from multivariate singular spectrum analysis—without imposing the normalization constraint—and multivariate singular spectrum analysis—when imposing the normalization constraint—, respectively, for leave (—), stay (—), and undecided (—).

Embedding. Let $\boldsymbol{X}_p = [\mathbf{x}_1^p \ \cdots \ \mathbf{x}_{K_p}^p]$, $\boldsymbol{X}_s = [\mathbf{x}_1^s \ \cdots \ \mathbf{x}_{K_s}^s]$ and $\boldsymbol{X}_u = [\mathbf{x}_1^u \ \cdots \ \mathbf{x}_{K_u}^u]$ be the trajectory matrices for $\mathbf{p}$, $\mathbf{s}$ and $\mathbf{u}$, with window lengths L_p, L_s and L_u, respectively. Next, we bind these matrices into a block Hankel matrix $\boldsymbol{X}$ as follows

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_p \\ \boldsymbol{X}_s \\ \boldsymbol{X}_u \end{bmatrix}. \tag{13}$$

Keeping in mind a theoretical result in Hassani & Mahmoudvand (2013, Section 4.1), on the window length achieving maximum rank, we suggest considering a single window length, given by $L = \lceil (N+1)/(D+1) \rceil$.

Singular value decomposition.[3] Let $\lambda_1, \ldots, \lambda_{3L_{sum}}$ be the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ sorted in decreasing order, and let $\mathbf{u}_1, \ldots, \mathbf{u}_{3L_{sum}}$ be the corresponding eigenvectors; we decompose the trajectory matrix

[3] As mentioned by a reviewer, it is important to note that the SVD for $\boldsymbol{X}$ and the transposed $\boldsymbol{X}$ are the same up to the interchange of left and right singular vectors, and thus it is not important how to stack the trajectory matrices of single series, vertically or horizontally, in (13) if to change L correspondingly.

as $\boldsymbol{X} = \sum_{i=1}^{L_{sum}} \boldsymbol{X}_i$, where $\boldsymbol{X}_i = \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^{\mathrm{T}}$, $\mathbf{v}_i = \boldsymbol{X}^{\mathrm{T}}\mathbf{u}_i/\sqrt{\lambda_i}$, and $L_{sum} = L_p + L_s + L_u$. Notice that in

$$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{X}_p\boldsymbol{X}_p^{\mathrm{T}} & \boldsymbol{X}_p\boldsymbol{X}_s^{\mathrm{T}} & \boldsymbol{X}_p\boldsymbol{X}_u^{\mathrm{T}} \\ \boldsymbol{X}_s\boldsymbol{X}_p^{\mathrm{T}} & \boldsymbol{X}_s\boldsymbol{X}_s^{\mathrm{T}} & \boldsymbol{X}_s\boldsymbol{X}_u^{\mathrm{T}} \\ \boldsymbol{X}_u\boldsymbol{X}_p^{\mathrm{T}} & \boldsymbol{X}_u\boldsymbol{X}_s^{\mathrm{T}} & \boldsymbol{X}_u\boldsymbol{X}_u^{\mathrm{T}} \end{bmatrix}, \tag{14}$$

the elements in the diagonal of the block matrix $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, relevant in the diagonal averaging step, correspond to the univariate SSA decomposition of each individual series.

Grouping, diagonal averaging, and projection onto the simplex. We allow for each trendline to be constructed from a different number of elementary components, that is

$$\widetilde{\mathbf{p}} = \mathbb{D}(\boldsymbol{X}^p_{I_p}), \quad \widetilde{\mathbf{s}} = \mathbb{D}(\boldsymbol{X}^s_{I_s}), \quad \widetilde{\mathbf{u}} = \mathbb{D}(\boldsymbol{X}^u_{I_u}),$$

where we learn about the sets I_p, I_s, and I_u by using the targeted grouping approach described in p. 13, and with

$$\begin{bmatrix} \boldsymbol{X}_{I_p^p} \\ \boldsymbol{X}_{I_p^s} \\ \boldsymbol{X}_{I_p^u} \end{bmatrix} = \sum_{i\in I_p} \boldsymbol{X}_i, \quad \begin{bmatrix} \boldsymbol{X}_{I_s^p} \\ \boldsymbol{X}_{I_s^s} \\ \boldsymbol{X}_{I_s^u} \end{bmatrix} = \sum_{i\in I_s} \boldsymbol{X}_i, \quad \begin{bmatrix} \boldsymbol{X}_{I_u^p} \\ \boldsymbol{X}_{I_u^s} \\ \boldsymbol{X}_{I_u^u} \end{bmatrix} = \sum_{i\in I_u} \boldsymbol{X}_i,$$

and $\mathbb{D}$ defined as in (10). To ensure that the triple of trendlines is in the unit simplex at every period, that is $\widetilde{\mathbf{p}} + \widetilde{\mathbf{s}} + \widetilde{\mathbf{u}} = \mathbf{1}$, we solve the optimization problem

$$\underset{\mathbf{y}_t\in\Delta_3}{\text{minimize}} \ \|\boldsymbol{\beta}_t - \mathbf{y}_t\|^2, \quad t \in \{1,\dots,N\}, \tag{15}$$

where $\boldsymbol{\beta}_t = (\tilde{p}_t, \tilde{s}_t, \tilde{u}_t)$ and Δ_3 is the unit simplex in $\mathbb{R}^3$. Wang & Carreira-Perpiñán (2013) propose an efficient way to compute the solution to the problem in (15), given by

$$y_{i,t} = ([\boldsymbol{\beta}_t]_i + \lambda_t, 0), \quad i \in \{1,2,3\}. \tag{16}$$

Here $\lambda_t = (1 - \sum_{i=1}^{\rho_t} \beta_{[i],t})/\rho_t$,

$$\rho_t = \left\{ 1 \le j \le 3 : \beta_{[j],t} + \frac{1}{j}\Big(1 - \sum_{i=1}^{j} \beta_{[i],t}\Big) > 0 \right\}, \quad t \in \{1,\dots,N\},$$

and $\beta_{[i],t}$ denotes the ith element after sorting the vector $\boldsymbol{\beta}_t$ in a decreasing way. In other words, our trendlines for leave (p_t), stay (s_t), and undecided (u_t) are given by

$$\widehat{p}_t = (\tilde{p}_t + \lambda_t, 0), \quad \widehat{s}_t = (\tilde{s}_t + \lambda_t, 0), \quad \widehat{u}_t = (\tilde{u}_t + \lambda_t, 0). \tag{17}$$

Parenthetically, we note that this step could be formulated as a single optimization problem, keeping in mind that diagonal averaging itself stems from a Frobenius norm minimization problem (Golyandina et al., 2001, Proposition 6.3).

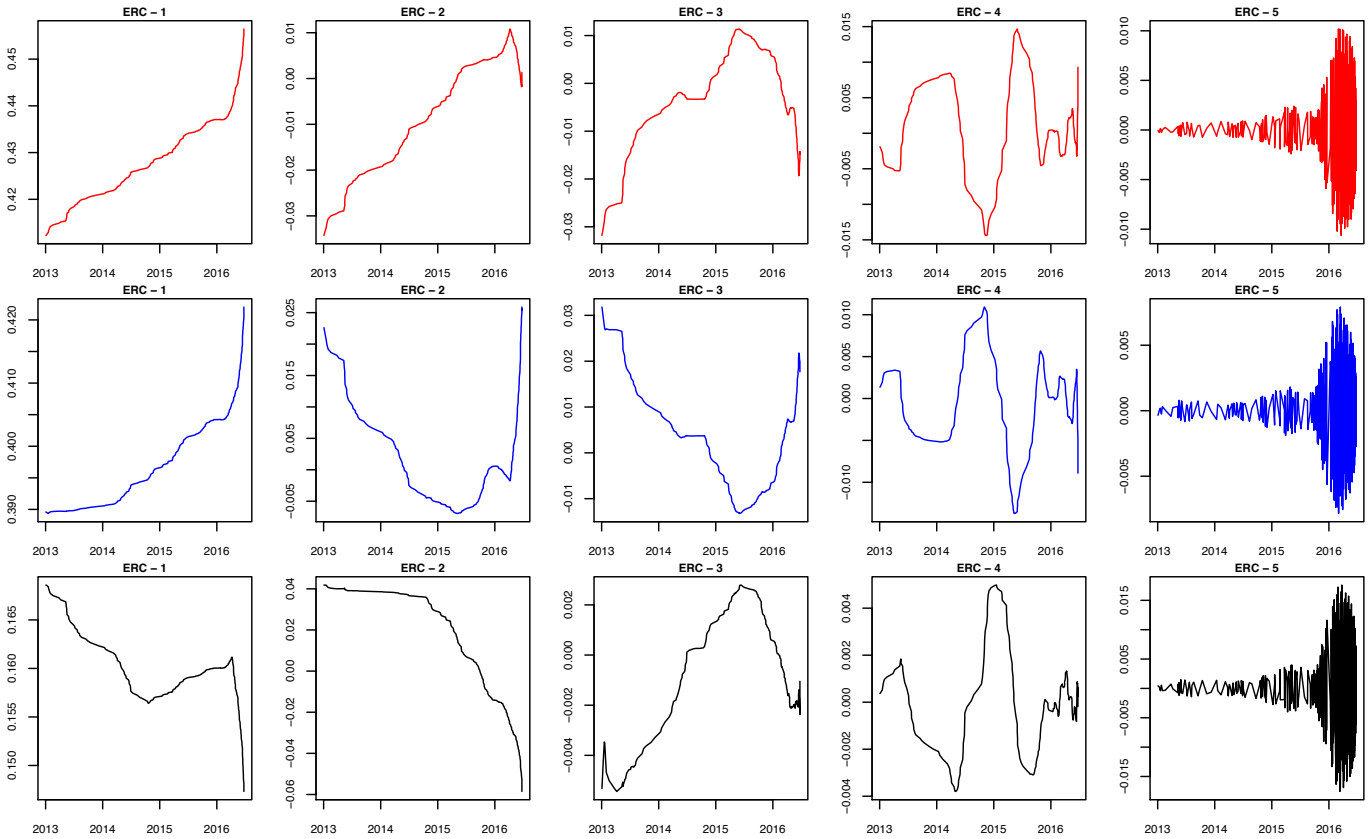


Figure 5: MSSA elementary reconstructed components (ERC) for leave (—), stay (—), and undecided (—).

4.4. Revisiting the dynamics of poll tracker data: Take II

In Figure 5, we plot the first 5 elementary reconstructed components resulting from applying the MSSA when $L = \lceil (N+1)/(D+1) \rceil = 69$, and using the cumulative periodogram-based criterion for selecting components discussed in Section 4.1, which yields $I_p = \{1, 2\}$, $I_s = \{1, 2, 3\}$, and $I_u = \{1, 2\}$. In Figure 4 we present the estimated trendline triplet $(\widetilde{\mathbf{p}}, \widetilde{\mathbf{s}}, \widetilde{\mathbf{u}})$ based on MSSA, the estimated dynamics in public opinion fits with the tendency of the polls in the same way as in the univariate case, but now the trends are constructed from a joint model, and obey the linear restriction $\mathbf{1} = \widetilde{\mathbf{p}} + \widetilde{\mathbf{s}} + \widetilde{\mathbf{u}}$, where $\mathbf{1}$ is a length N vector of ones.

5. Discussion

Political analysts and specialized media often rely on polls data—such as the Financial Times (FT) Brexit poll tracker discussed in Section 2—so to predict the outcome of an election or a referendum. Our MSSA-based method is tailored for the applied context of interest, and extends easily to settings where more than three options are available (Brexit, Bremain, undecided)—say for elections where a pool of candidates is available. We analyze data from the polls and assess whether the information contained in the battery of polls was indicative of whether the UK would choose to leave or to stay

in the EU. The conducted post-mortem puts forward the dynamics underlying the sentiment towards leaving the EU.

From a methodological outlook, a main goal of the article was on touching on some aspects of smoothing poll data—particularly on commenting on different ways of producing estimates on the unit simplex—and to highlight that multivariate singular spectrum analysis is a sturdy tool for disentangling poll data into a trendline and sampling error. The approaches discussed in this paper could be regarded as companions to Kalman smoothing (Green et al., 1999), which is another a popular approach for filtering and smoothing poll data. From a statistical viewpoint, the development of parametric statistical models supported on $\Delta_{D,\varepsilon}$, as defined in Equation (2), is a natural avenue for future research.

A main contribution of the article rests on a multivariate singular spectrum analysis method that yields trendlines on the unit simplex. The trendline derived resorting to multivariate singular spectrum analysis is shown to bear a resemblance with that of local polynomial smoothing, and it presents the nice feature of disentangling directly the dynamics into components that can be interpreted as changes in public opinion or sampling error. Although the current methods can be used for tracking and disentangling the dynamics of public opinion over time, there are natural opportunities for further developments that capitalize on other challenges posed by the data—and which are deliberately not tackled here. In particular, it seems natural to reweigh polls according to sample sizes, and to take on board the fact that polls conducted by different houses need not to be based on a similar survey methodology. In addition, as mentioned by a reviewer, loess smoothing can be used for non-equidistant measurements, while SSA is designed for equidistant measurements. The analysis conducted in the paper considers non-equidistant measurements as equidistant, applies SSA/MSSA, and returns to the initial time points. This trick is legal if SSA is used for smoothing and the density of time points is smoothly changed. Monte Carlo evidence reported in the supplementary materials suggests that the performance of the methods proposed in the article is tantamount regardless of whether measurements are equidistant or not, although performance tends to be better on regions of the domain where more data are available.

Acknowledgments

We thank the Editors and anonymous reviewers for helpful comments on an earlier version of the paper. The research was partially funded by the project INTERSTATA (**Inter**disciplinary **Stat**istics in **A**ction), by the International Research and Partnership Fund, Edinburgh, and by FCT (Fundação para a Ciência e a Tecnologia, Portugal), through the project UID/MAT/00006/2013.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52(2)*, 119–126.

Bergman, J., & Holmquist, B. (2014). Poll of polls: A compositional loess model. *Scandinavian Journal of Statistics*, *41(2)*, 301–310.

Brown, L. D., Cai, T. T., & DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, *30(1)*, 160–201.

Brockwell, P. J., & R. A. Davis. (1991). *Time Series: Data Analysis and Theory*. 2nd Ed. New York: Springer.

Cameron, D. (2013). EU speech at Bloomberg. URL: https://www.gov.uk/government/speeches/eu-speech-at-bloomberg.

Cressey, D. (2016). Academics across Europe join 'Brexit' debate. *Nature*, *530(7588)*, 15–15.

de Carvalho, M., & Rua, A. (2017). Real-time nowcasting the us output gap: Singular spectrum analysis at work. *International Journal of Forecasting*, *33(1)*, 185–198.

Dhingra, S., & Sampson, T. (2016). Life after Brexit: What are the UK's options outside the European union? CEP Brexit Analysis, CEPBREXIT01. London School of Economics and Political Scieence, CEP, London, UK.

Fan, J., & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Boca Raton, FL: Chapman & Hall/CRC.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*. New York: Wiley.

Golyandina, N. (2010). On the choice of parameters in singular spectrum analysis and related subspace-based methods. *Statistics and its Interface*, *3(3)*, 259–279.

Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. A. (2001). *Analysis of Time Series Structure: SSA and Related Techniques*. Boca Raton, FL. Chapman & Hall/CRC.

Golyandina, N., & Korobeynikov, A. (2014). Basic singular spectrum analysis and forecasting with R. *Computational Statistics & Data Analysis*, *71*, 934–954.

Golyandina N., Korobeynikov A., Shlemov A., & Usevich K. (2015). Multivariate and 2D extensions of singular spectrum analysis with the Rssa package. *Journal of Statistical Software*, *67(2)*, 1–78.

Green, D. P., Gerber, A. S., & de Boef, S. L. (1999). Tracking opinion over time: A method for reducing sampling error. *The Public Opinion Quarterly*, *63(2)*, 178–192.

Hassani, H., Heravi, S., & Zhigljavsky, A. (2009). Forecasting European industrial production with singular spectrum analysis. *International Journal of Forecasting*, *25(1)*, 103–118.

Hassani, H., Soofi, A. S., & Zhigljavsky, A. (2013). Predicting inflation dynamics with singular spectrum analysis. *Journal of the Royal Statistical Society, Ser.* A, *176(3)*, 743–760.

Hassani, H., & Mahmoudvand, R. (2013). Multivariate singular spectrum analysis: A general view and new vector forecasting approach. *International Journal of Energy and Statistics*, *1(1)*, 55–83.

Lin, X., & Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American statistical Association*, *95(450)*, 520–534.

Linzer, D. A. (2013). Dynamic Bayesian Forecasting of Presidential Elections in the States. *Journal of the American Statistical Association*, *108(501)*, 124–134.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, *9(1)*, 141–142.

Pasek, J. (2015). The polls—review predicting elections: Considering tools to pool the polls. *Public Opinion Quarterly*, *79(2)*, 594–619.

Pickup, M., & Johnston, R. (2008). Campaign trial heats as election forecasts: Measurement error and bias in 2004 presidential campaign polls. *International Journal of Forecasting*, *24(2)*, 272–284.

Qvortrup, M. (2016a). Last word: The EU referendum. *Political Insight*, *7(1)*, 40–40.

Qvortrup, M. (2016b). Referendums on membership and European integration 1972–2015. *The Political Quarterly*, *87(1)*, 61–68.

Rodrigues, P. C., & Mahmoudvand, R. (2018). The benefits of multivariate singular spectrum analysis over the univariate version. *Journal of the Franklin Institute*, *355(1)*, 544–564.

Rothschild, D. (2015). Combining forecasts for elections: Accurate, relevant, and timely. *International Journal of Forecasting*, *31(3)*, 952–964.

Thulin, M. (2014). The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics*, *8(1)*, 817–840.

Venables, W., & Ripley, B. D. (2002). *Modern Applied Statistics with S-PLUS*. 4th ed. New York: Springer.

Wand, M. P., & Jones, M. C. (1995). *Kernel Smoothing*. Boca Raton, FL: Chapmand & Hall/CRC.

Wang, W., & Carreira-Perpiñán, M. A., (2013). Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. arXiv preprint arXiv:1309.1541.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.

Yao, W., & Li, R. (2013). New local estimation procedure for a non-parametric regression function for longitudinal data. *Journal of the Royal Statistical Society, Ser.* B, *75(1)*, 123–138.

Supplementary materials to: "Brexit: Tracking and disentangling the sentiment towards leaving the EU"

Miguel de Carvalho[a*], Gabriel Martos[b]

[a] *University of Edinburgh, UK.* [b] *Universidad Torcuato di Tella, Argentina*

1. Monte Carlo evidence on target grouping for trendline extraction

1.1. Data generating processes and preliminary experiments

We now report numerical experiments illustrating the automatic criteria for learning about trendlines introduced in Section 4 of the paper. A Monte Carlo study is reported in Section 1.2; for now we concentrate on describing the data generating processes and illustrating the method on a single run experiment. We simulate data from the following stochastic processes:

1. $X_t = t^3 - 9t^2 + 23t + \varepsilon_t^X$, with $\varepsilon_t^X \sim N(0,1)$.
2. $Y_t = 10/t \sin(t/3) + \varepsilon_t^Y$, with $\varepsilon_t^Y \sim N(0,1)$.
3. (X_t, Y_t) with X_t and Y_t as in 1. and 2., and with ε_t^X independent of ε_t^Y.

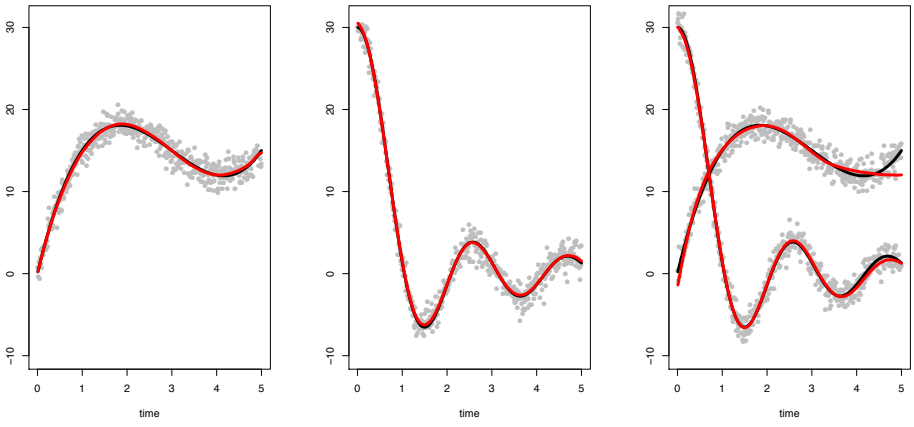


Figure 1: One-shot experiment for Scenario A. Trendlines (—) yield using SSA (left, middle) and MSSA (right), along with the true mean function (—) and data (•) generated from $\{X_{t_i}\}_{i=1}^N$ (left), $\{Y_{t_i}\}_{i=1}^N$ (middle), and $\{(X_{t_i}, Y_{t_i})\}_{i=1}^N$ (right). The number of components was selected via the automatic cumulative periodogram-based criterion from Section 4 in the paper.

*Corresponding author: School of Mathematics, James Clerk Maxwell Building, The King's Buildings, Peter Guthrie Tait Road, Edinburgh, EH9 3FD. Email: *Miguel.deCarvalho@ed.ac.uk*

We consider two scenarios. In Scenario A, we simulate data over a grid of equispaced points $t_i = 5(i/500)$, for $i = 1, \ldots, N = 500$; in Scenario B, we simulate non-equidistant measurements at times $\tau_i = 5(i/500)^2$, for $i = 1, \ldots, N = 500$. This results in the collections of data $\{X_{t_i}\}_{i=1}^N$, $\{Y_{t_i}\}_{i=1}^N$, and $\{(X_{t_i}, Y_{t_i})\}_{i=1}^N$ for Scenario A, and $\{X_{\tau_i}\}_{i=1}^N$, $\{Y_{\tau_i}\}_{i=1}^N$, and $\{(X_{\tau_i}, Y_{\tau_i})\}_{i=1}^N$ for Scenario B. Given that the intensity of polls is higher at the end of the observation period, Scenario B should be regarded as a better approximation to the setup underlying our Brexit case study.

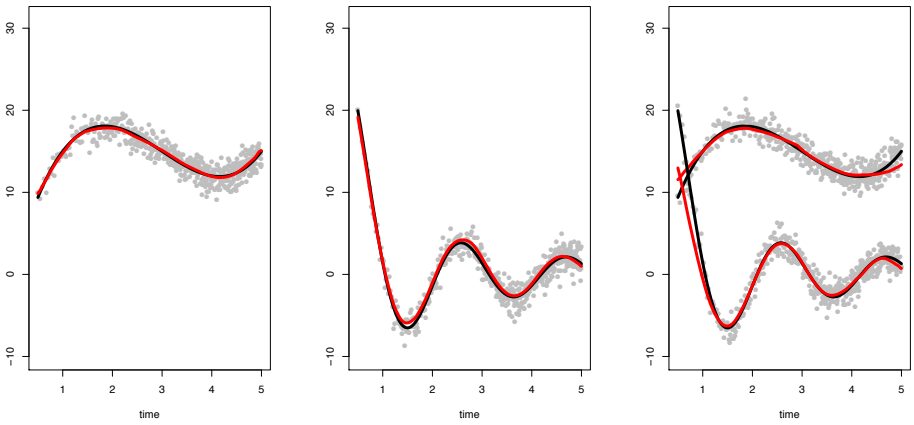


Figure 2: One-shot experiment for Scenario B. Trendlines (—) yield using SSA (left, middle) and MSSA (right), along with true mean function (—) and data (•) generated from $\{X_{\tau_i}\}_{i=1}^N$ (left), $\{Y_{\tau_i}\}_{i=1}^N$ (middle), and $\{(X_{\tau_i}, Y_{\tau_i})\}_{i=1}^N$ (right). The number of components was selected via the automatic cumulative periodogram-based criterion from Section 4 in the paper.

In Figures 1 and 2, we depict the estimated trendlines yield via SSA and MSSA, for Scenarios A and B, respectively, along with the true mean functions of the processes of interest, namely $\mu_t^X = t^3 - 9t^2 + 23t$, $\mu_t^Y = 10/t \sin(t/3)$, and (μ_t^X, μ_t^Y). To estimate trendlines in both scenarios, we set $L = \lceil (N+1)/2 \rceil$ for SSA and $L = \lceil (N+1)/3 \rceil$ for MSSA, where $\lceil \cdot \rceil$ is the ceiling function; in addition, we considered a 1% significance level to learn about the number of elementary components required to track the trendline, leading to three components for SSA and four components for MSSA (same result for Scenarios A and B). This one-shot experiment allows us to anticipate strengths and limitations with the methods. As can be seen from Figures 1 and 2, the criterion proposed in Section 4 seems to satisfactorily learn about the number of components required for extracting the trendlines, but there is some bias at the beginning and at the end of the observation period. Any conclusion on performance should however be regarded as tentative for now, as Figures 1 and 2 result from a single-run experiment. A full account on performance is given in the Monte Carlo experiment reported in Section 1.2.

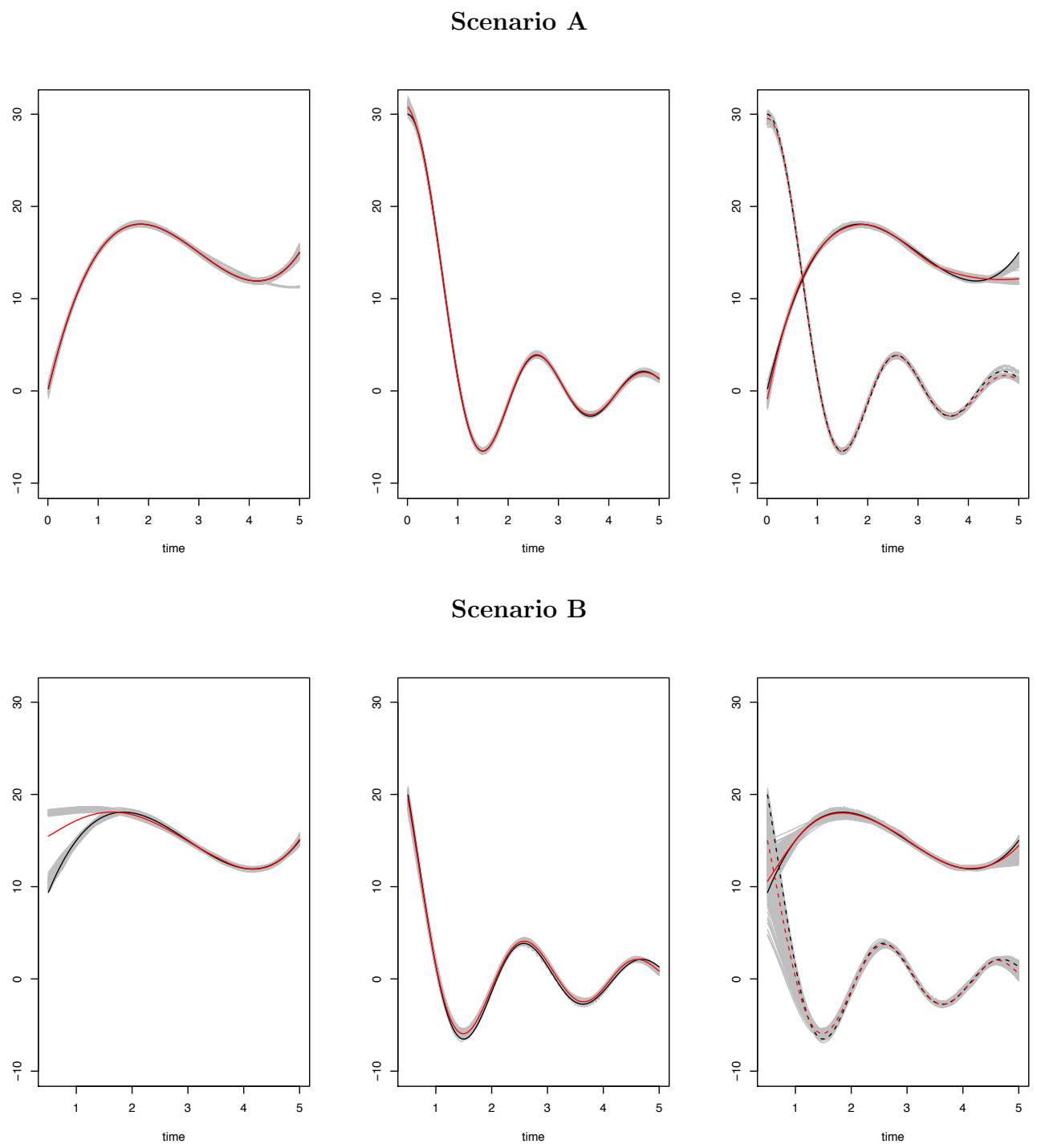


Figure 3: Monte Carlo evidence for Scenarios A and B. Trendlines (—) yield using SSA (left, middle) and MSSA (right) for each of the 500 simulated data sets, along with the true mean function (—), and Monte Carlo mean (—). The top and bottom pannels respectively correspond to Scenarios A and B. The number of components was selected via the automatic cumulative periodogram-based criterion from Section 4 in the paper.

1.2. Monte Carlo evidence

Here we report results from a Monte Carlo study where we repeat 500 times the one-shot experiment from Section 1.1. The results are presented in Figure 3, and in Table 1 we report the frequency of

elementary components recommended by the cumulative periodogram-based method over the Monte Carlo study. As can be seen from Figure 3, in both scenarios the estimated trajectories approximate reasonably well the true mean.

Interestingly, note that the bias appearing in Scenario A at the end of the observation period is no longer present in Scenario B; this is due to the fact that by construction τ_i puts more data at the end of observation period—a situation which mimics what we face in the poll tracker dataset. Note also that the 'reverse' also applies. That is, for Scenario B trendlines can at times be biased at the beginning of the observation period. This bias is explained by the fact that by construction τ_i puts less data at the beginning of the observation period—a situation which again mimics what we face in the poll tracker dataset.

Table 1: Frequency over the Monte Carlo study of number of elementary reconstructed components selected (ERC) via the automatic cumulative periodogram-based criterion from Section 4 in the paper, for Scenarios A and B.

			ERC				
Scenario			2	3	4	5	≥ 6
A	SSA on	X_t	4	494	2	0	0
	SSA on	Y_t	0	498	1	1	0
	MSSA on	(X_t, Y_t)	$\{0, 0\}$	$\{0, 0\}$	$\{442, 492\}$	$\{57, 5\}$	$\{1, 3\}$
B	SSA on	X_t	335	163	2	0	0
	SSA on	Y_t	0	496	2	2	0
	MSSA on	(X_t, Y_t)	$\{0, 0\}$	$\{0, 0\}$	$\{127, 489\}$	$\{372, 9\}$	$\{1, 2\}$

2. Sensitivity analysis

Following the practice recommended by Marron (2001, p. 533), we conducted inference over a range of bandwidths so to evaluate the sensitivity and reliability of the inference to the smoothing parameters. A snapshot of some of the experiments is presented in Figure 4 where we plot local linear regression trendline estimates obtained for the values of smoothing parameter in $\{0.01, 0.05, 0.20\}$. As expected, when h is small, the trendlines strive to interpolate between the points; and as h grows, trendlines approximate the mean of all polls. The smoothing parameter in the article was fixed according to the aggregated least square criterion discussed in Section 3.3.

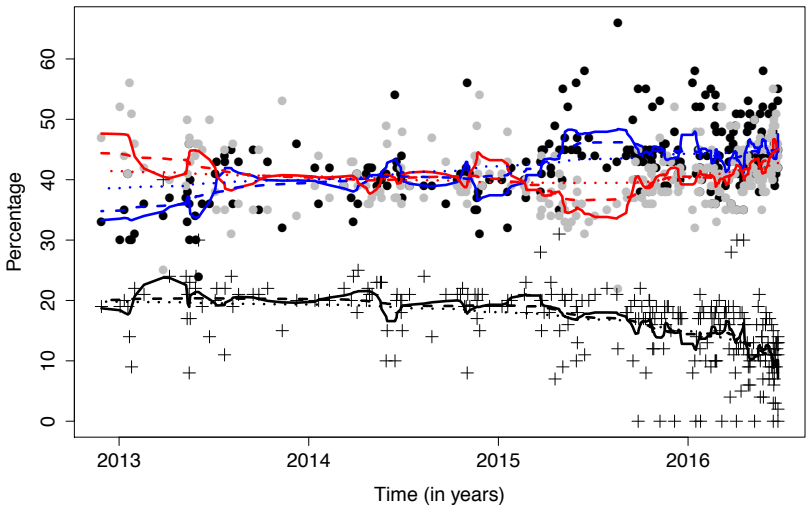


Figure 4: Bandwidth sensitivity analysis for leave (—), stay (—), and undecided (—): Solid line corresponds to $h = 0.01$, dashed line to $h = 0.05$, and dotted line to $h = 0.2$.

Following a suggestion by a reviewer, we also conducted similar analyses to the ones performed in the paper, but using monthly grouped data. In Figure 5 we report the Agresti–Coull confidence intervals for the proportion in favor Bremain. The obtained results show similar patterns in the evolution of the sentiment to Brexit as the one obtained with trendline estimations on Sections 3 and 4. As reported in Section 2, the grouped polls of June 2016 shows a small advantage in favor of Bremain.

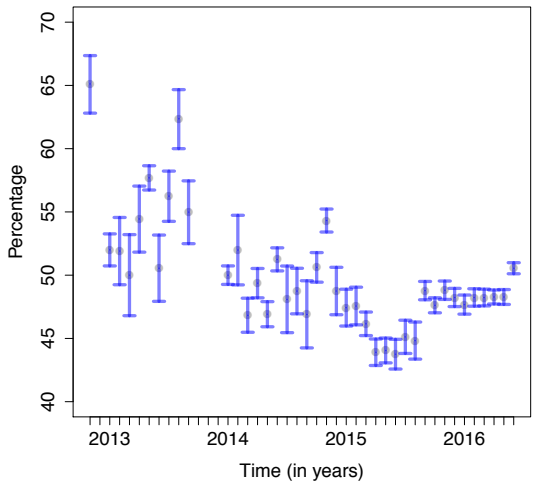


Figure 5: Agresti–Coull confidence intervals for the proportion in favor Bremain (data grouped monthly), focusing on decided voters only.

Rerunning the analysis of Section 4.3 with grouped data

In Figure 6 we plot the elementary reconstructed components resulting from applying multivariate

singular spectrum analysis. Figure 7 represents the trendlines for grouped data which were obtained by considering the elementary reconstruction components based on $I_p = \{1, 2\}$, $I_s = \{1, 2, 3\}$ and $I_u = \{1, 2\}$ for leave, stay and undecided, respectively; the choice of the latter sets was based on the recommendation from the cumulative periodogram-based criterion, graphical inspection of the components in Figure 6, and examination of the scree plot. In Figure 8 we present the estimated trendline triplet $(\tilde{\mathbf{p}}, \tilde{\mathbf{s}}, \tilde{\mathbf{u}})$ based on MSSA. The results are tantamount to those obtained in Section 4.2; compare with Figure 4 in the paper. Graphically, the differences between the trendline estimates obtained when we impose the normalization constraint (solid lines) and when the constraint is not imposed (dashed lines), are seemingly insignificant.

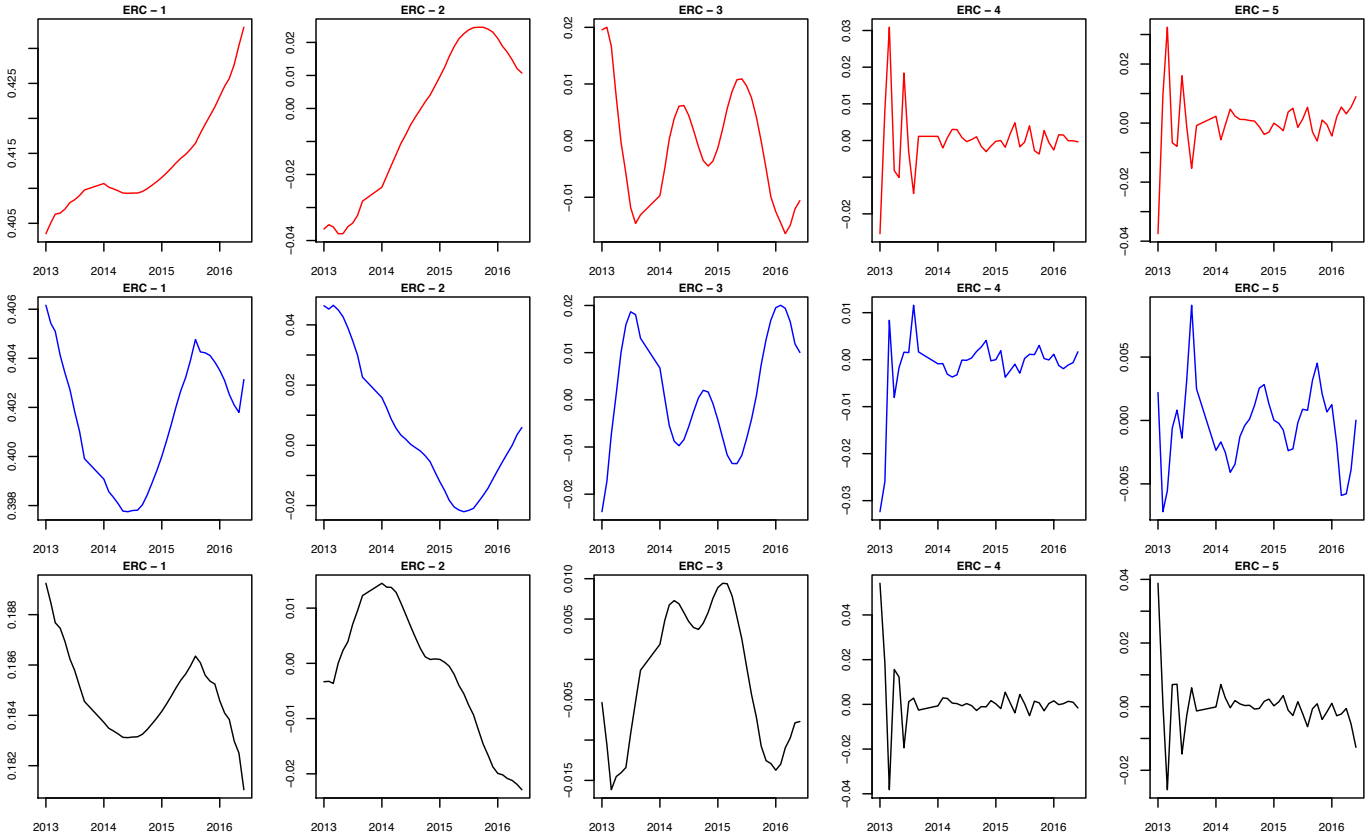


Figure 6: MSSA elementary reconstructed components obtained from monthly grouped data for leave (—), stay (—), and undecided (—).

3. Scree plot and cumulative periodograms

In Figure 9 we report the scree plots along the cumulative periodograms for the corresponding residuals of the SSA and MSSA models estimated in the article; see Section 4. The proposed criterion to determine the number of elementary reconstructed components suggests that in the case of SSA, trendlines should be constructed from elementary reconstruction components based on $I_p = \{1, 2\}$,

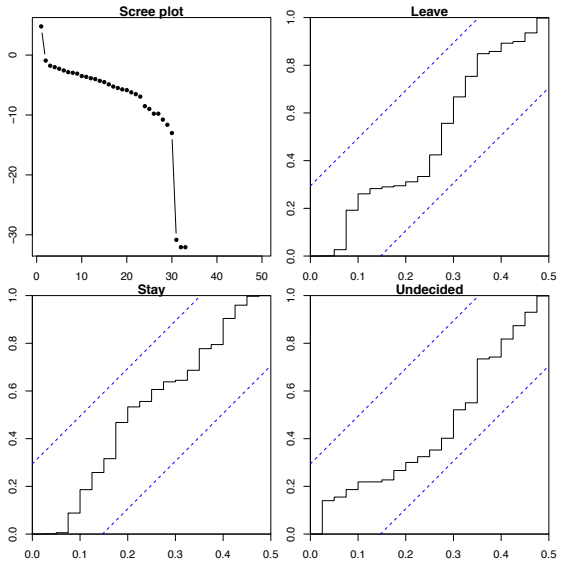


Figure 7: Scree plot for MSSA eigenvalues and cumulative periodograms for residuals of MSSA trendlines for leave, stay, and undecided (data grouped monthly).

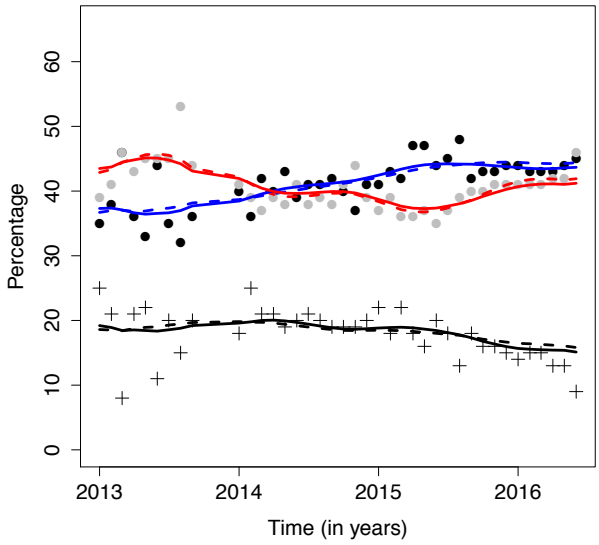


Figure 8: Trendlines based on monthly grouped data for the proportion in favor of leave (•), stay (•), and undecided (+); the dashed and solid lines represent smoothers obtained from MSSA—without imposing the normalization constraint—and MSSA—when imposing the normalization constraint—, respectively, for stay (—), leave (—), and undecided (—).

$I_s = \{1, 2\}$, and $I_u = \{1\}$ (for leave, stay, and undecided, respectively). In the case of MSSA, the elementary reconstruction components suggested by the cumulative periodogram-based criterion are those stemming from $I_p = \{1, 2\}$, $I_s = \{1, 2, 3\}$ and $I_u = \{1, 2\}$ for leave, stay and undecided, respectively.

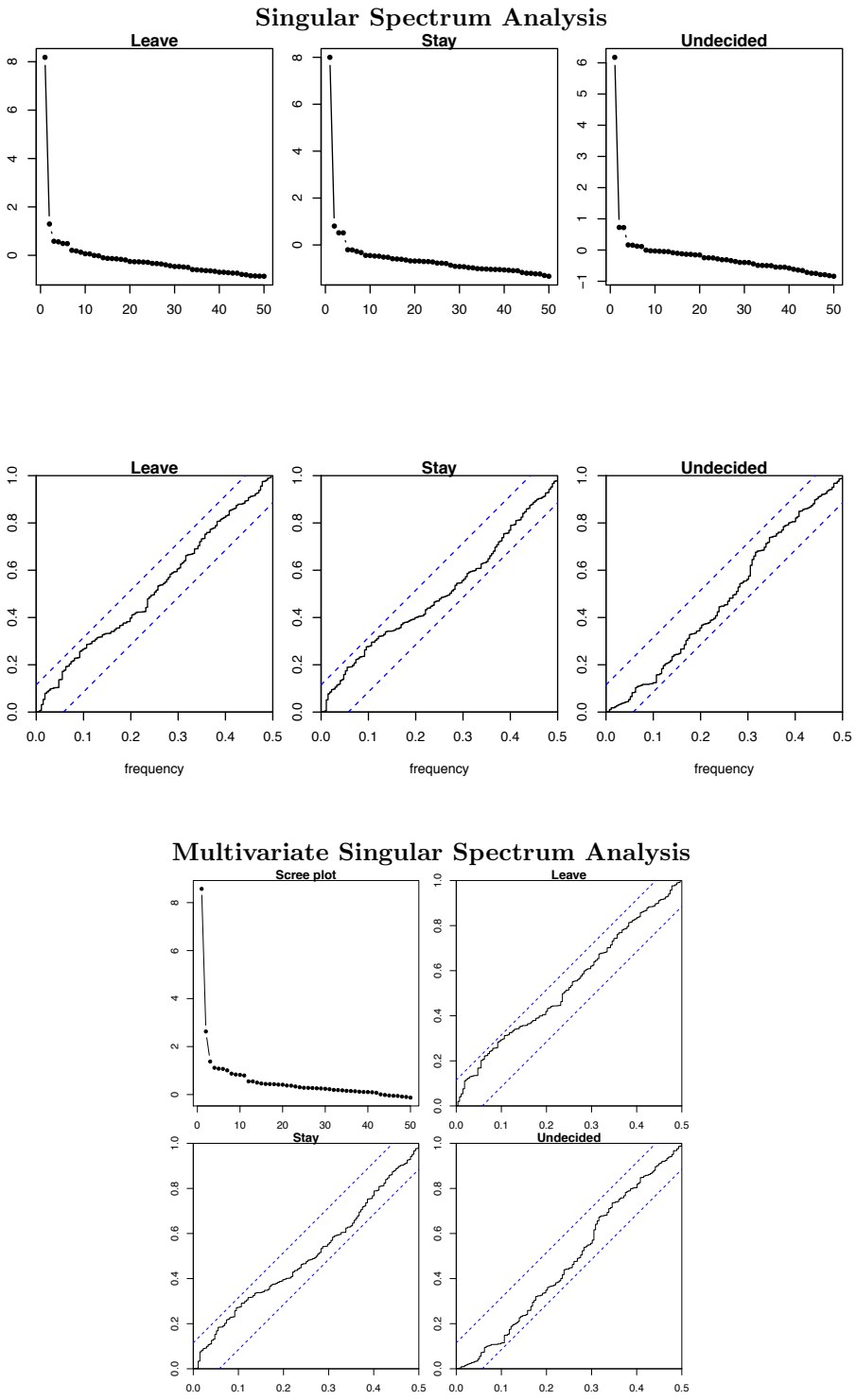


Figure 9: Above: SSA scree plot for eigenvalues and cumulative periodograms for residuals of SSA trendlines for leave, stay, and undecided. Below: Scree plot for MSSA eigenvalues and cumulative periodograms for residuals of MSSA trendlines for leave, stay, and undecided.

References

Marron J. S. (2001). Discussion of "Inference for density families using functional principal component analysis," *Journal of the American Statistical Association*, *96(454)*, 532–533.