I^{2}SDS - Decision Sciences
THE GEORGE WASHINGTON UNIVERSITY
SCHOOL OF BUSINESS

The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2017-3

From Least Squares to Signal Processing and Particle Filtering: An Incredible Journey

Nozer D. Singpurwalla
Department of Systems Engineering and Engineering Management
City University of Hong Kong

Nick Polson
Booth School of Business
University of Chicago, USA

Refik Soyer
Department of Decision Sciences
The George Washington University, USA

From Least Squares to Signal Processing and Particle Filtering

Nozer D. Singpurwalla
City University of Hong Kong [*]

Nicholas G. Polson
University of Chicago [†]

Refik Soyer
George Washington University [‡]

Monday 8th May, 2017

Abstract

De facto, signal processing is the interpolation and extrapolation of a sequence of observations viewed as a realization of a stochastic process. Its role in applied statistics ranges from scenarios in forecasting and time series analysis, to image reconstruction, machine learning, and the degradation modeling for reliability assessment. This topic, which has an old and honourable history dating back to the times of Gauss and Legendre, should therefore be of interest to readers of *Technometrics*. A general solution to the problem of filtering and prediction entails some formidable mathematics. Efforts to circumvent the mathematics has resulted in the need for introducing more explicit descriptions of the underlying process. One such example, and a noteworthy one, is the Kalman Filter Model, which is a special case of state space models or what statisticians refer to as Dynamic Linear Models. Implementing the Kalman Filter Model in the era of "big and high velocity non-Gaussian data" can pose computational challenges with respect to efficiency and timeliness. Particle filtering is a way to ease such computational burdens. The purpose of this paper is to trace the historical evolution of this development from its inception to its current state, with an expository focus on two versions of the particle filter, namely, the propagate first-update next and the update first-propagate next version.

By way of going beyond a pure review, this paper also makes transparent the importance and the role of a less recognized principle, namely, the *principle of conditionalization*, in filtering and prediction based on Bayesian methods. Furthermore, the paper also articulates the philosophical underpinnings of the filtering and prediction set-up, a matter that needs to be made explicit, and Yule's decomposition of a random variable in terms of a sequence of innovations.

Keywords: Dynamic linear models, Filtering likelihood, Genetic Monte Carlo, Kalman filter, Machine learning, Prediction, Smoothing likelihood, State space models, Reliability, Time series analysis.

[*] email: nsingpur@um.cityu.edu.hk
[†] email: ngp@chicagobooth.edu
[‡] email: soyer@gwu.edu

1 Antecedents to Signal Processing and Smoothing

It is fair to state that the genesis of signal processing is the work in 1795 of an 18 year-old Gauss on the method of least squares. The motivation for Gauss' work was astronomical studies on planet motion using telescopic data. Though this work was formally published only in 1809, Gauss laid out a general paradigm for all that has followed. In particular, he recognized that observations are merely approximations to the truth, that such errors of measurement call for more observations than the minimum required to determine the unknowns, that one needs to invoke dynamic models (such as Kepler's laws of motion) for estimating the unknowns, and that a minimization of a function of the residuals leads to their most accurate assessment. More importantly, Gauss also addressed the matter of suitable combination of observations that will provide the most accurate estimates. The above in turn gave birth to the design of filters as a linear or non-linear combination of observables. On page 269 of his *Theoria Motus Corporum Coelestium* (1809), Gauss predicted that his principle of least squares could spawn countless methods and devices by means of which numerical calculations can be expeditiously rendered. This opened the door for approaches like the Kalman Filter to thrive and to survive. Thus, in effect, the Kalman Filter is an efficient computational device to solve the least squares problem, and the particle filter enhances the efficiency of such computational algorithms by speeding them up and by allowing them to be applied in non-Gaussian contexts. But the journey from the ideas of Gauss to the currently popular particle filtering took over 200 years to complete, with the likes of Kolmogorov, Wiener, Bode, and Shannon in the driver's seat. Indeed, as suggested by a referee, a more appropriate title of this paper should have been "From Least Squares to Particle Filtering," but doing so could have detracted the attention of control theorists and signal processors who may view the topic of least squares as being predominantly statistical in nature.

It was almost 135 years after Gauss enunciated the key principles of estimation that Kolmogorov in 1939 provided his solution to the problem of interpolation and extrapolation with minimal assumptions. Specifically, Kolmogorov assumed that the underlying stochastic process is discrete in time, is stationary, and has finite second moments. This set the stage for all that is to follow, including Kolmogorov's 1940 paper which embeds the problem scenario in a Hilbert space and reduces his results of 1939 as a special case. Kolmogorov's 1940 paper is a tour de force in elegance and mathematical genius comprising of just 9 pages. One may wonder as to why problems of interpolation and extrapolation continue to persist despite its closure brought about by the above works. However, an examination of Kolmogorov's results, overviewed in Section 2 of this paper, reveals their formidable nature, and the difficulty in putting them to work.

At about the same time as Kolmogorov, circa 1942, Wiener working on the World War II problem of where to aim anti-aircraft guns at dodging airplanes arrived upon the continuous time formulation of the interpolation and extrapolation problem, now known as ”signal processing.” Here, interpolation got labeled as “filtering” (or smoothing) and extrapolation as ”prediction.“ Wiener’s work, presumed to be done independently of that by Kolmogorov, was for the National Defense Research Council, and remained classified until 1949, when it was reprinted as a book [Wiener (1949)]. Like Kolmogorov’s work, Wiener’s work was also mathematically formidable involving the notoriously famous Wiener-Hopf equation. In Section 3 we give an outline of Wiener’s work leading up to the above mentioned equation (which does not arise in the discrete case of a signal plus noise model). A noteworthy feature of Section 3, is Section 3.1, wherein the philosophical underpinnings of the Kolmogorov-Wiener setup are articulated, especially as they relate to the spirit and the excitement of the early 1920’s and 1940’s, namely quantum theory. It behooves those interested in filtering, extrapolation and machine learning, to be cognizant of what is it that spawned the models they engage with.

The material of Sections 2 and 3 gives the reader an appreciation for the need to develop efficient computational devices like the Kalman filter and the particle filter, which can now be seen as a computational device overlaid on another computational device in order to generalize and speed up the former. The remainder of this paper is organized as follows: Section 4 pertains to the genesis of the state space models via the works of Bode and Shannon (1950), and of Zadeh and Ragazzini (1950), based on electrical circuit theory. The important role played by these works in the development of the statistical theory of dynamic models seems to be unappreciated. Section 5 pertains to the Kalman Filter Model as prescribed by Kalman in 1960, and its (relatively less appreciated) relationship to Yule’s random innovations and the Box-Jenkins approach it spawned, and to Doob’s conditional expectation. Section 6 continues with the theme of Section 5 by providing an outline of the Bayesian prior to posterior iteration which is the essence of Kalman’s filtering algorithm. Whereas the material of Section 6 is well known (to most statisticians and signal processors), it is presented here to set the stage for the material of Section 7 on particle filtering whose exposition, albeit cursory, is a part of the objectives of this paper. Section 6 also highlights the routinely invoked, but less recognized, *principle of conditionalization*, implicit to Kalman filtering. Section 8 concludes the paper with some conjectures about the path forward.

The value of this paper rests on its expository character, vis a vis tracing the historical development from signal processing to particle filtering, articulating the principle of conditionalization, the philosophical underpinnings of the Kolmogorov-Wiener setup and the relationship between the

Kalman filter model and Yule's (1927) statistically motivated notion of random innovations, also known as ”random shocks”.

2 Kolmogorov's Interpolation and Extrapolation of a Sequence

Specifically, for a random variable $X(t)$, with t an integer and $-\infty < t < +\infty$, suppose that $E[X^2(t)] < \infty$, and that the sequence $\{X(t); -\infty < t < +\infty\}$ is stationary. Without loss of generality set $E[X(t)] = 0$, and note that $B(k) = E[X(t+k)X(t)] = B(-k)$, the autocorrelation at lag k, will not depend on t, for any integer $k \geq 0$. The problem of linear extrapolation is to select for any $n > 0$ and $m > 0$, real coefficients a_i, for which

$$L = a_1 X(t-1) + a_2 X(t-2) + \cdots + a_n X(t-n)$$

gives the closest approximation to $X(t+m)$. As a measure of accuracy of this approximation, Kolmogorov (1939) leans on the Gaussian paradigm of minimizing the error sum of squares and considers $\sigma^2 = E[(X(t+m) - L)^2]$ to seek values of a_i for which σ^2 is a minimum. If this minimum value is denoted by $\sigma^2_{\mathcal{E}}(n,m)$, then Kolmogorov shows that $\sigma^2_{\mathcal{E}}(n,m)$ has a limit, and he uses this limit to find the minimizing a_i's.

For the interpolation part, the estimation of $X(t)$ using $X(t \pm 1)$, $X(t \pm 2), \cdots, X(t \pm n)$ is considered, so that if

$$Q = a_1 X(t+1) + \cdots + a_n X(t+n) + a_{-1} X(t-1) + \cdots + a_{-n} X(t-n),$$

then the problem boils down to minimizing $\sigma^2 = E[(X(t) - Q)^2]$. If $\sigma^2_{\mathcal{I}}(n)$ denotes this minimum, then $\sigma^2_{\mathcal{I}}(n)$ cannot increase in n and so its limit, $\sigma^2_{\mathcal{I}}$, exists, and Kolmogorov finds this limit. In both of the above cases, Kolmogorov uses formidable mathematics pertaining to the spectral theory of stationary processes. This underscores the point made before that interpolation and extrapolation are difficult tasks.

3 Wiener's Theory: The Birth of Statistical Signal Processing

Whereas Kolmogorov's approach is cast in the language of probability, Wiener [cf. Wiener (1949)] casts his in the language of communications theory (and hence signal processing). More significantly, Wiener considers the continuous case, and endows the set-up with additional structure than that of Kolmogorov's. Specifically, an observable random sequence $y(t)$ is decomposed as the sum of a random signal $s(t)$ and perturbing noise $n(t)$, unrelated with $s(t)$; that is, $y(t) = s(t) + n(t)$. It is desired to operate on the $y(t)$'s in such a way so as to obtain, as well as is possible, the signal $s(t)$. The act of operating on the $y(t)$'s, became known as filtering, and a filter is a precise specification of the operation on $y(t)$. Wiener also considers the combining of a filtering operation with prediction. That is, operating on $y(t)$ to obtain a good approximation to $s(t + \alpha)$, for some $\alpha >$or < 0.

Underlying Wiener's approach are three assumptions. These are: that the stochastic processes generating the signal $s(t)$ and the noise $n(t)$ are stationary with finite second moments, that $s(t)$ is independent of $n(t)$, that the criterion for the error of approximation is mean square discrepancy, and that the operator on $y(t)$ for filtering and prediction is to be linear on the available information and be implementable (i.e. a computable function of the observed data assuming the availability of the data). In the language of communication theory, the filter is to be linear (in the observed data) and physically realizable (to be explained later). The available information is the past history of the perturbed signal $y(t)$. The assumptions of Wiener parallel those of Kolmogorov; the key differences between the two being a discrete t versus a continuous t, and a decomposition of the observable $y(t)$ into the form of a signal $s(t)$ and a noise $n(t)$. Even so, the probabilistic architecture underlying the two set-ups is identical.

3.1 Philosophical Underpinnings of the Kolmogorov-Wiener Setup

Predicting the future behavior of a signal based on a perturbed version of its present and past history is grounded in philosophical issues pertaining to causality, induction, and the nature of physical law. In general, prediction is based on the inductive premise that the observed patterns of the past will continue to be so in the future. This in turn is an assumption which implies that the past is the cause of the future. An assumption of causality like this one cannot be deduced mathematically. It can not be established empirically either, because empirical verification using statistical techniques entails the null hypothesis that the cause-effect relationship is true, and then an investigation to see if the evidence causes a rejection of the hypothesis. Indeed, the notion that the past is a guide to the future is a central postulate of all the empirical sciences. Classical physics attempted to describe the physical world via a set of (deterministic) causal laws whose role was to relate the past to the future. Examples are: Newton's Laws, Kepler's Law, Ohm's Law, and so on. Quantum physics denied such laws, and claimed them untenable for the microscopic world. Quantum physics claims that on an atomic scale, the laws of physics are only statistical, and that the only meaningful predictions are statistical.

The Kolmogorov-Wiener set-up adheres to the above quantum physics based view that all predictions are statistical, and so is the causal relationship between the past and the future. This viewpoint is asserted via two assumptions: stationarity, and the existence of second moments (i.e. correlations). Prediction is based on the existence of a correlation between the future values of the signal and the past values of the observables, and correlation is indeed the manifestation of a statistical relationship. Kolmogorov's requirement of a finite second moment of $X(t)$ is an assertion of the above thesis. Furthermore, the Kolmogorov-Wiener requirement that the filter be linear, is tantamount to the feature that the only type of relationship that needs to be considered, is a linear, and this manifests itself as a correlation.

To summarize, the routinely invoked Kolmogorov-Wiener assumptions of stationarity, finite second moments, and filter linearity are dictated by the philosophical considerations underlying causality and predictivity. Making this matter explicit is a feature of this paper, and one which enhances its expository character; also see Cox (1992).

3.2 Filtering, Prediction and the Wiener-Hopf Equation

We start with the Wiener-Hopf equation, and trace the steps that lead to it.

Suppose that $\varphi(x)$ is an unknown function of x, $0 \leq x < \infty$, and $K(\cdot)$ and $f(\cdot)$ are known functions with $K(\cdot)$ being monotone. Suppose that for $x > 0$,

$$\varphi(x) = -\int_0^\infty \varphi(y)K(x-y)\,dy + f(x), \tag{3.1}$$

and it is required that the solution to this equation be of the form

$$\varphi(x) \leq c < \infty, \text{ where } \lim{}_{x\to\infty}\varphi(x) = c.$$

(3.1) is the Wiener-Hopf equation, with equivalent representation:

$$\varphi(x) = -\int_{-\infty}^\infty \varphi(x-y)\,dK(y) + f(x), 0 \leq x < \infty. \tag{3.2}$$

(3.2) has been notoriously difficult to solve in general (for processes whose spectral densities are not rational), and attempts to get computationally efficient solutions have lead to approaches like the Kalman Filter. This is the topic of the next section. For now we outline the steps which led to it. The material here is abstracted from Davenport and Root (1958), p. 219.

A filter, $h(t)$, is a weighting function operating on $y(t)$ to give

$$\int_{-\infty}^{\infty} h(t-\tau)\, y(\tau)\, d\tau = \int_{-\infty}^{\infty} h(\tau)\, y(t-\tau)\, d\tau,$$

with the requirement of *physical realizability*, which means that $h(t) = 0$, for $t < 0$. One approach towards advocating the efficacy of the filter is to require that $h(\cdot)$ be chosen so that $\mathcal{E}$, the expected mean square, is minimized. Here, s is the process to be estimated, and y is the observable process:

$$\mathcal{E} = E\{ \left[s(t+\alpha) - \int_{-\infty}^{\infty} h(\tau)\, y(t-\tau)\, d\tau \right]^2 \}. \tag{3.3}$$

Since $s(t)$ and $n(t)$ are stationary, independent, and have finite second moments, their auto and cross-correlations exist, and are time invariant. Consequently,

$$\begin{aligned} \mathcal{E} = E\left[s^2(t+\alpha)\right] - 2\int_{-\infty}^{\infty} h(\tau)\, E\left[s(t+\alpha)\, y(t-\tau)\right] d\tau \\ + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau)h(\mu)E\left[y(t-\tau)\, y(t-\mu)\right] d\tau\, d\mu \\ = B_s(0) - 2\int_{-\infty}^{\infty} h(\tau)\, B_{sy}(\alpha+\tau)\, d\tau + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau)h(\mu)\, B_y(\tau-\mu) d\tau\, d\mu; \end{aligned}$$

where $B_s(k)$ is the autocorrelation at k of the signal process, $B_y(k)$ the autocorrelation of the observable process, and $B_{sy}(k)$ the cross-correlation at k of these processes.

It is shown [Davenport and Root (1958), p. 223-224] that a necessary and sufficient condition $h(t)$ must satisfy for $\mathcal{E}$ to be a minimum is

$$B_{sy}(\tau+\alpha) = \int_{0}^{\infty} h(\mu)\, B_y(\tau-\mu)\, d\mu,\ \tau \geq 0. \tag{3.4}$$

The above is an integral equation which relates a cross-correlation with an autocorrelation, and in the context of the philosophical material of Section 3.1, can be interpreted as a statistical law. The solution to (3.4) will yield an optimum smoothing and prediction filter, and the challenge here has been to find a solution. It has been shown that an exact solution to a realizable filter is based on the requirement that $S_y(f)$, the Fourier transform of $B_y(\tau)$, be rational (so that it can be easily factored), and the solution is expressed in terms of the factors of $S_y(f)$ and $S_{ys}(f)$, the cross spectral density of the $y(t)$ and $s(t)$. The solution therefore has been challenging to obtain, and this has spawned derivations alternate to the above, the pioneering ones being those by Bode and Shannon (April 1950) and by Zadeh and Ragazzini (July 1950). The underlying concept behind both the above

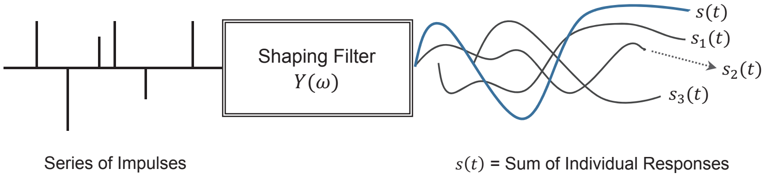

Figure 1: Impulses and Response of the Shaping Filter.

approaches was to give a more explicit description of the signal by introducing an additional filter, called the “shaping filter.” Whereas the statistics community (and possibly also the machine learning community) is well aware of the Kalman filter model, the pioneering works of Bode and Shannon and of Zadeh and Ragazzini which gave birth to *state-space models*, of which the Kalman filter is a special case, appears to be less recognized by the above community (communities). A purpose of this paper is to correct this possible skewness and highlight these overlooked historical footprints.

4 Precursors to Kalman Filtering: The Shaping and Matched Filters

The notion of introducing a shaping filter first appeared in Bode and Shannon (1950) whose aim was to develop a simplified approach for smoothing and filtering under Wiener's set up. The underpinnings of their approach, (which is a simple representation of white noise), was based on circuit design, and their discussion was cast in the language of communications theory entailing the notions of impulses and responses. Based on the first several readings of the Bode-Shannon paper, it is difficult to see as to how the material therein gave birth to the Kalman Filter Model and the other dynamic linear models which followed. But once the fog of terminology is cleared, the ideas become more transparent.

The starting point of the Bode-Shannon approach is a decomposition of a response $s(t)$, not necessarily the $s(t)$ of $y(t) = s(t) + n(t)$. This entails the introduction of a Shaping Filter, the inputs to which are a large number of closely spaced short impulses over time; see Figure 1. The Shaping Filter produces a response to each impulse, so that the response at time t spawned by impulse i is some function $s_i(t)$; see Figure 1. For a linear filter, the responses add up to produce $s(t) = \sum_i s_i(t)$, the total response of the shaping filter.

The shaping filter is characterized by its response to a unit impulse impressed on it at time 0. Thus, for example, if $K(t)$ is the response of a shaping filter at time $t > 0$, to a unit impulse at time 0, then $Y(\omega)$, the transfer function of the shaping filter, is the Fourier transform of $K(t)$; namely, the complex function

$$Y(\omega) = \int_{-\infty}^{\infty} K(t)\, e^{-j2\pi\omega t}\, dt.$$

Conversely, $K(t)$ is the inverse transform of $Y(\omega)$; see Figure 2. Thus

$$K(t) = \int_{-\infty}^{\infty} Y(\omega)\, e^{j2\pi\omega t}\, d\omega.$$

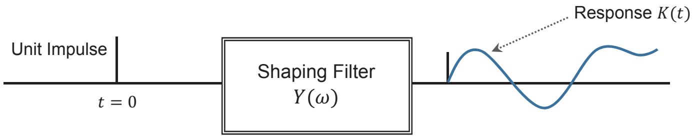

Figure 2: Response of Shaping Filter to a Unit Impulse.

Motivated by this line of thinking, the response of the shaping filter to any continuous input, say $Z(t)$, is obtained by breaking up $Z(t)$ into a large number of thin vertical slices and regarding each slice as an impulse of strength $Z(t)dt$. An impulse of strength $Z(t)dt$ impressed on the shaping filter at time t will produce a response $Z(t)dtK(t_1 - t)$ at t_1, so that $g(t_1)$, the total response of the filter is:

$$g(t_1) = \int_{-\infty}^{t_1} Z(t)K(t_1 - t)dt, \text{ or}$$

$$= \int_{-\infty}^{t_1} Z(t_1 - t)K(\tau)d\tau.$$

If realizability is a requirement, then $K(\tau) = 0$, for $\tau < 0$.

If the input function $Z(t)$ is deterministic, then so will be its output $g(t_1)$ for $t_1 > 0$. In Wiener's set-up, the signal $s(t)$ is assumed to be a stationary random process. The shaping filter which is presumed to generate $s(t)$ needs to have inputs that are impulses of random strength. To achieve the above Bode and Shannon assume that the closely spaced short impulses are independent, and have a common Gaussian distribution. The responses of these impulses add up to generate the stochastic process $s(t)$. It may be of interest to note that the Shaping Filter is merely a conceptual device introduced by Bode and Shannon to structure the input signal $s(t)$ of Wiener's set-up. As such, the Shaping Filter need not be realizable. The genesis of the Shaping Filter lies in circuit theory and pertains to the effects of a resistor on an electrical input.

The work of Zadeh and Ragazzini (1950) builds on the Bode-Shannon theme by generalizing it to assume that the signal $s(t)$ entails two parts, a stochastic process $\tilde{s}(t)$ on which is superimposed a deterministic part $N(t)$ which is a polynomial in t of degree n, but with coefficients that are unknown. Furthermore, Zadeh and Ragazzini require that $h(t)$, the weighting function of the smoothing and prediction filter vanish outside the range $0 \leq t \leq T$, for a specified T.

To recap, the effect of the Bode-Shannon, and the Zadeh-Ragazzini work is to expand the scope of Wiener set up by giving structure to the observed process via a shaping filter. As shown in Figure 3, the shaping filter (which is purely conceptual) precedes the desired filtering, and prediction filter, and this set-up constitutes a foundation for what are known as *state-space models*.

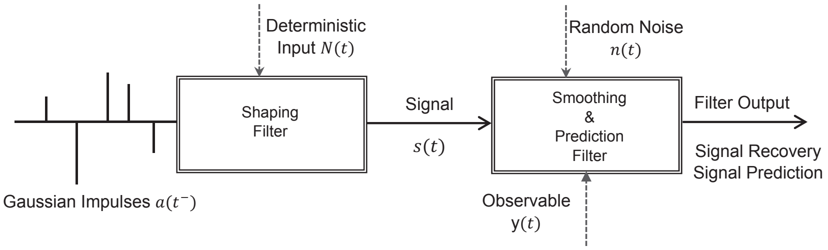

Figure 3: Tandem Architecture of Shaping and Smoothing Filter.

Since the $s(t)$'s share the impulse inputs, denoted by $a(t^-)$ in Figure 3, they will be dependent, and as a consequence, so will the $y(t)$'s. This is despite the fact that $a(t^-)$'s are independent. It is well known that a collection of dependent random variables can always be constructed by considering certain functions of a collection of independent variables; see for example, Singpurwalla et al. (2016). A way to mathematically encapsulate the architecture of Figure 3, ignoring the presence of the deterministic function $N(t)$, and discretizing t, as $t = 0, 1, \ldots$, is to write:

$$s(t) = \mathcal{F}[s(t-1)] + a(t), \text{ and} \tag{4.1.a}$$

$$y(t) = s(t) + n(t). \tag{4.1.b}$$

Here $\mathcal{F}$ is some function of $s(t)$, and the relationships above constitute the essence of a state space model of which the Kalman Filter Model [Kalman (1960)], with equation (4.1.a) constituting the dynamic part, is a special case. Linear cases are those in which the relationship between $y(t)$ and $s(t)$ is linear $-$as indicated in (4.1.b)$-$ and so is the relationship between $s(t)$ and $s(t-1)$. Otherwise, the cases are nonlinear.

Preceding the work of Bode and Shannon (1950), and that of Zadeh and Ragazzini (1950), is the unpublished work of North (1943), and the published work of van Vleck and Middleton (1946) on what is known as ”matched filters” [cf. Turin (1960)]. Underlying the idea of a matched filter is the requirement that a signal $s(t)$ be a deterministic and of known waveform, as opposed to a stochastic process. When such is the case, the smoothing filter $h(t)$ is easy to specify via an inverse Fourier transform. Such a filter is known as a matched filter because it is matched to $s(t)$, and its virtue is an enhanced ability to detect the presence or the absence of a signal $s(t)$. With $s(t)$ fully specified, the matched filter can be seen as a stepping stone to a structured stochastic process like the Kalman Filter.

There are many scenarios in signal processing wherein matched filters arise naturally [see Section VI of Turin (1960)]. They offer potential in non-signal processing applications, whenever a knowledge of $s(t)$ can be had either via the science of the scenario, or via empirical observations. A well illustrated case in point is the detection of cracks in a material via vibrothermography, discussed in good detail, by Li, Holland, and Meeker (2010). These authors consider the more complex scenario of filtering in three dimensions, and the three-dimensional signal can be specified using heat-dispersion theory, or via an empirical argument.

5 Relationship to Yule's Innovations and Doob's Conditional Expectations

Equation (4.1.a) of the Kalman Filter Model has a precedence and a parallel in the manner in which Yule (1927) conceptualized the autoregressive and the moving average processes of time series analysis, developed and popularized by Box and Jenkins (1970). Yule proposed the notion that a highly dependent series $s(t)$ is generated by an *innovation series* $a(t)$, where $a(t)$'s are independent and identically normally distributed with mean 0 and variance σ_a^2. Yule's *causal linear filter* transforms the process $a(t)$ to the process $s(t)$ via the linear operation

$$s(t) = \mu + \psi_0 a(t) + \psi_1 a(t-1) + \psi_2 a(t-2) + \cdots\cdots,$$

where μ, ψ_0, and the ψ_i's are unknown constants. Setting $\mu = 0$ and $\psi_0 = 1$, it can be seen that

$$s(t) = a(t) + \phi_1 s(t-1) + \phi_2 s(t-2) + \phi_3 s(t-3) + \cdots\cdots ,$$

where the ϕ_i's are related to the ψ_i's. Thus $s(t)$ is regressed on its previous values, and the resulting process is an autoregressive process. If the coefficients ψ_i are so chosen that $\phi_i = 0$ for $i \geq 2$, then the results is an autoregressive process of order one, which is equation (4.1.a). Observe the parallel between Yule's construction and the Bode-Shannon set-up as encapsulated in Figure 3. Yule's linear filter is Bode and Shannon's shaping filter; the latter has the advantage of time dependent weights whereas the former does not. With the Kalman Filter Model, we go an additional step beyond Yule's construction towards a smoothing and prediction filter. In effect, the ψ_i's of Yule's linear filter capture the essence of the shaping filter's $Y(\omega)$.

5.1 Filtering with Conditional Expectations: Martingales

It has been recognized [cf. Kalman (1960)] that the Wiener problem can also be approached from the point of view of conditional distributions and expectations. This perspective obviates the need to engage with circuit theory, whitening filters, and the language of signal processing. All that is needed is a knowledge of probability at the intermediate level, and a facility with manipulations that are mathematically cunning. The rewards are plenty, because now one need not be restricted to linear filters, and more importantly, one can lean on the powerful machinery of martingales.

We start by focusing on $(y(t) - y(t-1))$ the change experienced by the observable process $y(t)$, between $(t-1)$ and time t; assume for now that t is discrete. We then ask what is the "best" prediction of $(y(t) - y(t-1))$? A meaningful answer [cf. Kailath (1968)], it seems, would be the conditional expectation

$$E[y(t) - y(t-1) | y(1), \ldots, y(t-1)] = V(t). \tag{5.1}$$

That is, $V(t)$ is the predicted change in $y(t)$ at time t; it is based on a conditional expectation. Next, one considers the error in predicting $y(t)$ using $V(t)$. That is, the *innovation*

$$y(t) - E[y(t) | y(1), \ldots, y(t-1)]. \tag{5.2}$$

Let $U(t) = \sum_{j=1}^{t} V(j)$, the sum of the predicted changes, and

$$M(t) = \sum_{j=1}^{t} \left[y(j) - E[y(j) | y(1), \ldots, y(j-1)] \right],$$

the sum of prediction errors. It is now easy to see that

$$y(t) = U(t) + M(t). \tag{5.3}$$

This means that the sum of all changes in the $y(t)$'s, namely $y(t)$ itself, equals the sum of all the predicted changes $\sum_{j=1}^{t} V(j)$ in $y(t)$ plus $M(t)$, the sum of all the predicted errors. To achieve some semblance with Wiener's set-up, namely that $y(t) = s(t) + n(t)$, we invoke the relationship of equation (5.3) to write

$$y(t) - y(t-1) = V(t) + \big(M(t) - M(t-1)\big). \tag{5.4}$$

More about the quantity $M(t) - M(t-1)$ will be said later, but we first remark that equation (5.3) is known as *Doob's decomposition* of any observable process $y(t)$. Simple as it may seem, Doob's decomposition has some powerful implications, the first of which is that it gives birth to a martingale process.

Specifically, it can be verified—after some routine algebra—that $E[y(t)|y(1), \ldots, y(t-1)] = M(t-1)$, and this implies that $M(t)$ is a *martingale with respect to the process* $y(t)$. Furthermore, it can be shown that

$$E\big[M(t) - M(t-1)\big] = 0, \text{ and that} \tag{5.5a}$$

$$E\big[\big(M(t) - M(t-1)\big)\big(M(t-1) - M(t-2)\big] = 0. \tag{5.5b}$$

Thus if the martingale difference $\big(M(t) - M(t-1)\big)$ of equation (5.4) can be regarded as an error term, then its essence is that the errors have zero mean and are uncorrelated (but not necessarily independent). Equation (5.5b) is the orthogonal increments property of martingales, and is a weakening of the independent increments property assumed in set-ups like classical regression.

Equation (5.3) is quite general and entails practically no assumptions, save for the existence of conditional distributions and the thesis that conditional expectations are "reasonable" or "meaningful" as predictors of unknowns. A *dynamic statistical model* builds upon the theme of equation (5.3) by parameterizing the $V(t)$ process. One such parameterization is to let $V(t) = \alpha y(t-1)$, for some constant $\alpha > 0$. This parameterization states that the expected change in $y(t)$, namely $y(t) - y(t-1)$ is proportional to $y(t-1)$, with $\alpha > 0$ as the constant of proportionality. With this in place it is easy to see that

$$y(t) - y(t-1) = \alpha y(t-1) + \big(M(t) - M(t-1)\big),$$

or that

$$y(t) = (1+\alpha)\, y(t-1) + \big(M(t) - M(t-1)\big), \tag{5.6}$$

an autoregressive process of order 1, with orthogonal errors having mean zero (the latter property is known as *colored noise*). Note that $y(t)$ is a stationary process only when $-2 < \alpha < 0$.

Since a simplified version of Kalman's state space model of equation (4.1) is of the form

$$y(t) = s(t) + n(t), \text{ and}$$

$$s(t) = s(t-1) + a(t), \tag{5.7}$$

a correspondence between the above and the model based on Doob's decomposition—equation (5.3)—is easy to identify. Specifically, iterating on equation (5.7), it is easy to see that for any $n \leq t$,

$$y(t) = s(t-n) + n(t) + \sum_{j=t-n+1}^{t} a(j),$$

so that for $n = t$,

$$y(t) = \big(s(0) + n(t)\big) + \sum_{j=1}^{t} a(j).$$

The desired correspondence holds if $\big(s(0) + n(t)\big)$ is identified with $M(t)$, and $a(j)$ identified with $V(j)$. It is assumed that at $t = 0$, the value of the signal $s(0)$ is known.

There exists a continuous version of the Doob composition, known as the *Doob-Meyer Decomposition*, which spawns a martingale process $\{M(t);\, t \geq 0\}$ with respect to the process $\{y(t);\, t \geq 0\}$. A consequence of the martingale process is an ability to use Levy's Theorem [cf. Doob (1953), Theorem 11.9], which asserts that a martingale process with variance t is a Brownian motion process (also known as Wiener process). Results such as these, expand the scope of Wiener's theory by enabling filtering under more general Gaussian processes, non-Gaussian, and discontinuous processes. For example, Kara, Mandrekar, and Park (1974) discuss recursive least-squares estimation when the noise is a martingale, and Mandrekar and Naik-Nimbalkar (2009) consider estimation when the noise is a fractional Brownian motion. The recent books by Mandrekar and Rudinger (2015), and Mandrekar and Gawarecki (2015) outline the theory and provide a source of references.

Whereas all of the above is conceptually natural, implementation poses a challenge. As a consequence, the filtering algorithm which bears Kalman's name, continues to be used and discussed.

5.2 Antecedents to Kalman's Filtering Algorithm

The algorithm proposed by Kalman (1960), even for a simplified linear version of the state-space model, is cumbersome to describe. The essential features of this algorithm are: all data available up to some time are employed to estimate the state parameter at that time; at any given time one does not retain the whole record of all observations up to that time, their effect being encapsulated in the estimate of the state vector at that time; new data are optimally combined with the most recent state vector. In Section 6 we overview a Bayesian prior to posterior iterative approach for addressing the filtering, smoothing, and prediction as prescribed by equation (5.7). Many find the Bayesian perspective easier to digest. But before doing so we outline below some antecedents that may have lead Kalman to develop his algorithm [cf. Sorenson (1970)].

For the set-up of (5.7), a filter's weighting function for signal $s(t)$, can be easily developed via the method of least squares based on n previous observations $y(t)$. However, a new solution needs to be generated for each new observation and this could be demanding. The idea that upon the receipt of $y(n+1)$, an estimate of $s(n+1)$ can be based on an estimate of $s(n)$ obtained via $y(1), \ldots, y(n)$, is due to Folin in 1955 [cf. Bucy (1968), p. 129]. The notion of recursive filtering and prediction is also present in the works of Swerling (Jan. 1958) and Blum (March 1958), though Swerling's set up is deterministic and there is no mention by him of state space models. Swerling's work was motivated by applications to estimating orbits of earth satellites and space vehicles. A comparison of Swerling's and Kalman's formulation is in Swerling (1998). In the statistical sciences the method of stochastic approximation by Robbins (1951) and by Kiefer and Wolfowitz (1952), were also being studied. Thus it appears that the time was ripe for the recursive approach to state-space estimation proposed by Kalman in 1960—albeit almost after 9 years since the works of Robbins and Kiefer and Wolfowitz. According to Sorenson (1970), Swerling in 1968 wrote a letter to the AIAA Journal claiming priority for the Kalman filter equations based on his 1958 work described in a RAND memorandum on orbit determination.

Of noteworthy mention here is also the striking work (in the former USSR) of Stratonovich (1959, 1960a, 1960b). Stratonovich was the first to emphasize the importance of Markov processes in signal detection in continuous time, and in the sequel, the development of the theory of Conditional Markov Processes.

6 Bayesian Learning and (Kalman) Filtering

Bayesian learning via a prior to posterior iteration can be seen as an implementation of the conditional expectation principle, which is the basis of Doob's decomposition. When the underlying distributions are assumed to be Gaussian (or more generally spherically symmetric) and admit a state-space representation, the principle of least squares and conditional expectation yield identical answers. To appreciate this and related matters, we find it convenient to re-cast the state-space model of equation (5.7) in a notation palatable to statisticians [eg. Meinhold and Singpurwalla (1983)] as:

$$Y_t = \theta_t + v_t \tag{6.1.a}$$

$$\theta_t = \theta_{t-1} + w_t, \tag{6.1.b}$$

where θ_t is an unknown (dynamic) parameter whose value changes with $t = 0, 1, 2, \ldots,$ and v_t and w_t are errors. The v_t's are assumed to be uncorrelated and identically normally distributed with mean 0, and variance σ_v^2; this is denoted as $v_t \sim \mathcal{N}(0, \sigma_v^2)$, with v_t independent of w_t, and $w_t \sim \mathcal{N}(0, \sigma_w^2)$. If $\mathbf{Y_t} = (Y_1, Y_2, \ldots, Y_t)$, then the prediction problem boils down to assessment of $P(Y_t|\mathbf{Y_{t-1}})$, the conditional distribution of a future Y_t, were (supposing that) $\mathbf{Y_{t-1}}$ be known. Note the emphasis on the word "were." By contrast Wiener's prediction problem boils down to an assessment of $P(Y_t|\mathbf{y_{t-1}})$, the distribution of a future Y_t having actually observed $\mathbf{y_{t-1}} = (y_1, \ldots, y_{t-1})$, where y_τ is an observed realization of Y_τ; note the emphasis on the word "actually." The distinction between $P(Y_t|\mathbf{Y_{t-1}})$ and $P(Y_t|\mathbf{y_{t-1}})$ is philosophical and subtle. The mechanics leading to an assessment of both could be the same, but this need not be so; see Section 6.2. Similarly with filtering and smoothing, which entail assessments of $P(\theta_t|\mathbf{Y_t})$ and $P(\theta_j|\mathbf{Y_t})$, respectively, for any $j = 1, 2, \ldots, (t-1)$. What follows next is merely an application of the calculus of probability to achieve the desired assessments. Specifically:

$$P(Y_t|\mathbf{Y_{t-1}}) = \int_{\theta_t} P(Y_t|\theta_t, \mathbf{Y_{t-1}})\, P(\theta_t|\mathbf{Y_{t-1}})\, d\theta_t = \int_{\theta_t} P(Y_t|\theta_t)\, P(\theta_t|\mathbf{Y_{t-1}})\, d\theta_t, \tag{6.2}$$

by law of total probability, and assuming Y_t independent of $\mathbf{Y_{t-1}}$.

From (6.1.a), $(Y_t|\theta_t) \sim \mathcal{N}(\theta_t, \sigma_v^2)$, so that to complete the assessment of $P(Y_t|\mathbf{Y_{t-1}})$ we need to know $P(\theta_t|\mathbf{Y_{t-1}})$. By extending the conversation to θ_{t-1}, and then assuming, given θ_{t-1}, θ_t independent of $\mathbf{Y_{t-1}}$

$$P(\theta_t|\mathbf{Y_{t-1}}) = \int_{\theta_{t-1}} P(\theta_t|\theta_{t-1})\, P(\theta_{t-1}|\mathbf{Y_{t-1}})\, d\theta_{t-1}. \tag{6.3}$$

To obtain $P(\theta_t|\theta_{t-1})$, we lean on (6.1.b) to assert that $(\theta_t|\theta_{t-1}) \sim \mathcal{N}(\theta_{t-1}, \sigma_w^2)$. With the above in place, suppose that $P(\theta_{t-1}|\mathbf{Y_{t-1}})$ is governed by $(\theta_{t-1}|\mathbf{Y_{t-1}}) \sim \mathcal{N}(m_{t-1}, C_{t-1})$, then by the properties of the Gaussian distribution, $(\theta_t|\mathbf{Y_{t-1}}) \sim \mathcal{N}(m_{t-1}, C_{t-1} + \sigma_w^2)$. Using an analogous argument $P(Y_t|\mathbf{Y_{t-1}})$ of equation (6.2) is governed by $\mathcal{N}(m_{t-1},\ C_{t-1} + \sigma_w^2 + \sigma_v^2)$.

Were we to receive the next observation Y_t, then we would be required to assess $P(Y_{t+1}|\mathbf{Y_t})$, and to do so we would need to know $P(\theta_t|\mathbf{Y_t})$. By Bayes' Law,

$$P(\theta_t|\mathbf{Y_t}) = P(\theta_t|Y_t, \mathbf{Y_{t-1}}) \propto \mathcal{L}(\theta_t; Y_t) P(\theta_t|\mathbf{Y_{t-1}}), \tag{6.4}$$

where $\mathcal{L}(\theta_t; Y_t)$ is the likelihood of θ_t with Y_t fixed (and assumed to depend on Y_t alone). The last term is $(\theta_t|\mathbf{Y_{t-1}}) \sim \mathcal{N}(m_{t-1}, C_{t-1} + \sigma_w^2)$. Assuming that the likelihood $\mathcal{L}(\theta_t; Y_t)$ is induced by the feature that $(Y_t|\theta_t) \sim \mathcal{N}(\theta_t, \sigma_v^2)$−equation (6.1.a)−it follows from routine Bayesian prior to posterior calculations that $(\theta_t|\mathbf{Y_t}) \sim \mathcal{N}(m_t, C_t)$.

Wiener's problem of *smoothing* boils down to an assessment of θ_t, were we to know $\mathbf{Y_{t+1}}$. For this we assess $P(\theta_t, \theta_{t+1}|\mathbf{Y_{t+1}})$ and integrate out θ_{t+1}. However, $P(\theta_t, \theta_{t+1}|\mathbf{Y_{t+1}})$ can be obtained as a conditional distribution of $P(\theta_t, \theta_{t+1}, Y_{t+1}|\mathbf{Y_t})$, and this can be assessed via $P(\theta_t|\mathbf{Y_t})$, $P(Y_{t+1}|\mathbf{Y_t})$, and $P(\theta_{t+1}|\mathbf{Y_t})$, all of which we are able to obtain via the discussions of the previous paragraphs. Smoothing pertains to making revised probabilistic assessments of θ_t given all the currently observed information. The intuition here is that better estimates of θ_t are obtained when data subsequent to Y_t is also at hand.

Thus under the simple set-up of equation (6.1), Wiener's filtering, smoothing and prediction problems can be solved in closed form via the principle of conditional expectation, implemented via the mechanics of Bayesian learning. The process of predict and update provides an optimal Bayesian solution for the linear Gaussian state-space model. Matters become computationally challenging when the error distributions are non-Gaussian, have non-constant variances, are correlated, or when (6.1) entails non-linearities. When such is the case, one resorts to *Gibbs sampling* which is a *Markov Chain Monte Carlo* (MCMC) method; it is outlined in Section 6.1. The efficacy of MCMC depends on the convergence of a Markov Chain to an equilibrium distribution. The essence of the Gibbs sampling as applied to a linear Gaussian state-space model (the Kalman filter model) is described below. Our aim is to set the stage for a discussion of the particle filtering algorithm as an alternative to the Gibbs sampling. But before doing so, it may be helpful to remark that if the observed process is "invertible" in the sense of Box and Jenkins (1970), then a naive approach for overcoming the obstacle of a growing dimension is to *filter out* (i.e. eliminate) observations that have occurred in the remote past. When the process is not invertible, then the naive approach will lead to misleading answers. An archetypal example of a non-invertible process is a moving average process of order one, whose coefficient is greater than or equal to one in absolute value. Non-invertibility can arise due to over differencing; see for example, Abraham and Ledolter (1983, pp. 233- 236).

Prior to the advent of Gibbs sampling, the matter of nonlinearity (i.e. an inability to write the evolution of the state variable and/or the observed process as a linear model) was treated by variants of the procedure described above, via what is known as an *extended Kalman Filter* (EKF). This entailed a local linearization of the nonlinear equations by a Taylor series approximation [cf. Anderson and Moore (1979)]. Since Swerling's (1959) original formulation included the nonlinear case as well, it may be claimed that the EKF is the original Swerling filter. However, the EKF was found to be credible only under scenarios wherein the underlying nonlinearities were almost linear, and thus an alternative, namely, the *unscented Kalman filter* (UKF) was proposed by Julier and Uhlmann (1977); also see van der Merwe et al. (2000). The UKF which is not restricted to the requirement of Gaussian distributions is based on the intuition that it is easier to approximate a Gaussian distribution than it is to approximate an arbitrary nonlinear function. Accordingly, a set of points that are deterministically selected from the Gaussian approximation to $P(\theta_t|\mathbf{Y_t})$ are propagated through the underlying nonlinearity, and the points thus propagated used to obtain a Gaussian approximation to the transformed distribution [cf. Arulampalam, Maskell, Gordon and Clapp (2002)]. If the underlying density is bimodal or heavily skewed, then a Gaussian will not approximate it well spawning the need for *robustifying* the Kalman filter using influence functions or thick tailed distributions, such as the Student's−t [cf. Meinhold and Singpurwalla (1989)], or by Monte Carlo based approaches such as Gibbs sampling or particle filtering.

6.1 The Gibbs Sampling Algorithm for Kalman Filtering

As mentioned in the previous section, the Gibbs sampling algorithm for Kalman filtering becomes germane under non-Gaussianity and non-linearity of the Kalman filter model for which a closed form solution exists when otherwise. A fundamental step in the Kalman filter algorithm is the recursive transitioning from $P(\theta_{t-1}|\mathbf{Y_{t-1}})$ to $P(\theta_t|\mathbf{Y_t})$. This operation entails a likelihood and an application of Bayes' law. The specifics of the operation were outlined in the paragraph following (6.4). There is an important aspect of this operation, which is germane to particle filtering. Specifically, to transition from $P(\theta_{t-1}|\mathbf{Y_{t-1}})$ to $P(\theta_t|\mathbf{Y_t})$, one first propagates from θ_{t-1}, to θ_t via (6.1.b), and then brings in the effect of Y_t via the likelihood and Bayes' Law. An exercise like this is legislated by a factorization of the form

$$P(Y_{t+1}, \theta_{t+1}|\theta_t) = P(Y_{t+1}|\theta_{t+1})P(\theta_{t+1}|\theta_t),$$

assuming that given θ_{t+1}, Y_{t+1} is independent of θ_t. Thus with conventional filtering, the motto is: "*propagate first−update next,*" a meaningful thing to do when a real time decision is to be made at t, based on knowledge about θ_t at time t; for example, in automatic control. Here all that matters is $P(\theta_t|\mathbf{Y_t})$.

However, were the scenario be such that a decision based on knowledge about θ_t can be delayed until time $(t+1)$ with $\mathbf{Y_{t+1}}$ at hand, then a smoothed assessment of θ_t based on $\mathbf{Y_{t+1}}$ would be preferable to one based on $\mathbf{Y_t}$ alone. Such delayed decision scenarios arise in the context of statistical inference. Such an exercise is legislated by the factorization

$$P(Y_{t+1}, \theta_{t+1}|\theta_t) = P(\theta_{t+1}|\theta_t, Y_{t+1})P(Y_{t+1}|\theta_t),$$

where the motto would be to: "*update first—propagate next.*" This motto does not entail any assumption of conditional independence. Either motto can be implemented in both the Gibbs sampling algorithm, or the particle filter mechanism (discussed in Section 7). The expression $P(Y_{t+1}, \theta_{t+1}|\theta_t)$ which arises in the context of transitioning from $P(\theta_t|\mathbf{Y_t})$ to $P(\theta_{t+1}|\mathbf{Y_{t+1}})$ will be motivated in Section 7.1.

As a synopsis of the Gibbs sampling algorithm for the model of (6.1), we focus attention on the case $t = 2$, and suppose that Y_1 and Y_2 are observed as y_1 and y_2, respectively. Consider the 4-tuple $(\theta_1, \theta_2, y_1, y_2)$. The set of 2 conditional distributions spawned by this 4-tuple have the distributions:

$$(\theta_1|\theta_2, y_1, y_2) \sim (\theta_1|\theta_2, y_1)$$
$$(\theta_2|\theta_1, y_1, y_2) \sim (\theta_2|\theta_1, y_2).$$

Setting $\theta_1^{(0)}$ and $\theta_2^{(0)}$ as starting values of θ_1 and θ_2, we update $\theta_2^{(0)}$ to $\theta_2^{(1)}$ by generating a sample (indeed, a particle) from $(\theta_2|\theta_1^{(0)}, y_2)$. To do so, we note that

$$P(\theta_2|\theta_1^{(0)}, y_2) \propto P(y_2|\theta_2, \theta_1^{(0)})\, P(\theta_2|\theta_1^{(0)}) = P(y_2|\theta_2)\, P(\theta_2|\theta_1^{(0)}),$$

and the last two probabilities are specified by the assumed structure of (6.1).

Next, we generate $\theta_1^{(1)}$ from $P(\theta_1|\theta_2^{(1)}, y_1) \propto P(\theta_2^{(1)}|\theta_1, y_1)\, \mathcal{L}(\theta_1; y_1)\, P(\theta_1) = P(\theta_2^{(1)}|\theta_1)\mathcal{L}(\theta_1; y_1)\, P(\theta_1)$, where $\mathcal{L}$ is the likelihood. The $\theta_1^{(1)}$ is an update of $\theta_1^{(0)}$.

The above process repeats, so that after k iterations we have

$$(\theta_1^{(0)},\, \theta_2^{(0)}),\, (\theta_1^{(1)},\, \theta_2^{(1)}),\, (\theta_1^{(2)},\, \theta_2^{(2)}), \ldots,\, (\theta_1^{(k)},\, \theta_2^{(k)})$$

based on the starting values $\theta_1^{(0)}$ and $\theta_2^{(0)}$, and given values y_1 and y_2. Under some mild regularity conditions, as $k \to \infty$, the distribution of $(\theta_1^{(k)},\, \theta_2^{(k)})$ converges to the posterior distribution $P(\theta_1, \theta_2|\, y_1, y_2)$; see Gelfand and Smith (1990). Alternatively, samples from the posterior distribution $P(\theta_1, \theta_2|\, y_1, y_2)$ can also be generated using the *forward filtering backward sampling algorithm*; see Fruhwirth-Schnatter (1994), and Carter and Kohn (1994).

Cleary, each new observation increases the size of the tuple by two, and calls for the generation of a new set of k variates. This is computationally burdensome which the particle filter avoids. But first some words about the caveat of conditioning.

6.2 Filtering, Smoothing, and the Principle of Conditionalization

An important, but underemphasized, point pertains to be the subjunctive nature of the discussion up until now. This has to do with the feature that all of probability, to include conditional probability and Bayes' Law, is in the subjunctive mood. That is, the discussion up until now is based on the premise that "were Y_i to be observed as $y_i,\ i = 1, 2, \ldots, t$," and not on the premise that Y_i is actually observed as y_i. The above difference is encapsulated in the claim that all of probability is in the irrealis (or subjunctive) mood, whereas with actual data at hand, inference has to be in the indicative mood; see, for example, Singpurwalla (2016). The development of Section 6.1 and the ensuing histograms therein are meaningful for some assumed value y_i of Y_i. What happens to this development if when the y_i's are the actual observed values of Y_i's ?

Our answer is that everything that has been said before continues to be valid, but only if the philosophical *principle of conditionalization* is adopted [cf. Diaconis and Zabell (1982), or Singpurwalla (2007)]. This means that underlying the current practice in signal processing, forecasting, and control theory, there is an implicit adherence to the principle of conditionalization. Making this point explicit to the engineering and the statistical communities is a feature of this paper which goes beyond a mere review. The principle of conditionalization is best exposited via a subjectivistic interpretation of probability.

Suppose that for two uncertain events A and B, one is able to specify the conditional probability $P(A|B)$. In the subjectivistic context $P(A|B)$ denotes a two-sided bet on the occurrence of event A, under the supposition that event B has occurred. Under the principle of conditionalization, the above bet must continue to hold even when one is informed that event B has actually occurred. In other words, under conditionalization, one's disposition towards betting on event A is indifferent as to whether B is supposed to have occurred or has actually occurred. Several individuals starting with Ramsey (1931) have questioned the universality of this principle. They have claimed that the actual occurrence of B could change one's disposition to bets on event A made under the supposition that event B has occurred. In statistical inference, using Bayes' Law, the principle of conditionalization manifests itself whenever the likelihood is specified by interchanging the roles of parameters and the variables in an assumed probability model. This practice is so routinely followed that its philosophical underpinnings are almost forgotten.

Were the principle of conditionalization not adopted, then the likelihood of (6.4) would not necessarily be induced by the feature that $(Y_t|\theta_t) \sim \mathcal{N}(\theta_t, \sigma_v^2)$, and the commonly used Kalman filter equations for filtering, smoothing, and extrapolations would not follow ! Under such circumstances, the likelihood—which is a weighting function—will be an arbitrary function, specified via judgmental considerations, and the ensuing relationships different from those used in current practice.

7 The Particle Filter

As stated, a closed form solution to the Kalman Filter Model assuming that the underlying distributions are Gaussian entails the inversion of matrices whose size grows with the number of observations. This, in the current era of big data can be forbidding. Gibbs sampling can come to the rescue here, but the Gibbs sampler can be computationally burdensome and its success rests on the convergence of the underlying Markov Chain to an equilibrium distribution [cf. Smith and Roberts (1993)]. By contrast the particle filter mechanism mimics the Bayesian prior to posterior learning step by step, and leans on the law of large numbers to ensure convergence. Indeed, the particle filter (also known as a *genetic Monte Carlo* algorithm) better exploits the Markovian nature of equation (6.1.a) than the Gibbs sampling algorithm, and in so doing it:

i) Circumvents the computational burden spawned by the growing size of the MCMC tuple—see section 6.1, and

ii) Obviates the need for large storage memory by not requiring that all observations prior to the current y_t be retained. This feature of particle filtering truly embodies the essential spirit of recursive estimation as enunciated by Folin in 1955. But as will be pointed out later, the particle filter is not without its drawbacks.

Particle filters (PF) work online and use a discrete set of values called *particles*, each with a weight, to represent the distribution of a state at time t, and to update this distribution at each subsequent time by changing the particle weights according to their likelihoods. There are several versions of the PF, and several surveys and tutorials about it, one of the most comprehensive one being that by Arulampalam, Maskell, Gordon, and Clapp (2002), and one of the most recent one being that by Doucet and Johansen (2011). Also noteworthy is the exhaustive treatise by Chen (2003), and the expository set of lecture notes by Pollock (2010) and by Turner (2013). Chen's (2003) paper is all inclusive with a very thorough set of references; it is written from the perspective of a control theorist with an emphasis on engineering mathematics which statisticians may find challenging to decipher. The current paper may serve as a good prelude to Chen's paper for those who are interested in digesting the material therein.

Importance sampling (IS) and its variants are the key tools which drive the PF; thus it seems appropriate to give below a broad based overview of the essentials of IS.

7.1 Importance Sampling and its Variants

IS was originally introduced in statistical physics as a variance reducing technique. The essence of the idea here is that a mathematical expectation with respect to a target distribution is well approximated by a weighted average of random draws from another distribution called the *importance distribution*. That is, if a random variable θ has a probability density $p(\theta)$, then

$$\mu_f = E_p[f(\theta)] = \int f(\theta)p(\theta)d\theta,$$

and if $q(\theta)$ is some other probability density of θ, with the property that $q(\theta) > 0$, whenever $f(\theta)p(\theta) \neq 0$, then $\mu_f = E_q[\omega(\theta)f(\theta)]$, where $\omega(\cdot) = p(\cdot)/q(\cdot)$. The presumption here is that it is possible to sample from $q(\theta)$ but not from $p(\theta)$.

Thus, if we draw a sample $\theta^{(1)}, \ldots, \theta^{(m)}$ from $q(\theta)$, then $\sum_i^m f(\theta^{(i)})\omega(\theta^{(i)})$ will (by the strong law of large numbers) converge almost surely to μ_f.

The merit of IS is clearly apparent in a Bayesian context wherein the posterior, $p(\theta|y) \propto \mathcal{L}(\theta; y)q(\theta)$, is known only up to a normalizing constant, so that it is possible to sample from the prior $q(\theta)$ but not from $p(\theta|y)$. When such is the case, an estimate of $\mu_f = \int_\theta f(\theta; y)p(\theta|y)d\theta$ is given by

$$\hat{\mu}_f = \sum_i f(\theta^{(i)}; y)\omega(\theta^{(i)}),$$

where $\omega(\theta^{(i)}) = \frac{\mathcal{L}(\theta^{(i)};y)}{\sum_j \mathcal{L}(\theta^{(j)};y)}$ and $\theta^{(1)}, \theta^{(2)}, \ldots$ is a random draw from $q(\theta)$.

In state-space models θ is high dimensional and $p(\theta)$ leads to a chain like decomposition of θ. This enables the sequential construction of the importance density, and now one is able to sequentially update the posterior density at some time t, without modifying the previously simulated states. This is the idea behind *sequential importance sampling* (SIS) discussed by Liu (2001). A common problem encountered with SIS is the *degeneracy* phenomenon, where after few iterations all but a few particles will have negligible weights. Indeed it is shown by Liu (2001) that the weight sequence forms a martingale leading to the feature that the variance of the importance weights increase over time. Consequently, a very small portion of the draws carry most of the weight making the SIS procedure computationally inefficient. Details about the above matters are given in Kong, Liu, and Wong (1994), who among other things propose an approach for overcoming the problem of degeneracy.

Resampling is another approach by which the effects of degeneracy can be reduced. The idea here is to eliminate particles having a small weight and concentrate on particles with a large weight by picking a particle with a probability proportional to its weight. Such a particle filtering process was proposed by Gordon, Salmond, and Smith (1993) in their classic and ground breaking paper; it is known as *sampling importance resampling* (SIR). Whereas the SIR filter resamples particles at the end of an iteration, say at time $(t-1)$ before an observation y_t at t is taken, the *auxiliary particle filter* (APF) introduced by Pitt and Shephard (1999), employs the knowledge about y_t before resampling at $(t-1)$. This ensures that particles that are likely to be compatible with y_t have a good chance of surviving, and in so doing makes the particle filtering process computationally efficient.

Collectively, the process of using a discrete set of weighted particles to represent the distribution of a state, and to update this distribution by changing the particle weights, as is done under the SIS, SIR, and APF algorithms is also known as *sequential Monte Carlo* (SMC), a term coined by Liu and Chen (1998). The PF methods mentioned above suffer from the ”curse of dimensionality” [cf. Bengtsson, Bickel, and Li (2008)]. This happens when $p-$ the dimension of the state space, and $q-$ the dimension of the observation vector are very large in relation to n, where $t = 1, 2, \dots, n$. When such is the case, which arises in the context of climate modeling, dimension reduction techniques which entail a decomposition of the state and observation vectors into many overlapping patches, are invoked. The *ensemble Kalman filter* (EnKF), which is a combination of SMC and the Kalman filter, works under the decomposition scheme mentioned above, whereas the PF does not [cf. Lei and Bickel (2009)].

7.2 Architecture(s) of the Particle Filter Algorithm

Figures 4 and 5 encapsulate the architecture of two versions of the particle filtering mechanism, the former subscribing to the motto of propagate first−update next, and the latter to that of update first−propagate next; see Section 6.1. These figures can be construed as a graphical appreciation of the particle filter mechanism. The essential import of the mechanism of Figures 4 and 5 pertains to the process of transitioning from the distribution of $(\theta_t|\mathbf{Y_t})$ to the distribution of $(\theta_{t+1}|\mathbf{Y_{t+1}})$ upon the receipt of new data. For convenience and ease of exposition, we assume that $(\theta_t|\mathbf{Y_t}) \sim \mathcal{N}(m_t, C_t)$; the notation used here is that of Section 6. The mechanics of the particle filter algorithm is general enough to accommodate distributions other than the Gaussian, and that is another virtue.

As a matter of historical note, even though the recent impetus in particle filtering has been triggered by the 1993 paper of Gordon, Salmond and Smith, the core of the underlying idea goes back to

Galton (1877) [cf. Stigler (2011)].

To discuss the transitioning from a specified $P(\theta_t|\mathbf{Y_t})$ to a $P(\theta_{t+1}|\mathbf{Y_{t+1}})$ upon receipt of Y_{t+1}, we start by considering $P(\theta_{t+1}|\mathbf{Y_{t+1}}) = P(\theta_{t+1}|\mathbf{Y_t}, Y_{t+1}) \propto P(\theta_{t+1}, Y_{t+1}|\mathbf{Y_t})$, and observe that

$$P(\theta_{t+1}, Y_{t+1}|\mathbf{Y_t}) = \int_{\theta_t} P(\theta_{t+1}, Y_{t+1}|\theta_t, \mathbf{Y_t}) P(\theta_t|\mathbf{Y_t}) d\theta_t.$$

Since $P(\theta_t|\mathbf{Y_t})$ is assumed known as $\mathcal{N}(m_t, C_t)$, we focus attention on $P(\theta_{t+1}, Y_{t+1}|\theta_t, \mathbf{Y_t})$ to see how it could be simplified by factorization. There are two factorizations of this joint conditional distribution each leading to a protocol for updating. The first factorization leads to the protocol of *propagate first - update next*; the second to the *update first - propagate next* protocol; see Section 6.1.

7.2.1 The Propagate First—Update Next Protocol

The entity $P(\theta_{t+1}, Y_{t+1}|\theta_t, \mathbf{Y_t})$ of the equation above can be factored as $P(Y_{t+1}|\theta_{t+1}, \theta_t, \mathbf{Y_t}) \times P(\theta_{t+1}|\theta_t, \mathbf{Y_t})$. If Y_{t+1} is assumed independent of θ_t and $\mathbf{Y_t}$ given θ_{t+1}, and θ_{t+1} assumed independent of $\mathbf{Y_t}$ given θ_t, then this factorization simplifies as:

$$P(\theta_{t+1}, Y_{t+1}|\theta_t, \mathbf{Y_t}) = P(Y_{t+1}|\theta_{t+1}) P(\theta_{t+1}|\theta_t). \tag{7.1}$$

Equation (7.1) is the basis of the ”propagate first-update next” protocol. By this it is meant that in moving from the right to left of this equation, one starts by propagating θ_t to θ_{t+1} via equation (6.1.b), and then upon the receipt of Y_{t+1} updates θ_t to θ_{t+1} [using the expression (7.2) given below].

Plugging the simplified factorization of equation (7.1) in the expression for $P(\theta_{t+1}|\mathbf{Y_{t+1}})$ discussed before, we have

$$P(\theta_{t+1}|\mathbf{Y_{t+1}}) \propto \int_{\theta_t} P(Y_{t+1}|\theta_{t+1}) P(\theta_{t+1}|\theta_t) P(\theta_t|\mathbf{Y_t}) d\theta_t. \tag{7.2}$$

The essence of particle filtering under this propagate first-update next protocol is an implementation of equation (7.2), via a simulation exercise, wherein one starts by generating a sample of size N from the distribution of $(\theta_t|\mathbf{Y_t})$, which for purposes of discussion has been assumed Gaussian, and works one's way from right to left. Denote these generated values, known as *particles*, by $\theta_t^{(i)}, i = 1, \ldots, N$; these particles get propagated to $\theta_{t+1}^{(i)}$ via the mechanism driving $P(\theta_{t+1}|\theta_t)$, namely, equation (6.1). With Y_{t+1} observed as y_{t+1}, $P(Y_{t+1}|\theta_{t+1})$ gets replaced by $\mathcal{L}(\theta_{t+1}; y_{t+1}) = P(y_{t+1}|\theta_{t+1})$, the ”filtering” likelihood of θ_{t+1} under an observed y_{t+1}. The rest follows from the schematics of Figure 4, with equation (7.2) in perspective. The *importance weights* $w_{t+1}^{(i)} = \frac{P(y_{t+1}|\theta_{t+1}^{(i)})}{\sum_{j=1}^{N} P(y_{t+1}|\theta_{t+1}^{(j)})}$ modulate the propagated particles $\theta_{t+1}^{(i)}$ by emphasizing those which are meaningful, and diffusing those which are skewed (i.e.

outliers); these importance weights add to 1. Despite the introduction of these modulating weights, there remains the possibility of degeneracy, because the generated particles could be concentrated around a few values causing a collapse of the process. As mentioned, this is a drawback of all such simulation exercises. Recall, the importance weights sum to one, and each weight is proportional to the likelihood of the $\theta_{t+1}^{(i)}$ which spawns it [cf. Carvalho et al. (2010)]. Observe that the flow of actions depicted in Figure 4 mimics the architecture of equation (7.2) as one moves from its right to its left.

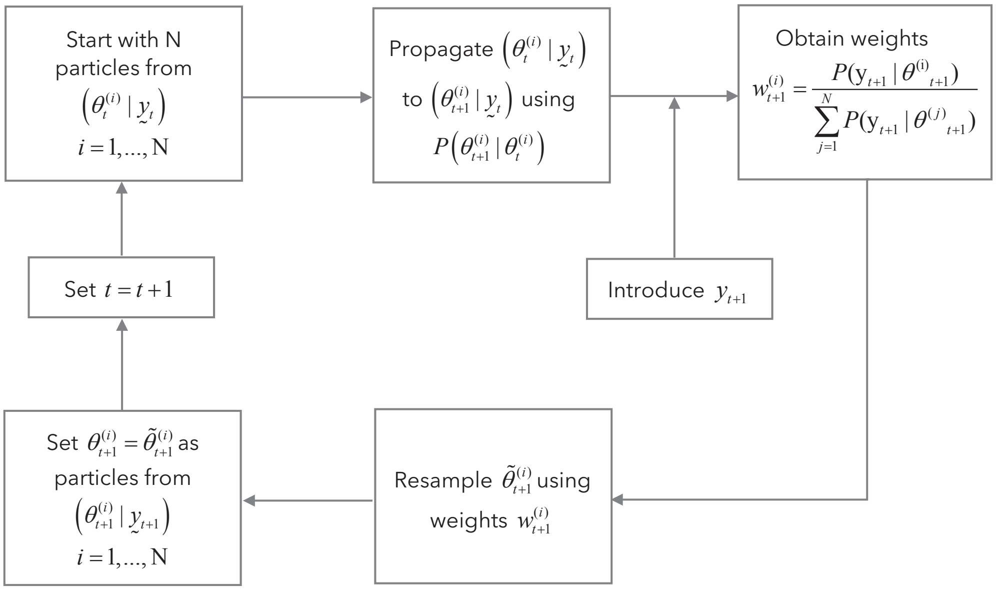
Start with N particles from $\left(\theta_t^{(i)} \mid \underset{\sim}{y}_t\right)$ $i=1,\ldots,\mathrm{N}$
Propagate $\left(\theta_t^{(i)} \mid \underset{\sim}{y}_t\right)$ to $\left(\theta_{t+1}^{(i)} \mid \underset{\sim}{y}_t\right)$ using $P\left(\theta_{t+1}^{(i)} \mid \theta_t^{(i)}\right)$
Obtain weights $w_{t+1}^{(i)} = \frac{P(\mathrm{y}_{t+1} \mid \theta^{(\mathrm{i})}{}_{t+1})}{\sum_{j=1}^{N} P(\mathrm{y}_{t+1} \mid \theta^{(j)}{}_{t+1})}$
Set $t=t+1$
Introduce y_{t+1}
Set $\theta_{t+1}^{(i)} = \tilde{\theta}_{t+1}^{(i)}$ as particles from $\left(\theta_{t+1}^{(i)} \mid \underset{\sim}{y}_{t+1}\right)$ $i=1,\ldots,\mathrm{N}$
Resample $\tilde{\theta}_{t+1}^{(i)}$ using weights $w_{t+1}^{(i)}$

Figure 4: Particle Filtering under propagate First−Update Next Protocol.

The one open question pertains to N the number of cycles that the algorithm needs to execute. Barring the prospect of degeneracy, the law of large numbers will, for large N, ensure convergence to a stationary distribution. This stationary distribution represents the updated (posterior) distribution $\mathcal{N}(m_{t+1}, C_{t+1})$ in our assumed case.

7.2.2 The Update First−Propagate Next Protocol

The entity $P(\theta_{t+1}, Y_{t+1}|\theta_t, \mathbf{Y_t})$ of the previous section has an alternate factorization, and this factorization forms the basis of the "update first−propagate next protocol" for the particle filter. Thus, the two protocols of particle filtering discussed here are motivated by the two factorizations of $P(\theta_{t+1}, Y_{t+1}|\theta_t, \mathbf{Y_t})$. Specifically, $P(\theta_{t+1}, Y_{t+1}|\theta_t, \mathbf{Y_t})$, can also be factored as follows:

$$P(\theta_{t+1}, Y_{t+1}|\theta_t, \mathbf{Y_t}) = P(\theta_{t+1}|\theta_t, Y_{t+1}, \mathbf{Y_t})P(Y_{t+1}|\theta_t, \mathbf{Y_t}) = P(\theta_{t+1}|\theta_t, Y_{t+1})P(Y_{t+1}|\theta_t, \mathbf{Y_t}), \tag{7.3}$$

if θ_{t+1} is assumed independent of $\mathbf{Y_t}$, given θ_t and Y_{t+1}.

Incorporating the factorization of (7.3) into the relationship

$$P(\theta_{t+1}|\mathbf{Y_{t+1}}) \propto \int_{\theta_t} P(\theta_{t+1}, Y_{t+1}|\theta_t, \mathbf{Y_t})P(\theta_t|\mathbf{Y_t})d\theta_t$$

given before, we have

$$P(\theta_{t+1}|\mathbf{Y_{t+1}}) \propto \int_{\theta_t} P(\theta_{t+1}|\theta_t, Y_{t+1})P(Y_{t+1}|\theta_t, \mathbf{Y_t})\, P(\theta_t|\mathbf{Y_t})d\theta_t, \tag{7.4}$$

where

$$P(Y_{t+1}|\theta_t, \mathbf{Y_t}) = \int_{\theta_{t+1}} P(Y_{t+1}|\theta_{t+1}, \theta_t, \mathbf{Y_t})P(\theta_{t+1}|\theta_t, \mathbf{Y_t})d\theta_{t+1}$$

by law of total probability, by conditioning on θ_{t+1}. Assuming θ_{t+1} is independent of $\mathbf{Y_t}$ given θ_t, and Y_{t+1} is independent of θ_t and $\mathbf{Y_t}$ given θ_{t+1}, we have

$$P(Y_{t+1}|\theta_t, \mathbf{Y_t}) = \int_{\theta_{t+1}} P(Y_{t+1}|\theta_{t+1})P(\theta_{t+1}|\theta_t)d\theta_{t+1} = P(Y_{t+1}|\theta_t).$$

Thus, (7.4) simplifies as:

$$P(\theta_{t+1}|\mathbf{Y_{t+1}}) \propto \int_{\theta_t} P(\theta_{t+1}|\theta_t, Y_{t+1})P(Y_{t+1}|\theta_t)P(\theta_t|\mathbf{Y_t})d\theta_t. \tag{7.5}$$

Equation (7.5) parallels (7.2) and is an alternate to it. It encapsulates the "update first-propagate next" protocol. Note that the key difference between (7.2) and (7.5) pertains to the feature that the former entailed $P(Y_{t+1}|\theta_{t+1})$ whereas the latter entails $P(Y_{t+1}|\theta_t)$. With Y_{t+1} observed as y_{t+1}, $P(Y_{t+1}|\theta_{t+1})$ spawns the *filtering likelihood* $\mathcal{L}(\theta_{t+1}; y_{t+1})$ whereas $P(Y_{t+1}|\theta_t)$ spawns the *smoothing likelihood* $\mathcal{L}(\theta_t; y_{t+1}) = P(y_{t+1}|\theta_t)$. An advantage of the smoothing likelihood over the filtering likelihood is that were y_t an outlier but y_{t+1} not, then a consideration of a likelihood based on y_{t+1} would diminish the ill effects of y_t.

Filtering under the update first-propagate next protocol is an implementation of equation (7.5) via a simulation starting with the generation of N particles $\theta_t^{(i)}$, $i = 1, \ldots, N$ from the distribution of $(\theta_t|\mathbf{Y_t})$ and using each $\theta_t^{(i)}$ to specify a smoothing likelihood $\mathcal{L}(\theta_t^{(i)}; y_{t+1})$ and the ensuing importance weights $w_{t+1}^{(i)} \propto \mathcal{L}(\theta_t^{(i)}; y_{t+1})$. Proceeding as above, going from right to left of (7.5) we have

$$P(\theta_{t+1}|\mathbf{Y_{t+1}}) \approx \sum_{i=1}^{N} P(\theta_{t+1}^{(i)}|\theta_t^{(i)}, y_{t+1})w_{t+1}^{(i)},$$

where $P(\theta_{t+1}^{(i)}|\theta_t^{(i)}, y_{t+1})$ is evaluated via Bayes' Law as

$$P(\theta_{t+1}^{(i)}|\theta_t^{(i)}, Y_{t+1}) \propto P(Y_{t+1}|\theta_{t+1}^{(i)})P(\theta_{t+1}^{(i)}|\theta_t^{(i)}),$$

assuming that Y_{t+1} is independent of $\theta_t^{(i)}$ given $\theta_{t+1}^{(i)}$. Thus with Y_{t+1} observed as y_{t+1}, we have

$$P(\theta_{t+1}^{(i)}|\theta_t^{(i)}, y_{t+1}) \propto \mathcal{L}(\theta_{t+1}^{(i)}; y_{t+1})P(\theta_{t+1}^{(i)}|\theta_t^{(i)}).$$

The schematics of Figure 5 illustrates the above operations. Before closing this sub-section, it is appropriate to cite the recent paper by Sukhavasi and Hassibi (2013) which describes the mechanics of filtering when the observation space (as opposed to the state-space) is quantized by particles.

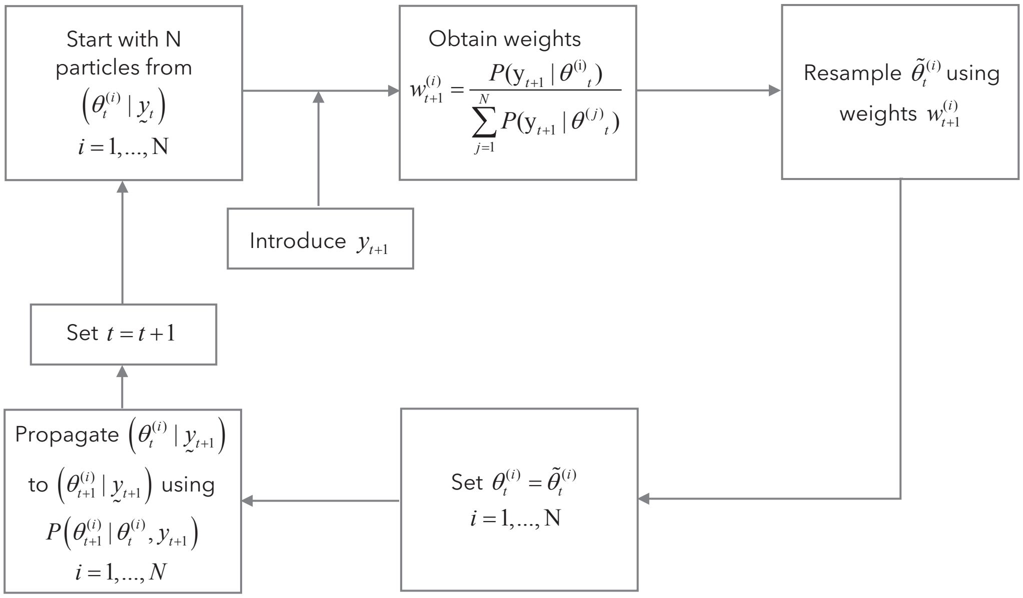
Start with N particles from $\left(\theta_t^{(i)} \mid \underset{\sim}{y}_t\right)$ $i=1,\ldots,\mathrm{N}$
Obtain weights $w_{t+1}^{(i)} = \frac{P(\mathrm{y}_{t+1} \mid \theta^{(\mathrm{i})}{}_t)}{\sum_{j=1}^{N} P(\mathrm{y}_{t+1} \mid \theta^{(j)}{}_t)}$
Resample $\tilde{\theta}_t^{(i)}$ using weights $w_{t+1}^{(i)}$
Introduce y_{t+1}
Set $t=t+1$
Propagate $\left(\theta_t^{(i)} \mid \underset{\sim}{y}_{t+1}\right)$ to $\left(\theta_{t+1}^{(i)} \mid \underset{\sim}{y}_{t+1}\right)$ using $P\left(\theta_{t+1}^{(i)} \mid \theta_t^{(i)}, y_{t+1}\right)$ $i=1,\ldots,N$
Set $\theta_t^{(i)} = \tilde{\theta}_t^{(i)}$ $i=1,\ldots,\mathrm{N}$

Figure 5: Particle Filtering under Update First−propagate Next Protocol.

8 Summary, Conclusions, and the Path Forward

This paper has been primarily written for an audience of applied statisticians, applied probabilists, econometricians, engineers, and time-series analysts, many of whom are familiar with state-space models, but who may not be fully cognizant of the genesis, the evolution and the mathematical underpinnings of such models. The several references citing the work of Mandrekar and his colleagues are given here to provide the reader some sense of what appeals to theoretical probabilists in this arena. Control theorists may find little that is new to them. An exception could be the material of Section 3.1 on the philosophical basis of the Kolmogorov-Wiener setup in the context of quantum theory, and the material of Section 6.2 on the role of the less recognized principle of conditionalization on the commonly used results in filtering. An adherence to this principle is philosophically not mandatory, and when not adhered to, it could fundamentally change the nature of some well known results and the algorithms which produce these.

Besides the philosophical material of Sections 3.1 and 6.2, what distinguishes this paper from other surveys and reviews on filtering is its encompassiveness. Rather than focussing solely on computational or simulation issues, the paper gravitates towards the underlying ideas, and traces the key mileposts of the subject which constitute the core of its foundations. See Figure 6 whose title is inspired by term ”the quark jungle of particle physics.” It starts with the work of Gauss, who laid out a general paradigm for all that is to follow, and then moves on to that of Kolmogorov who put forth a mathematical framework to operationalize Gauss’ paradigm. Wiener enters the picture, presumably independently of Kolmogorov, and ends up adding some structure to Kolmogorov’s very general setup. But this was not enough; ease of implementation continued to be a problem. This first motivated North to propose the ”matched filter,” which in turn was followed up by Bode and Shannon, and Zadeh and Ragazzini to push the envelope further by adding structure to Wiener’s setup, so that now Kolmogorov’s setup had an enhancement in two tiers, the first due to Wiener, and the second due to North, Bode-Shannon, and Zadeh-Ragazzini. These have paved the path towards development of hidden Markov and state-space models. Whereas Shannon and Zadeh have been acknowledged as the originators of information and fuzzy set theory, respectively, the signal role played by them in the development of state-space models warrants a more emphatic recognition. The dates shown in Figure 6 are accurate to the best of our knowledge.

Not to be forgotten is the role of statisticians like Robbins, Kiefer, and Wolfowitz in the enhancement and development of state-space models. Noteworthy is the landmark paper of Lindley and Smith (1972), who gave Kalman’s algorithm a Bayesian prior to posterior interpretation, and in so doing opened the floodgate for statisticians to join the party. This has been a fortuitous development, because statisticians and applied probabilists have developed powerful computational and simulation tools that have advanced the state of the art of filtering by increasing its efficiency. In exchange, state-space models and filtering techniques enhanced the scope of regression models by making them dynamic, and have enhanced the scope of statistical modeling vis a vis graphical modeling and causal analysis. A recent paper by Smith and Freeman (2011) provides a striking perspective on this recip-

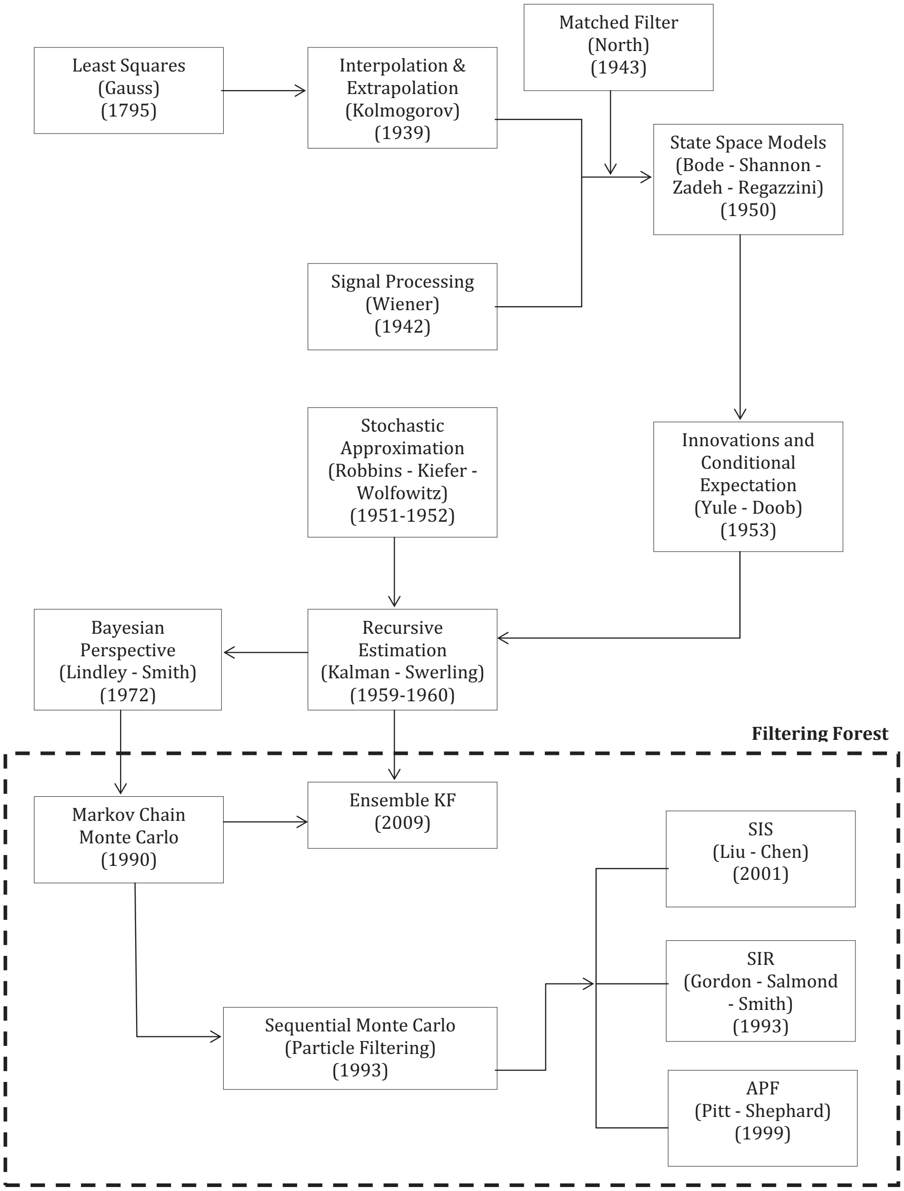

Figure 6: The Journey: From Least Squares to the Filtering Forest.

rocal relationship.

Traditionally, state-space models have primarily been used in signal processing, image analysis, target tracking, astronomical studies, and time series analysis. The era of big data has opened the door to other applications as well, and this is what we mean by path forward. An inkling of this possibility is the work of Li, Holland, and Meeker (2010), which pertains to a problem in reliability. Big data tends to be high dimensional because it is often generated by an array of sensors that are spatially placed and which generate volumes of information in real time. In the application by Li et al. (2010) filtering is done in three dimensions via a matched filter, and the challenge of doing so is addressed by the Fast Fourier Transform. It has often been claimed that the future of reliability and maintainability will be driven by an ability to anticipate failure and to take timely preventive measures by the real time tracking of degradation and wear; see, for example, Lu and Meeker (1993). Sentiments such as these have spawned efforts such as those by Qian and Yan (2015) for using the particle filter to predict useful life of bearings, by Wang, Miao, Zhou, and Zhow (2015) for gear, by Sun, Zou, Wang, and Pecht (2012) for gas turbines, and by Zio and Peloni (2011) for tracking fatigue crack growth. An overview of prognostics based on particle filter methods is given by Jouin, Gouriveau, Hissel, Pera and Zerhouni (2016). More recently, with the advent of self driving cars and airplanes such as the ”Dreamliner”, filtering techniques have been used to predict the residual lifetimes of rechargeable batteries. Here, degradation is often described by a Brownian motion process with an adaptive drift, and is tracked by a particle filter; see, for example, Wang, Carr, Xu, and Kobbacy (2011), Dalal, Ma, and He (2011), Xing, Ma, Tsui, and Pecht (2013), and Si (2015).

With big data, one may also need to engage with stochastic processes in high dimensions. The theoretical foundation for doing the above was initiated by Wiener and Masani (1957,1958), and by Masani (1960). The material there is technically demanding, and is mentioned here mainly for sake of historical completeness.

ACKNOWLEDGEMENTS

The authors thank the editors and referees for their thorough and constructive reviews which enhanced the scope and quality of the paper.

One of the authors (NS) was greatly inspired by Professor Thomas Kailath’s (1991) review paper cited below and benefited greatly from its comprehensive and exhaustive coverage. He thanks Tom for this contribution to his learning, and also his comments on an earlier draft of this paper. He also thanks Professor William Meeker who made several comments pertaining to scope, coverage, and projecting to the future, and for raising awareness about ”matched filters.” Finally, Viyadhar (Atma) Mandrekar took many pains alerting him to the more recent mathematical work in filtering.

Nozer Singpurwalla’s work was supported in part by the US Office of Naval Research - Global Grant N62909-15-1-2026 with The City University of Hong Kong, and The City University of Hong Kong Project Numbers 9380068 and 9042083.

References

[1] Abraham, B. and Ledolter, J. (1983). *Statistical Methods for Forecasting*, John Wiley: New York, NY.

[2] Anderson, B. D.O. and Moore, J. B. (1979). *Optimal Filtering*, Prentice-Hall: Englewood Cliffs, NJ.

[3] Arulampalam, M. S., Maskell, S., Gordon, N. and Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, Vol. 50: 174-188.

[4] Bengtsson, T., Bickel, P. and Li, B. (2008). Curse of dimensionality revisited: The collapse of importance sampling in very large scale systems. *IMS Collections: Probability and Statistics*, Vol. 2: 316-334.

[5] Blum, M. (1958). Recursion formulas for growing memory digital filters. Information Theory, *IRE Transactions*, Vol. 4: 24-30.

[6] Bode H. W. and Shannon, C. E. (1950). A simplified derivation of linear least squares smoothing and prediction theory. *Proceedings of the IRE*, Vol. 38(4): 417-425.

[7] Box, G.E.P. and Jenkins, G.M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.

[8] Bucy, R. S. and Joseph, P. D. (1968). *Filtering for Stochastic Processes with Applications to Guidance*. Interscience, New York.

[9] Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, Vol. 81: 541-553.

[10] Carvalho, C., Johannes, M. S., Lopes, H. F. and Polson, N. G. (2010). Particle learning and smoothing. *Statistical Science*, Vol. 25:88-106.

[11] Chen, Z. (2003). Bayesian filtering: From Kalman filters to particle filters, and beyond. *Statistics*, Vol. 182: 1-69.

[12] Cox, D. R. (1992).Causality: Some statistical aspects. *Journal of the Royal Statistical Society, Series A,* Vol. 155: 291-301

[13] Dalal, M., Ma, J. and He, D. (2011). Lithium-ion battery life prognostic health management system using particle filtering framework. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, Vol. 225:81-90.

[14] Davenport, W. B. and Root, W. L. (1958). *Random Signals and Noise*. New York: McGraw-Hill.

[15] Diaconis, P. and Zabell, S. L. (1982). Updating subjective probability. *Journal of the American Statistical Association*, Vol. 77: 822-830.

[16] Doob, J. L. (1953). *Stochastic Processes*. Wiley: New York.

[17] Doucet, A. and Johansen, A. M. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering, Eds.,D. Crisan and B. Rozovsky*, 656-704, Oxford University Press.

[18] Fruhwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models, *Journal of Time Series Analysis*, Vol. 15:183-202.

[19] Galton, F. (1877). Typical laws of heredity. *Nature*, Vol. 15:492–495, 512–514, 532–533. Also published in Proceedings of the Royal Institution of Great Britain 8:282–301.

[20] Gauss, C. F. (1809). Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss. sumtibus Frid. Perthes et IH Besser.

[21] Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-Based approaches to calculating marginal densities", *Journal of the American Statistical Association*, Vol. 85:398-409.

[22] Gordon, N., Salmond, D. and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc.* F, Vol. 140(2): 107-113.

[23] Jouin, M., Gouriveau, R., Hissel, D., Pera, M.-C. and Zerhouni, N. (2016). Particle filter-based prognostics: Review, discussion and perspectives. *Mechanical Systems and Signal Processing*, Vol. 7273:2-31.

[24] Julier, S. and Uhlmann, J. (1997). A new extension of the Kalman filter to nonlinear systems. in *Proc. AeroSense: 11th Int. Symp. Aerosp./Defense Sensing, Simulat. Contr.*

[25] Kailath, T. (1968). An innovations approach to least-squares estimation–Part I: Linear filtering in additive white noise. *IEEE Transactions on Automatic Control*, Vol. 13: 646-655.

[26] Kailath, T. (1991). From Kalman filtering to innovations, martingales, scattering and other nice things. *Mathematical System Theory: The Influence of RE Kalman*, 55-88.

[27] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, Vol. 82: 35-45.

[28] Kara, H., Mandrekar, V. and Park, G. L. (1974). Wide-sense martingale approach to linear optimal estimation. *SIAM Journal of Applied Mathematics*, Vol. 27:293-302.

[29] Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, Vol. 23(3): 462-466.

[30] Kolmogorov, A. N. (1939). Sur l'interpolation et extrapolation des suites stationaries. *C. R. Acad. Sci. Paris*, 208: 2043-2045.

[31] Kong, A., Liu, J. S. and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.*, Vol. 89:278-288.

[32] Lei, J. and Bickel, P. (2009). Ensemble filtering for high dimensional non-linear state space models. Technical Report.

[33] Li, M., Holland, S. M. and Meeker, W. Q. (2010). Statistical methods for automatic crack detection based on vibrothermography sequence of images of data. *Applied Stochastic Models in Business and Industry*, Vol. 26:481-495.

[34] Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for linear models. *Journal of the Royal Statistical Society, Series B*, Vol. 34: 1-41.

[35] Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. J. *Amer. Statist. Assoc.*, Vol. 93: 1032-1044.

[36] Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

[37] Lu, C. J., and Meeker, W. Q. (1993). Using degradation measures to estimate a time to failure distribution. *Technometrics*, Vol. 35: 161-174.

[38] Mandrekar, V. and Naik-Nimbalkar, U. V. (2009). Identification of a Markovian system with observations corrupted by fractional Brownian motion. *Statistics and Probability Letters*, Vol. 79:965-968.

[39] Mandrekar, V. and Rudiger, B. (2015). *Stochastic Integration in Banach Spaces*. Springer Verlag.

[40] Mandrekar, V. and Gawaracki, L. (2015). *Stochastic Analysis for Gaussian Random Processes and Fields: With Applications*. CRC Press, Boca Raton.

[41] Masani, P. (1960). The prediction theory of multivariate stochastic processes III. *Acta Mathematica*, Vol. 104:141-162.

[42] Meinhold, R. J. and Singpurwalla. N. D. (1983). Understanding the Kalman filter. *The American Statistician*, Vol. 37(2): 123-127.

[43] Meinhold, R. J. and Singpurwalla. N. D. (1989). Robustification of Kalman filter models. *Journal of the American Statistical Association*, Vol. 84:479-486.

[44] North, D. O. (1943). Analysis of factors which determine signal/noise discrimination in radar. *Rept. PTR-6C*, RCA Laboratories, Princeton, N.J.

[45] Pitt, M. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Amer. Statist. Assoc.*, Vol. 94: 590-599.

[46] Pollock, M. (2010). Introduction to particle filtering: Discussion. Technical Report.

[47] Qian, Y. and Yan, R. (2015). Remaining useful life prediction of rolling bearings using an enhanced particle filter. *IEEE Transactions on Instrumentation and Measurement*, Vol. 64:2696-2707.

[48] Ramsey, F. P. (1931). Truth and probability. in *The Foundations of Mathematics and Other Logical Essays*, Ed. R. G. Braithwaite, 156-198. London: Routledge and Kegan Paul.

[49] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400-407.

[50] Si, X. (2015). An adaptive prognostic approach via nonlinear degradation modeling: Application to battery data. *IEEE Transactions on Industrial Electronics*, Vol. 62: 5082-5096.

[51] Singpurwalla, N. D. (2007). Betting on residual life: The caveats of conditioning. *Statistics & Probability Letters*, Vol. 77(12): 1354-1361.

[52] Singpurwalla, N. D. (2017). Mood transitions in Bayesian inference. Technical Report.

[53] Singpurwalla, N. D., Arnold, B. C., Gastwirth, J. L., Gordon, A. S. and Ng, H. K. T. (2016). Adversarial and amiable inference in medical diagnosis, reliability, and survival analysis. *International Statistical Review*, Vol. 84: 390-412.

[54] Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo Methods, *Journal of the Royal Statistical Society, Series B*, Vol. 55: 3-23.

[55] Smith, J. Q. and Freeman, G. (2011). Distributional Kalman filters for Bayesian forecasting and closed form recurrences. *Journal of Forecasting*, Vol. 30: 210-224.

[56] Sorenson, H. W. (1970). Least-squares estimation: from Gauss to Kalman. *Spectrum*, IEEE, Vol. 7(7): 63-68.

[57] Stigler, S. M. (2011). Galton visualizing Bayesian inference. *Chance*, Vol. 24(1): 8-10.

[58] Stratonovich, R. L. (1959). On the theory of optimal non-linear filtration of random functions. *Th. Prob. and Appl.*, Vol. 4: 223-225.

[59] Stratonovich, R. L. (1960a). Application to the theory of Markov processes for optimum filtration of signals. *Radio Eng. Electron. Phys. (USSR)*, Vol. 1: 1-19.

[60] Stratonovich, R. L. (1960b). Conditional Markov processes. *Th. Prob. and Appl.*, Vol. 5: 156-178.

[61] Sukhavasi, R. T. and Hassibi, B. (2013). The Kalman-like particle filter: Optimal estimation with quantized innovations/measurements. *IEEE Transactions on Signal Processing*, Vol. 61: 131-136.

[62] Sun, J., Zuo, H., Wang, W. and Pecht, M. G. (2012). Application of a state space modeling technique to system prognostics based on a health index for condition-based maintenance. *Mechanical Systems and Signal Processing*, Vol. 28:585-596.

[63] Swerling, P. (1959). A proposed stagewise differential correction procedure for satellite tracking and prediction. *J. Astronautic. Sci.*, Vol 6: 46-59.

[64] Swerling, P. (1998). Comparison of Swerling's and Kalman's formulations of Swerling-Kalman filters. In *Tracking and Kalman Filtering Made Easy*, Eli Brookner, John Wiley and Sons, Inc.

[65] Turin, G. L. (1960). An introduction to matched filters. *IRE Transactions on Information Theory*, Vol. 6:311-329.

[66] Turner, L. (2013). An introduction to particle filtering. Technical Report.

[67] van der Merwe, R., Doucet, A., de Freitas, J . F. G. and Wan, E. (2000). The unscented particle filter. *Advances in Neural Information Processing Systems*, Vol. 13: 584-590.

[68] van Vleck, I. and Middleton, D. (1946). A theoretical comparison of the visual, aural, and meter perception of pulsed signals in the presence of noise. *Journal of Applied Physics*, Vol. 17:940-971.

[69] Wang, M., Miao, Q., Zhou, Q. and Zhow, G. (2015). An intelligent prognostic system for gear performance degradation assessment and remaining useful life estimation. *Journal of Vibration and Acoustics*, Vol. 137:21004-1-21004-11.

[70] Wang, W., Carr, M., Xu, W., and Kobbacy, K. (2011). A model for residual life prediction based on Brownian motion with an adaptive drift. *Microelectronics Reliability*, Vol. 51: 285-293.

[71] Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Cambridge, MA: MIT press.

[72] Wiener, N. and Masani, P. (1957). The prediction theory of multivariate stochastic processes I. *Acta Mathematica*, Vol. 98:111-150.

[73] Wiener, N. and Masani, P. (1958). The prediction theory of multivariate stochastic processes II. *Acta Mathematica*, Vol. 99:93-137.

[74] Xing, Y, Ma, E. W., Tsui, K.-L. and Pecht, M. (2013). An ensemble model for predicting the remaining useful performance of lithium-ion batteries. *Microelectronics Reliability*, Vol. 53: 811-820.

[75] Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London*. Series A, 267-298.

[76] Zadeh L. A. and Ragazzini, J. R. (1950) An extension of Wiener's theory of prediction. *Journal of Applied Physics*, Vol. 21(7): 645-655.

[77] Zio, E. and Peloni, G. (2011). Particle filtering prognostic estimation of the remaining useful life of nonlinear components. *Reliability Engineering and System Safety*, Vol. 96:403-409.