Master Thesis. Meteorological time series imputation using Kalman filters

Simone Massaro

January 2023

Contents

1 Introduction **2**

2 Methods **2**
2.1 Math 2
2.1.1 Probabilistic Machine Learning 2
2.1.2 Notation 2
2.1.3 Kalman Filter Introduction 3
2.1.4 Filter 3
2.1.5 Kalman Smoother 5
2.1.6 Predictions 5
2.2 Filter Implementation 5
2.2.1 Parameter constraints 5
2.2.2 Numerical stability 5
2.3 Loss Function 5
2.3.1 Joint distribution of the gap 5
2.3.2 Joint distribution state for gaps 6
2.3.3 Joint distribution state - partial observations 8

3 Results **8**

4 Discussion **8**

1 Introduction

- problem
- state of the art
- our approach
  - uncertainties
  - combination of multiple sources of information [picture?]
  - custom implementation of Kalman filter imputation library
- why is relevant

2 Methods

2.1 Math

2.1.1 Probabilistic Machine Learning

- probability
- Conditional probability
- Bayes theorem
- Gaussian Inference

2.1.2 Notation

[TODO data format maybe a picture]

- t Number of time steps
- observations
  - n Number of variables observed
  - $y_{:,t}$ or y_t vector of all the n variables at time t, $\in \mathbb{R}^n$
  - $y_{n,:}$ vector of the nth variable at for time steps in t, $\in \mathbb{R}^T$
  - $y_{n,t}$ nth variable at time t, $\in \mathbb{R}$
  - $Y_M = [x_{:,1}, ...x_{:,t}]$ Matrix with all the n variables at all time steps, $\in \mathbb{R}^{n \times t}$
  - Y is a vector obtained by "flattening" X_M, by putting next to each other all variable at time t, $\in \mathbb{R}^{(n \cdot t)}$
  - y_t^{ng} vector of variable that are not missing (ng = not gap)) at time t, $\in \mathbb{R}^{n_{ng}}$. Note at different times the shape of this vector can change

- Y^{ng} all observations

- latent state
  - k Number of variables in latent state
  - $x_{:,t}$ or x_t vector of all the k state variables at time t, $\in \mathbb{R}^k$
  - $x_{k,:}$ vector of the kth variable at for time steps in t, $\in \mathbb{R}^t$
  - $x_{k,t}$ kth variable at time t, $\in \mathbb{R}$
  - $X_M = [x_{:,1}, ...x_{:,t}]$ Matrix with all the k variables at all time steps, $\in \mathbb{R}^{k \times t}$
  - X is a vector obtained by "flattening" X_M, by putting next to each other all variable at time t, $\in \mathbb{R}^{(k \cdot t)}$

2.1.3 Kalman Filter Introduction

- why Kalman filter

- picture of Kalman filter state

Description The latent state (x) is modelled using a Markov chain. Which means that the state at time t depends only on the state at time $t-1$ and not the states at previous times

Basic equations

$$p(x_t|x_{t-1}) = \mathcal{N}(Ax_{t-1} + b, Q) \tag{1}$$

The observation are derived from the state using a linear map plus random noise

$$p(y_t|x_t) = \mathcal{N}(Hx_t + d, R) \tag{2}$$

2.1.4 Filter

Filter prediction The probability distribution of state at time t is computed using the state a time $t-1$
The state at time $t-1$ has a distribution

$$p(x_{t-1}) = \mathcal{N}(m_{t-1}, P_{t-1})$$

Combining this equation with equation 1 and using the properties of a linear map of a Gaussian distribution we obtain:

$$p(x_t) = \mathcal{N}(x_t; m_t^-, P_t^-) \tag{3}$$

where:

- predicted state mean: $m_t^- = Am_{t-1} + Bc_t + d$

- predicted state covariance: $P_t^- = AP_{t-1}A^T + Q$

The mean and the covariance of the state at time 0 are parameters of the models that are learned

Filter correct Probability of state at time ‘t is corrected using the observations at time t
This uses equation 2 and the formula for posterior distributions for Gaussian distributions.

$$p(x_t|y_t) = \mathcal{N}(x_t; m_t, P_t) \tag{4}$$

where:

- predicted obs mean: $z_t = Hm_t^- + d$
- predicted obs covariance: $S_t = HP_t^- H^T + R$
- Kalman gain $K_t = P_t^- H^T S_t^{-1}$
- corrected state mean: $m_t = m_t^- + K_t(y_t - z_t)$
- corrected state covariance: $P_t = (I - K_t H)P_t^-$

Missing observations If all the observations at time t are missing the correct step is skipped and the filtered state at time t (equation 4) is the same of the filtered state.
If only some observations are missing a variation of equation 4 can be used.
y_t^{ng} is a vector containing the observations that are not missing at time t.
It can be expressed as a linear transformation of y_t

$$y_t^{ng} = My_t$$

where M is a mask matrix that is used to select the subset of y_t that is observed. $M \in \mathbb{R}^{n_{ng} \times n}$ and is made of rows which are made of all zeros but for an entry 1 at column corresponding to the of the index non-missing observation.
For example if $y_t = [y_{0,t}, y_{1,t}, y_{2,t}]^T$ and $y_{0,t}$ is the missing observation then

$$M = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

hence:

$$p(y_t^{ng}) = \mathcal{N}(M\mu_{y_t}, M\Sigma_{y_t}M^T)$$

from which you can derive

$$p(y_t^{ng}|x_t) = p(MHx_t + Mb, MRM^T) \tag{5}$$

Then the posterior $p(x_t|y_t^{ng})$ can be computed similarly of equation 4 as:

$$p(x_t|y_t^{ng}) = \mathcal{N}(x_t; m_t, P_t) \tag{6}$$

where:

- predicted obs mean: $z_t = MHm_t^- + Md$

- predicted obs covariance: $S_t = MHP_t^-(MH)^T + MRM^T$
- Kalman gain $K_t = P_t^-(MH)^T S_t^{-1}$
- corrected state mean: $m_t = m_t^- + K_t(My_t - z_t)$
- corrected state covariance: $P_t = (I - K_t MH)P_t^-$

2.1.5 Kalman Smoother

- Kalman smoothing gain: $G_t = P_t A^T (P_{t+1}^-)^{-1}$
- smoothed mean: $m_t^s = m_t + G_t(m_{t+1}^s - m_{t+1}^-)$
- smoothed covariance: $P_t^s = P_t + G_t(P_{t+1}^s - P_{t+1}^-)G_t^T$

2.1.6 Predictions

The prediction at time t (y_t) are computed from the state (x_t) using:

$$p(y_t|x_t) = \mathcal{N}(Hx_t + d, R + HP_t^s H^T)$$

2.2 Filter Implementation

The filter has been implemented as a PyTorch module

- gradients
- batch support

2.2.1 Parameter constraints

- posdef
- diag posdef

2.2.2 Numerical stability

- min value of R?
- average

2.3 Loss Function

2.3.1 Joint distribution of the gap

The goal is to obtain the joint distribution of the variables in the gap Y^g, which is $[y_t^g, y_{t+1}^g ... y^g(t + t_g)]$ for a gap that goes from t to $t + t_g$. $Y^g \in \mathbb{R}^{t_g \times n_g}$, where n_g is the number of variables missing in the gap.
For simplicity we are assuming for now that during the gap the variables missing don't change.

The goal is to obtain $p(Y^g|Y^ng)$
From the Kalman smoother it's easy to obtain $p(y_t^g|Y^{ng}) = \mathcal{N}(\mu_t, \Sigma_t)$
However, the problem is that y_t^g and y_{t+1}^g are not independent so it gets more complex. Assuming that $p(y_t^g|y_{t+1}^g) = \mathcal{N}(\mu_{t,t+1}, \Sigma_{t,t+1})$ the joint distribution has the form:

$$p(Y^g|Y^{ng}) = \mathcal{N}\left(\begin{matrix} \mu_t \\ \mu_{t+1} \\ \cdots \\ \mu_{t+t_g} \end{matrix}, \begin{matrix} \Sigma_t & \Sigma_{t,t+1} & \cdots & \Sigma_{t,t+t_g} \\ \Sigma_{t+1,t} & \Sigma_{t+1} & \cdots & \Sigma_{t+1,t+t_g} \\ \vdots & \vdots & \ddots & \cdots \\ \Sigma_{t+t_g,t} & \Sigma_{t+t_g,t+1} & \cdots & \Sigma_{t+t_g} \end{matrix}\right)$$

$p(Y_g|Y_{ng}) = \int p(Y_g|X_g)p(X_g|Y)dX_g$

2.3.2 Joint distribution state for gaps

Two states For simplicity, I am starting with the joint distribution of the filter on a gap where there are no observations and are interested only on the joint distribution of two consecutive states. The aim is to find $p(x_t, x_{t+1} \mid x_t, Y_{1:t})$
The starting point is:

- $x_{t+1} = Ax_t + \varepsilon_{t+1}$
- $p(x_t \mid Y_{1:t}) = \mathcal{N}(x_t; m_t, P_t)$
- $p(\varepsilon_t) = \mathcal{N}(\varepsilon_t; 0, Q)$

Since all distributions are Gaussian, the joint distribution is also Gaussian

$$\begin{aligned} &p(x_t, x_{t+1}|x_t) = \mathcal{N}\left(\begin{bmatrix} x_t \\ x_{t+1} \end{bmatrix}; \begin{bmatrix} m_t \\ Am_t \end{bmatrix}, \Sigma_{x_t,x_{t+1}}\right) \\ &\Sigma_{x_t,x_{t+1}} = \begin{bmatrix} \langle (x_t - \mu_{x_t})(x_t - \mu_{x_t})^T) \rangle & \langle (x_t - \mu_{x_t})(x_{t+1} - \mu_{x_{t+1}})^T \rangle \\ \langle (x_{t+1} - \mu_{x_{t+1}})(x_t - \mu_{x_t})^T \rangle & \langle (x_{t+1} - \mu_{x_{t+1}})(x_{t+1} - \mu_{x_{t+1}})^T \rangle \end{bmatrix} \end{aligned} \tag{7}$$

we can compute the covariance using the expectation operator and its properties.

Second element on the diagonal

$$\begin{aligned} &\langle (x_{t+1} - \mu_{x_{t+1}})(x_{t+1} - \mu_{x_{t+1}})^T \rangle = \\ &= \langle (Ax_t + \varepsilon_{t+1} - Am_t)(Ax_t + \varepsilon_{t+1} - Am_t)^T \rangle = \\ &= \langle (A(x_t - m_t) + \varepsilon_{t+1})(A(x_t - m_t) + \varepsilon_{t+1})^T \rangle = \\ &= \langle A(x_t - m_t)(x_t - m_t)^T A^T + \varepsilon_{t+1}(x_t - m_t)^T A^T + A(x_t - m_t)\varepsilon_{t+1}^T + \varepsilon_{t+1}\varepsilon_{t+1}^T \rangle = \\ &= \langle A(x_t - m_t)(x_t - m_t)^T A^T \rangle + \langle \varepsilon_{t+1}(x_t - m_t)^T A^T \rangle + \langle A(x_t - m_t)\varepsilon_{t+1}^T \rangle + \langle \varepsilon_{t+1}\varepsilon_{t+1}^T \rangle = \\ &= A\langle (x_t - m_t)(x_t - m_t)^T \rangle A^T + 0 + 0 + \langle \varepsilon_{t+1}\varepsilon_{t+1}^T \rangle = \\ &= AP_tA^T + Q \end{aligned} \tag{8}$$

off-diagonal element

$$\begin{aligned}
&\langle (x_{t+1} - \mu_{x_{t+1}})(x_t - \mu_{x_t}^T\rangle = \langle (Ax_t + \varepsilon_{t+1} - Am_t)(x_t - Am_t)^T\rangle = \\
&= \langle A(x_t - m_t)(x_t - m_t)^T + \varepsilon_{t+1}(x_t - m_t)^T\rangle = \\
&= \langle A(x_t - m_t)(x_t - m_t)^T\rangle + \langle \varepsilon_{t+1}(x_t - m_t)^T A^T\rangle = \\
&= A\langle (x_t - m_t)^T\rangle + 0 = \\
&= AP_t
\end{aligned} \tag{9}$$

Joint distribution state substituting in equation 7:

$$p(x_t, x_{t+1} \mid x_t, Y_{1:t}) = \mathcal{N}\left(\begin{bmatrix} x_t \\ x_{t+1} \end{bmatrix}; \begin{bmatrix} m_t \\ Am_t \end{bmatrix}, \begin{bmatrix} P_t & AP_t \\ AP_t & AP_tA^T + Q \end{bmatrix}\right) \tag{10}$$

Multiple States A similar reasoning can be applied to more than two states, but the equations become more complex
To obtain $p(x_t, x_{t+1}, x_{t+2} \mid x_t, Y_{1:t})$ we also need to compute $\langle x_t x_{t+2}^T\rangle$ and $\langle x_{t+2} x_{t+2}^T\rangle$

Covariance diagonal

$$\begin{aligned}
&\langle (x_{t+2} - \mu_{x_{t+2}})(x_{t+2} - \mu_{x_{t+2}})^T\rangle = \\
&= \langle (A(Ax_t + \varepsilon_{t+1}) + \varepsilon_{t+2} - AAm_t)(A(Ax_t + \varepsilon_{t+1}) + \varepsilon_{t+2} - AAm_t)^T\rangle = \\
&= \langle (AAx_t + A\varepsilon_{t+1} + \varepsilon_{t+2} - AAm_t)(AAx_t + A\varepsilon_{t+1} + \varepsilon_{t+2} - AAm_t)^T\rangle = \\
&= \langle AA(x_t - m_t)(x_t - m_t)^T A^T A^T\rangle + \langle A\varepsilon_{t+1}\varepsilon t + 1^T A^T\rangle + \langle \varepsilon_{t+2}\varepsilon_{t+2}^T\rangle = \\
&= AAP_t(AA)^T + AQA^T + Q
\end{aligned} \tag{11}$$

which (probably) can be generalized as: [TODO actually need to prove this and check that notation is correct]

$$\langle (x_t - \mu_{x_t})(x_{t+k} - \mu_{x_{t+k}})^T\rangle = A^k P_t (A^k)^T + \sum_{i=0}^{k-1} A^i Q (A^i)^T \tag{12}$$

Covariance off-diagonal

$$\langle (x_{t+k} - \mu_{x_{t+k}})(x_{t+k} - \mu_{x_{t+k}})^T\rangle = A^k P_t (A^k)^T \tag{13}$$

Mean

$$\langle x_{t+k}\rangle = A^k m_t \tag{14}$$

Joint distribution state In this way it is possible to obtain $P(X)$ for any number of states.

$$p(X_{t:t+k} \mid x_t, Y_{1:t}) = \mathcal{N}\left(\begin{bmatrix} x_t \\ \vdots \\ x_{t+k} \end{bmatrix}; \begin{bmatrix} m_t \\ \vdots \\ A^k m_t \end{bmatrix}, \begin{bmatrix} P_t & \cdots & A^k P_t (A^k)^T \\ \vdots & \ddots & \vdots \\ A^k P_t (A^k)^T & \cdots & A P_t (A^k)^T + \sum_{i=0}^{k-1} A^i Q (A^i)^T \end{bmatrix}\right) \tag{15}$$

2.3.3 Joint distribution state - partial observations

In the case the there are partial observations to the reasoning of the previous paragraph cannot be applied as by combining equations 3 and 6

$$\begin{aligned} m_t^- &= A m_{t-1} + B c_t + d \\ P_t^- &= A P_{t-1} A^T + Q \\ z_t &= M H m_t^- + M d \\ S_t &= M H P_t^- (MH)^T + M R M^T \\ K_t &= P_t^- (MH)^T S_t^{-1} \\ m_t &= m_t^- + K_t (M y_t - z_t) \\ P_t &= (I - K_t M H) P_t^- \\ p(x_t | x_{t-1}, y_t^{ng}) &= \mathcal{N}(x_t; m_t, P_t) \end{aligned} \tag{16}$$

From this equation is not possible to write x_t and linear map of x_{t-1} plus another random variable, since the mean of x_t depends on the covariance of x_{t-1}

For the same reason this approach cannot be applied for the smoother.

3 Results

4 Discussion

- comparison other approaches: MDS, GPFA
- performance