Signal extraction and prediction of pure trend time series using a state-space models with an application to the creation of an emergence index

Sonia Mazzi *Data Science Campus - Office for National Statistics - UK*

This document provides a summary of the theory and technical details underpinning signal extraction and forecasting using a Local Linear Model (LLM-SSM) based on its formulation as a state-space model. It also derives an emergence index and explains how the LLM-SSM can be used to compute it. Examples are provided using US patent data.

Keywords: Local Linear Model, State-Space Model, Kalman Filter, Diffuse Kalman Filter, smoothing filter

1 Introduction

Suppose that we have $y = \{y_i\}_{i=1}^n$ and $y_i = \mu_i + \epsilon_i$, where μ_i is a stochastic process (with linear dynamic) and ϵ_i's are independent and identically distributed (iid) with mean zero and constant variance. Note that y_i can be a vector. This is a signal-plus-noise model. $\mu = \{\mu_i\}$ is the signal, which is unobserved and the feature that we wish to extract from $y = \{y_i\}$. The process $\epsilon = \{\epsilon_i\}$ is the noise, which has no significant stochastic structure, such as autocorrelation, and has expected value zero. We assume that $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed (i.i.d.) random variables with mean 0 and variance σ^2.

The stochastic specification of μ_i allows to model y in different ways. We will use a particular state-space model to determine the stochastic structure of μ, and therefore of y.

For more general details on state-space models and signal-extraction algorithms see de Jong (1991).

2 The local linear trend model

We say that y is pure-trend data if, besides from white noise, the main source of variation of y is trend. If we do not consider covariates for y we can say that the trend is the only source of variation of y. Note that we do not wish to explain the data generating process and we do not wish to explain what makes y vary. We wish to extract a smooth signal indicating the local trend and predict its behaviour in terms of trend.

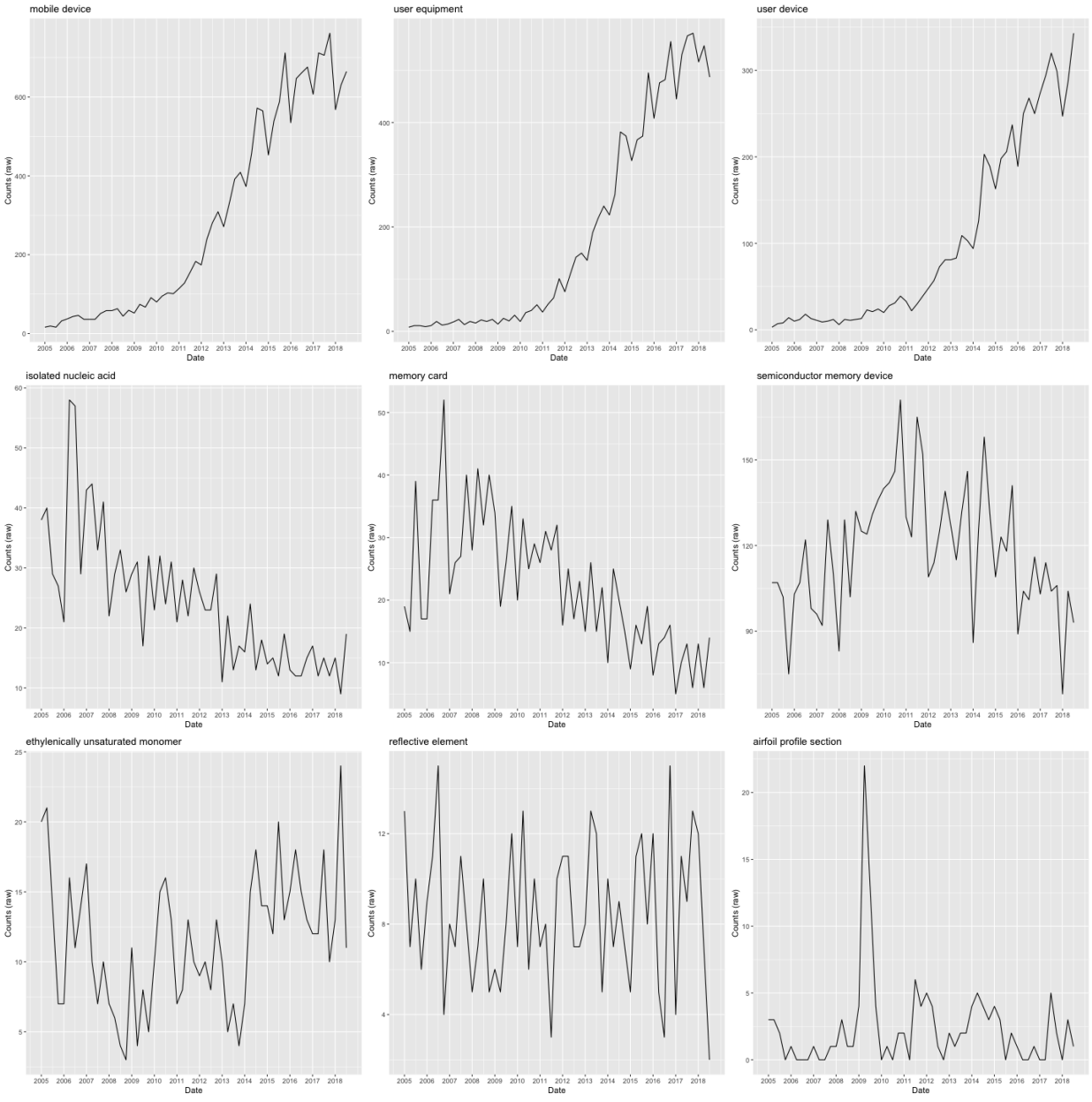
EXAMPLE. *The emerging technologies time series data*.

Nowadays, attention to emerging technologies is increasing. Indicators of technological emergence promise valuable intelligence to those determining R&D priorities, in charge of portfolio management, carrying out technology opportunity analysis and managing innovation. Indicators can address varied subjects: breakthrough science, novel technology, commercial innovation, etc.

One such indicator can be based on extracting “relevant" terms from patent applications, and assessing the counts of the occurrence of such terms at different time points. The relevance of a term is quantified by using natural language processing criteria, namely TF-IDF, which is a way

to weigh the frequency with which a term appears in a set of documents. For a given period of time, this process creates time series of counts for each relevant term.

The time plots below are for quarterly counts of nine terms between 2005 and the third quarter of 2018 extracted from US patent applications. The plots below aim to illustrate that the shape of the general trend of each series can be quite varied. In what follows we will use the data to further illustrate the quantification of emergence.



For the given period, has the usage of the term been increasing (emergent term), decreasing (declining term) or been kept more or less constant (static term)? If two or more terms are emergent terms for the considered time period, how can we rank their emergence? What will happen with the usage of these terms in the future? We shall try to answer these questions using a model-based approach.

In the LLM-SSM it is assumed that $y_i = \mu_i + \epsilon_i$, $\mu_{i+1} = \mu_i + d_i + \nu_i$, and $d_{i+1} = \delta d_i + \eta_i$, where $\{\epsilon_i\}$, $\{\nu_i\}$, $\{\eta_i\}$ are mutually independent white noise processes with variances σ^2, σ_ν^2 and σ_η^2.

The rationale for the model stems from initially assuming that the signal or trend, μ_i, follows a fully deterministic specification: $\mu_i = a + d\,i$, with a and d constant. Whilst this specification would be clearly unsatisfactory for most time series, if we assume that $\mu = \{\mu_i\}$ is "smooth", μ could be described locally, rather than globally, by a linear model (Taylor's theorem). The idea is to let a and d vary with time to allow μ_i to adapt to the evolution of the series.

If $\mu_i = a + d\,i$ then $\mu_{i+1} = \mu_i + d$.

To allow for flexibility in the model we could establish that

$$\begin{aligned} \mu_{i+1} &= \mu_i + d_i + \nu_i, \\ d_{i+1} &= \delta d_i + \eta_i. \end{aligned}$$

where $\nu = \{\nu_i\}$ and $\eta = \{\eta_i\}$ are independent white noise processes with mean zero and variance σ_ν^2 and σ_η^2 respectively. ϵ is independent of ν and η.

The full model is

$$\begin{aligned} y_i &= \mu_i + \epsilon_i, \\ \mu_{i+1} &= \mu_i + d_i + \nu_i, \\ d_{i+1} &= \delta d_i + \eta_i. \end{aligned} \tag{1}$$

Note that

$$\begin{aligned} y_{i+1} - y_i &= \mu_{i+1} - \mu_i + \epsilon_{i+1} - \epsilon_i, \\ &= d_i + \nu_i + \epsilon_{i+1} - \epsilon_i. \end{aligned}$$

So the model is ARIMA(0,1,1) plus a stochastic term, d_i, which has an AR(1) specification. In ARIMA(0,1,1) d_i would be a constant.

3 State-space form of the LLM-SSM

To cast the model into state-space form we write

$$\begin{aligned} y_i &= Z\alpha_i + Gu_i, \\ \alpha_{i+1} &= T_i\alpha_i + Hu_i. \end{aligned}$$

where $u_0, u_1, \ldots, u_n$ are i.i.d with $Var(u_i) = \sigma^2 I_3$, where I_n denotes the $n \times n$ identity matrix, and

$$\alpha_i = \begin{pmatrix} \mu_i \\ d_i \end{pmatrix},\ Z = (1\ 0),\ T_i = \begin{pmatrix} 1 & 1 \\ 0 & \delta \end{pmatrix},\ u_i = \begin{pmatrix} \epsilon_i \\ \eta_i \\ \nu_i \end{pmatrix},\ G = (1\ 0\ 0),\ H = \begin{pmatrix} 0 & \sigma_\nu/\sigma & 0 \\ 0 & 0 & \sigma_\eta/\sigma \end{pmatrix},$$

with $\alpha_1 = W_0\beta$, $\beta = b + B\gamma$, $\gamma \sim (c, \sigma^2 C)$ and in this case $W_0 = B = I_2$, $b = 0_2$, where 0_n denotes a vector of zeroes of length n.

If we think of μ_i as $\mu_i = a_i + d_i t_i$, α_1 represents $(a_1 + d_1 t_1 \quad d_1)^t$. v^t means the transpose of v.

The Diffuse Kalman Filter (DKF) is a generalisation of the Kalman Filter which allows to deal with initial conditions in the same paradigm that estimation, signal extraction and prediction take place.

4 The Diffuse Kalman Filter

The DKF in this case is the recursion

$$\begin{aligned} E_i &= (0,\, y_i) - ZA_i, \\ D_i &= ZP_iZ^t + GG^t, \\ K_i &= T_iP_iZ^tD_i^{-1}, \\ A_{i+1} &= T_iA_i + K_iE_i, \\ P_{i+1} &= (T_i - K_iZ)P_iT_i^t + HH^t, \\ Q_{i+1} &= Q_i + E_i^tD_i^{-1}E_i, \end{aligned} \qquad i = 1, \ldots, n,$$

with starting conditions $A_1 = (-I_2, 0_2)$, $P_1 = 0$, $Q_1 = 0$.

After the DKF is run, obtain Q_{n+1} and partition it

$$Q_{n+1} = \begin{pmatrix} S & s \\ s^t & q \end{pmatrix},$$

so that if Q_{n+1} has c columns and r rows, S is a $(q-1) \times (c-1)$ matrix.

The MLE of α_1 is $S^{-1}s$, with covariance matrix $\sigma^2 S^{-1}$, the MLE of σ^2 is $(q - s^tS^{-1}s)/n$ and the log-likelihood (maximised with respect to σ^2 and α_1) is $-\frac{1}{2}\left[n\log(\hat{\sigma}^2) + \sum_{i=1}^{n}\log(|D_i|)\right]$.

About hyperparameter estimation

The parameters $\delta, \sigma_\epsilon, \sigma_\eta, \sigma_\nu$ need to be estimated. We can do this via maximum likelihood. Note that the DKF provides an estimate of σ_ϵ. In the DKF the unknown parameters, to be estimated, appear as $\sigma_\nu^* = \sigma_\nu/\sigma_\epsilon$, $\sigma_\eta^* = \sigma_\eta/\sigma_\epsilon$ and δ. σ_ν^* and σ_η^* are "signal-to-noise ratio" parameters. Since we are only interested in smooth signals, we restrict the parameter space to search for σ_ν^*, and also for σ_η^*, to be the interval $[0, 0.5]$. For the same reason, the search for a smooth signal, the parameter δ is restricted to the interval $[0.85, 1]$.

5 Signal extraction

In the context of a pure trend model we wish to extract the smooth signal. That is we would like to predict $E(\mu_t|y)$, $t = 1, \ldots, n$. Since $\mu_t = Z\alpha_t$, $E(\mu_t|y) = ZE(\alpha_t|y)$, it suffices to predict $E(\alpha_t|y)$.

The DKF gives us the predictions of $E(\alpha_{t+1}|y_1, \ldots, y_t)$. We use the smoothing filter, a backwards recursion, to obtain the smoothed valued of the series.

The Smoothing Filter (SF) is the recursion

$$\begin{aligned} N_{i-1} &= Z^tD_i^{-1}E_i + L_i^tN_i, \\ R_{i-1} &= Z^tD_i^{-1}Z + L_i^tR_iL_i, \end{aligned}$$

with $N_n = 0$, $R_n = 0$, $L_i = T_i - K_iZ$, $i = 1, \ldots, n$, and all other quantities as in the DKF.

Then,

$$\begin{aligned} \tilde{\alpha}_i &= E(\alpha_i|y) = (A_i + P_iN_{i-1})(-S^{-1}s;\, 1) \\ mse(\tilde{\alpha}_i) &= \sigma^2\left(P_i - P_iR_{i-1}P_i + N_{i-1,\gamma}S^{-1}N_{i-1,\gamma}^t\right), \end{aligned}$$

where $N_{i-1,\gamma}$ denotes all but the last column of $A_i + P_iN_{i-1}$.

The state vector is $\alpha_i = (\mu_i \;\; d_i)^t$. The second entry of α_i can be interpreted as we usually interpret the first derivative. Negative values indicate a decline in the levels of μ, positive values

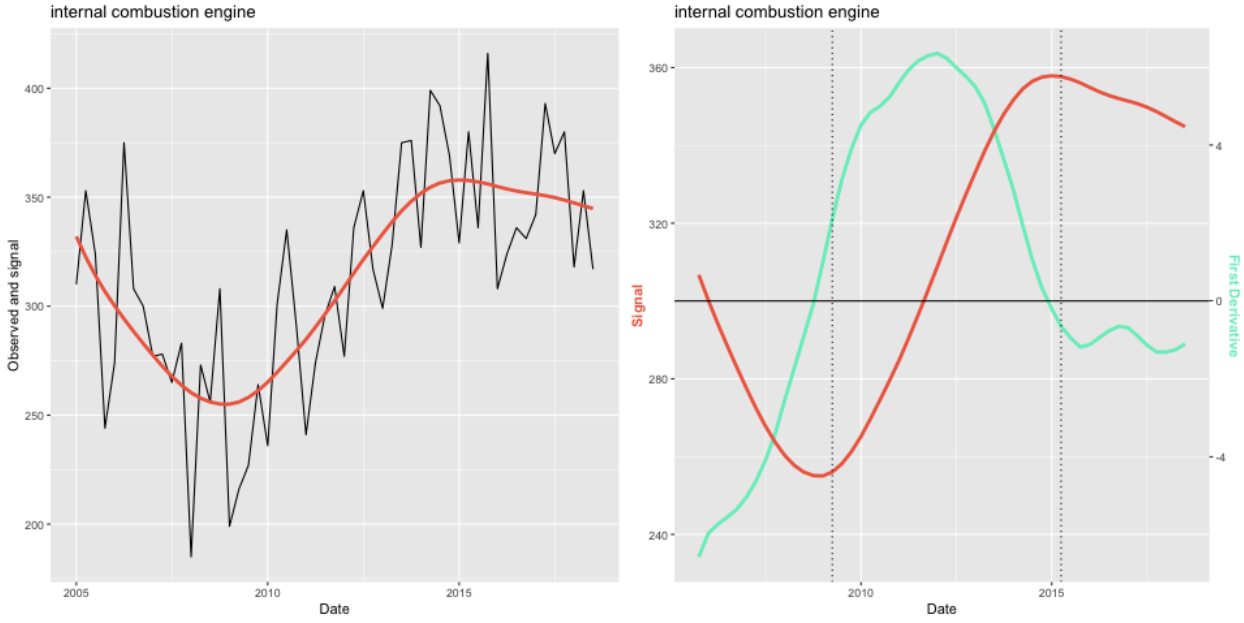
of d_i indicate incremental values of μ_i, values of d_i near zero may indicate a change point in μ_i or simply very small change or constancy of the signal.

From the point of view of forecasting, it might be beneficial not just to forecast values of the series but also predict values of the first derivative to predict the general future trend: upwards, downwards, constant and/or approaching a change point.

EXAMPLE. *Emerging technologies data (cont.)*

In the case of emerging technologies data, we are not only interested in the local level of the series but also on how the local level changes. Therefore, the LLM-SSM, which provides a way to focus on these two features of a time series, seems appropriate for the study of the quantification of emergence of technological terms.

On the left panel below is the time plot of quarterly counts of the term "internal combustion engine" between 2005 and the second quarter of 2018 together with the extracted signal. On the right panel we plot the extracted signal (red) and first derivative (blue). We mark the time points at which the first derivative crosses zero as this marks the time from when the trend is increasing (first derivative positive) to decreasing (first derivative negative) or viceversa.



6 Constructing an emergence index

Now that we can extract the level or trend and the first derivative of a pure trend series, how do we use that information to quantify and compare the emergence of terms?

$y = \{y_i\}_{i=1}^n$ represents one time series of quarterly counts of the term during a given time period. We assume that y follows the LLM-SSM (1). $\mu = \{\mu_i\}$ represents the signal, level or trend and σ_ϵ is the noise standard deviation. Also, $d = \{d_i\}_{i=1}^n$ represents the rates of change, or first derivatives, of the signal at each time point.

Positive values of d_i indicate that the signal is increasing, negative values of d_i indicate that the signal is decreasing. If $d_i = 0$ then that indicates a point of maximum, minimum or perhaps an inflection point of the signal.

6.1 The index

During a time period, say consecutive times $j_1, \ldots, j_m$, we define

$$E_1 = \sum_{k=1}^{m} d_{j_k}$$

as a measure of emergence during that given time period. Positive values of E indicate that growth was dominant in the time period. Negative values of E_1 indicate that decline was dominant in the time period. Values of E_1 near zero, positive or negative, could arise in many ways. One way is if, for example, the d_{j_k}'s are mostly very small and so the series is not changing level greatly. Or, it could be that there are approximately an equal number of positive and negative d's, and so there was growth and decline in the same period.

Now, how do we compare total emergence of different terms? The derivative measures the rate of change of a function. Given a smooth function f and two points x and y, the derivative of f at x, is the instantaneous rate of change of f at x, is

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

So, we see that if f takes on larger values than g, f' will be larger than g'.

For this reason, we define another measure of emergence, which we propose to adopt as an emergence index, which considers the rate of growth relative to the value of the signal, or the ratio of the first derivative and the signal. This is equivalent to what in financial econometrics is called the rate of returns (in continuous time) or simply the returns. Define E_2 to be

$$E_2 = \sum_{k=1}^{m} \frac{d_{j_k}}{\mu_{j_k}}.$$

where μ_i denotes the signal at time i and d_i denotes the derivative of the signal at time i.

Note that E_1 and E_2 have the same sign.

Normalising the emergence index

It should be noted that E_1 and E_2 may increase or decrease simply because the considered time span is enlarged. For this reason, and to allow a comparison of emergence for the same series during different time periods, we will consider the measures $\bar{E}_1$ and $\bar{E}_2$

$$\bar{E}_1 = \frac{E_1}{m}, \quad \text{and} \quad \bar{E}_2 = \frac{E_2}{m},$$

where m is the number of time periods used to compute E_1 and/or E_2. Using $\bar{E}_1$ or $\bar{E}_2$ doesn't change emergence rankings produced for different series during the same time period. Another advantage of using $\bar{E}_2$ instead of E_2 is that the absolute value of $\bar{E}_2$ is generally less than 1. This simplifies comparison across series.

Computing E_2

Another important issue with the computation of E_2 is the fact that if some of the counts in y are zero or positive but small, this may inflate the value of E_2. For this reason, we propose that, for computational purposes, E_2 is defined as

$$E_2 = \sum_k \xi_{j_k}$$

where

$$\xi_{j_k} = d_{j_k}/\mu_{j_k}, \quad \text{if} \quad \mu_{j_k} > 3, \qquad \text{and} \qquad \xi_{j_k} = 0, \quad \text{if} \quad \mu_{j_k} \leq 3.$$

This doesn't affect the quantification of emergence as we assume that a series with very low counts, fluctuating between 0 and 3, can be deemed to have negligible emergence. This last assumption must therefore be verified for the particular application being considered.

EXAMPLE. *Comparing E_1 and E_2.*

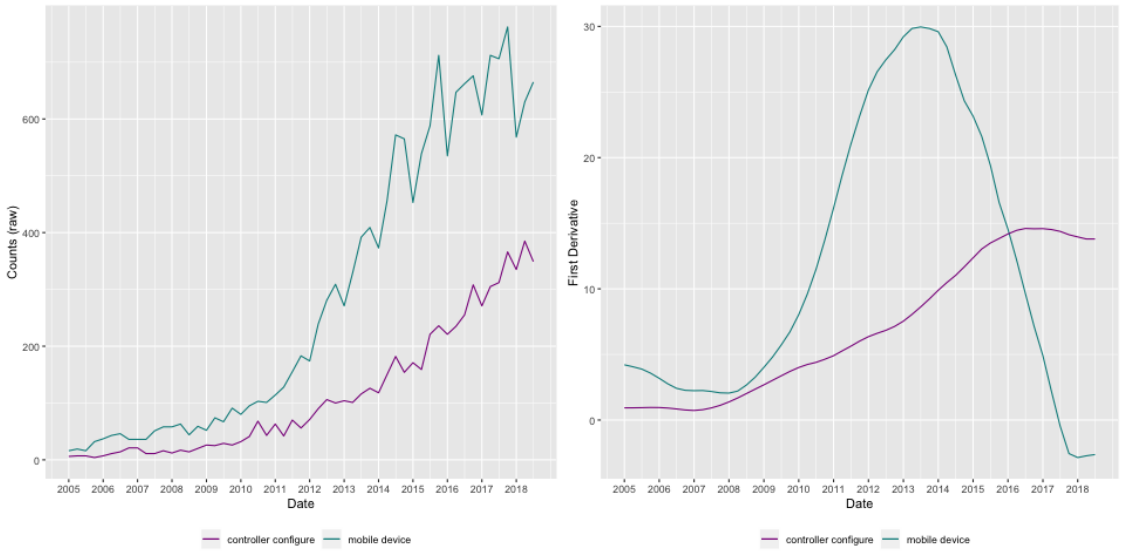
Consider the terms "mobile device" and "controller configure". For the time period 1st quarter 2005 - 3rd quarter 2018 the raw counts of "mobile device" are higher than those of "controller configure". See the top left panel of the plots below. The bottom left panel shows the extracted trends or smooth signals for both series of counts. Both series appear to be emergent and, to the naked eye, "mobile device" seems to be "more emergent" than "controller device". The top right panel shows the extracted first derivatives for both series and again we see the dominance of "mobile device". When we consider net growth for both series at each time point (the first derivative divided by the signal) we see now that the dominance of "mobile device" is not so clear. In the first considered year, 2005, both series show an elevated net growth, which is common for early stages of emergence.

When we add net growth at each time point, whether we include the first year or not, it becomes clear that although "mobile device" has larger E_1 , $\sum d_i$, "controller configure" has a longer, more steady and persistent growth over the time period than "mobile device". So, it seems right that "controller configure" should have a larger emergence index than "mobile device", as reflected by E_2, $\sum d_i/\mu_i$.

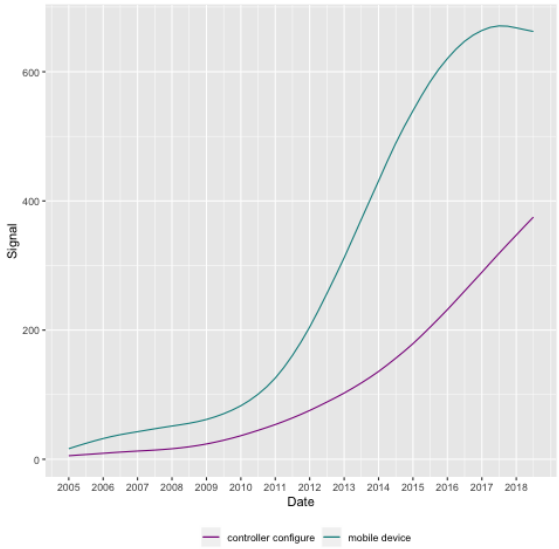
Our assertions above are summarised and quantified in the following table. Note how E_2 allows to focus on the whole time period or in a sub-period.

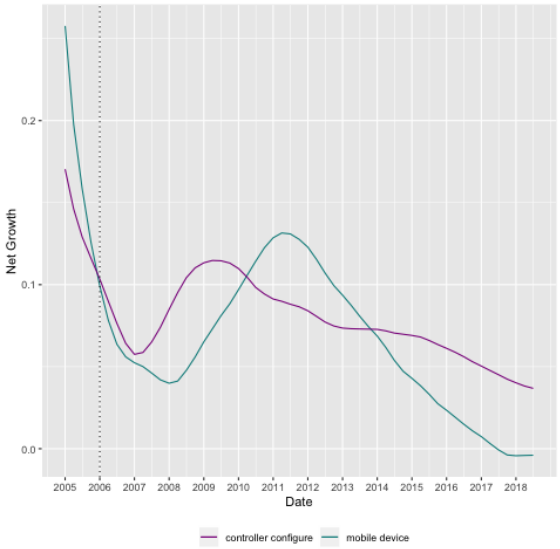
Term	σ_ϵ	$E_1 = \sum_i d_i$	$E_2 = \sum_i d_i/\mu_i$	$E_2 = \sum_{i\geq 5} d_i/\mu_i$
mobile device	38.838	778.173	3.894	3.155
controller configure	13.567	511.429	4.455	3.893

Term	$\bar{E}_1 = \frac{1}{55}\sum_i d_i$	$\bar{E}_2 = \frac{1}{55}\sum_i d_i/\mu_i$	$\bar{E}_2 = \frac{1}{51}\sum_{i\geq 5} d_i/\mu_i$
mobile device	14.149	0.071	0.062
controller configure	9.299	0.081	0.076


Counts (raw)
600
400
200
0
2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018
Date
controller configure
mobile device

First Derivative
30
20
10
0
2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018
Date
controller configure
mobile device

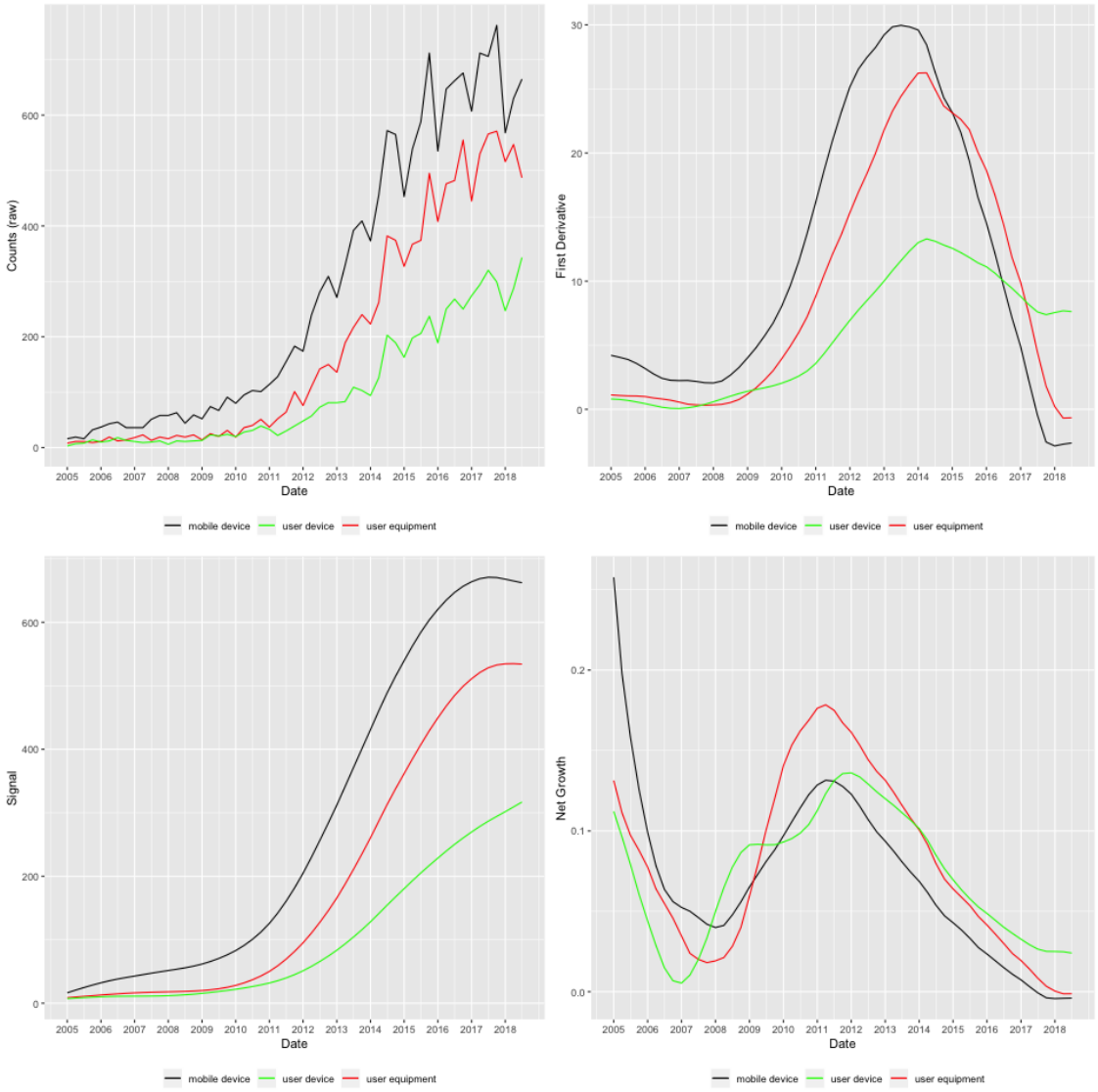

Signal
600
400
200
0
2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018
Date
controller configure
mobile device


Net Growth
0.2
0.1
0.0
2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018
Date
controller configure
mobile device

EXAMPLE. E_2 *vs.* $\bar{E}_2$

We consider the series of counts for "mobile device", "user equipment" and "user device".

Note in the plot in the upper top left corner the black curve, "mobile device", definitely grows the fastest of the three at the very start of the considered time period. The plot in the upper right corner with the first derivatives at each time point ranks emergence according to the magnitude of the signal. The plot in the lower right corner, depicts the ratio of the first derivative by the signal (net growth). In my opinion this point-wise measure of emergence captures the features we would like to focus on: it distinguishes not only the main features of emergence at the beginning of the series but also doesn't give weight to the particular magnitude or scale of the signal and tracks the emergence well across periods.



If we consider the entire time period, E_2 ranks "user equipment" as most emergent, followed by "user device" and the least emergent series is "mobile device".

Ignoring the first two years, when the terms first started to emerge, the ranking stays the same as when considering the whole time period. Note that for the term "user equipment", $\bar{E}_2$ remains the same whether we consider the first two years or not. On the other hand E_2 is larger when considering the entire time period than when discarding the first two years. This is an example

where E_2 increases as the considered time period expands due to the fact that there are more time periods but doesn't really indicate that there was more emergence in the wider time span.

If we consider only the time from 2014, then the ranking changes to "user device" as being the most emergent series, followed by "user equipment" and "mobile device".

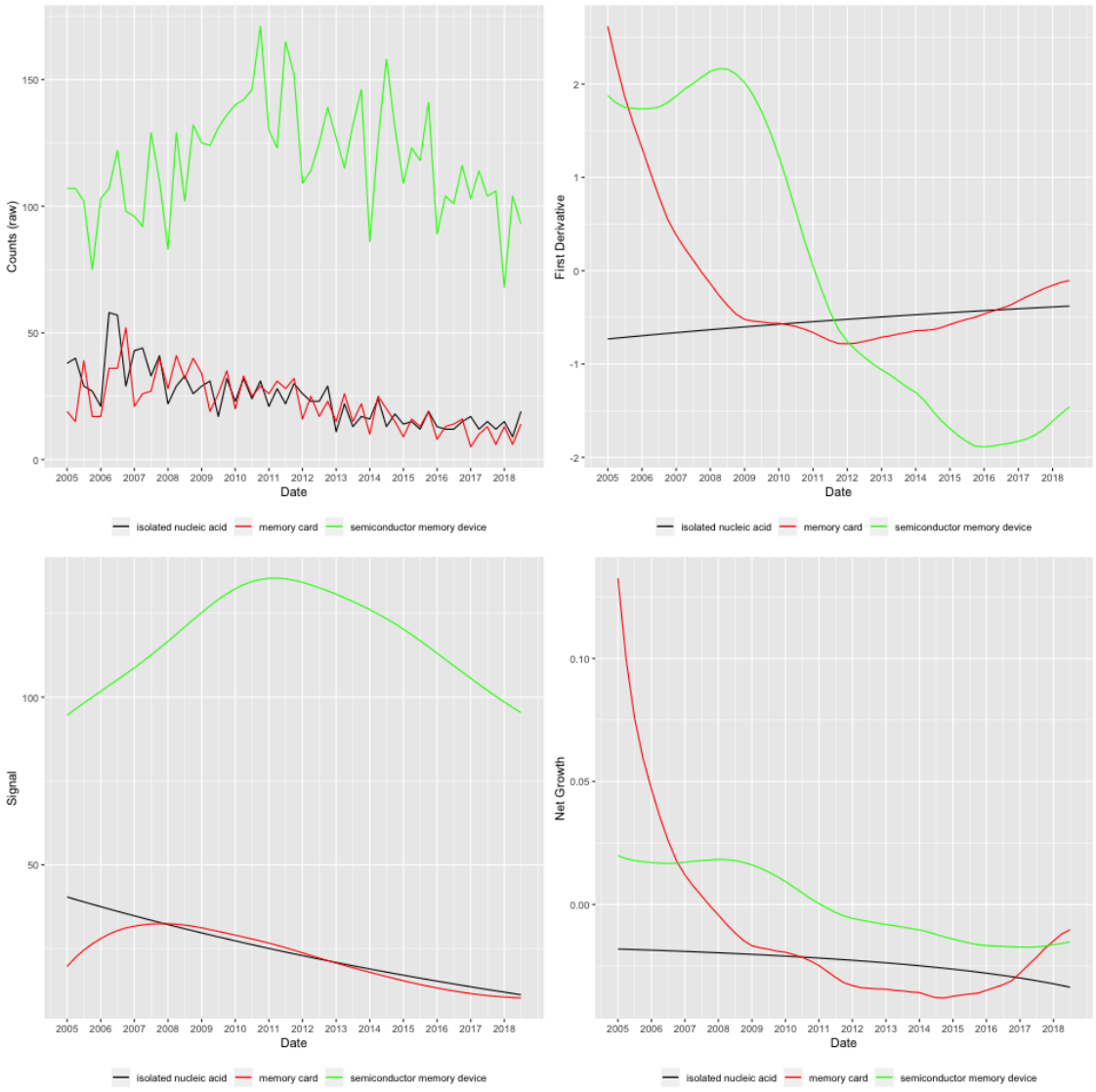
Our assertions above are summarised in the following tables.

Term	σ_ϵ	$E_1 = \sum_i d_i$	$E_2 = \sum_i d_i/\mu_i$	$E_2 = \sum_{i\geq 9} d_i/\mu_i$	$E_2 = \sum_{i\geq 37} d_i/\mu_i$
mobile device	38.838	643.448	3.894	2.858	0.436
user equipment	26.169	524.989	4.371	3.702	0.739
user device	16.673	317.314	3.984	3.542	0.957

Term	$\bar{E}_1 = \frac{1}{55}\sum_i d_i$	$\bar{E}_2 = \frac{1}{55}\sum_i d_i/\mu_i$	$\bar{E}_2 = \frac{1}{47}\sum_{i\geq 9} d_i/\mu_i$	$\bar{E}_2 = \frac{1}{19}\sum_{i\geq 37} d_i/\mu_i$
mobile device	11.699	0.071	0.061	0.023
user equipment	9.545	0.079	0.079	0.039
user device	5.769	0.072	0.075	0.050

EXAMPLE. *Comparing declining series*

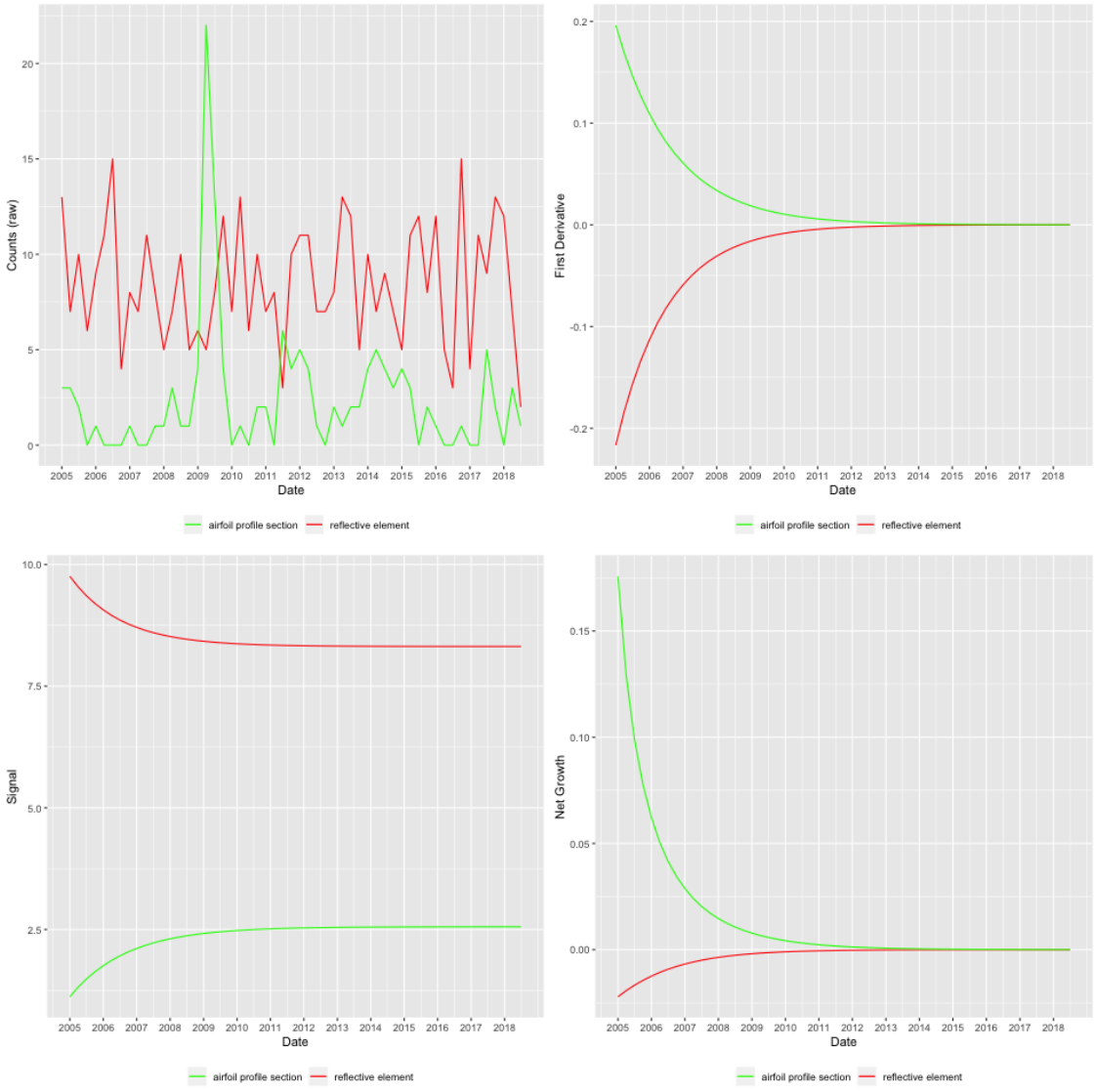
We consider the series of counts for "isolated nucleic acid", "memory card" and "semiconductor memory device". The terms are mostly declining in counts. For a few years at the start of the considered time period (2005 - 3rd quarter 2018) "memory card" and "semiconductor memory device" had increasing counts."semiconductor memory device" has the largest counts throughout the period and also the largest E_2 and $\bar{E}_2$ values. Without considering the first two years, "memory card" and "isolated nucleic acid" have the same $\bar{E}_2$, not E_2, scores. The same holds for the time period from 2014.



Term	σ_ϵ	$E_1 = \sum_i d_i$	$E_2 = \sum_i d_i / \mu_i$	$E_2 = \sum_{i \geq 9} d_i / \mu_i$	$E_2 = \sum_{i \geq 37} d_i / \mu_i$
isolated nucleic acid	6.801	-29.518	-1.294	-1.146	-0.545
memory card	6.760	-8.838	-0.578	-1.111	-0.546
semiconductor memory device	16.522	-0.634	-0.001	-0.142	-0.292

Term	$\bar{E}_1 = \frac{1}{55} \sum_i d_i$	$\bar{E}_2 = \frac{1}{55} \sum_i d_i / \mu_i$	$\bar{E}_2 = \frac{1}{47} \sum_{i \geq 9} d_i / \mu_i$	$\bar{E}_2 = \frac{1}{19} \sum_{i \geq 37} d_i / \mu_i$
isolated nucleic acid	-0.537	-0.024	-0.024	-0.029
memory card	-0.161	-0.011	-0.024	-0.029
semiconductor memory device	-0.012	0.000	-0.003	-0.015

EXAMPLE. *Static series*
We consider the series of counts for "reflective element" and "airfoil profile section".



Term	σ_ϵ	$E_1 = \sum_i d_i$	$E_2 = \sum_i d_i/\mu_i$	$E_2 = \sum_{i\geq 9} d_i/\mu_i$	$E_2 = \sum_{i\geq 37} d_i/\mu_i$
refl.element	3.104	-1.444	-0.159	-0.046	-0.0005
airf.prof.sect.	3.466	1.439	0.863	0.193	0.0027

Term	$\bar{E}_1 = \frac{1}{55}\sum_i d_i$	$\bar{E}_2 = \frac{1}{55}\sum_i d_i/\mu_i$	$\bar{E}_2 = \frac{1}{47}\sum_{i\geq 9} d_i/\mu_i$	$\bar{E}_2 = \frac{1}{19}\sum_{i\geq 37} d_i/\mu_i$
refl.element	-0.026	-0.003	-0.001	-2e-05
airf.prof.sect.	0.026	0.016	0.004	1e-04

References

de Jong, Piet. 1991. "The diffuse Kalman filter." *The Annals of Statistics* 19(2):1073–1083.
URL: *https://projecteuclid.org/download/pdf_1/euclid.aos/1176348139*