Entropy **2014**, *16*, 1047-1069; doi:10.3390/e16021047

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

The Kalman Filter Revisited Using Maximum Relative Entropy

Adom Giffin [1,*] **and Renaldas Urniezius** [2]

[1] Department of Mathematics, Clarkson University, Potsdam, New York, USA

[2] Department of Control Technologies, Kaunas University of Technology, Lt-51367 Kaunas, Lithuania

* Author to whom correspondence should be addressed; E-Mail: physics101@gmail.com.

Received: 12 September 2013; in revised form: 5 February 2014 / Accepted: 7 February 2014 / Published: 19 February 2014

Abstract: In 1960, Rudolf E. Kalman created what is known as the Kalman filter, which is a way to estimate unknown variables from noisy measurements. The algorithm follows the logic that if the previous state of the system is known, it could be used as the *best guess* for the current state. This information is first applied *a priori* to any measurement by using it in the underlying dynamics of the system. Second, measurements of the unknown variables are taken. These two pieces of information are taken into account to determine the current state of the system. Bayesian inference is specifically designed to accommodate the problem of updating what we think of the world based on partial or uncertain information. In this paper, we present a derivation of the general Bayesian filter, then adapt it for Markov systems. A simple example is shown for pedagogical purposes. We also show that by using the Kalman assumptions or "constraints", we can arrive at the Kalman filter using the method of maximum (relative) entropy (MrE), which goes beyond Bayesian methods. Finally, we derive a generalized, nonlinear filter using MrE, where the original Kalman Filter is a special case. We further show that the variable relationship can be any function, and thus, approximations, such as the extended Kalman filter, the unscented Kalman filter and other Kalman variants are special cases as well.

Keywords: Kalman filter; extended Kalman; unscented Kalman; Bayes; Bayesian; complexity; relative entropy; dynamical systems

PACS Classification: 02.50.Cw,05.45.-a,89.70.Eg,89.70.Cf

---

1. Introduction

In 1960, Rudolf E. Kalman demonstrated an ingenious way to estimate unknown variables from noisy measurements [1]. He did this by including information about the underlying dynamical system that governed the variables under consideration. With this, the optimal state of the system was inferred. To do this, he had two main assumptions: (1) all noise was Gaussian or normal and linearly additive; (2) the dynamical system was linear. The result was what is known as the Kalman filter.

Essentially, the algorithm follows the logic that if the previous state of the system is known, it could be used as the *best guess* for the current state. This information is used in two ways, the first is that *prior* to any measurement, the underlying dynamics of the system may be known. Given this knowledge and the previous state, the new state could be determined. Second, measurements of the unknown variables are taken. These two ways may conflict. Which solution should we believe? The answer is that we should believe them *both*, with some uncertainty. They should both be taken into account to determine what our new belief is for the state or what the values of the variables are after the measurements.

Bayesian inference is specifically designed to accommodate the problem of updating what we think of the world based on partial or uncertain information. It is well remarked that the Kalman filter is a special case of Bayesian inference [2]. We present our own derivation of the general Bayesian filter, then adapt it for Markov systems. A simple example is shown for pedagogical purposes with emphasis on the construction of the Kalman gain. Besides offering a greater pedagogical understanding of the algorithm, this also offers an insight into what to do when the Kalman assumptions are not valid or no longer apply. This allows the enhancement of sophisticated solutions, such as the extended Kalman, unscented Kalman, *etc*. [3].

However, Bayes rule does not assign probabilities; it is only a rule to manipulate them. The MaxEnt method [4,5] was designed to assign probabilities. This method has evolved to a more general method, the method of maximum (relative) entropy (MrE) [6,7], which has the advantage of not only assigning probabilities but *updating* them when new information is given in the form of constraints on the family of allowed posteriors. The main purpose of this paper is to show both general and specific examples of how the MrE can be applied using data and moment constraints. It should also be noted that Bayes' rule is a special case of MrE. This implies that MrE is *capable of producing every aspect of orthodox Bayesian inference* and proves the complete compatibility of the Bayesian and entropy methods. Further, it opens the door to tackling problems that could not be addressed by either the MaxEnt or orthodox Bayesian methods individually; problems in which one has data and moment constraints. Thus, Kalman filters can be seen as a special case of filters developed using the MrE methods with Kalman assumptions.

In this paper, we will show several things; first, the derivation of the general Bayesian filter as used for problems that are of the nature that the Kalman filter is intended for, *i.e.*, Markov systems. Second, we will show a simple example illustrating that the Kalman filter is a special case of the general Bayesian filter for pedagogical purposes. Third, we show that using the Kalman assumptions or "constraints", we can arrive at the Kalman filter from MrE directly. Finally, we will show how the same Kalman logic can be applied to non-linear dynamical systems using Bayes rule and avoid approximations that are usually applied in extended Kalman filter and the unscented Kalman filter.

2. Bayesian Filter

Here, we will build the Bayesian filter. We start with Bayes rule,

$$p(x_k|Y_k) = \frac{p(Y_k|x_k)\, p(x_k)}{p(Y_k)} \tag{1}$$

where the k is a temporal index, $p(x_k)$ is our prior, $p(Y_k|x_k)$ is our likelihood, $p(x_k|Y_k)$ is the posterior, $Y_k = \{y_1,..., y_k\}$ are our measurements and x_k is some unknown variable that we would like to infer. We can split Y_k into two sets, y_k and Y_{k-1}, where $Y_{k-1} = \{y_{k-1},..., y_1\}$, which would give us,

$$p(x_k|y_k,Y_{k-1}) = \frac{p(y_k,Y_{k-1}|x_k)\, p(x_k)}{p(y_k,Y_{k-1})} \,. \tag{2}$$

Using the product rule [8] (which is used to derive Bayes theorem), we attain,

$$p(x_k|y_k,Y_{k-1}) = \frac{p(y_k|x_k, Y_{k-1})\, p(Y_{k-1}|x_k)\, p(x_k)}{p(y_k|Y_{k-1})p(Y_{k-1})} \,.$$

Further, we can recognize that we can use Bayes Rule to rewrite part of the equation as,

$$p(x_k|Y_{k-1}) = \frac{p(Y_{k-1}|x_k)\, p(x_k)}{p(Y_{k-1})} \,. \tag{3}$$

This can be seen as the prior, for x_k, given all the other measurements; in other words, all the Bayesian updating on x_k prior to measuring y_k. This yields,

$$p(x_k|y_k,Y_{k-1}) = \frac{p(y_k|x_k, Y_{k-1})\, p(x_k|Y_{k-1})}{p(y_k|Y_{k-1})} \,, \tag{4}$$

which is sometimes referred to as a recursive Bayesian filter, because each subsequent solution step (posterior) is the prior for the new step.

At this point, we come to our first key *assumption* for Kalman; if x_k is "complete" [9], then y_k is not conditionally dependent on Y_{k-1} or $p(y_k|x_k, Y_{k-1}) = p(y_k|x_k)$. In other words, if we have, x_k, then we do not need Y_{k-1} to determine the probability of y_k. This then yields,

$$p(x_k|y_k,Y_{k-1}) = \frac{p(y_k|x_k)\, p(x_k|Y_{k-1})}{p(y_k|Y_{k-1})} \,. \tag{5}$$

Often, the form for the prior, $p(x_k|Y_{k-1})$, is not known. However, it can be seen as a marginal,

$$p(x_k|Y_{k-1}) = \int p(x_k|x_{k-1})\, p(x_{k-1}|Y_{k-1}) dx_{k-1} \tag{6}$$

where x_{k-1} is the previous state and $p(x_{k-1}|Y_{k-1})$ is the previous *state* posterior. This completes the typical recursive Bayesian filter with the "complete" or Markov assumption.

3. Kalman Filter

The second key assumption of the Kalman filter is that we assume that we do not have the past measurements, Y_{k-1}, when trying to determine our belief for x_k. This means that we need a form for our prior that allows us not to use past measurements. However, the previous value for the state, x_{k-1}, is known.

Now, we will include the main Kalman assumptions above, first that all noise is Gaussian and linearly additive. Therefore, we will use Gaussians for our density distributions,

$$p(y_k|x_k) = \frac{1}{\sqrt{2\pi\sigma_{y_k}^2}} \exp\left[-\frac{1}{2\sigma_{y_k}^2}\left(y_k - \langle y_k\rangle\right)^2\right] \tag{7}$$

and,

$$p(x_k|Y_{k-1}) = \frac{1}{\sqrt{2\pi\sigma_{x_k}^2}} \exp\left[-\frac{1}{2\sigma_{x_k}^2}\left(x_k - \langle x_k\rangle\right)^2\right] \tag{8}$$

where $\langle y_k\rangle$ and $\langle x_k\rangle$ are the means of y_k and x_k and σ_y^2 and σ_x^2 are the variances of each variable, respectively. Note, for illustration purposes, we limit ourselves to one variable. In later sections, we will include multiple variables. Thus, the posterior that we are looking for is,

$$p(x_k|y_k,Y_{k-1}) \propto \frac{1}{\sqrt{2\pi\sigma_{y_k}^2}} \exp\left[-\frac{1}{2\sigma_{y_k}^2}\left(y_k - \langle y_k\rangle\right)^2\right] \frac{1}{\sqrt{2\pi\sigma_{x_k}^2}} \exp\left[-\frac{1}{2\sigma_{x_k}^2}\left(x_k - \langle x_k\rangle\right)^2\right] . \tag{9}$$

where the constant of proportionality is $1/p(y_k|Y_{k-1})$.

The next question is deciding on the value of the means, since we are not inferring those, as can be seen from the posterior. For this, we have to look at the "forward" problems for each density function. For the prior, the forward problem is $x_k = F_{k,k-1}x_{k-1} + \eta_{k-1}$, where $F_{k,k-1}$ is called the "transition matrix" and for the likelihood, it is $y_k = G_k x_k + \varepsilon_k$, where G_k is called the "measurement matrix" function [10]. Each have Gaussian noise, η_{k-1} and ε_k, with zero means, respectively. Therefore, for the prior, we have,

$$\langle x_k\rangle = \int x_k \, p(x_k|Y_{k-1})dx_k \, , \tag{10}$$

where substituting yields,

$$\langle x_k\rangle = \int (F_{k,k-1}x_{k-1} + \eta_{k-1})p(x_k|Y_{k-1})dx_k \, , \tag{11}$$

and finally, we have,

$$\langle x_k\rangle = F_{k,k-1}x_{k-1} \, . \tag{12}$$

Here, it should be noted that a critical point for the construction of Kalman is that we wish to determine Equation (1). By definition, the prior, Equation (3), is conditional on previous measurements. However, this is replaced with a function that is purely based on the dynamics of the system, x, *i.e.*, there are no previous measurements explicitly in Equation (12).

For the likelihood we have,

$$\langle y_k\rangle = \int y_k \, p(y_k|x_k)dy_k \, , \tag{13}$$

where substituting yields,

$$\langle y_k \rangle = \int (G_k x_k + \varepsilon_k) p(y_k|x_k) dy_k \ . \tag{14}$$

and finally, we have,

$$\langle y_k \rangle = G_k x_k \tag{15}$$

Therefore, the posterior is now,

$$p(x_k|y_{k,}Y_{k-1}) \propto \frac{1}{\sqrt{2\pi\sigma_{y_k}^2}} \exp\left[-\frac{1}{2\sigma_{y_k}^2}\left(y_k - G_k x_k\right)^2\right] \frac{1}{\sqrt{2\pi\sigma_{x_k}^2}} \exp\left[-\frac{1}{2\sigma_{x_k}^2}\left(x_k - F_{k,k-1}x_{k-1}\right)^2\right] \tag{16}$$

Note, this is similar to a least squares (which itself is a special case of Bayes). There is one more obvious question to be answered: while this may be a solution for the density function in regards to x_k, how do we get x_{k-1}? We need a single number. The answer depends on the what is considered the "best guess" or point estimate for x_{k-1}. There are many choices, such as the mean, median or mode. However, since we are dealing with a symmetric solution, they are one in the same. Therefore, the easiest point estimate to get is the mode, $\hat{x}$, where,

$$\frac{\partial p}{\partial x}|_{x=\hat{x}} = 0, \tag{17}$$

for any x. Sometimes, this is called the maximum *a posteriori* (MAP) solution. Thus, the answer to the question noted above is that the point estimate that is used in Equation (12) is $\hat{x}_{k-1}$, which is the mode from the previous step.

3.1. A Simple Example

To show its processing workflow, we show a very simple example. We wish to know our $1D$ location given the known dynamical system and a measurement at each time step. First, we let $G_k = 1$. Then, we apply the dynamical equation,

$$x_k = x_{k-1} + v_0 \Delta t + \eta_{k-1} \ , \tag{18}$$

where v_0 is a known velocity constant and Δt is a known time step. It should noted, as well, that we can write this in terms of the noise,

$$\eta_{k-1} = x_k - \left(x_{k-1} + v_0 \Delta t\right) \ , \tag{19}$$

and thus, what we are modeling with our Gaussian is the *noise*. However, we eventually wish to know x_k, so we will write for our case,

$$\langle x_k \rangle = x_{k-1} + v_0 \Delta t \ , \tag{20}$$

since, as before, the mean of the noise is zero. Thus, our posterior is,

$$p(x_k|y_{k,}Y_{k-1}) \propto \frac{1}{\sqrt{2\pi\sigma_{y_k}^2}} \exp\left[-\frac{1}{2\sigma_{y_k}^2}\left(y_k - x_k\right)^2\right] \frac{1}{\sqrt{2\pi\sigma_{x_k}^2}} \exp\left[-\frac{1}{2\sigma_{x_k}^2}\left(x_k - x_{k-1} - v_0\Delta t\right)^2\right] \tag{21}$$

Determining the mode is trivial,

$$\hat{x}_k = \frac{\sigma_{y_k}^2 x_{k-1} + \sigma_y^2 v_0 \Delta t + \sigma_x^2 y_k}{\sigma_{y_k}^2 + \sigma_{x_k}^2} \tag{22}$$

and where the variance is,

$$\hat{\sigma}_k^2 = \frac{1}{\frac{1}{\sigma_{y_k}^2} + \frac{1}{\sigma_{x_k}^2}} \,. \tag{23}$$

It is presented in this form to show that the the new variance, $\hat{\sigma}_k^2$, is the inverse of the sum of the inverse individual variances, which is true in general for this type of problem; for example, if we had more than the bivariate case.

One last question that needs to be addressed is what is the value of x_{k-1}? The last assumption of the Kalman filter is that the MAP estimate of the *last* estimate is the best estimate for this value. This is a key assumption and not trivial, as it means that with each iterative step, information would be lost generally. This is not the case for the Kalman filter, since the Gaussian is assumed, and therefore, the mode and the variance uniquely identify the distribution.

These solutions can be manipulated and written in the other following form, as well,

$$\hat{x}_k = x_{k-1} + v_0 \Delta t + K_k (y_k - x_{k-1} - v_0 \Delta t) \tag{24}$$

where K_k is what is known as the Kalman "gain", which for this example is,

$$K_k = \frac{1}{\frac{1}{\sigma_{y_k}^2} + \frac{1}{\sigma_{x_k}^2}} = \frac{\sigma_{y_k}^2 \sigma_{x_k}^2}{\sigma_{x_k}^2 + \sigma_{y_k}^2} \,. \tag{25}$$

This is an especially important pedagogical result, as it shows that Kalman "gain" Equation (25) is simply the new, *inferred* variance Equation (23).

4. Maximum Relative Entropy

First, we present a review of maximum relative entropy. For a more detailed discussion and several examples, please see [6,11]. Our first concern when using the MrE method to update from a prior to a posterior distribution is to define the space in which the search for the posterior will be conducted. We wish to infer something about the values of one or several quantities, $\theta \in \Theta$, on the basis of three pieces of information: prior information about θ (the prior), the known relationship between x *and* θ (the model) and the observed values of the data $x \in \mathcal{X}$. Since we are concerned with both x *and* θ, the relevant space is neither $\mathcal{X}$ nor Θ, but the product $\mathcal{X} \times \Theta$, and our attention must be focused on the joint distribution, $P(x, \theta)$. The selected joint posterior, $P_{\text{new}}(x, \theta)$, is that which maximizes the entropy (in MrE terminology, we "maximize" the negative relative entropy, S, so that $S \leq 0$. This is the same as minimizing the relative entropy),

$$S[P, P_{\text{old}}] = -\int P(x, \theta) \log \frac{P(x, \theta)}{P_{\text{old}}(x, \theta)} dx d\theta \,, \tag{26}$$

subject to the appropriate constraints. $P_{\text{old}}(x, \theta)$ contains our prior information, which we call the *joint prior* . To be explicit,

$$P_{\text{old}}(x, \theta) = P_{\text{old}}(\theta) P_{\text{old}}(x|\theta) \,, \tag{27}$$

where $P_{\text{old}}(\theta)$ is the traditional Bayesian prior and $P_{\text{old}}(x|\theta)$ is the likelihood. It is important to note that they *both* contain prior information. The Bayesian prior is defined as containing prior information.

However, the likelihood is not traditionally thought of in terms of prior information. Of course, it is reasonable to see it as such, because the likelihood represents the model (the relationship between θ and x) that has already been established. Thus, we consider both pieces, the Bayesian prior and the likelihood to be *prior* information.

The new information is the *observed data*, x', which in the MrE framework must be expressed in the form of a constraint on the allowed posteriors. The family of posteriors that reflects the fact that x is now known to be x' is such that,

$$P(x) = \int P(x,\theta)\, d\theta = \delta(x - x') \ , \tag{28}$$

where $\delta(x - x')$ is the Dirac delta function. This amounts to an *infinite* number of constraints: there is one constraint on $P(x,\theta)$ for each value of the variable, x, and each constraint will require its own Lagrange multiplier, $\lambda(x)$. Furthermore, we impose the usual normalization constraint,

$$\int P(x,\theta)\, dxd\theta = 1 \ , \tag{29}$$

and include additional information about θ in the form of a constraint on the expected value of some function $f(\theta)$,

$$\int P(x,\theta)\, f(\theta)\, dxd\theta = \langle f(\theta)\rangle = F \ . \tag{30}$$

Note that an additional constraint in the form of $\int P(x,\theta)g(x)dxd\theta = \langle g\rangle = G$ could only be used when it does not contradict data constraint Equation (28). Therefore, it is redundant, and the constraint would simply get absorbed when solving for $\lambda(x)$. We also emphasize that constraints imposed at the level of the prior need not be satisfied by the posterior. What we do here differs from the standard Bayesian practice in that we *require* the constraint to be satisfied by the posterior distribution.

We proceed by maximizing Equation (26) subject to the above constraints. The purpose of maximizing the entropy is to determine the value for P when $S = 0$, meaning that we want the value of P that is closest to P_{old} given the constraints. The calculus of variations is used to do this by varying $P \to \delta P$, *i.e.*, setting the derivative with respect to P equal to zero. The Lagrange multipliers, α, β and $\lambda(x)$ are used so that the P that is chosen satisfies the constraint equations. The actual values are determined by the value of the constraints themselves. We now provide the detailed steps in this maximization process.

First we setup the variational form with the Lagrange multipliers,

$$\delta P(x,\theta)\left\{\begin{array}{c} S[P,P_{\text{old}}] + \alpha\left[\int P(x,\theta)\, dxd\theta - 1\right] \\ +\beta\left[\int P(x,\theta)\, f(\theta)\, dxd\theta - F\right] \\ +\int \lambda(x)\left[\int P(x,\theta)\, dxd\theta - \delta(x - x')\right] \end{array}\right\} = 0 \ . \tag{31}$$

We expand the entropy function Equation (26),

$$\delta P(x,\theta)\left\{\begin{array}{c} -\int P(x,\theta)\log P(x,\theta)\, dxd\theta \\ +\int P(x,\theta)\log P_{\text{old}}(x,\theta)\, dxd\theta \\ +\alpha\left[\int P(x,\theta)\, dxd\theta - 1\right] \\ +\beta\left[\int P(x,\theta)\, f(\theta)\, dxd\theta - F\right] \\ +\int \lambda(x)\left[\int P(x,\theta)\, dxd\theta - \delta(x - x')\right] \end{array}\right\} = 0 \ . \tag{32}$$

Next, vary the functions with respect to $P(x,\theta)$,

$$\left\{\begin{array}{c} -\int \delta P(x,\theta)\log P(x,\theta)\,dxd\theta - \int P(x,\theta)\,\frac{1}{P(x,\theta)}\delta P(x,\theta)\,dxd\theta \\ +\int \delta P(x,\theta)\log P_{\text{old}}(x,\theta)\,dxd\theta + 0 \\ +\alpha\left[\int \delta P(x,\theta)\,dxd\theta\right] \\ +\beta\left[\int \delta P(x,\theta)\,f(\theta)\,dxd\theta\right] \\ +\int \lambda(x)\left[\int \delta P(x,\theta)\,dxd\theta\right] \end{array}\right\} = 0\,, \tag{33}$$

which can be rewritten as

$$\int \{-\log P(x,\theta) - 1 + \log P_{\text{old}}(x,\theta) + \alpha + \beta f(\theta) + \lambda(x)\}\,\delta P(x,\theta)\,dxd\theta = 0\,.$$

The terms inside the brackets must sum to zero, therefore we can write,

$$\log P(x,\theta) = \log P_{\text{old}}(x,\theta) - 1 + \alpha + \beta f(\theta) + \lambda(x) \tag{34}$$

or

$$P_{\text{new}}(x,\theta) = P_{\text{old}}(x,\theta)\,e^{(-1+\alpha+\beta f(\theta)+\lambda(x))} \tag{35}$$

In order to determine the Lagrange multipliers, we substitute our solution Equation (35) into the various constraint equations. The constant α is eliminated by substituting Equation (35) into Equation (29),

$$\int P_{\text{old}}(x,\theta)\,e^{(-1+\alpha+\beta f(\theta)+\lambda(x))}dxd\theta = 1\,. \tag{36}$$

Dividing both sides by the constant $e^{(-1+\alpha)}$,

$$\int P_{\text{old}}(x,\theta)\,e^{\beta f(\theta)+\lambda(x)}dxd\theta = e^{(1-\alpha)}\,. \tag{37}$$

Then substituting back into Equation (35) yields

$$P_{\text{new}}(x,\theta) = P_{\text{old}}(x,\theta)\,\frac{e^{\lambda(x)+\beta f(\theta)}}{Z}\,, \tag{38}$$

where

$$Z = e^{1-\alpha} = \int e^{\beta f(\theta)+\lambda(x)}P_{\text{old}}(x,\theta)\,dxd\theta\,. \tag{39}$$

In the same fashion, the Lagrange multipliers $\lambda(x)$ are determined by substituting Equation (38) into Equation (28)

$$\int P_{\text{old}}(x,\theta)\,\frac{e^{\lambda(x)+\beta f(\theta)}}{Z}d\theta = \delta(x-x') \tag{40}$$

or

$$e^{\lambda(x)} = \frac{Z}{\int e^{\beta f(\theta)}P_{\text{old}}(x,\theta)\,d\theta}\delta(x-x')\,. \tag{41}$$

The posterior now becomes

$$P_{\text{new}}(x,\theta) = P_{\text{old}}(x,\theta)\,\delta(x-x')\frac{e^{\beta f(\theta)}}{\zeta(x,\beta)}\,, \tag{42}$$

where $\zeta(x,\beta) = \int e^{\beta f(\theta)} P_{\text{old}}(x,\theta)\, d\theta$.

The Lagrange multiplier β is determined by first substituting Equation (42) into Equation (30),

$$\int \left[P_{\text{old}}(x,\theta)\, \delta(x - x') \frac{e^{\beta f(\theta)}}{\zeta(x,\beta)} \right] x f(\theta)\, dx d\theta = F \,. \tag{43}$$

Integrating over x yields,

$$\frac{\int e^{\beta f(\theta)} P_{\text{old}}(x',\theta) f(\theta)\, d\theta}{\zeta(x',\beta)} = F \,, \tag{44}$$

where $\zeta(x,\beta) \to \zeta(x',\beta) = \int e^{\beta f(\theta)} P_{\text{old}}(x',\theta) d\theta$. Now β can be determined rewriting Equation (44) as

$$\frac{\partial \ln \zeta(x',\beta)}{\partial \beta} = F \,. \tag{45}$$

The final step is to marginalize the posterior, $P_{\text{new}}(x,\theta)$ over x to get our updated probability,

$$P_{\text{new}}(\theta) = P_{\text{old}}(x',\theta) \frac{e^{\beta f(\theta)}}{\zeta(x',\beta)} \tag{46}$$

Additionally, this result can be rewritten using the product rule ($P(x,\theta) = P(x)\, P(\theta|x)$) as

$$P_{\text{new}}(\theta) = P_{\text{old}}(\theta)\, P_{\text{old}}(x'|\theta) \frac{e^{\beta f(\theta)}}{\zeta'(x',\beta)} \,, \tag{47}$$

where $\zeta'(x',\beta) = \int e^{\beta f(\theta)} P_{\text{old}}(\theta)\, P_{\text{old}}(x'|\theta) d\theta$. The right side resembles Bayes theorem, where the term $P_{\text{old}}(x'|\theta)$ is the standard Bayesian likelihood and $P_{\text{old}}(\theta)$ is the prior. The exponential term is a *modification* to these two terms. In an effort to put some names to these pieces we will call the standard Bayesian likelihood the *likelihood* and the exponential part the *likelihood modifier* so that the product of the two gives the *modified likelihood*. The denominator is the normalization or *marginal modified likelihood*. Notice when $\beta = 0$ (no moment constraint) we recover Bayes' rule. For $\beta \neq 0$ Bayes' rule is modified by a "canonical" exponential factor.

5. Maximum Relative Entropy and Kalman

There are works where entropy maximization is being used in Kalman filtering [12–14]. For example, [14] uses entropy maximization as one of the approximation tools to reduce uncertainty. Here, we show that if the same assumptions are taken into account, the *explicit closed form solution is derived*. So, a numerical comparison, as in [14], becomes unnecessary and even limited. To the best of our knowledge, there is no work that shows a direct link between the original Kalman filter and maximization of the relative entropy and produces the closed form solution. We will now present a more complicated example illustrating the maximum relative entropy (MrE) solution and discuss the assumptions and constraints that lead to the same closed form Kalman filter solution.

This example consists of analyzing a linear system composed of two equations that represent linear motion with constant acceleration $c_{a,k}$. The dynamics of the velocity, v_k, and the position, x_k, are encoded in the following two relationships, which we assume are relevant for predicting the linear motion of the next state,

$$x_k = x_{k-1} + \triangle t\, v_{k-1} + \frac{c_{a,k}\, \triangle t^2}{2} \,, \tag{48}$$

$$v_k = v_{k-1} + \triangle t \; c_{a,k} \; , \tag{49}$$

where $\triangle t$ is the discretization interval between the subsequent measurements, and the index, k, represents the temporal discretization interval.

Here, we will derive the so-called "prediction step", which will be the posterior or the following optimization criterion or entropy, which has the form,

$$S\left(\bar{P}_k, P_{prior,k}\right) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \bar{P}_k\left(x_k, v_k\right) \ln \frac{\bar{P}_k\left(x_k, v_k\right)}{P_{prior,k}\left(x_k, v_k\right)} dx_k dv_k \; , \tag{50}$$

where x_k is the position, v_k is the velocity, $P_{prior,k}\left(x_k, v_k\right)$ is the prior probability distribution function (which is sometimes a uniform distribution for the first sample) and $\bar{P}_k\left(x_k, v_k\right)$ is the posterior distribution to be found as a result of the first Kalman filter step, which is also called a "prediction" step [9]. In fact, it is the object that we are deriving, $\bar{P}_k\left(x_k, v_k\right)$, that will be the *prior* for our Bayesian filter and in the special case, the Kalman filter, such as Equation (8).

All constraints come from the same Kalman filter assumptions. We derive the first constraint using Equation (48), which is the variance,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(x_k - \langle x_k \rangle\right)^2 \bar{P}_k\left(x_k, v_k\right) dx_k dv_k = 0 \; , \tag{51}$$

where,

$$x_k = x_{k-1} + \triangle t \; v_{k-1} + \frac{c_{a,k} \; \triangle t^2}{2} . \tag{52}$$

Notice that there is no noise term in Equation (52). This is why the variance is zero. There is no noise, and thus, the variables of interest could be determined explicitly from the model. However, obviously, all variables (and constants, if any) have some noise (η) variables, which in the Kalman filter are assumed to be additive and linear. Then, Equation (51) can be rewritten to,

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\Psi_k\right)^2 \bar{P}_k\left(x_k, v_k, \eta_{x,k-1}, \eta_{v,k-1}, \eta_{a,k-1}\right) \Omega_k = 0 \; , \tag{53}$$

where $\Omega_k = \mathrm{d}x_k \mathrm{d}v_k \mathrm{d}\eta_{x,k-1} \mathrm{d}\eta_{v,k-1} \mathrm{d}\eta_{a,k-1}$, and where,

$$\Psi_k = x_k - \left[\left(\hat{x}_{k-1} + \eta_{x,k-1}\right) + \triangle t \; \left(\hat{v}_{k-1} + \eta_{v,k-1}\right) + \frac{\left(c_{a,k} + \eta_{a,k-1}\right) \; \triangle t^2}{2}\right] \; , \tag{54}$$

where Ψ_k is the noise term of the model. Note that this is effectively following the note mentioned in conjunction with Equation (19) and where $\hat{x}_{k-1}$ and $\hat{v}_{k-1}$ are estimates of our variables from the previous discretization interval and $\eta_{x,k-1}, \eta_{v,k-1}, \eta_{a,k-1}$ are multivariate normal distribution additive noise variables, which have means of zero. Frequently, the joint prior distribution of noise variables is defined by four main assumptions:

(1) The means of all noise variables are zero;

(2) The joint distribution function is a multivariate normal distribution;

(3) The covariance matrix is not only valid for the previous posterior distribution discretization interval, but also for the *current* posterior distribution discretization interval, which in our case is

$\bar{P}_k\left(x_k, v_k, \eta_{x,k-1}, \eta_{v,k-1}, \eta_{a,k-1}\right)$. In other words, it is implied that in our specific case, we have the following equalities,

$$\begin{aligned}&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\eta_{x,k-1}^2\bar{P}_k\left(x_k, v_k, \eta_{x,k-1}, \eta_{v,k-1}, \eta_{a,k-1}\right)\Omega_k = \\ &\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\eta_{x,k-1}^2 P_{k-1}\left(x_{k-1}, v_{k-1}, \eta_{x,k-1}, \eta_{v,k-1}, \eta_{a,k-1}\right)\Omega_{k-1} = \sigma_{x,k-1}^2,\end{aligned} \tag{55}$$

where $\Omega_{k-1} = \mathrm{d}x_{k-1}\mathrm{d}v_{k-1}\mathrm{d}\eta_{x,k-1}\mathrm{d}\eta_{v,k-1}\mathrm{d}\eta_{a,k-1}$,

$$\begin{aligned}&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\eta_{v,k-1}^2\bar{P}_k\left(x_k, v_k, \eta_{x,k-1}, \eta_{v,k-1}, \eta_{a,k-1}\right)\Omega_k = \\ &\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\eta_{v,k-1}^2 P_{k-1}\left(x_{k-1}, v_{k-1}, \eta_{x,k-1}, \eta_{v,k-1}, \eta_{a,k-1}\right)\Omega_{k-1} = \sigma_{v,k-1}^2,\end{aligned} \tag{56}$$

$$\begin{aligned}&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\eta_{x,k-1}\eta_{v,k-1}\bar{P}_k\left(x_k, v_k, \eta_{x,k-1}, \eta_{v,k-1}, \eta_{a,k-1}\right)\Omega_k = \\ &\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\eta_{x,k-1}\eta_{v,k-1}P_{k-1}\left(x_{k-1}, v_{k-1}, \eta_{x,k-1}, \eta_{v,k-1}, \eta_{a,k-1}\right)\Omega_{k-1} = cov_{x,v,k-1},\end{aligned} \tag{57}$$

$$\begin{aligned}&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\eta_{a,k-1}^2\bar{P}_k\left(x_k, v_k, \eta_{x,k-1}, \eta_{v,k-1}, \eta_{a,k-1}\right)\Omega_k = \\ &\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\eta_{a,k-1}^2 P_{k-1}\left(x_{k-1}, v_{k-1}, \eta_{x,k-1}, \eta_{v,k-1}, \eta_{a,k-1}\right)\Omega_{k-1} = \sigma_{a,k-1}^2,\end{aligned} \tag{58}$$

where the variances and covariances are usually taken from the inference result of the previous discretization interval or are set with initial guesses.

(4) The last, but not the least, assumption in Kalman Filtering is that our noise variables are independent from our main state variables, *i.e.*, x_k and v_k. The main benefit of this assumption is that we can manipulate Equation (53) by applying the constraints in Equations (55)–(58). Keep in mind that the means of noise variables are zeros, so many additive terms will zero out after integrations. Therefore, we finally get Equation (53) in the form of,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x_k - \langle x_k\rangle\right)^2\bar{P}_k\left(x_k, v_k\right)dx_k dv_k \tag{59}$$

$$= \sigma_{x,k-1}^2 + 2cov_{x,v,k-1}\triangle t + \sigma_{v,k-1}^2\triangle t^2 + \frac{\sigma_{a,k-1}^2\,\triangle t^4}{4}, \tag{60}$$

where,

$$\langle x_k\rangle = \hat{x}_{k-1} + \triangle t\;\hat{v}_{k-1} + \frac{c_{a,k}\;\triangle t^2}{2}. \tag{61}$$

Similarly, we can construct two other MrE constraints based on Kalman filter assumptions as,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(v_k - \langle v_k\rangle\right)^2\bar{P}_k\left(x_k, v_k\right)dx_k dv_k = \sigma_{v,k-1}^2 + \sigma_{a,k-1}^2\;\triangle t^2\;, \tag{62}$$

where,

$$\langle v_k\rangle = \hat{v}_{k-1} + \triangle t\;c_{a,k}\;, \tag{63}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x_k-\langle x_k\rangle\right)\left(v_k-\langle v_k\rangle\right)\bar{P}_k\left(x_k,v_k\right)dx_kdv_k \tag{64}$$
$$= \; cov_{x,v,k-1}+\sigma_{v,k-1}^2\triangle t+\frac{\sigma_{a,k-1}^2\cdot\triangle t^3}{2}.$$

With these equations, along with a normalization constraint, like Equation (29), we get the posterior distribution function of the prediction step of the Kalman filter using MrE. For illustration purposes, we assume that $P_{prior,k}\left(x_k,v_k\right)$ is a uniform distribution. In other words, the distribution that maximizes Equation (50) with constraint Equations (59), (62) and (64) is the prediction step function below. This solution yields the same answer as the prediction step as in the Kalman filter, where $\bar{P}_k\left(x_k,v_k\right)$ is an unknown posterior distribution function being constrained. This distribution is simply a multivariate normal distribution with a covariance matrix containing elements defined by scalar values on the right side of Equations (59), (62) and (64), and the means defined by our mathematical model, which are,

$$x_k=\hat{x}_{k-1}+\triangle t\;\hat{v}_{k-1}+\frac{c_{a,k}\;\triangle t^2}{2}\;, \tag{65}$$

$$v_k=\hat{v}_{k-1}+\triangle t\;c_{a,k}\;. \tag{66}$$

For simplicity's sake, we will introduce those scalar coefficients as,

$$\bar{\sigma}_{x,k}^2=\sigma_{x,k-1}^2+2cov_{x,v,k-1}\triangle t+\sigma_{v,k-1}^2\triangle t^2+\frac{\sigma_{a,k-1}^2\;\triangle t^4}{4}, \tag{67}$$

$$\bar{\sigma}_{v,k}^2=\sigma_{v,k-1}^2+\sigma_{a,k-1}^2\;\triangle t^2, \tag{68}$$

$$\overline{cov}_{x,v,k}=cov_{x,v,k-1}+\sigma_{v,k-1}^2\triangle t+\frac{\sigma_{a,k-1}^2\;\triangle t^3}{2}\;. \tag{69}$$

Then, the posterior multivariate normal distribution of this inference step (the prediction step of the Kalman filter) is,

$$\bar{P}_k\left(x_k,v_k\right)=\left(2\pi\sqrt{\bar{\sigma}_{x,k}^2\bar{\sigma}_{v,k}^2-\overline{cov}_{x,v,k}^2}\right)^{-1}\exp\left(\begin{array}{c}\frac{1}{2\left(\overline{cov}_{x,v,k}^2-\bar{\sigma}_{x,k}^2\bar{\sigma}_{v,k}^2\right)}\left(\bar{\sigma}_{v,k}^2\left(x_k-\langle x_k\rangle\right)^2+\bar{\sigma}_{x,k}^2\left(v_k-\langle v_k\rangle\right)\right)\\ +2\overline{cov}_{x,v,k}\left(x_k-\langle x_k\rangle\right)\left(v_k-\langle v_k\rangle\right)\end{array}\right). \tag{70}$$

To be clear, this "posterior" Equation (70) is effectively the "prior" Equation (8) that would be in the Kalman filter.

5.1. Kalman Filter's Updating Step

Our focus here is the traditional updating step of the Kalman filter and its reproduction by MrE. The measurement distribution needed would be obtained in a similar manner as in the predictive step using the following constraints,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(y_k-\langle y_k\rangle\right)^2P_{\text{likelihood},k}\left(y_k,v_k\right)dy_k\;dv_k=\sigma_{y,k}^2, \tag{71}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}y_kP_{\text{likelihood},k}\left(y_k,v_k\right)dy_kdv_k=\langle y_k\rangle\,, \tag{72}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P_{\text{likelihood},k}\left(y_k, v_k\right) dy_k dv_k = \delta\left(y_k - y'_k\right) \ , \tag{73}$$

with a normalization constraint like Equation (29) and where,

$$\langle y_k \rangle = G_k x_k \ , \tag{74}$$

for this example, where y'_k is the observation value that is analogous to Equation (28). Using these pieces, we get,

$$P_{\text{likelihood},k}(x_k, v_k) = \frac{1}{\sqrt{2\pi\sigma_{y,k}^2}} \exp\left[-\frac{1}{2\sigma_{y,k}^2}\left(y'_k - G_k x_k\right)^2\right] \ . \tag{75}$$

To avoid confusion, we make this note: this indeed is a function of x_k and v_k; the reason being that following the previous section, the y_k in $P_{\text{likelihood},k}\left(y_k, v_k\right)$ was replaced with the observed value, y'_k through the data constraint (Dirac delta function) and the mean of y_k is a function of x_k. Therefore, the joint posterior that is analogous to Equation (16) would be,

$$P_k\left(x_k, v_k\right) \propto \bar{P}_k\left(x_k, v_k\right) \ P_{\text{likelihood},k}(x_k, v_k) \ . \tag{76}$$

To get the estimates of state variables for discretization interval k, we select the modes as in Equation (17),

$$\frac{\partial P_k\left(x_k, v_k\right)}{\partial x_k}|_{x_k=\hat{x}_k} = 0, \text{ and } \frac{\partial P_k\left(x_k, v_k\right)}{\partial v_k}|_{v_k=\hat{v}_k} = 0. \tag{77}$$

This yields the final estimates for the discretization interval, k, as follows,

$$\hat{x}_k = y'_k + \frac{2\sigma_{y,k}^2\left(2\hat{x}_{k-1} - 2y'_k + \triangle t\left(2\hat{v}_{k-1} + \triangle t\ c_{a,k}\right)\right)}{4\left(\sigma_{y,k}^2 + \bar{\sigma}_{x,k}^2\right)}, \tag{78}$$

$$\hat{v}_k = \hat{v}_{k-1} + \triangle t \cdot c_{a,k} + \frac{\left(y'_k - \hat{x}_{k-1} - \frac{\triangle t}{2}\left(2\hat{v}_{k-1} + \triangle t\ c_{a,k}\right)\right)\overline{cov}_{x,v,k}}{4\left(\sigma_{y,k}^2 + \bar{\sigma}_{x,k}^2\right)}. \tag{79}$$

The inferred estimates of the covariance and variance elements of the covariance matrix are as follows,

$$\sigma_{x,k}^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x_k - \hat{x}_k\right)^2 P_k\left(x_k, v_k\right) dx_k dv_k = \sigma_{y,k}^2 - \frac{\sigma_{y,k}^4}{\sigma_{y,k}^2 + \bar{\sigma}_{x,k}^2}, \tag{80}$$

$$\begin{aligned} cov_{x,v,k} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x_k - \hat{x}_k\right)\left(v_k - \hat{v}_k\right) P_k\left(x_k, v_k\right) dx_k dv_k \\ &= \left(2\left(\sigma_{y,k}^2 + \bar{\sigma}_{x,k}^2\right)\right)^{-1}\sigma_{y,k}^2\left(2cov_{x,v,k-1} + 2\sigma_{v,k-1}^2\triangle t + \sigma_{a,k-1}^2\triangle t^3\right), \end{aligned} \tag{81}$$

$$\begin{aligned} \sigma_{v,k}^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(v_k - \hat{v}_k\right)^2 P_k\left(x_k, v_k\right) dx_k dv_k \\ &= \left(4\left(\sigma_{y,k}^2 + \bar{\sigma}_{x,k}^2\right)\right)^{-1}\left(\begin{array}{c}\triangle t^2\sigma_{a,k-1}^2\left(4\sigma_{x,k-1}^2 + 4cov_{x,v,k-1}\triangle t + \sigma_{v,k-1}^2\triangle t^2 + 4\sigma_{y,k}^2\right) + \\ 4\left(\sigma_{v,k-1}^2\left(\sigma_{x,k-1}^2 + \sigma_{y,k}^2\right) - cov_{x,v,k-1}^2\right)\end{array}\right). \end{aligned} \tag{82}$$

These covariance values and estimates are then used in the next discretization interval step by repeating the same procedure from the beginning, *i.e.*, starting with Equation (50), we have the next discretization interval's Kalman filter's first step as,

$$S\left(\bar{P}_{k+1}, P_{\text{prior},k+1}\right) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \bar{P}_{k+1}\left(x_{k+1}, v_{k+1}\right) \ln \frac{\bar{P}_{k+1}\left(x_{k+1}, v_{k+1}\right)}{P_{\text{prior},k+1}\left(x_{k+1}, v_{k+1}\right)} dx_{k+1} dv_{k+1} \ , \tag{83}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(x_{k+1} - \langle x_{k+1}\rangle\right)^2 \bar{P}_{k+1}\left(x_{k+1}, v_{k+1}\right) dx_{k+1} dv_{k+1} \tag{84}$$

$$= \ \sigma_{x,k}^2 + 2cov_{x,v,k}\triangle t + \sigma_{v,k}^2 \triangle t^2 + \frac{\sigma_{a,k}^2 \cdot \triangle t^4}{4},$$

where,

$$\langle x_{k+1}\rangle = \hat{x}_k + \triangle t \ \hat{v}_k + \frac{c_{a,k} \ \triangle t^2}{2} \tag{85}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(v_{k+1} - \langle v_{k+1}\rangle\right)^2 \bar{P}_{k+1}\left(x_{k+1}, v_{k+1}\right) dx_{k+1} dv_{k+1} = \sigma_{v,k}^2 + \sigma_{a,k}^2 \ \triangle t^2 \ , \tag{86}$$

where,

$$\langle v_{k+1}\rangle = \hat{v}_k + \triangle t \cdot c_{a,k}. \tag{87}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(x_{k+1} - \langle x_{k+1}\rangle\right)\left(v_{k+1} - \langle v_{k+1}\rangle\right) \bar{P}_{k+1}\left(x_{k+1}, v_{k+1}\right) dx_{k+1} dv_{k+1} \tag{88}$$

$$= \ cov_{x,v,k} + \sigma_{v,k}^2 \triangle t + \frac{\sigma_{a,k}^2 \cdot \triangle t^3}{2},$$

where we assumed that $\sigma_{a,k}^2 = \sigma_{a,k-1}^2$ and $c_{a,k} = c_{a,k-1}$, *i.e.*, acceleration is constant in this problem. The iterative nature of Kalman filter comes into effect by assuming that $P_{\text{prior},k+1}\left(x_{k+1}, v_{k+1}\right) = P_{\text{prior},k+1}\left(x_k, v_k\right) = P_k\left(x_k, v_k\right)$. In other words, the previous discretization interval's posterior function is the prior for the current discretization interval.

5.2. Kalman Filter Revisited

We will now present the solution of the Kalman filter that is the same closed form solution as in the previous subsection. First, we need to construct our problem in matrix form.

The mathematical model or our state space system, as in Equation (48) and Equation (49), has the following matrix representation:

$$\bar{\mathbf{x}}_k = F_k \bar{\mathbf{x}}_{k-1} + h_k, \tag{89}$$

where $\bar{\mathbf{x}}_k$ is a vector that has coordinates representing our position and velocity variables. The transition matrix and additive term matrices are,

$$F_k = \begin{bmatrix} 1 & \triangle t \\ 0 & 1 \end{bmatrix}, \tag{90}$$

$$h_k = \begin{bmatrix} \frac{1}{2} c_{a,k-1} \triangle t^2 \\ c_{a,k-1} \triangle t \end{bmatrix}. \tag{91}$$

The covariance matrix of our state space variables is,

$$\Phi_{k-1} = \begin{bmatrix} \sigma^2_{x,k-1} & cov_{x,v,k-1} \\ cov_{x,v,k-1} & \sigma^2_{v,k-1} \end{bmatrix}. \tag{92}$$

Then, the covariance matrix of the so-called noise, based on our mathematical model, is,

$$Q_k = \begin{bmatrix} \frac{1}{4}\sigma^2_{a,k} \triangle t^4 & \frac{1}{2}\sigma^2_{a,k} \triangle t^3 \\ \frac{1}{2}\sigma^2_{a,k} \triangle t^3 & \sigma^2_{a,k} \triangle t^2 \end{bmatrix}. \tag{93}$$

The prediction step of the Kalman filter for calculating the covariance matrix is,

$$\bar{\Phi}_k = F_k \Phi_{k-1} F_k^T + Q_k, \tag{94}$$

and this equation produces the same (the exact closed form solution expressed by elementary algebraic functions) covariance values between variables as in Equations (67), (68) and (69). Therefore, there is a one-to-one match between the prediction step of the Kalman filter and the MrE solution. If we do not measure velocity, but just position (as in the simple example above), then our observation model, G_k, its covariance matrix, R_k, and the measurement vector matrix, y_k, are,

$$G_k = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ and } R_k = \left[\sigma^2_{y,k}\right], \text{ and } y_k = \left[y'_k\right]. \tag{95}$$

The updating step of the Kalman filter requires the calculation of the Kalman gain,

$$K_k = \bar{\Phi}_k G_k^T \left(G_k \bar{\Phi}_k G_k^T + R_k\right)^{-1}, \tag{96}$$

and the covariance matrix of the Kalman filter is,

$$\Phi_k = \left(I - K_k G_k\right) \bar{\Phi}_k \left(I - K_k G_k\right)^T + K_k R_k K_k^T, \tag{97}$$

where I is a unit matrix. Matrix Equation (97) has elements that are exactly the same as the closed form solutions in Equations (80), (81) and (82). The expectation values of the Kalman filter are,

$$\mathbf{x}_k = \bar{\mathbf{x}}_k + K_k \left(y_k - G_k \bar{\mathbf{x}}_k\right), \tag{98}$$

which produces exactly the same closed form solution as Equation (78) and Equation (79). Thus, the Kalman filter produces the same solution MrE. In other words, if our assumptions are the same, both approaches return the same answer.

Sometimes, there are discussions [15] about the numerical stability of the Kalman filter and a different, and simplified version of covariance matrix Equation (97) is used [9],

$$\Phi_k = \left(I - K_k G_k\right) \bar{\Phi}_k. \tag{99}$$

However, this equation sometimes produces numerically unstable results [16], but the context is important. On page 151 of [16] it states: "However, Joseph's stabilized version has more computations

than the form given by Equation 3.44. Hence, a filter designer must trade off computational workload *versus* potential round off errors". We think that there is some confusion here. If mathematical reduction is done in the estimates' expressions (after those are obtained by applying the matrix operations as defined in the original Kalman filter), then by definition, there are no rounding errors in the remaining irreducible expressions (which are already reduced to the lowest terms), which might result in a filter's instability. By reduction to the expression's lowest terms, we also solve the computational workload issue and avoid the round off error problem. In other words, we must do our homework offline to avoid computational workload and potential round off errors when solving practical engineering problems online. Joseph's stabilized version or the original matrix form discussion is an artifact of matrix manipulation, but is not an artifact of the filter itself.

Summarizing, if our state space system and its corresponding transition matrix is of such a size and/or sparsity that we can get its inverse matrix analytically in its reduced solution without numerical iterations, then selecting which version (Joseph's stabilized or original) does not matter, because the closed form and the numeric answer would be the same. From a practical point of view, the maximum relative entropy method might be particularly useful for loosely coupled systems (not necessarily small), because its complexity (the total number of Lagrange multipliers) is equal to the total number of transition equations, and it has no explicit difficulty in calculating inverse matrices, because of the variational techniques used.

The fact is that both Equation (97) and Equation 99 return exactly the same closed form solution as MrE does in Equations (80)–(82), and by definition, these expressions are numerically stable, always. However, if one applies Kalman filter matrix operations and does not continue with further reductions of the final estimates' expressions into their irreducible forms, then, yes, expression Equation (97) might be better to use compared to Equation (99). This shows one more benefit of MrE: the closed form solutions allow one to avoid numeric instability in certain situations.

6. Nonlinear Filter

The original Kalman filter has an assumption that the relationships between the state space system's variables are linear. This assumption allows it to be expressed in a matrix form. Therefore, by definition, the Kalman filter is a linear filter, and nonlinear relationships have no explicit representations in transition matrix Equation (90). For that reason, variants of the Kalman filter were invented. One is called the extended Kalman filter, where any function is approximated locally by calculating a Jacobian at the approximate estimated location. Another variant is the unscented Kalman filter, which locally restricts the data sampling to a set of $2n+1$ sigma points, where n is the dimension of the state space. It allows the avoidance of calculating the Jacobian and has the benefit that the nonlinear transition can be locally approximated by a cubic characteristic, if we looked from a single variable's point of view. In the next section, we will derive the Kalman filter expression by a probabilistic approach by applying a one-to-one transition between the state variables, where the transition can be monotonic increasing or decreasing and not necessarily linear, as in original Kalman filter assumptions. More complex and generalized formulae might be found in [17].

6.1. Generalized Univariate Nonlinear Filter for Monotonic Transitions

In this section, we construct a general transformation of variables. While this can be found in undergraduate texts and advanced literature on Kalman filtering [18,19], we feel the need to include it, as it allows us to further point directly to why Kalman assumptions are necessary. Assume we have a single random variable, X(thus, a univariate approach). A random variable is a function whose value is subject to variations, due to some randomness. A value of this function (which we will call a random variable from now on) is associated with some probability (discrete case) or with probability density (continuous case). In this section, we deal with continuous real-valued data values only, but the approach itself is not restricted to them only.

The cumulative distribution function (cdf) of X is the function given by,

$$F_X(x) = P(X \leq x). \tag{100}$$

If probability density function (pdf) f of a random variable, X, is known, then the cdf is given by,

$$F_X(x) = \int_{-\infty}^{\infty} f_X(\acute{x})\, d\acute{x} \text{ and } f_X(x) = \frac{dFx}{dx}, \tag{101}$$

where $\acute{x}$ and x are values of the measurable space.

Assume we are measuring the traveled distance by a robot in meters (random variable X) and we can learn how many kilometers the robot has traveled (random variable Y); then, there is a physical relationship in the measurable space, because we know the ratio between kilometers and meters (A = 1000; note that this parameter is known, *i.e.*, 1000, is a constant), as,

$$y = A \cdot x. \tag{102}$$

One of the very first Kalman filter assumptions is that the relationships in Equation (102) have the same configuration in the random variables space, too, *i.e.*, by definition, we can write this as,

$$Y \equiv A \cdot X. \tag{103}$$

Or more generally, there is an invertible function, $g(X)$ (whose inverse function is $X = g^{-1}(Y)$), with which we can transform the random variable, X, to Y. In other words, the variable, Y, has complete probabilistic information for this transformation, and we can therefore get all the probabilistic characteristics of X. We could assume some nonlinear assumptions here, where there are one-to-many links between variables [17], but for simplicity's sake, we avoid this here.

Current one-to-one assumptions are still more general than the original Kalman filter assumptions, because we allow not only the linear equation system of variables, but also the system of any continuously increasing or decreasing functions. Then, the definition or the meaning of Equation (103) can be represented by the following expressions:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P\left(g_{\text{increasing}}(X) \leq y\right) = \\ &= P\left(X \leq g^{-1}_{\text{increasing}}(y)\right) = F_X\left(g^{-1}_{\text{increasing}}(y)\right), \end{aligned} \tag{104}$$

$$\begin{gathered} F_Y(y) = P(Y \leq y) = P\left(g_{\text{decreasing}}(X) \leq y\right) = \\ = P\left(X \geq g^{-1}_{\text{decreasing}}(y)\right) = 1 - P\left(X \leq g^{-1}_{\text{decreasing}}(y)\right) = \\ = 1 - F_X\left(g^{-1}_{\text{decreasing}}(y)\right), \end{gathered} \tag{105}$$

where Equation (104) is the case when $g(X)$ is a continuous increasing function $g_{\text{increasing}}(x)$ and Equation (105) is the case when $g(X)$ is a continuous decreasing function $g_{\text{decreasing}}(X)$. The main parts of these inequalities are $P(Y \leq y) = P\left(g_{\text{increasing}}(X) \leq y\right)$ and $P(Y \leq y) = P\left(g_{\text{decreasing}}(X) \leq y\right)$, *i.e.*, by definition, we can apply the transformation of random variables when constructing the probabilities. In other words, if we know the cdf of the random variable, X (the right side of Equation (104) and Equation (105)), then we can construct the cdf of Y (the left side of Equation (104) and Equation (105). Extending Equation (101) by reusing the expression of the cumulative distribution from Equation (104) and applying the Fundamental Theorem of Calculus and the Chain rule, we attain,

$$\begin{gathered} f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X\left(g^{-1}_{\text{increasing}}(y)\right)}{dy} \\ = f_X\left(g^{-1}_{\text{increasing}}(y)\right) \frac{d\left(g^{-1}_{\text{increasing}}(y)\right)}{dy}. \end{gathered} \tag{106}$$

Applying the same derivation for 105 yields,

$$f_Y(y) = -f_X\left(g^{-1}_{\text{decreasing}}(y)\right) \frac{d\left(g^{-1}_{\text{decreasing}}(y)\right)}{dy}. \tag{107}$$

The sign of $f_Y(y)$ depends on the sign of $\frac{d\left(g^{-1}_{\text{decreasing}}(y)\right)}{dy}$, *i.e.*, when it is increasing (the derivative is non-zero), the sign is positive, and when decreasing, the sign is negative. Therefore, in the general case, when function $g(X)$ is invertible and monotonic, we can write the final transformation expression as,

$$f_Y(y) = f_X\left(g^{-1}_{\text{decreasing}}(y)\right) \left| \frac{d\left(g^{-1}_{\text{decreasing}}(y)\right)}{dy} \right|. \tag{108}$$

This is simply known as the Method of Transformations in most elementary probability books or in dynamical systems, the Perron–Frobenius operator. The Perron–Frobenius operator for surjective maps is shown on page 187 in [20]. It is noted in [21]; however, it misses the original proof. We go one step further and show its link to the Kalman filter. We see that expression Equation (108) holds for any continuous monotonic increasing or decreasing function (so, it is not restricted to just the linear functions, as in the Kalman filter). Therefore, we can safely remove the main Kalman filter assumption that all random variables link to each other through linear relationships at the first Kalman stage, called the prediction step. Moreover, we can extend Equation (108) to the cases when the relationships between x and y are not one-to-one. Thus, in general, such a filter could be a nonlinear filter.

6.2. Generalized Multivariate Nonlinear Filter for Monotonic Transitions

We begin the generation for a multivariate, nonlinear filter. A system of transformations in the measurement space is,

$$\begin{cases} y_1 = g_1(x_1, x_2, \cdots, x_n); \\ y_2 = g_2(x_1, x_2, \cdots, x_n); \\ \vdots \\ y_n = g_n(x_1, x_2, \cdots, x_n), \end{cases} \tag{109}$$

where there are n-dimensional vectors, and the transformation function, g, is a multivariate continuous function representing the one-to-one representation between measurement spaces of vectors $\mathbf{x} = (x_1, x_2, \cdots x_n)$ and $\mathbf{y} = (y_1, y, \cdots y_n)$. Again, this is not necessary to assume that $g(\mathbf{x})$ is a system of linear functions, like in the Kalman filter. Therefore, the Kalman filter is a special case of the approach that we outline here. The investigation of situations when we have many-to-one (or other combinations) of relationships between vectors $\mathbf{x}$ and $\mathbf{y}$ is out of scope of this work. Therefore, we have an invertible one-to-one transformation:

$$\begin{cases} x_1 = g_1^{-1}(y_1, y_2, \cdots, y_n); \\ x_2 = g_2^{-1}(y_1, y_2, \cdots, y_n); \\ \vdots \\ x_n = g_n^{-1}(y_1, y_2, \cdots, y_n). \end{cases} \tag{110}$$

Again, the definition of the expression $\mathbf{Y} \equiv g(\mathbf{X})$ between random vector variables $\mathbf{Y}$ and $\mathbf{X}$ assumes that,

$$\begin{gathered} F_Y(y_1, y_2, \cdots, y_n) = P(Y_1 \leq y_1, \cdots, Y_n \leq y_n) = \\ = P(g_1(x_1, x_2, \cdots, x_n) \leq y_1, \cdots, g_n(x_1, x_2, \cdots, x_n) \leq y_n). \end{gathered} \tag{111}$$

Following the previous subsection derivation Equation (106) yields the relationship in the pdf:

$$f_Y(y_1, y_2, \cdots, y_n) = f_X\left(g_1^{-1}(y_1, y_2, \cdots, y_n), \cdots, g_n^{-1}(y_1, y_2, \cdots, y_n)\right) \left| |J(x_1, x_2, \cdots, x_n)|^{-1} \right|. \tag{112}$$

where the Jacobian is defined as,

$$J(x_1, x_2, \cdots, x_n) \equiv \begin{bmatrix} \frac{\partial g_1(x_1, x_2, \cdots, x_n)}{\partial x_1} & \cdots & \frac{\partial g_1(x_1, x_2, \cdots, x_n)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n(x_1, x_2, \cdots, x_n)}{\partial x_1} & \cdots & \frac{\partial g_n(x_1, x_2, \cdots, x_n)}{\partial x_n} \end{bmatrix}. \tag{113}$$

The expression in Equation (112) outlines the basis of the first step of the Kalman filter. This step includes modification of the covariance matrix. In other words, based on our knowledge about the prediction model in Equation (109), we can infer the pdf of the next sample. In a previous subsection, the constant, A, represented a relationship between distinct dimensions of the same measured variable, x, where y was the value in another physical dimension. In the Kalman filter, however, x represents a certain measurement value happening at the time, t_k, and y represents the subsequent measurement value, y, happening at the time $t_{k+1} = t_k + \triangle t$, where index k ranges from one to n and is the discretization interval of our measurement signal, $x(t)$. Therefore, in the Kalman filter context, we should rewrite Equation (102) as,

$$x(t_{k+1}) = g(x(t_k)). \tag{114}$$

Alternatively, in n-dimensional space, we could write,

$$\begin{cases} x_1\left(t_{k+1}\right) = g_1\left(x_1\left(t_k\right), x_2\left(t_k\right), \cdots, x_n\left(t_k\right)\right); \\ x_2\left(t_{k+1}\right) = g_2\left(x_1\left(t_k\right), x_2\left(t_k\right), \cdots, x_n\left(t_k\right)\right); \\ \vdots \\ x_n\left(t_{k+1}\right) = g_n\left(x_1\left(t_k\right), x_2\left(t_k\right), \cdots, x_n\left(t_k\right)\right). \end{cases} \tag{115}$$

The predicted time moment sample from the pdf of the previous sample can be obtained by utilizing Equation (115) in the probabilistic space, *i.e.*, we can rewrite Equation (112) and Equation (113) for the prediction step of the Kalman filter as,

$$\begin{aligned} & f_Y\left(x_1\left(t_{k+1}\right), x_2\left(t_{k+1}\right), \cdots, x_n\left(t_{k+1}\right)\right) \\ = & f_X\begin{pmatrix} g_1^{-1}\left(x_1\left(t_{k+1}\right), x_2\left(t_{k+1}\right), \cdots, x_n\left(t_{k+1}\right)\right) \\ \vdots \\ g_n^{-1}\left(x_1\left(t_{k+1}\right), x_2\left(t_{k+1}\right), \cdots, x_n\left(t_{k+1}\right)\right) \end{pmatrix} \left|\left|J\left(x_1\left(t_k\right), x_2\left(t_k\right), \cdots, x_n\left(t_k\right)\right)\right|^{-1}\right|, \end{aligned} \tag{116}$$

The expression in Equation (116) and covariance matrix change fully describe the prediction step of the nonlinear filter, where we assume that there is a one-to-one relation between subsequent time measurements. The original Kalman filter is a special case of probabilistic inference, when we assume that the transformation function is nothing more than the linear transformation, *i.e.*,

$$\begin{aligned} x_k\left(t_{k+1}\right) &= g_1\left(x_1\left(t_k\right), x_2\left(t_k\right), \cdots, x_n\left(t_k\right)\right) \\ &= A_1 x_1\left(t_k\right) + A_2 x_2\left(t_k\right) + \cdots + A_n x_n\left(t_k\right). \end{aligned} \tag{117}$$

There is a way to deal with surjective transformations, when the transforming function is nonlinear. In such cases, we should also develop the relationship between such a surjective transform and the extensions of the Kalman filter, such as the extended Kalman filter and the unscented Kalman filter. However, such a development is not in the scope of this paper, and we are going to include them in future discussions.

6.3. Kalman Filter Revisited Using a Jacobian

In this subsection, we will revisit the Kalman filter example from the previous section. Our state space system with transformation function $g\left(x\right)$ stays the same,

$$\begin{cases} x_k = x_{k-1} + \triangle t\, v_{k-1} + \frac{c_{a,k} \cdot \triangle t^2}{2}; \\ v_k = v_{k-1} + \triangle t\, c_{a,k}. \end{cases} \tag{118}$$

The inverse function, $g^{-1}\left(\cdots\right)$, of the transformation is,

$$\begin{cases} x_{k-1} = x_k - \triangle t \cdot v_{k-1} - \frac{c_{a,k} \cdot \triangle t^2}{2}; \\ v_{k-1} = v_k - \triangle t \cdot c_{a,k}. \end{cases} \tag{119}$$

Putting Equation (118) into Jacobian expression Equation (113) yields,

$$J\left(x_{k-1}v_{k-1}\right) = \begin{bmatrix} \frac{\partial\left(x_{k-1}+\triangle t\cdot v_{k-1}+\frac{c_{a,k}\cdot\triangle t^2}{2}\right)}{\partial x_{k-1}} & \frac{\partial\left(x_{k-1}+\triangle t\cdot v_{k-1}+\frac{c_{a,k}\cdot\triangle t^2}{2}\right)}{\partial v_{k-1}} \\ \frac{\partial\left(v_{k-1}+\triangle t\cdot c_{a,k}\right)}{\partial x_{k-1}} & \frac{\partial\left(v_{k-1}+\triangle t\cdot c_{a,k}\right)}{\partial v_{k-1}} \end{bmatrix} \tag{120}$$

$$= \begin{vmatrix} 1 & \triangle t \\ 0 & 1 \end{vmatrix} = 1.$$

Placing the modulus of the inverse Jacobian from Equation (120) and the inverse transform from Equation (119) to Equation (112) yields:

$$f_k\left(x_k, v_k\right) = f_{k-1}\left(x_k - \triangle t\ v_{k-1} - \frac{c_{a,k}\ \triangle t^2}{2}, v_k - \triangle t\ c_{a,k}\right). \tag{121}$$

The prior distribution, $f_{k-1}\left(\cdots\right)$, is constructed exactly the same as in previous subsection by eliminating the random noise variables. In other words, it is a multivariate normal distribution with a covariance matrix of,

$$\bar{\Sigma}_{k-1} = \begin{bmatrix} \bar{\sigma}^2_{x,k-1} & \overline{cov}_{x,v,k-1} \\ \overline{cov}_{x,v,k-1} & \bar{\sigma}^2_{v,k-1} \end{bmatrix}, \tag{122}$$

and the form of,

$$f_{k-1}\left(x_{k-1}, v_{k-1}\right) = P_{k-1}\left(x_{k-1}, v_{k-1}\right)$$

$$= \left(2\pi\sqrt{\bar{\sigma}^2_{x,k-1}\bar{\sigma}^2_{v,k-1} - \overline{cov}^2_{x,v,k-1}}\right)^{-1} \exp\left(\frac{\bar{\sigma}^2_{v,k-1}\left(x_{k-1}-\langle x_{k-1}\rangle\right)^2 + \bar{\sigma}^2_{x,k-1}\left(v_{k-1}-\langle v_{k-1}\rangle\right)^2 + 2\overline{cov}_{x,v,k-1}\left(x_{k-1}-\langle x_{k-1}\rangle\right)\left(v_{k-1}-\langle v_{k-1}\rangle\right)}{2\left(\overline{cov}^2_{x,v,k-1}-\bar{\sigma}^2_{x,k-1}\bar{\sigma}^2_{v,k-1}\right)}\right) \tag{123}$$

After inserting Equation (123) into Equation (121), we recover the exact closed form solution as in Equation (70). Therefore, our initial assumptions are relevant and match the assumptions we made in the maximization of relative entropy or the assumptions of the original Kalman filter.

7. Summary and Final Remarks

Kalman demonstrated an ingenious way to estimate unknown variables from noisy measurements, in part by making various assumptions. In this paper, we derive the Bayesian filter and, then, show that by applying the Kalman assumptions, we arrive at a solution that is consistent with the original Kalman filter for pedagogical purposes; explicitly showing that the "transition" or "predictive" step is the prior information and the "measurement" or "updating" step is the likelihood of Bayes' rule. Further, we showed that the well-known Kalman gain is the new uncertainty associated with the posterior distribution.

Recently, a paper [22] used maximum relative entropy (MrE) with the Kalman assumptions, but did not explicitly state that there is a direct link between these two approaches. Here, we showed that the method of maximum relative entropy (MrE) explicitly produces the same mathematical solutions as the Kalman filter, and thus, Kalman is a special case of MrE. We also showed that the closed form

solutions after the application of MrE are immune to real-life numeric problems, which might occur when manipulating the Kalman filter matrix operations.

By applying and manipulating pure probabilistic definitions and techniques used in signal analysis theory, we derived a general, nonlinear filter, where constraining the variables of interest in the form of continuous monotonic increasing or decreasing functions and not necessarily a linear set of functions, like in the original Kalman filter. Thus, we can include more information and extend approximation approaches, such as the extended Kalman filter and unscented Kalman filter techniques and other hybrid variants.

In the end, we derived general distributions using MrE for use in Bayes' Theorem for the same purposes as the original Kalman filter and all of its offshoots. However, MrE can do even more. An important future work will be to include *simultaneous* constraints on the posterior that Bayes cannot do easily alone, such as including macroscopic or course-grained relationships between the various variables of interest. This has been demonstrated in [11].

Acknowledgments

We would like to acknowledge valuable discussions with Julian Center and Peter D. Joseph.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic. Eng.* **1960**, *82*, 35–45.
2. Meinhold, R.J.; Singpurwalla, N.D. Understanding the Kalman Filter. *Am. Statis.* **1983**, *37*, 123–127.
3. Gibbs, B.P. *Advanced Kalman Filtering, Least-Squares and Modeling: A Practical Handbook*; Wiley: New York, NY, USA, 2011.
4. Rosenkrantz, R.D., Ed. *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*; Dordrecht: Reidel, Holand, 1983.
5. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
6. Giffin, A; Caticha, A. Updating Probabilities with Data and Moments. In AIP Conference Procedings Bayesian Inference and Maximum Entropy Methods in Science and Engineering; Knuth, K., Caticha, A., Center, J. L., Giffin, A., Rodríguez, C.C.,Eds.; American Institute of Physics: Melville, NY, USA, 2007; Volume 954, p. 74.
7. Caticha, A.; Giffin, A. Updating Probabilities. In Conference Procedings of Bayesian Inference and Maximum Entropy Methods in Science and Engineering; Mohammad-Djafari, A., Ed.; American Institute of Physics: Melville, NY, USA, 2006; Volume 872, p. 31.
8. Cox, R.T. Probability, frequency, and reasonable expectation. *Am. J. Phys.* **1946**, *14*, 1–13.
9. Thrun, S.; Burgard, W.; Fox, D. *Probabilistic Robotics*. MIT Press: Cambridge, MA, USA, 2006.

10. Chen, Z. Bayesian filtering: From Kalman filters to particle filters, and beyond. *Statistics* **2003**, *182*, 1–69.
11. Giffin, A. From physics to economics: An econometric example using maximum relative entropy. *Physica A* **2009**, *388*, 1610–1620.
12. Mitter, S.K.; Newton, N.J. A Variational Approach to Nonlinear Estimation. *SIAM J. Contr.* **2004**, *42*, 1813–1833.
13. Mitter, S.K.; Newton, N.J. Information and Entropy Flow in the Kalman-Bucy Filter. *J. Stat. Phys.* **2005**, *118*, 145–176.
14. Eyink, G.; Kim, S. A maximum entropy method for particle filtering. *J. Stat. Phys.* **2006**, *123*, 1071–1128.
15. Joseph, P.D. Kalman Filter Lessons. Available online: http://home.earthlink.net /˜pdjoseph/id11.html (accessed on 18 February 2014).
16. Crassidis, J.L.; Junkins, J.L. *Optimal Estimation of Dynamical Systems*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2004.
17. Poularikas, A. D. *The Handbook of Formulas and Tables for Signal Processing*; CRC Press: Boca Raton, FA, USA, 1998.
18. Jazwinski, A. H. *Stochastic Processes and Filtering Theory*; Academic Press: New York, NY, USA, 1970.
19. Crisan, D.; Rozovskii, B. L. *The Oxford Handbook of Nonlinear Filtering*; Oxford University Press: Oxford, UK, 2011.
20. Katok,A.; Hasselblatt, B. *Introduction to the Modern Theory of Dynamical Systems*; Cambridge University Press: Cambridge, UK, 1996.
21. Beck, C.; Schögl, F. *Thermodynamics of Chaotic Systems: An Introduction*; Cambridge University Press, Cambridge, UK, 1995.
22. Urniezius, R. Online robot dead reckoning localization using maximum relative entropy optimization with model constraints. In AIP Conference Proceedings of Bayesian Inference and Maximum Entropy Methods in Science and Engineering; Mohammad-Djafari, A., Ed.; American Institute of Physics: Melville, NY, USA, 2011; Volume 1305, p. 274.