# Deep Kalman Filters

Alex Martinez, Ronjon Nag, Alex Tamkin, Pol Rosello, Pranav Sriram

November 16, 2016

## 1 Introduction

In *Deep Kalman Filters*, Krishnan et al. attempt to apply a modified version of Kalman filters to determine what the best course of treatment for a patient might be, even for patients with only an incomplete medical history. Kalman filters are generative probabilistic graphical models for time series data, used in fields ranging from finance to robotics. Despite the broad success of Kalman filters, they are limited in their ability to model complex sequential data due to their assumptions of linear relationships between successive latent states and between latent states and observed states. This paper extends classical Kalman filters by replacing linear transformations with nonlinear transformations parameterized by neural networks, yielding "Deep Kalman Filters" [1]. The increased complexity of the model necessitates novel training methods and allows for new applications. The technique was first applied to perturbed MNIST data, which the authors refer to as a "Healing MNIST database," and then secondly applied to a diabetes dataset.

## 2 Context and Related Work: Classical Kalman Filter

The classical Kalman Filter aims to model latent states $z_t \in \mathbb{R}^s$, using a sequence of observations $x_t \in \mathbb{R}^s$ and actions $u_t \in \mathbb{R}^c$, as follows:

$$z_t \sim \mathcal{N}\left(G_t z_{t-1} + B_t u_{t-1}, \Sigma_t\right) \text{ (action-transition)}, \qquad x_t \sim \mathcal{N}\left(F_t z_t, \Gamma_t\right) \text{ (observation)} \tag{1}$$

At each time step, there's subsequently an update to the subsequent values of $x_t$ and $\Sigma_t$:

$$\begin{aligned} K &= \Sigma_t F_t^T \left(F_t \Sigma_t F_t^T + R_t\right)^{-1} \text{ (Kalman Gain)} \\ z_t &:= z_t + K\left(x_t - F_t z_t\right) \\ \Sigma_t &:= \left(I + K F_t\right)\Sigma_t \end{aligned} \tag{2}$$

where $\Sigma_t$ and $\Gamma_t$ are co-variance matrices with possible time-variance (e.g. $\Sigma_t = G_t \Sigma_{t-1} G_t^T + Q_t$ and $\Gamma_t = F_t \Gamma_{t-1} F_t^T$). $G_t \in \mathbb{R}^{s \times s}$ is a state-transition matrix, $B_t \in \mathbb{R}^{s \times c}$ is a control-input matrix, $F_t \in \mathbb{R}^{d \times s}$ is an observation matrix, and $R_t$ and $Q_t$ are sensor and ambient noise matrices. Note that our predictions and updates currently rely on linear transformations.

## 3 Context and Related Work: Application

Other researchers have used linear Kalman filters for biomedical applications. A common application is control algorithms for medical devices. For example, Parker, et al use linear Kalman filters to create a control algorithm for a glucose pump for diabetes patients [3].

In addition to control algorithms, Kalman filters are also utilized in biomedical domains for their denoising properties, as they are in other fields. For instance, Mneimneh, et al apply Kalman filters to ECG measurements in order to remove baseline wandering. The Kalman filter approach performs favorably relative to other approaches like moving averages or cubic spline techniques [4].

Kalman filters have also seen direct application to electronic medical record (EMR) data. Caballero and Akella use Dynamic Linear Models (DLMs) to predict the probability of mortality for patients using

EMR data[5]. The authors employ linear Kalman filters to update the latent state of their model, which outperforms other widely-used models.

As the authors of this paper note, the research on counterfactual inference for administering medication is neither plentiful nor advanced. For example, Lu, et al use Kaplan-Meier survival curves and Cox regression models to evaluate the effects of a policy change on Medicaid patients with bipolar disorder[6]. The difficulties in modeling more complex, higher-dimensionality situations are one factor that motivates this paper.

# 4 Discussion of Model

The paper next describes how to extend the Kalman filter with nonlinear transformations. Our training step involves optimizing our choices of $F$, $G$, and $B$ such that they maximize the conditional likelihood of the observations $x_t$ given external actions $u_t$ i.e. $\max_\theta \log p_\theta(x_1, .., x_t | u_1, .., u_{t-1})$. This can be done tractably, if not analytically, for linear transformations. However, non-linear transformations render the posterior distribution $p(\vec{z}|\vec{x}, \vec{u})$ intractable to compute. To combat this, we leverage recent work in variational autoencoders to use an approximation $q_\phi$ of the posterior that's trained using a *recognition network*. This network can vary widely in terms of architecture and complexity. Out of the two multi-layer perceptron and two neural network models the paper evaluated, the bi-directional rNN performed best as a recognition network.

We now switch from the original "engineering" perspective to a Bayesian interpretation of the objective our recognition network will help maximize. Let $p(x, z|\, u) = p_0(z|u)\, p_\theta(x|z, u)$ be a generative model for the set of observations $x$ and the intractable posterior distribution $p_\theta(z_t|x_1, .., x_t) \propto p_\theta(x_t|z_t, x_1, .., x_{t-1})\, p_\theta(z_t|x_1, .., x_{t-1})$. We now use the approximate posterior distribution $q_\phi(z|x)$, modeled by the *recognition network*, to obtain a lower bound on the marginal likelihood $\log p_\theta(x)$ as follows:

$$\begin{aligned}
\log p_\theta(\vec{x}|\, \vec{u}) &= \log \int_z \frac{q_\phi(\vec{z}|\vec{x}, \vec{u})}{q_\phi(\vec{z}|\vec{x}, \vec{u})} p_\theta(\vec{x}|\vec{z}, \vec{u})\, p_0(\vec{z}|\vec{u})\, dz \\
&\geq \int_z q_\phi(\vec{z}|\vec{x}, \vec{u}) \log \left( \frac{p_\theta(\vec{x}|\vec{z}, \vec{u})\, p_0(\vec{z}|\vec{u})}{q_\phi(\vec{z}|\vec{x}, \vec{u})} \right) dz \\
&\geq \mathop{\mathbb{E}}_{q_\phi(\vec{z}|\vec{x}, \vec{u})} [\log p_\theta(\vec{x}|\vec{z}, \vec{u})] - KL(q_\phi(\vec{z}|\vec{x}, \vec{u}) \,\|\, p_0(\vec{z}|\vec{u})) = \mathcal{L}(x; (\theta, \phi))
\end{aligned} \tag{3}$$

Utilizing assumptions such as that $(x_t \perp x_{\neg t})\,|\vec{z}$, we finally have the expression for $\mathcal{L}$ as:

$$\begin{aligned}
\mathcal{L}(x; (\theta, \phi)) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_\phi(z_t|\vec{x}, \vec{u})} [\log p_\theta(x_t|z_t)] - KL(q_\phi(z_1|\vec{x}, \vec{u}) \,\|\, p_0(z_1)) - \\
\sum_{t=2}^{T} \mathop{\mathbb{E}}_{q_\phi(z_{t-1}|\vec{x}, \vec{u})} KL(q_\phi(z_t|z_{t-1}, \vec{x}, \vec{u}) \,\|\, p_0(z_t|z_{t-1}, u_{t-1}))
\end{aligned} \tag{4}$$

Since we can use Monte Carlo estimates of the gradients of $\mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)]$ and $KL(q_\phi(z|x) \,\|\, p_o(z))$ w.r.t. $q_\phi(z_t)$, we can use **stochastic back-propagation** in a variational autoencoder whose objective is to maximize the lower bound $\mathcal{L}(x; (\theta, \phi))$ using an inference-update algorithm in which we: infer $\vec{z}$ from input $\vec{x}$, $\vec{u}$ via $q_\phi$; reconstruct the input using the current estimates of the posterior via $p_\theta$; estimate gradients of the likelihood w.r.t. $\theta$ and $\phi$; update model parameters.

Upon training the generative temporal model, we can then perform *counterfactual inference*. First, we perform inference using a history of observations and actions $x_1, ..., x_t, u_1, ...u_{t-1}$ using the learned $q_\phi$ to get an estimate for $z_t$. We can then forward sample from this latent state and varying choices of actions $u_t$ in order to contrast their outcomes. The generative model for the deep Kalman filter is given by:

$$\begin{aligned}
z_1 &\sim \mathcal{N}(\mu, \Sigma_0) \\
z_t &\sim \mathcal{N}(G_\alpha(z_{t-1}, u_{t-1}, \Delta_t), S_\beta(z_{t-1}, u_{t-1}, \Delta_t)) \\
x_t &\sim \Pi(F_\kappa(z_t))
\end{aligned} \tag{5}$$

where $\Delta_t$ is the period between $t-1$ and $t$. The functions $G_\alpha, S_\beta, F_\kappa$ are parameterized by deep neural networks, making $\theta = \{\alpha, \beta, \kappa\}$ the parameters of the generative model.

The model used by the researchers is interesting from a theoretical perspective for extending Kalman filters to use nonlinear dynamics. It successfully overcomes the hurdle of training the neural network using stochastic back-propagation, and is a novel application of neural networks outside of their more common usage as classifiers. The core idea behind the model is likely to be useful in many fields and is certainly not limited to bioinformatics.

The model's neural network itself is not particularly novel, and with only two layers, it is definitely not "deep", as the title of the paper implies. An interesting extension of this work would be to use a deeper, more modern network architecture and compare its performance to the presented model.

# 5 Discussion of Experiments and Applications

The authors evaluate their model on two separate datasets: a synthetic dataset derived from the MNIST Handwritten Digits dataset, and a healthcare dataset from diabetic and pre-diabetic patients. The MNIST dataset is modified as a noisy sequence of rotated digits which they attempt to model with deep Kalman filters. The authors claim the sequence could represent the "temporal evolution of patients" and offer several parallels between the two (high-dimensional, noisy sequential data), but the argument is not very convincing given the nature of the problem and data remains very far from their intended goal of learning a time-varying, generative model of patients from electronic health records.

To that end, the authors also train their model on a dataset of 8000 diabetic and pre-diabetic patients, attempting to assess the effect of anti-diabetic drugs on a patient's A1c and glucose levels. They compare the log-likelihood of their model's predictions on the test set with those of traditional Kalman filters with linear emission and transition dynamics, and find that their deep Kalman filter model (with both non-linear emission and transition dynamics) performs best. They also perform counterfactual inference on their results to conclude that patients who do not receive anti-diabetic drugs are much more prone to having high glucose and A1c levels. Overall, we found this to be a more convincing experiment given that it directly tackled the original stated purpose of the paper, but the analysis of the results was not particularly thorough, and the performance of their model was not compared to models unrelated to Kalman filters. It is not very surprising that a more expressive version of the same model performed better, in the same way that a three-layer neural net is likely to perform better than a two-layer neural net. We would have liked to see an experiment on a larger, more challenging dataset, where the conclusions reached by their model cannot already be inferred by humans, and where a favorable comparison with other unrelated models would have been more impressive.

# References


[1] Rahul G. Krishnan, Uri Shalit, and David Sontag. *Deep Kalman Filters.* https://arxiv.org/pdf/1511.05121v2.pdf.

[2] Matthew James Johnson, David Duvenaud, Alexander B. Wiltschko, Sandeep R. Datta, Ryan P. Adams. *Composing graphical models with neural networks for structured representations and fast inference.* https://arxiv.org/pdf/1603.06277v3.pdf.

[3] Robert S. Parker, Francis J. Doyle, III, and Nicholas A. Peppas *A Model-Based Algorithm for Blood Glucose Control in Type I Diabetic Patients* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.383.9794&rep=rep1&type=pdf

[4] MA Mneimneh, EE Yaz, MT Johnson, RJ Povinelli *An Adaptive Kalman Filter for Removing Baseline Wandering in ECG Signals* http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4511836

[5] Karla Caballero and Ram Akella *Dynamically Modeling Patient's Health State from Electronic Medical Records: A Time Series Approach* https://dl.acm.org/citation.cfm?id=2783289

[6] Christine Y Lu, Alyce S Adams, Dennis Ross-Degnan, Fang Zhang, Yuting Zhang, Carl Salzman, and Stephen B Soumerai *Association Between Prior Authorization for Psychiatric Medications and Use of Health Services Among Medicaid Patients With Bipolar Disorder* http://europepmc.org/articles/pmc3053119