On Data Fusion Methodologies For Spontaneous and Solicited Safety Data Evaluation

Hal Li
Merck Research Laboratories
North Wales, PA, USA
hal.li@merck.com

William Wang
Merck Research Laboratories
North Wales, PA, USA
william_wang@merck.com

Abstract *— The volume of adverse events reported has been increasing at a rapid rate for over the past 15 years. Though these data are collected from many sources, all clinical safety data have been treated as equally important to patient safety. However, there is evidence that growth in data volume may negatively affect safety monitoring. Here, we examine a method of incorporating different safety data sources for risk identification.*

Specifically, we examine the use of data fusion analytical methodologies on safety data. We validate that the Kalman filter methodology is most effective when used to fuse low level noisy data. We then assessed this method on real world data with simulated noise in safety databases. We find that the Kalman filter effectively suppresses noise and creates smooth estimates when the SNR is not too large. When the SNR is high, the estimation will not represent the characteristic of the true data.

Keywords *— Data fusion, Real world data, Simulation, Solicited safety data, Data stream mining*

I. Introduction

The purpose of safety signal detection is to differentiate between true target signals and interfering noise. Therefore, the best signal detection systems are able to both correctly identify all real targets (minimize false negatives) and exclude all erroneous non-signals (minimize false positives). Currently, the canonical method for post marketing detection of safety signals involves disproportionality analyses. Despite the prevalence of these methods, they are fundamentally designed to be used on spontaneous report databases where patients or healthcare providers have reported events to health authorities or marketing authorization holders. In general, these reports are of high quality and therefore the data can be mined effectively to find the signals of interest.

A challenge to these systems that has emerged recently is the increasing volume of adverse event reports that have been collected. Specifically, these reports have increased by around 10-20% every year resulting in a compounding 400% increase over a decade[1]. An additional challenge with using this data is that much of it consists of reports from patient support programs (PSP), patient assistant programs (PAP), market research programs (MRP), and social media accounts. Unfortunately, the data from these programs may not be as valuable for the purpose of monitoring safety because the reports were not intended to collect information relating to medication use. For example, a PSP is a structured initiative for caregivers to assist patients in managing their disease and include programs such as homecare, compliance adherence programs, and disease management programs. Similarly, PAPs are charitable in design and assist patients through reimbursement, compensation, or access programs. Together, the relative proportion of these solicited reports have increased compared to classic spontaneous reports.

Interestingly, when the impact of including these PSP reports in signal detection analyses were examined, Klein et al. found that numerous spurious false positive signals were generated. Furthermore, these false results could be eliminated by removing solicited data from the database. In addition, they found that by increasing the background reporting rate for the drug of interest, solicited reports also increased the threshold for detecting signals of disproportionate reporting (SDR). Therefore, it is possible that some true signals will fall under the elevated threshold and be missed. The errors that arise from pooling spontaneous data along with solicited data were called precautionary reporting bias.

Given that the quantity of spontaneous data will likely continue to increase over time, a pressing need exists to find better methods for harmonizing solicited reports with classical spontaneous reports. These methods will need to account for the disparate levels of data quality from spontaneous healthcare provider reports on one extreme to unverified social media reports on the other. Any successful method should be able to incorporate some of this new data without compromising the quality of the analyses and without adversely impacting the rate of false positives and false negatives in the analysis.

Here, we apply the Kalman filter technique on safety data as a method of incorporating solicited data with spontaneous reports. We believe that this technique provides several advantages over existing techniques and describe our methodology for adapting the method to safety data. After describing the advantages of the technique, we proceed to analyze data on alcohol-related deaths in Finland as a proof of concept.

II. Methods

A. Data Fusion

Currently available data fusion techniques can be classified into three categories that are non-exclusive from each other: (i) data association, (ii) state estimation, and (iii) decision fusion. The goal of state estimation techniques is to determine the final target state from available observations and measurements. In general, these methods assume that some of the observations actually come from the target signal while others simply represent only noise. The estimation problem involves

determining the values of a vector state such as size or position that is able to fit with observed data as much as possible. That means that given a set of redundant observations, the goal is to determine the set of parameters that provide the best possible fit to the observed data.

State estimation methods can either be linear or nonlinear. When several conditions are met, namely that the equations of the object state is linear, the measurements are linear, and the noise follows the Gaussian distribution, a recursive Kalman filter will provide the optimal theoretical solution for statistical estimations.

B. The Kalman Filter

The Kalman filter was originally proposed by Kalman in 1960 [2] as a novel approach to solving problems of linear filtering and accurate prediction. It has since become a very popular method of data fusion and has been widely applied and studied in numerous contexts. At its core, the Kalman filter works by using the space-time model shown below to estimate the state x of a discrete time process.

$$x\,(k+1) = \Phi(k)\,x\,(k) + G(k)\,u\,(k) + w(k)$$

Within this model, the measurements and observations z at a given time k of the state x can be represented as

$$z\,(k) = H(k)\,x\,(k) + \mathrm{V}\,(k)$$

Furthermore, the parameter $\Phi(k)$ is the state transition matrix, the parameter $G(k)$ represents the input matrix transition, the parameter $u\,(k)$ is the input vector, and the parameter $H(k)$ represents the measurement matrix. Two additional variables w and V are assigned to be random Gaussian variables that have the properties of zero mean and covariance matrices that correspond respectively with $Q(k)$ and $R(k)$.

Within this model, the estimation of $x(k)$, represented here as $\hat{x}(k)$, as well as the prediction $x(k+1)$, represented here as $\hat{x}(k+1 \mid k)$ can be easily determined. They are based on measurements as well as systemic parameters and are given by the following equations:

$$\hat{x}\,(k) = \hat{x}\,(k \mid k+1) + K(k)\,[z\,(k) - H(k)\hat{x}\,(k \mid k-1)]$$

$$\hat{x}\,(k+1 \mid k) = \Phi(k)\hat{x}\,(k \mid k) + G(k)\,u\,(k)$$

In addition, the gain of the filter can be calculated and is determined by the following equation:

$$K(k) = P(k \mid k-1)H^{\mathrm{T}}(k) \times [H(k)\,P\,(k \mid k-1)H^{\mathrm{T}}\,(k) + R(k)]^{-1}$$

Here, the prediction covariance matrix is represented as $P(k \mid k-1)$ and this matrix can be determined by the following equations:

$$P\,(k+1 \mid k) = \Phi(k)\,P\,(k)\Phi^{\mathrm{T}}\,(k) + Q(k)$$

$$P\,(k) = P(k \mid k-1) - K(k)H(k)\,P\,(k \mid k-1)$$

C. Uses of the Kalman Filter

The Kalman filter is most effective when used to fuse low level data. In particular, it is able to obtain optimal statistical estimations when the system can appropriately be modeled as a linear model and when the error can appropriately be modeled as Gaussian noise. Modifications to the basic filter can also be performed to address nonlinear measurements and nonlinear dynamic models. One example modification is the extended Kalman filter, which extends the approach to be an optimal method for implementing a nonlinear recursive filter. It is thus commonly used for fusing data in some applications such as robotics. Importantly, the Kalman filter exists as an R package known as KFAS, which stands for Kalman filtering and smoothing.

The Gaussian state space modeling exists as a special case within the KFAS filtering method. For a linear Gaussian state space model featuring continuous states as well as discrete time intervals from t = 1, . . . , n, we can define an observation equation and a state equation as follows:

$$y_t = Z_t\alpha_t + \varepsilon_t,$$

$$\alpha_{t+1} = T_t\alpha_t + R_t\,\eta_t,$$

where we can independently define the variables $\varepsilon_t \sim N(0,H_t)$, $\eta_t \sim N(0,Q_t)$ and $\alpha_1 \sim N(\alpha_1, P_1)$ Furthermore, we also assume that y_t is a $p \times 1$ vector, α_{t+1} is an $m \times 1$ vector and η_t is a $k \times 1$ vector.

Next we create and denote $\alpha = (\alpha_1{}^{\mathrm{T}},\ldots, \alpha_n{}^{\mathrm{T}})^{\mathrm{T}}$ and $y = (y_1{}^{\mathrm{T}},\ldots, y_n{}^{\mathrm{T}})^{\mathrm{T}}$. Within this model, α_t is a vector of the latent state process at a particular time point t and y_t contains the observations found at said time t. In contrast, system matrices Z_t, T_t, and R_t, as well as covariance matrices H_t and Q_t are usually time invariant but may depend on the definition ascribed by a particular model. In most cases, at least a few of these matrices will contain unknown parameters that will be estimated.

In KFAS filtering method, the predominant goal of state space modeling is to gain knowledge of a latent state a given certain observations y. This goal can be achieved through the use of two recursive algorithms, namely Kalman filtering and smoothing.

With Kalman filtering method, we are able to obtain predictions as well as prediction errors for one step ahead as follows:

$$\alpha_{t+1} = \mathrm{E}(\alpha_{t+1} \mid y_t,\ldots, y_1),$$
$$v_t = y_t - Z_t\,\alpha_t$$

Furthermore, the related covariance matrices can be defined as follows:

$$P_{t+1} = \mathrm{VAR}(\alpha_{t+1} \mid y_t,\ldots, y_1),$$
$$F_t = \mathrm{VAR}(v_t) = Z_tP_tZ_t{}^{\mathrm{T}} + H_t.$$

By using the results from this Kalman filtering method, we are able to subsequently establish state smoothing equations that run backwards in time, which yields

$$\hat{\alpha_t} = \mathrm{E}(\alpha_t \mid y_n,\ldots, y_1),$$
$$V_t = \mathrm{VAR}(\alpha_t \mid y_n,\ldots, y_1).$$

Additional smoothed estimates can be similarly calculated for the disturbance terms ε_t and η_t, as well as for the signal

$\theta_t = Z_t\ \alpha_t$. Additional details of these algorithms have been described by Durbin and Koopman (2012).

Finally, we note that the prior distribution of an initial state vector α_1. is defined as a multivariate Gaussian distribution that features a mean α_1 as well as a covariance matrix P_1. It can be proven that given all parameters included in the system matrices, the results of the Kalman filter and smoother will be equivalent with Bayesian analysis that is conducted using the same prior distribution for α_1.

D. Application of the Kalman Filter on Safety Data

For our analysis, we will be applying the methods described above on a dataset of alcohol-related deaths that were tabulated per 100,000 persons on a yearly basis in Finland. Data were included for individuals between the ages of 40-49 from the years 1969–2007. Data was available from Statistics Finland and are used as an example to illustrate how to use the Kalman Filter in the evaluation of safety data.

For observations $y_1, \ldots, y_n$, we can assume that $y_t \sim N(\mu_t, \sigma_t^2)$ for t = 1, . . . , n, where we set μ_t as a random walk with the following drift process: $\mu_{t+1} = \mu_t + \nu + \eta_t$, with $\eta_t \sim N(0, \sigma_\eta^2)$, assuming that there is no prior information regarding the initial state μ_1 or constant slope ν.

From the Kalman filter algorithm we are therefore able to obtain predictions for one step ahead of the states $\alpha_t = (\mu_t, \nu_t)^T$. Note that when a new observation y_t is available at a certain time point t, the estimate of ν will be updated in order to take the new information provided by y_t into account. Therefore, when Kalman filtering is completed, α_{n+1} will provide the final estimate of the constant slope term after accounting for all available data. Similarly, the Kalman filter will compute predictions at one step ahead for μ_t, which, after smoothing will give us estimates of μ_t for t = 1, . . . , n given all available data.

Finally, the Kalman filter computes prediction errors at one step ahead for $\nu_t = y_{t+1} - \mu_t$, and can use this information as well as previous predictions in order to correct predictions for the next time point. The smoothing algorithm will therefore take both past and future values into account at each time point in order to produce a smoother estimate of the latent process.

III. Results and Discussions

A. The smoothed estimates from the Kalman filter

In order to implement the previously described analysis strategy, we applied a Kalman filter we applied on a Gaussian distribution with mean μ_t and variance u_t. As mentioned above, the specific dataset used was the compilation of alcohol-related deaths that were tabulated per 100,000 persons on a yearly basis in Finland. Figure 1 shows the observations with the smoothed estimates of the random walk process μt. As we pointed out earlier, the smoothing algorithm takes into account both the past and the future values at each time point, thus producing more smoothed estimates of the latent process. This can easily be seen at the beginning of the series where the predictions seem to be lagging the observations by one time step. Because the Kalman filter takes into account both the past and the future values at each time point, the noise that is added to the true signal can be filtered out so that the true signal will be amplified. For this reason, as we mentioned earlier, the Kalman filter data fusion method is most effective when used to fuse low level noisy data.

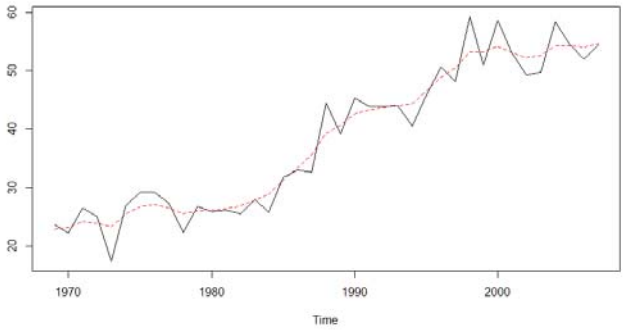

Figure 1: Alcohol-related deaths in Finland in the age group of 40–49 years with smoothed estimates.

B. Applying the Kalman filter on noisy data

As we discussed, solicited safety data could be noisy. Specifically, safety data consist of whether or not an adverse side effect occurs in patients who are taking a certain medication. Therefore, safety data are believed to follow a binomial distribution since the outcomes are boolean. For common safety signals that occur often, the central limit theorem states that this binomial distribution can be well approximated by a normal distribution. Therefore, we decided to use a normal distribution to generate noise for application to the data.

Applying the Kalman filter on these noisy data to discovering novel hidden patterns is infact a data stream mining process. In order to assess the Kalman filter's effectiveness on filtering the noise and estimating the true signal, we next generated 1000 sets of data with normally distributed noise. To simplify the problem, we assume the noisy signal has the same mean as the true signal, while its standard deviation is proportional to the magnitude of the true signal:

$Sn_t \sim \text{Normal}(y_t, \sigma_t^2)$

With $\sigma_t = k\ y_t$. In this example, we chose k=0.25.

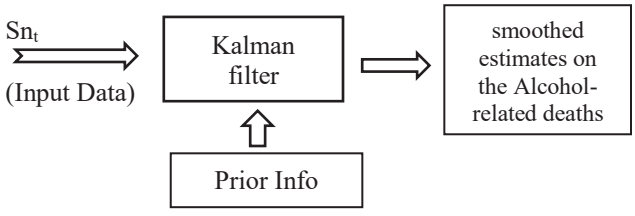

Figure 2: Illustration on applying the Kalman filter on noisy data that was generated by adding normally distributed noise to create the smoothed estimates of alcohol-related deaths

We further assume that the observed alcohol-related deaths that were tabulated per 100,000 persons on a yearly basis in

Finland are the true accurate observations. The mean and 95% confident interval (CI) were calculated each year from these and we generated 1000 sets of data with normally distributed noise. These are the input data to the Kalman filter algorithm. The mean and CI during each year are demonstrated in Figure 3. As we expected, the mean equals the observed alcohol-related deaths.

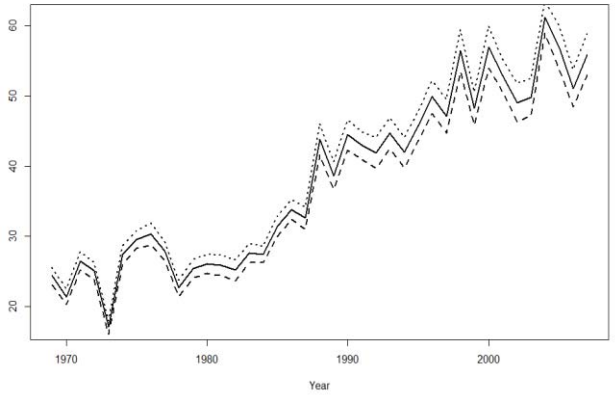

Figure 3: Mean and 95% CI from the generated normally distributed noise based on the Alcohol-related deaths (Input Data).

Next, the same Kalman filter algorithm which was described above was applied to each of the 1000 sets of noisy data. As a consequence, 1000 sets of smoothed estimates of alcohol-related deaths were generated from this data fusion Kalman filter algorithm. These are the output data from the Kalman filter algorithm. Once again, the mean and 95% confident intervals (CI) were calculated for each year from these 1000 sets of Kalman filter algorithm smoothed estimates on the noisy alcohol-related death data.

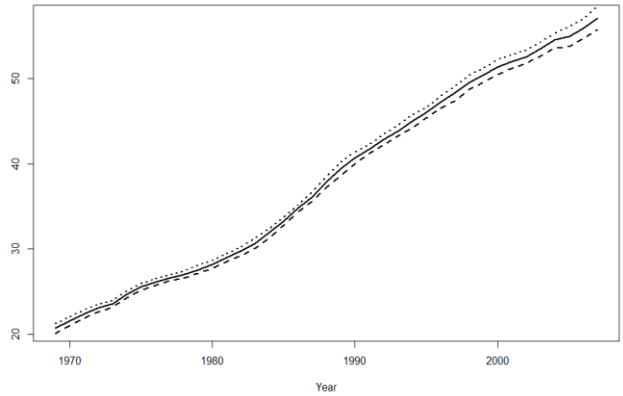

Figure 4: Mean and 95% CI from the Kalman filter algorithm smoothed estimates on the noisy Alcohol-related death data.

We can visually compare the width of the confidence interval from the generated data with normally distributed noise (Illustrated in Figure 3) and the width of the confidence interval from the Kalman filter algorithm smoothed estimates (Illustrated in Figure 4). From these visual comparisons, we can observe that the width of the CI from the output datasets are only half of those from the input datasets. This result demonstrated, as we expected, that the Kalman filter data fusion method is effective when used to fuse low level noisy data in order to generate a smoothed estimation of the true data.

C. The impact of noisy data on safety monitoring

In order to assess impact of noisy data on the process of safety monitoring, we examined the smoothed Kalman filter estimates on the noisy alcohol-related death data. Similar to the previous section, we assume the noisy signal has the same mean as the true signal, while its standard deviation is proportional to the magnitude of the true signal:

$Sn_t \sim \text{Normal}(y_t, \sigma_t^2)$

Where $\sigma_t = k\, y_t$.

This time; however, we chose k=0.25 (Figure 5), k=0.33 (Figure 6) and k=0.5 (Figure 7) for our model. We then conducted an analysis and examined the output from the Kalman filter algorithm. From our examples, one can observe with small amount of noise (k=0.25), the Kalman filter smoothed estimates are pretty similar to those from the ture signal (Figure 1). As the amount of random noise included in the input safety data increases (k=0.33 or 0.5), the Kalman filter smoothed estimates lose both the trend and the accuracy of prediction (Figure 6 and 7). That means that at these levels of noise, the noisy data are negatively affecting safety signal detection.

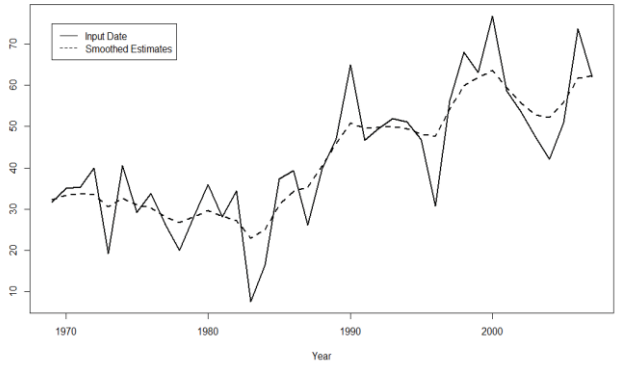

Figure 5: Noise ($\sigma_t = 0.25\, y_t$) and the observed Alcohol-related deaths in Finland in the age group of 40–49 years with smoothed estimates.

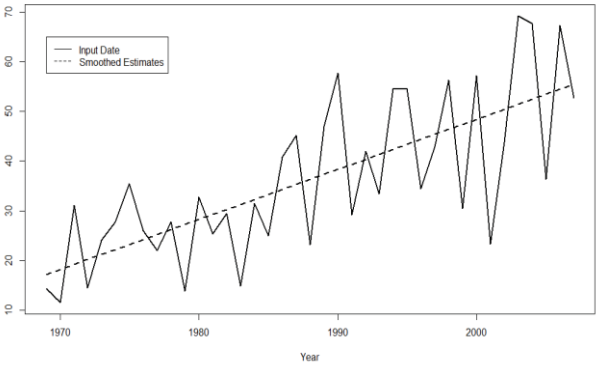

Figure 6: Noise (σ_t=0.33 y_t) and the observed Alcohol-related deaths in Finland in the age group of 40–49 years with smoothed estimates.

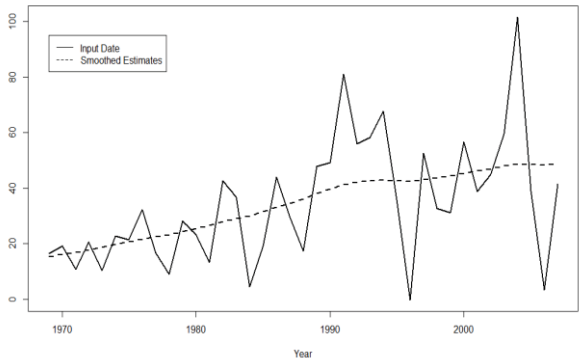

Figure 7: Noise (σ_t=0.5 y_t) and the observed Alcohol-related deaths in Finland in the age group of 40–49 years with smoothed estimates.

IV. Conclusions

As the quantity of solicited data of variable quality increases over time, the ability to extract safety signals from noisy data will become increasingly important. One popular and exciting method for filtering data is known as the Kalman filter, which we have applied in our case study to a set of data on alcohol related deaths in Finland. Upon generating 1000 sets of data with incorporation of normally distributed random noise, we find that the output from the Kalman filter is more precise than the input. Visually, this finding can be determined because the width of the interval is roughly half that of the input. These findings remained robust when incorporating roughly 25% noise into the data. However, when the k was increased to 0.33 or 0.5, we found that the performance of the Kalman filter suffered due to decreased accuracy. Given this experience, we believe that using the Kalman filter will be an effective method for incorporating noisy data into safety evaluations so long as the noise does not exceed too high a threshold.

V. Reference

[1] Food and Drug Administration. FDA Adverse Event Reporting System Public Dashboard. 2018 [cited 2018 November 28].
[2] R. E. Kalman, "A new approach to linear filtering and prediction problems," Journal of Basic Engineering, vol. 82, no. 1, pp. 35–45, 1960.