10-708: Probabilistic Graphical Models, Spring 2015

11 : Factor Analysis and State Space Models

Lecturer: Eric P. Xing *Scribes: Yilin He*

1 Overview

In this lecture, we introduced **Factor Analysis** and **Space State Model**.

Factor Analysis is a latent variable model with continuous random vector as the latent variable. For Factor Analysis, we assume the measured data vector lies near a lower-dimensional manifold. We then model the data in two-stages: First, generate a point in the manifold according to a probability density. Then observed data is generated from another density based on the point we generated in the first step. In this sense, Factor Analysis is very similar to a mixture model except its latent variable is continuous.

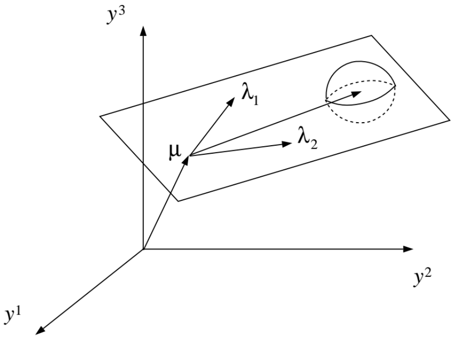
Space State Models (SSM) can be considered as a generalization to the traditional HMM model where the latent variable has continuous values. Therefore SSM and HMM share the same inference problem: calculate the conditional probability of latent variable given observed data. The inference problem can then be divided into two type of problems: filtering(forward) inference and smoothing(backward) inference.

In this notes, we will first discuss Factor Analysis in section 2. Then in section 3, we will talk about SSM in detail. Section 4 is an appendix which contains the mathematical background information for the notes.

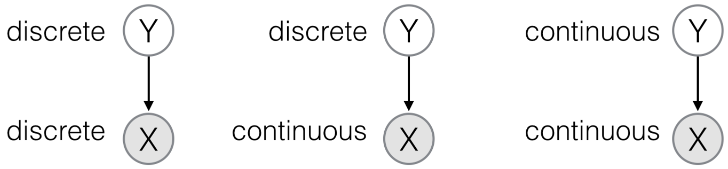
2 Factor Analysis

2.1 Introduction

We can think of Factor Analysis as generating a point x on a linear subspace based on some Gaussian distribution. Then we observe the data y which is conditionally generated from a Gaussian distribution centered at point x with some noise.



Factor analysis is more complex graphical model compared to mixture models. In mixture model, the latent variables are discrete variables while observed variables can be either discrete or continuous. Mixture models can be later built into HMMs while Factor Analysis leads to State Space Models.



Mixture model
e.g., mixture of multinomials

Mixture model
e.g., mixture of Gaussians

Factor analysis

2.2 Parameterization

In Factor Analysis model, variable X is a latent Gaussian variable with p dimensions and Y is the observed variable with q dimensions. We assume $p < q$. And the model is parameterized as follows, let X be a marginal Gaussian distribution with mean 0 and identity covariance matrix. Let the conditional distribution Y be a Gaussian distribution with mean $\mu + \Lambda x$ and diagonal covariance matrix Ψ

$$
\begin{gathered}
X \sim \mathcal{N}(0, I) \\
Y|X \sim \mathcal{N}(\mu + \Lambda x, \Psi)
\end{gathered}
$$

Since Both X and Y are Gaussian distribution, we may conclude that their joint distribution and the marginal distribution of Y is also Gaussian. Therefore we can calculate the marginal of Y and the conditional distribution of X given Y by calculating their mean and variance.

In order to calculate the marginal distribution, we can express Y as a sum:

$$Y = \mu + \Lambda x + W$$

where W is a distribution as $\mathcal{N}(0, \Psi)$ and it's independent of X.

$$
\begin{aligned}
\text{E(Y)} &= E(\mu + \Lambda X + W) \\
&= \mu + \Lambda EX + EW \\
&= \mu \\
Var(Y) &= \text{E}\big[(\mu + \Lambda X + W - \mu)(\mu + \Lambda X + W - \mu)^T\big] \\
&= E\left[(\Lambda X + W)(\Lambda X + W)^T\right] \\
&= \Lambda E\left(XX^T\right)\Lambda^T + E(WW^T) \\
&= \Lambda\Lambda^T + \Psi
\end{aligned}
$$

Another way of calculating the mean and variance for marginal distribution is by law of total variance and law of total expectation

$$\begin{aligned} E\left(Y\right) &= E\left(\mu + \Lambda X\right) \\ &= \mu \\ Var\left(Y\right) &= Var\left(\mu + \Lambda X\right) + E\Psi \\ &= E\left[\left(\Lambda X\right)\left(\Lambda X\right)^T\right] + \Psi \\ &= \Lambda\Lambda^T + \Psi \end{aligned}$$

In addition, we also need to calculate the covarince between X and Y

$$\begin{aligned} Cov(X,Y) &= E\left[X\left(\mu + \Lambda X + W - \mu\right)\right] \\ &= E\left[X\left(\Lambda X + W\right)^T\right] \\ &= \Lambda^T \end{aligned}$$

Finally we have the joint distribution for X, Y:

$$X, Y \sim \mathcal{N}(\begin{bmatrix} 0 \\ \mu^T \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix})$$

We now have the conditional distribution of latent variable X given observed variable Y.

$$p(X|y) \sim \mathcal{N}(\left(I + \Lambda^T\Psi^{-1}\Lambda\right)^{-1}\Lambda^T\Psi^{-1}\left(y - \mu\right), I + \Lambda^T\Psi^{-1}\Lambda)^{-1})$$

2.3 Inference

We will calculate the conditional distribution of $X|Y$:

$$E(X|y) = \Lambda^T\left(\Lambda\Lambda^T + \Psi\right)^{-1}\left(y - \mu\right)$$

Since this calculation requires inverting a $q \times q$ matrix, we can rewrite this equation to the following way so we only need to inver a $p \times p$ dimensional matrix. This is preferred because $p < q$.

$$E(X|y) = \left(I + \Lambda^T\Psi^{-1}\Lambda\right)^{-1}\Lambda^T\Psi^{-1}\left(y - \mu\right)$$

We then compute the variance of X. For similar reason we prefer the second form.

$$\begin{aligned} Var(X|y) &= I - \Lambda^T\left(\Lambda\Lambda^T + \Psi\right)^{-1}\Lambda \\ &= (I + \Lambda^T\Psi^{-1}\Lambda)^{-1} \end{aligned}$$

Now we can solve the problem of finding maximum likelihood estimates of the parameters for our model by looking at the likelihood of marginal probability. The log likelihood is a Gaussian log likelihood:

$$l\left(\theta|D\right) = -\frac{N}{2}log|\Lambda\Lambda^T + \Psi| - \frac{1}{2}\left\{\sum_n\left(y_n - \mu\right)^T\left(\Lambda\Lambda^T + \Psi\right)^{-1}\left(y_n - \mu\right)\right\}$$

and we can obtain the usual maximum likelihood estimate which is sample mean:

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_n y_n$$

2.4 Learning with EM

In this section we will derive the EM algorithm for Factor Analysis. Suppose we have the complete data, then the estimation of X reduce to a Gaussian density estimation and Y is a linear function of x with additive white Gaussian noise W. Therefore In E step, we will "fill in" X, then in M step, we will estimate Λ and Ψ using linear regression.

E step

First we compute the complete likelihood, which is a product of Gaussian distribution. Then we take its logarithm,

$$\begin{aligned} l_c\left(\theta_c|D\right) &= -\frac{N}{2}log|\Psi| - \frac{1}{2}\sum_n tr\left(x_n^T x_n\right) - \frac{1}{2}\sum_n tr\left[\left(y_n - \Lambda x_n\right)^T \Psi^{-1}\left(y_n - \Lambda x_n\right)\right] \\ &= -\frac{N}{2}log|\Psi| - \frac{1}{2}\sum_n tr\left(x_n^T x_n\right) - \frac{1}{2}\sum_n tr\left[\left(y_n - \Lambda x_n\right)^T \left(y_n - \Lambda x_n\right)\Psi^{-1}\right] \\ &= -\frac{N}{2}log|\Psi| - \frac{N}{2}tr\left(S\Psi^{-1}\right) \end{aligned}$$

where we define

$$S = \frac{1}{N}\sum_n \left(y_n - \Lambda x_n\right)\left(y_n - \Lambda x_n\right)^T$$

We now take the conditional expectation of the complete log likelihood.

$$Q\left(\theta|\theta^{(t)}\right) = -\frac{N}{2}log|\Psi| - \frac{N}{2}tr\left(\langle S\rangle \Psi^{-1}\right)$$

where,

$$\begin{aligned} \langle S\rangle &= \frac{1}{N}\sum_n \left\langle y_n y_n^T - y_n X_n^T \Lambda^T - \Lambda X_n y_n^T + \Lambda X_n X_n^T \Lambda^T\right\rangle \\ &= \frac{1}{N}\sum_n \left(y_n y_n^T - y_n\left\langle X_n^T\right\rangle \Lambda^T - \Lambda\left\langle X_n\right\rangle y_n^T + \Lambda\left\langle X_n X_n^T\right\rangle \Lambda^T\right) \end{aligned}$$

From previous section we have already obtained these expectations.

$$\begin{aligned} \langle X_n\rangle &= E\left(X_n|Y_n\right) \\ \left\langle X_n X_n^T\right\rangle &= Var\left(X_n|Y_n\right) + E\left(X_n|Y_n\right)E\left(X_n|Y_n\right)^T \end{aligned}$$

M step

In order to update Λ and Ψ, we compute the derivative of Q with respect to each variable:

$$\begin{aligned}
Q\left(\Lambda|\theta^{(t)}\right) &= \frac{1}{N}\sum_n tr\left(y_n y_n^T - y_n \left\langle X_n^T\right\rangle \Lambda^T - \Lambda \left\langle X_n\right\rangle y_n^T + \Lambda \left\langle X_n X_n^T\right\rangle \Lambda^T \Psi^{-1}\right)\\
\frac{\partial Q}{\partial \Lambda} &= \sum_n \Psi^{-1} y_n \left\langle X_n^T\right\rangle - \sum_n \Psi^{-1}\Lambda \left\langle X_n X_n^T\right\rangle = 0\\
\Lambda^{t+1} &= \left(\sum_i y_i \left\langle x_i^T\right\rangle\right)\left(\sum_i \left\langle x_i x_i^T\right\rangle\right)^{-1}\\
Q\left(\theta|\theta^{(t)}\right) &= -\frac{N}{2} log|\Psi| - \frac{N}{2} tr\left(\left\langle S\right\rangle \Psi^{-1}\right)\\
\frac{\partial Q}{\partial \Psi} &= \frac{N}{2}\Psi - \frac{N}{2}\left\langle S\right\rangle = 0\\
\Psi^{t+1} &= \left\langle S\right\rangle
\end{aligned}$$

3 State Space Model (SSM)

3.1 Introduction

State space model can be considered as a sequential FA or a continuous state HMM. It's structured identical to HMM with real-valued nodes and linear-Gaussian probability model. In the notes we will develop the inference method which is Kalman filter. To represent the transition between nodes, we can allow the mean of the state at time $t+1$ to be a linear function of the state at time t. So,

$$\begin{aligned}
x_{t+1} &= AX_t + Gw_t\\
y_t &= CX_t + v_t
\end{aligned}$$

where w_t and v_t are Gaussian noise which is independent from the noise in previous states.

$$\begin{aligned}
w &\sim \mathcal{N}(0, Q)\\
v &\sim \mathcal{N}(0, R)
\end{aligned}$$

Then we have x_{t+1} as a Gaussian distribution

$$x_{t+1}|x_t \sim \mathcal{N}(Ax_t, GQG^T)$$

In State Space Model, the mean of observed value is a linear function of the state. Therefore we have,

$$y_t|x_t \sim \mathcal{N}(Cx_t, R)$$

Finally, the initial state x_0 also follows a Gaussian distribution:

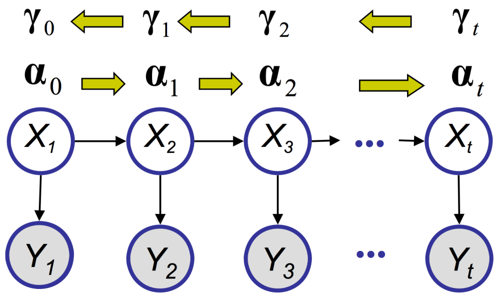
$$x_0 \sim \mathcal{N}(0, \Sigma_0)$$

3.2 Inference Problems

- Filtering: given $y_1, y_2, ..., y_t$, estimate x_t. This is same as calculating $P(x_t|y_{1:t})$

Kalman filter: perform exact online inference (sequential Bayesian updating) in an LDS. It is the Gaussian analog of the forward algorithm for HMMs.

$$p(X_t = i|y_{1:t}) = \alpha_t^i \propto p(y_t|X_t = i) \sum_j P(X_t = i|X_{t-1} = j)\alpha_{t-1}^j$$



- Smoothing: given $y_1, y_2, ..., y_t$, estimate $x_t (t < T)$

  The Rauch-Tung-Strievel smoother is a way to perform exact off-line inference in an LDS. It is the Gaussian analog of the forwards-backwards (alpha-gamma) algorithm:

$$p(X_t = i|y_{1:T}) = \gamma_t^i \propto \sum_j \alpha_t^i P(X_{t+1}^j|X_i^j)\gamma_{t+1}^j$$

3.3 Kalman Filtering Derivation

Assumption

Let t denote time. Assume we have a linear dynamic model for latent states and a observations are derived from latent state linearly, e.g.

$$\begin{aligned} \boldsymbol{x}_t &= A\boldsymbol{x}_{t-1} + G\boldsymbol{w}_{t-1} \\ \boldsymbol{y}_t &= C\boldsymbol{x}_t + \boldsymbol{v}_t \end{aligned}$$

where $\boldsymbol{x}_t \in \mathbb{R}^n$ denote the latent state and $\mathbf{y}_t \in \mathbb{R}^m$ denote the observation at time t, $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times m}$, $G\boldsymbol{w} \sim \mathcal{N}(0; Q)$, $\boldsymbol{v}_t \sim \mathcal{N}(0; R)$, and $\boldsymbol{x}_0 \sim \mathcal{N}(0; \Sigma_0)$.

Two Steps

Kalman filtering is a recursive procedure to update the latent state $\boldsymbol{x}_t$. In each iteration, there are two steps:

Predict Step Compute predicted latent state distribution $p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t})$ from prior belief $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ and dynamic model $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$. This step is also called time update.

Update Step Compute new belief of the latent state distribution $p\left(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t+1}\right)$ from prediction $p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t})$ and observation $\boldsymbol{y}_{t+1}$ by using the observation model $p\left(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}\right)$. The step is also called measurement update since it's using the measured information $\boldsymbol{y}_{t+1}$.

Since all distributions are normal, their linear combinations are also normal. Hence we just need the expectation and variance to describe each distribution.

Derivation

The goal is to compute $p(x_{t+1}|y_{1:t+1})$. We will do so by first compute the prediction $p\left(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}\right)$ by using the dynamic model. Then, we use the observation model to find the joint distribution $p\left(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right)$, note the joint distribution is also multivariate Gaussian. Finally compute $p\left(\boldsymbol{x}_{t+1}|\mathbf{y}_{t+1}, \boldsymbol{y}_{1:t}\right)$ with formula of conditional Gaussian distributions. The final term is just our goal.

First we need to compute $p\left(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}\right)$. Since

$$p\left(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}\right) = p\left(\boldsymbol{x}_{t+1}|\boldsymbol{x}_{t}\right) p\left(\boldsymbol{x}_{t}|\boldsymbol{y}_{1:t}\right) = p\left(A\boldsymbol{x}_{t} + G\boldsymbol{w}_{t+1}|\boldsymbol{y}_{1:t}\right),$$

we have

$$\hat{x}_{t+1|t} = \mathbb{E}\left(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}\right) = \mathbb{E}\left(A\boldsymbol{x}_{t} + G\boldsymbol{w}_{t+1}|\boldsymbol{y}_{1:t}\right) = A\mathbb{E}\left(\boldsymbol{x}_{t}|\boldsymbol{y}_{1:t}\right) + \mathbb{E}\left(G\boldsymbol{w}_{t+1}|\boldsymbol{y}_{1:t}\right) = A\hat{x}_{t|t} + \mathbf{0} = A\hat{x}_{t|t}$$

and

$$\begin{aligned}
P_{t+1|t} &= \mathbb{E}\left(\left(\boldsymbol{x}_{t+1} - \hat{x}_{t+1|t}\right)\left(\boldsymbol{x}_{t+1} - \hat{x}_{t+1|t}\right)^{T} |\boldsymbol{y}_{1:t}\right) \\
&= \mathbb{E}\left(\left(A\boldsymbol{x}_{t} + G\boldsymbol{w}_{t} - \hat{x}_{t+1|t}\right)\left(A\boldsymbol{x}_{t} + G\boldsymbol{w}_{t} - \hat{x}_{t+1|t}\right)^{T} |\boldsymbol{y}_{1:t}\right) \\
&= \mathbb{E}\left(\left(A\boldsymbol{x}_{t} + G\boldsymbol{w}_{t} - A\hat{x}_{t|t}\right)\left(A\boldsymbol{x}_{t} + G\boldsymbol{w}_{t} - A\hat{x}_{t|t}\right)^{T} |\boldsymbol{y}_{1:t}\right) \\
&= \mathbb{E}\left(\left(A\boldsymbol{x}_{t} - A\hat{x}_{t|t}\right)\left(A\boldsymbol{x}_{t} - A\hat{x}_{t|t}\right)^{T} + G\boldsymbol{w}_{t}\left(A\boldsymbol{x}_{t} - A\hat{x}_{t|t}\right)^{T} + \left(A\boldsymbol{x}_{t} - A\hat{x}_{t|t}\right) G\boldsymbol{w}_{t}^{T} + G\boldsymbol{w}_{t}G\boldsymbol{w}_{t}^{T}|\boldsymbol{y}_{1:t}\right) \\
&= A\mathbb{E}\left(\left(\boldsymbol{x}_{t} - \hat{x}_{t|t}\right)\left(\boldsymbol{x}_{t} - \hat{x}_{t|t}\right)^{T} |\boldsymbol{y}_{1:t}\right) A^{T} + G\boldsymbol{w}_{t}\mathbb{E}\left(\boldsymbol{x}_{t} - \hat{x}_{t|t}\right)^{T} A^{T} + A\mathbb{E}\left(\boldsymbol{x}_{t} - \hat{x}_{t|t}|\boldsymbol{y}_{1:t}\right)\boldsymbol{w}_{t}^{T} G^{T} + G\boldsymbol{w}_{t}\boldsymbol{w}_{t}^{T} G^{T} \\
&= AP_{t|t}A^{T} + \mathbf{0} + \mathbf{0} + GQG^{T} \\
&= AP_{t|t}A^{T} + GQG^{T}.
\end{aligned}$$

where $\hat{x}_{t+1|t}$ and $P_{t+1|t}$ denote the expectation and variance of $p\left(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}\right)$ and $\hat{x}_{t|t}$ and $P_{t|t}$ denote the expectation and variance of $p\left(\boldsymbol{x}_{t}|\boldsymbol{y}_{1:t}\right)$, respectively.

Next, we will find the joint distribution $p\left(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right)$, or equivalently the expectation $\boldsymbol{m}_{t+1|t}$ and variance $V_{t+1|t}$ where

$$\hat{x}_{\boldsymbol{x}_{t+1},\boldsymbol{y}_{t+1}|t} = \begin{bmatrix} \hat{x}_{\boldsymbol{x}_{t+1}|t} \\ E[\boldsymbol{y}_{t+1}|t] \end{bmatrix}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{x}_{t+1},\boldsymbol{y}_{t+1}|t} = \begin{bmatrix} P_{t+1|t} & \operatorname{Cov}\left(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right) \\ \operatorname{Cov}\left(\boldsymbol{y}_{t+1}, \boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}\right) & \operatorname{Var}\left(\boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right) \end{bmatrix}.$$

So

$$E[\boldsymbol{y}_{t+1}|t] = \mathbb{E}\left(\boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right) = \mathbb{E}\left(C\boldsymbol{x}_{t+1} + \boldsymbol{v}_{t+1}|\boldsymbol{y}_{1:t}\right) = C\hat{x}_{t+1|t}$$

$$\begin{aligned}
\operatorname{Cov}\left(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right) &= \mathbb{E}\left(\left(\boldsymbol{x}_{t+1} - \hat{x}_{t+1|t}\right)\left(\boldsymbol{y}_{t+1} - \hat{x}_{\boldsymbol{y}_{t+1|t}}\right)^{T} |\boldsymbol{y}_{1:t}\right) \\
&= \mathbb{E}\left(\left(\boldsymbol{x}_{t+1} - \hat{x}_{t+1|t}\right)\left(C\boldsymbol{x}_{t+1} + \mathbf{v}_{t+1} - C\hat{x}_{t+1|t}\right)^{T} |\boldsymbol{y}_{1:t}\right) \\
&= \mathbb{E}\left(\left(\boldsymbol{x}_{t+1} - \hat{x}_{t+1|t}\right)\left(\boldsymbol{x}_{t+1} - \hat{x}_{t+1|t}\right)^{T} |\boldsymbol{y}_{1:t}\right) C^{T} \\
&= P_{t+1|t}C^{T}
\end{aligned}$$

$$\operatorname{Cov}\left(\boldsymbol{y}_{t+1}, \boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}\right) = \operatorname{Cov}\left(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right)^{T} = CP_{t+1|t}^{T} = CP_{t+1|t}$$

$$
\begin{aligned}
\operatorname{Var}\left(\boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right) &= \mathbb{E}\left(\left(\boldsymbol{y}_{t+1} - \hat{x}_{\boldsymbol{y}_{t+1|t}}\right)\left(\boldsymbol{y}_{t+1} - \hat{x}_{\boldsymbol{y}_{t+1|t}}\right)^T |\boldsymbol{y}_{1:t}\right) \\
&= \mathbb{E}\left(\left(C\boldsymbol{x}_{t+1} + \boldsymbol{v}_{t+1} - C\hat{x}_{t+1|t}\right)\left(C\boldsymbol{x}_{t+1} + \mathbf{v}_{t+1} - C\hat{x}_{t+1|t}\right)^T |\boldsymbol{y}_{1:t}\right) \\
&= CP_{t+1|t}C^T + R
\end{aligned}
$$

The last derivation is similar to the one for $P_{t+1|t}$ so it's not repeated. Let

$$
\begin{aligned}
K &= \operatorname{Cov}\left(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right)\operatorname{Var}\left(\boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right)^{-1} \\
&= P_{t+1|t}C^T\left(CP_{t+1|t}C^T + R\right)^{-1} \\
&= AP_{t|t}A^TC^T\left(CAP_{t|t}A^TC^T + R\right)^{-1}
\end{aligned}
$$

and we call K the Kalman gain matrix. Note that the computation of K doesn't require a new observation, so it can be precomputed in next observation is acquired. Next, to find $p\left(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t+1}\right) = p\left(\boldsymbol{x}_{t+1}|\mathbf{y}_{t+1}, \boldsymbol{y}_{1:t}\right)$, by the formula for conditional Gaussian distribution, we have

$$
\begin{aligned}
\hat{x}_{\boldsymbol{x}_{t+1|t+1}} &= \hat{x}_{t+1|t} + \operatorname{Cov}\left(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right)\operatorname{Var}\left(\boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right)^{-1}\left(\boldsymbol{y}_{t+1} - \hat{x}_{\boldsymbol{y}_{t+1|t}}\right) \\
&= A\hat{x}_{t|t} + K\left(\boldsymbol{y}_{t+1} - CA\hat{x}_{t|t}\right)
\end{aligned}
$$

$$
\begin{aligned}
P_{t+1|t+1} &= P_{t+1|t} + \operatorname{Cov}\left(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right)\operatorname{Var}\left(\boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t}\right)^{-1}\operatorname{Cov}\left(\boldsymbol{y}_{t+1}, \boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}\right) \\
&= AP_{t|t}A^T + Q + KCAP_{t|t}A^T.
\end{aligned}
$$

We are done here for the derivation.

Complexity

Let's look at the complexity of one Kalman Filtering step.

Let $x \in R^{N_x}$, $y \in R^{N_y}$ and assume dense matrix P and dense matrix A, computing

$$P_{t+1|t} = AP_{t|t}A^T + GQG^T$$

takes $O(N_x^2)$ time. And pre-computing the Kalman Gain Matrix

$$k_{t+1} = P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1}$$

takes $O(N_y^3)$ time. Therefore the total complexity of Kalman Filter is $\max(O(N_x^2), O(N_y^3))$. Because of its high complexity, Kalman Filter is not widely used today.

3.4 Rauch-Tung-Striebel smoother

RTS smoother is an algorithm to compute optimal latent state distributions using all observations from time 0 to time T. That is we want to compute $p\left(\boldsymbol{x}_t|\boldsymbol{y}_{1:T}\right)$ for all $0 < t \leq T$.

The algorithm has two steps

- Forward inference (exactly same process as Kalman Filter)
- Backward inference

The backward step starts from time T. That is we start with $p\left(\boldsymbol{x}_T|\boldsymbol{y}_{1:T}\right)$, which is the last latent state distribution computed by forward inference. Next we update previous latent states with the following recursive relationship:

$$\hat{x}_{t|t} = \hat{x}_{t|t} + L_t\left(\hat{x}_{t+1|t} - \hat{x}_{t+1|t}\right)$$

$$P_{t|t} = P_{t|t} + L_t\left(P_{t+1|t} - P_{t+1|t}\right)L_t^T$$

where

$$L_t = P_{t|t}A^T P_{t+1|t}^{-1}.$$

We are basically trying to get better knowledge of the current latent state by using information in the future up to time T. We treat $p\left(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}\right)$ and $p\left(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}\right)$ as the base and hope to improve the first distribution by introducing $p\left(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:T}\right)$, a "better truth" than $\boldsymbol{y}_{t+1}$, which is used in the forwarding inference. $p\left(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:T}\right)$ is better because it's a snapshot of all future information giving the linear Markov process assumption, while $\boldsymbol{y}_{t+1}$ is just one observation.

Derviation of Backward Inference

First, we calculate the joint distribution of x_t and x_{t+1}. Since we have $\hat{x}_{t+1|t} = A\hat{x}_{t|t}$, we know that

$$\mathrm{Cov}\left(\boldsymbol{x}_{\mathrm{t}}, \boldsymbol{x}_{\mathrm{t+1}}|\boldsymbol{y}_{\mathrm{0:t}}\right) = \mathbb{E}\left[\left(\boldsymbol{x}_t - \hat{x}_{t|t}\right)\left(\boldsymbol{x}_{t+1} - \hat{x}_{t+1|t}\right)|\boldsymbol{y}_{0:t}\right] = \Sigma_{t|t}A^T$$

Therefore we have

$$p\left(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}|\boldsymbol{y}_{0:t}\right) \sim \mathcal{N}\left(\begin{bmatrix}\hat{x}_{t|t}\\ \hat{x}_{t+1|t}\end{bmatrix}, \begin{bmatrix}P_{t|t} & P_{t|t}A^T\\ AP_{t|t} & P_{t+1|t}\end{bmatrix}\right)$$

We can obtain all the variables after a forward Kalman filtering pass. Now we move to the backward computation. where we want to compute $p\left(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}, \boldsymbol{y}_{0:t}\right)$.

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{x}_t|\boldsymbol{x}_{t+1|\boldsymbol{y}_{0:t}}\right] &= \hat{x}_{x_{t|t}} + P_{t|t}A^T P_{t+1|t}^{-1}\left(\boldsymbol{x}_{t+1} - \hat{x}_{t+1|t}\right)\\ &= \hat{x}_{x_{t|t}} + L_t\left(\boldsymbol{x}_{t+1} - \hat{x}_{t+1|t}\right)\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\left[\boldsymbol{x}_t|\boldsymbol{x}_{t+1|\boldsymbol{y}_{0:t}}\right] &= P_{t|t} - P_{t|t}A^T P_{t+1|t}^{-1}AP_{t|t}\\ &= P_{t|t} - L_t P_{t+1|t}^{-1}L_t^T\end{aligned}$$

Since x_t is conditional independent of $y_{t+1}, ..., y_T$ given x_{t+1}, we have

$$\begin{aligned}E\left[x_t, x_{t+1}|y_0, \ldots, y_T\right] &= E\left[x_t, x_{t+1}|y_0, \ldots, y_t\right]\\ &= \hat{x}_{t|t} + L_t\left(x_{t+1} - \hat{x}_{t+1|t}\right)\end{aligned}$$

$$\begin{aligned}Var\left[x_t, x_{t+1}|y_0, \ldots, y_T\right] &= Var\left[x_t, x_{t+1}|y_0, \ldots, y_t\right]\\ &= P_{t|t} - L_t P_{t+1|t}^{-1}L_t^T\end{aligned}$$

By law of total expectation and law of total variance, we can get:

$$\begin{aligned}
\hat{x}_{t|T} &= E\left[x_t|y_0,\ldots,y_T\right] \\
&= E\left[E\left[x_t|x_{t+1},y_0,\ldots,y_T\right]|y_0,\ldots,y_T\right] \\
&= E\left[\hat{x}_{t|t} + L_t\left(x_{t+1} - \hat{x}_{t+1|t}\right)|y_0,\ldots,y_T\right] \\
&= \hat{x}_{t|t} + L_t\left(x_{t+1|T} - \hat{x}_{t+1|t}\right) \\
P_{t|T} &= P_{t|t} + L_t\left(P^{-1}_{t+1|T} - P^{-1}_{t+1|t}\right)L^T
\end{aligned}$$

4 Review of Mathematical Background

4.1 Multivariate Gaussian

Multivariate Gaussian Density: Let's recall the pdf for a Gaussian distribution is of the following form:

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

If we represent a multivariate Gaussian distribution in the following block form:

$$p(\begin{bmatrix}\mathbf{x_1}\\\mathbf{x_2}\end{bmatrix}|\mu,\Sigma) = N(\begin{bmatrix}\mathbf{x_1}\\\mathbf{x_2}\end{bmatrix}|\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix},\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\\Sigma_{21} & \Sigma_{22}\end{bmatrix})$$

Then we can represent the marginal probability and conditional probability with μ and Σ:

$$\begin{aligned}
p(\mathbf{x}_2) &= N(\mathbf{x}_2|\mathbf{m}_2^m,\mathbf{V}_2^m) \\
\mathbf{m}_2^m &= \mu_2 \\
\mathbf{V}_2^m &= \Sigma_{22}
\end{aligned}
\qquad
\begin{aligned}
p(\mathbf{x}_1|\mathbf{x}_2) &= N(\mathbf{x}_1|\mathbf{m}_{1|2},\mathbf{V}_{1|2}) \\
\mathbf{m}_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\mu_2) \\
\mathbf{V}_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}
\end{aligned}$$

4.2 Matrix Inversion

It is also useful to remember the matrix inversion for block matrix. In particular, if we consider the following block matrix M:

$$M = \begin{bmatrix}E & F\\G & H\end{bmatrix}$$

We can derve the inverse of matrix M^{-1} as follows:

$$\begin{aligned}
M^{-1} = \begin{bmatrix}E & F\\G & H\end{bmatrix}^{-1} &= \begin{bmatrix}I & 0\\-H^{-1}G & I\end{bmatrix}\begin{bmatrix}(M/H)^{-1} & 0\\0 & H^{-1}\end{bmatrix}\begin{bmatrix}I & -FH^{-1}\\0 & I\end{bmatrix} \\
&= \begin{bmatrix}E^{-1}+E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1}\\G-(M/E)^{-1}GE^{-1} & (M/E)^{-1}\end{bmatrix}^{-1}
\end{aligned}$$

We also get the matrix inverse lemma:

$$\left(E-FH^{-1}G\right)^{-1} = E^{-1} + E^{-1}F\left(H-GE^{-1}F\right)^{-1}GE^{-1}$$

4.3 Some Matrix Algebra

Finally we look at some useful tricks about trace and derivative of the matrix. The trace of a matrix is defined as follows:

$$tr\left[A\right] = \sum_i a_{ii}$$

Cyclical permutations:

$$tr\left[ABC\right] = tr\left[CAB\right] = tr\left[BCA\right]$$

Taking derivatives of a trace:

$$\frac{\partial tr\left[BA\right]}{\partial A} = B^T$$
$$\frac{\partial tr\left[x^T Ax\right]}{\partial A} = \frac{\partial tr\left[xx^T A\right]}{\partial A} = xx^T$$

In addition to trace, it is also important to know how to take the derivatives of the determinants:

$$\frac{\partial log|A|}{\partial A} = A^{-1}$$