10-708: Probabilistic Graphical Models 10-708, Spring 2016

11 : Factor Analysis and State Space Models

Lecturer: Eric P. Xing *Scribes: Rahul Nallamothu, Syed Zahir Bokhari, Yu Zhang*

1 Background Review

In this section, we will review some of mathematical concepts that will be used later in the material.

1.1 Multivariate Gaussian

The pdf of joint Gaussian distribution of x_1, x_2 can be written in block form as

$$p\left(\begin{bmatrix}x_1\\x_2\end{bmatrix} \mid \begin{bmatrix}\mu\\\Sigma\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}; \begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\\Sigma_{21} & \Sigma_{22}\end{pmatrix}\right)$$

The joint probability can also be written as:

$$p(x_1, x_2) = \frac{1}{(2\pi)^{N/2}|E|^{1/2}} exp\Big([(x_1-\mu_1)^T, (x_2-\mu_2)^T]\begin{bmatrix}\Omega_{11} & \Omega_{12}\\\Omega_{21} & \Omega_{22}\end{bmatrix}[(x_1-\mu_1), (x_2-\mu_2)]^T\Big)$$

where

$$\Omega = \Sigma^{-1} = \begin{bmatrix}\Omega_{11} & \Omega_{12}\\\Omega_{21} & \Omega_{22}\end{bmatrix}$$

$$\begin{aligned}
\Omega_{11} &= \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^T\Sigma_{11}^{-1}\\
\Omega_{22} &= \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}^T(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)^{-1}\Sigma_{12}\Sigma_{22}^{-1}\\
\Omega_{12} &= \Omega_{21}^T = -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1}
\end{aligned}$$

We can show that $p(x_1, x_2)$ can also be written as:

$$\begin{aligned}
p(x_1, x_2) &= \mathcal{N}(x_1; \mu_1, \Sigma_{11})\mathcal{N}(x_2; m_{2|1}, V_{2|1})\\
&= \mathcal{N}(x_2; \mu_2, \Sigma_{22})\mathcal{N}(x_1; m_{1|2}, V_{1|2})
\end{aligned}$$

where,

$$\begin{aligned} m_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \\ V_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ m_{2|1} &= \mu_2 + \Sigma_{12}^T\Sigma_{11}^{-1}(x_1 - \mu_1), \\ V_{2|1} &= \Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12} \end{aligned}$$

Now, using the above form we can write the marginal and conditional probabilities as the following:

$$\begin{aligned} p(x_1) &= \mathcal{N}(x_1|m_1^M, v_1^M) \\ m_1^M &= \mu_1 \\ v_1^M &= \Sigma_{11} \end{aligned}$$

$$\begin{aligned} p(x_2) &= \mathcal{N}(x_2|m_2^M, v_2^M) \\ m_2^M &= \mu_2 \\ v_2^M &= \Sigma_{22} \end{aligned}$$

$$\begin{aligned} p(x_1|x_2) &= \mathcal{N}(x_1|m_{1|2}, V_{1|2}) \\ m_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ V_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

1.2 Matrix Inversion

It is also useful to remember the inversion of matrices written in block form. Consider such a matrix M to be:

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

Then we can write the inverse of Matrix M as

$$\mathbf{M} = \begin{bmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{bmatrix}$$

The following matrix inversion lemma will also be used further in this material:

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

1.3 Matrix Algebra

In this section we will look at some formulae involving, traces, determinants and derivatives

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

The cyclical property of trace:

$$tr[ABC] = tr[CAB] = tr[BCA]$$

Derivatives involving trace:

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

Derivatives of determinants:

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

2 Factor Analysis

Factor analysis is a latent variable model where the latent variable is a continuous random vector. So, the model essentially is $X \to Y$ where X is continuous, hidden and Y is continuous, observed. Geometrically, it can be interpreted as sampling X from a Gaussian in low-dimensional subspace and then generating Y by sampling a normal distribution conditioned on X. The following figure from slides illustrates that.

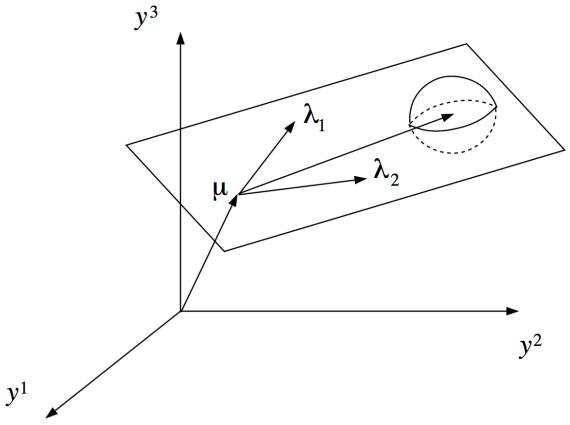

Figure 1: Illustration of Factor Analysis Model

The problem of estimating X based on Y is a dimensionality reduction problem.

2.1 Inference

X is a p-dimensional variable, Y is a q-dimensional variable where $p < q$ and we begin with X and $Y|X$.

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

Since, the distributions of X and $Y|X$ are both gaussian, the marginals, conditional and joint probabilities are all gaussian. So, these distributions are characterized by their mean and variance. To calculate the marginals:

$$\begin{aligned} Y &= \mu + \Lambda X + W \\ E[Y] &= E[\mu + \Lambda X + W] \\ &= E[\mu] + \Lambda E[X] + E[W] \\ &= \mu + 0 + 0 \\ &= \mu \\ Var[Y] &= E[(Y - \mu)(Y - \mu)^T] \\ &= E[(\mu + \Lambda X + W - \mu)(\mu + \Lambda X + W - \mu)^T] \\ &= E[(\Lambda X + W)(\Lambda X + W)^T] \\ &= E[\Lambda X X^T \Lambda^T + W W^T] \\ &= \Lambda E[X X^T] \Lambda^T + E[W W^T] \\ &= \Lambda \Lambda^T + \Psi \end{aligned}$$

To write the joint distribution we also need the covariance of X and Y.

$$\begin{aligned} Cov[X, Y] &= E[(X - 0)(Y - \mu)^T] \\ &= E[(X)(\mu + \Lambda X + W - \mu)^T] \\ &= E[X X^T \Lambda^T + X W^T] \\ &= \Lambda^T \end{aligned}$$

Simillarly, Cov[Y,X] = Λ. Therfore the joint probability can be written as:

$$p\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} X \\ Y \end{bmatrix}; \begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}\right)$$

Applying Gaussian Conditioning formulae shown in section 1, we get the following result for the posterior of latent variable X, given Y.

$$p(X|Y) = \mathcal{N}(X|m_{1|2}, V_{1|2})$$

where,

$$\begin{aligned} m_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y - \mu_2) \\ &= \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(Y - \mu) \end{aligned}$$

$$\begin{aligned} V_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &= I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda \end{aligned}$$

Inverting $(\Lambda\Lambda^T + \Psi)$ involves inverting a $|y| \times |y|$ matrix, So we use matrix lemma to replace it by $(I + \Lambda^T\Psi^{-1}\Lambda)^{-1}$.

So, the final equations for $m_{1|2}$ and $V_{1|2}$ are:

$$\begin{aligned} V_{1|2} &= (I + \Lambda^T\Psi^{-1}\Lambda)^{-1} \\ m_{1|2} &= V_{1|2}\Lambda^T\Psi^{-1}(Y - \mu) \end{aligned}$$

2.2 Learning

In this section, learning strategy of Factor Analysis will be discussed. In previous sections we have known that there are three parameters to be learned:

- Loading matrix Λ
- Manifold center μ
- Variance Ψ

So we are able to formalize the problem as a log likelihood function:

$$[\Lambda^*, \mu^*, \Psi^*] = argmax(Loglikelihood(Y))$$

The incomplete log likelihood If we consider the incomplete data log likelihood function, which in factor analysis is the marginal density of y, we have:

$$\begin{aligned} l(\theta|D) &= -\frac{N}{2}log|\Lambda\Lambda^T + \Psi| - \frac{1}{2}\{\sum_n (y_n - \mu)^T(\Lambda\Lambda^T + \Psi)^{-1}(y_n - \mu)\} \\ &= -\frac{N}{2}log|\Lambda\Lambda^T + \Psi| - \frac{1}{2}tr[(\Lambda\Lambda^T + \Psi)^{-1}S], \text{where } S = \sum_n (y_n - \mu)^T(y_n - \mu) \end{aligned}$$

Obviously, estimating μ is trivial, but parameters Λ and Ψ are still coupled non-linearly in the expression.

To decouple the parameters and obtain a simple algorithm for MLE, we consider EM algorithm in the following part.

EM algorithm As we have learned a few classes earlier, complete log likelihood would be the objective function that we consider after taking the expectation. Suppose here that we have "complete data", which means X and Y are both observed, it's clear that the estimation of the distribution of X would reduce to a Gaussian density estimation problem. So, in E step, we will try to fill in X by calculating the expected complete log likelihood and identify the expected sufficient statistics. Then, in M step, we will reduce to just estimating Λ and Ψ using linear regression.

E step The complete likelihood is simply a product of Gaussian distributions.

$$\begin{aligned} l_c(\theta|D_c) &= -\frac{N}{2}log|\Psi| - \frac{1}{2}\sum_n x_n^T x_n - \frac{1}{2}\sum_n (y_n - {}_n)^T \Psi^{-1}(y_n - \Lambda_n) \\ &= -\frac{N}{2}log|\Psi| - \frac{1}{2}\sum_n tr[x_n x_n^T] - \frac{1}{2}\sum_n tr[(y_n - \Lambda x_n)(y_n - \Lambda x_n)^T \Psi^{-1}] \\ &= -\frac{N}{2}log|\Psi| - \frac{N}{2}tr(S\Psi^{-1}) \end{aligned}$$

where we have:

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

Take the expectation

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

Here the conditional expectation $\langle s\rangle$ is

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

To draw a conclusion, the expected sufficient statistics that we need are actually conditional expectations $\langle X_n\rangle$ and $\langle X_n x_n^T\rangle$. We've already have these expectation derived in the previous sections. Thus

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

M step Since we have "filled in" X in the E step by calculating sufficient statistics, we are able to compute parameters by means of taking the derivative of expected complete log likelihood Q with respect to parameters.

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

Here we have $\Psi^{t+1} = \langle s\rangle$

$$\begin{aligned}\frac{\partial}{\partial \Lambda}\langle l_c\rangle &= \frac{\partial}{\partial \Lambda}(-\frac{N}{2}log|\Psi| - \frac{1}{2}\sum_n tr[x_n x_n^T] - \frac{1}{2}\sum_n tr[(y_n - \Lambda x_n)(y_n - \Lambda x_n)^T\Psi^{-1}]) \\ &= -\frac{N}{2}\Psi^{-1}\frac{\partial}{\partial \Lambda}\langle S\rangle \\ &= -\frac{N}{2}\Psi^{-1}\frac{\partial}{\partial \Lambda}(\frac{1}{N}\sum_n \langle y_n y_n^T - y_n\langle X_n^T\rangle\Lambda^T - \Lambda\langle X_n\rangle y_n^T + \Lambda\langle X_n X_n^T\rangle\Lambda^T)) \\ &= \Psi^{-1}\sum_n y_n\langle X_n^T\rangle - \Psi^{-1}\Lambda\sum_n\langle X_n X_n^T\rangle\end{aligned}$$

Here we have $\Lambda^{t+1} = (\sum_n y_n\langle X_n^T\rangle)(\sum_n\langle X_n X_n^T\rangle)^{-1}$

Model Invariance and Identifiability Since Λ only appear as outer product $\Lambda\Lambda^T$, the model is invariant to rotation and axis flips of the latent space:

$$(\Lambda Q)(\Lambda Q)^T = \Lambda(QQ^T)\Lambda^T = \Lambda\Lambda^T$$

This means there is no optimal solution of parameter estimations. Such models are called un-identifiable since multiple sets of parameters would be obtained when filling in same set of parameters.

3 State Space Models

3.1 Introduction

We have just learned Factor Analysis, whose latent and observed variables are both continuous Gaussians. If we connect multiple factor analysis models as what we do to mixture model, we can get a HMM-like graphical model, which is called State Space Model as shown in Fig. 2

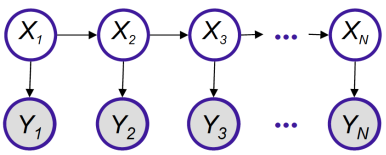

Figure 2: Graphical model for SSM

Here we have:

$$
\begin{aligned}
x_t &= Ax_{t-1} + Gw_t \\
y_t &= Cx_{t-1} + v_t \\
w_t &\sim N(0; Q), v_t \ N(0; R) \\
x_0 &\sim N(0; \Sigma_0)
\end{aligned}
$$

3.2 Inference

There are two interesting inference problems worth mentioning in SMM model: filtering and smoothing.

Filtering is a way to perform exact inference in an Linear Dynamic System, to infer the current latent variable based on current as well as previous observed variables. The problem of filtering is formalized as computing $P(x_t|y_{1:t})$:

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

Fig. 3 is an example graph for this problem.

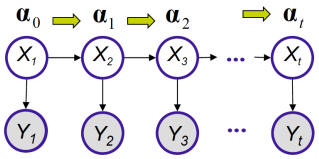

Figure 3: Graphical model for SSM

Smoothing is another inference problem in which we compute the current latent variable given observables at all time steps, which can be formalized as computing $P(x_t|y_{1:T})$:

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

Fig. 4 is an example graph for this problem.

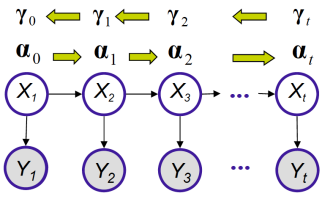

Figure 4: Graphical model for SSM

4 Kalman Filtering

4.1 Overview of derivation

Kalman filtering is an online filtering algorithm for use on state space models. It is widely used and particularly efficient. Since the state space models we are dealing with all have conditional probability distributions that are linear Gaussian, the system defines a large multivariate Gaussian.

This means that all the marginals are Gaussian, and we can represent the belief state $p(X|y_{1:t})$ as a Gaussian with mean

$$\mathbf{E}[X_t|y_{1:t}] = \mu_{t|t}$$

and covariance

$$\mathbf{E}[(X_t - \mu_{t|t})(X_t - \mu_{t|t})^T] = P_{t|t}$$

It is common to work with the inverse of the covariance matrix, called the precision matrix. This is known as the information form.

Kalman filter is a recursive procedure to update the belief state. it has two main phases: the **predict** step, and the **update** step. Essentially, instead of trying to solve for $p(X_{t+1}|y_{1:t+1})$ directly, we break it into two parts. In the predict step, we seek to compute $p(X_{t+1}|y_{1:t})$ from the prior belief $p(X_t|y_{1:t})$ and the dynamics model $p(X_{t+1}|X_t)$. This is called the time update. In the update step, we compute our goal $p(X_{t+1}|y_{1:t+1})$ from the prediction $p(X_{t+1}|y_{1:t})$, the observation y_{t+1}, and the observation model $p(y_{t+1}|X_{t+1})$. This is called the measurement update.

The advantage of this process is that, since the variables are Gaussian, then everything else ends up being Gaussian. Recall from the beginning of the notes that if we have

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

This means that, given that the marginal z_1 is Gaussian, then the joint $z_1 z_2$ is Gaussian, the marginal z_2 is Gaussian, and finally the conditional $z_2|z_1$ must also be Gaussian.

The Kalman filter essentially follows this process. Given $p(X_t|y_{1:t})$ is Gaussian, we can get $p(X_{t+1}|y_{1:t})$, with $x_{t+1} = f(x_t) = Ax_t + w$. Then from this result and the observation model $y_{t+1} = Cx_{t+1} + v$, we can finally get $p(X_{t+1}, y_{t+1}|y_{1:t})$.

4.2 Predict Step

For the **dynamical model**, we have that

$$x_{t+1} = Ax_t + Gw_t$$

where $w_t \sim \mathcal{N}(0; Q)$. We wish to find the parameters of the distribution of $p(X_{t+1}|y_{1:t})$. Since everything here is a Gaussian, this means we want the mean and covariance of this distribution. For one step ahead

prediction of state:

$$\begin{aligned}
\mathbf{E}[X_{t+1}|y_{1:t}] =& \mathbf{E}[Ax_t + Gw_t] \\
=& A\mu_{t|t} + 0 \\
=& \hat{x}_{t+1|t} \\
\hat{x}_{t+1|t} =& A\mu_{t|t} \\
\mathbf{E}[(X_{t+1} - \hat{x}_{t+1|t})(X_{t+1} - \hat{x}_{t+1|t})^T|y_{1:t}] =& \mathbf{E}[(AX_t + Gw_t - A\mu_{t|t})(AX_t + Gw_t - A\mu_{t|t})^T|y_{1:t}] \\
=& \mathbf{E}[(AX_t - A\mu_{t|t})(AX_t - A\mu_{t|t})^T|y_{1:t}] + \mathbf{E}[(AX_t - A\mu_{t|t})w_t^T G^T|y_{1:t}] \\
& \mathbf{E}[Gw_t(AX_t - A\mu_{t|t})|y_{1:t}] + \mathbf{E}[Gw_t w_t^T G^T|y_{1:t}] \\
=& \mathbf{E}[(AX_t - A\mu_{t|t})(AX_t - A\mu_{t|t})^T|y_{1:t}] + 0 + 0 + \mathbf{E}[Gw_t w_t^T G^T|y_{1:t}] \\
=& A\mathbf{E}[(X_t - \mu_{t|t})(X_t - \mu_{t|t})^T|y_{1:t}]A^T + G\mathbf{E}[w_t w_t^T|y_{1:t}]G^T \\
P_{t+1|t} =& AP_{t|t}A^T + GQG^T
\end{aligned}$$

And so the prediction for the dynamical model is the mean $\hat{x}_{t+1|t} = A\mu_{t|t}$ and covariance is $P_{t+1|t} = AP_{t|t}A^T + GQG^T$.

For the **observation model**, have have that

$$y_t = Cx_t + v_t$$

where $v_t \sim (0; R)$. We now wish to find the parameters of the model of the observation. Once again, it is Gaussian, since the prior parts are all Gaussian. For one step ahead prediction of observation:

$$\begin{aligned}
\mathbf{E}[Y_{t+1}|y_{1:t}] =& \mathbf{E}[Cx_t + w_t] \\
=& C\hat{x}_{t+1|t} + 0 \\
\hat{y}_{t+1|t} =& C\hat{x}_{t+1|t} \\
\mathbf{E}[(Y_{t+1} - \hat{y}_{t+1|t})(Y_{t+1} - \hat{y}_{t+1|t})^T|y_{1:t}] =& \mathbf{E}[(CX_{t+1} + v_t - C\hat{x}_{t+1|t})(CX_{t+1} + v_t - C\hat{x}_{t+1|t})^T|y_{1:t}] \\
=& C\mathbf{E}[(X_{t+1} - \hat{x}_{t+1|t})(X_{t+1} - \hat{x}_{t+1|t})^T]C^T + \mathbf{E}[v_t v_t^T] \\
=& CP_{t+1|t}C^T + R \\
\mathbf{E}[(Y_{t+1} - \hat{y}_{t+1|t})(X_{t+1} - \hat{x}_{t+1|t})^T|y_{1:t}] =& \mathbf{E}[(CX_{t+1} + v_t - C\hat{x}_{t+1|t})(X_{t+1} + \hat{x}_{t+1|t})^T|y_{1:t}] \\
=& C\mathbf{E}[(X_{t+1} - \hat{x}_{t+1|t})(X_{t+1} - \hat{x}_{t+1|t})^T] + \mathbf{E}[v_t(X_{t+1} - \hat{x}_{t+1|t})] \\
=& CP_{t+1|t}
\end{aligned}$$

And so finally, for the observation variable we have mean $C\hat{x}_{t+1|t}$ and variance $CP_{t+1|t}C^T + R$, as well as covariance with state variable $CP_{t+1|t}$.

4.3 Update Step

Will be continued in the next lecture