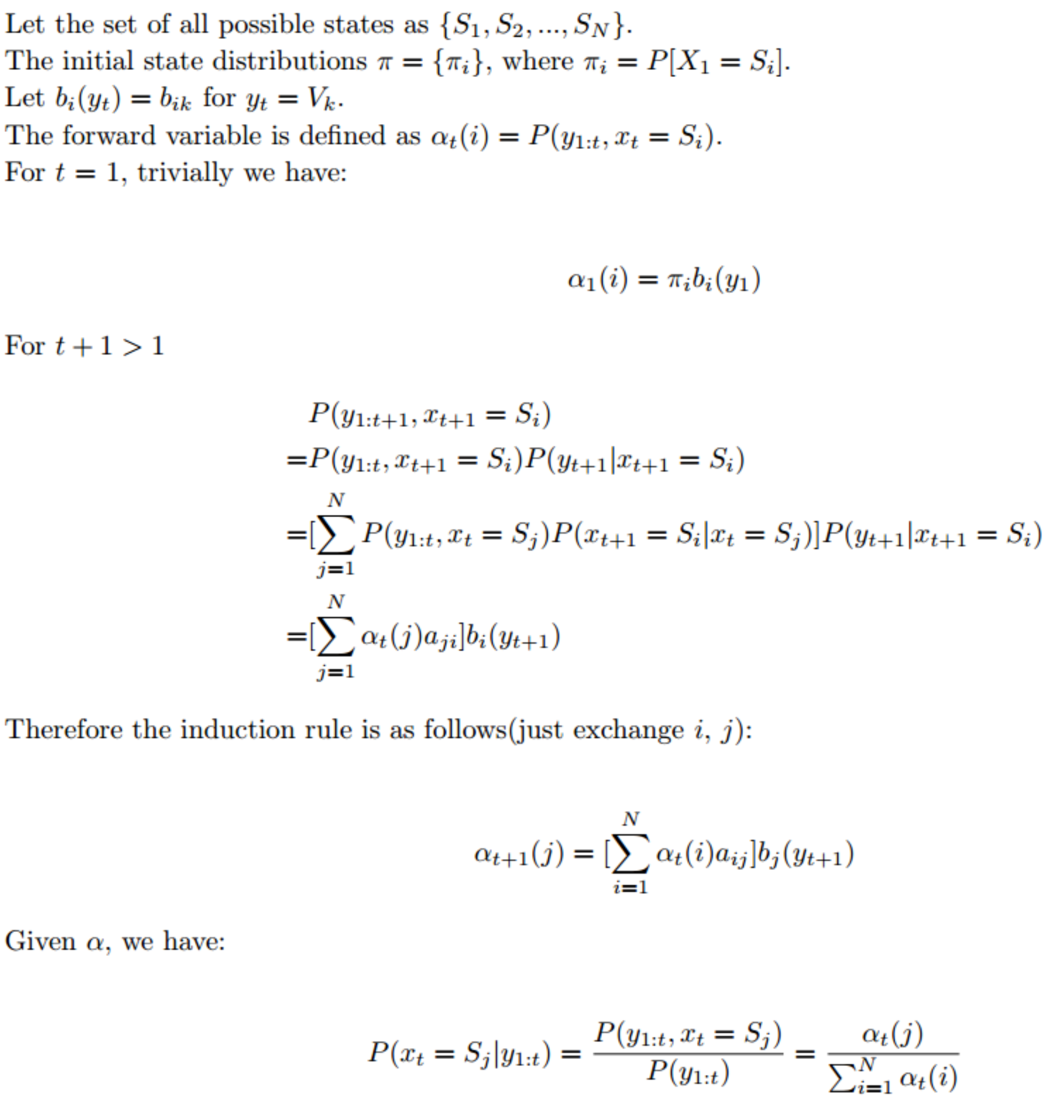
HW 11 Solutions

ATTRIBUTION: PROBLEM 1.2 CITES PROF. ERIC XING'S LECTURE NOTES FROM HIS PROBABILISTIC GRAPHIC MODELS CLASS (10-708). http://www.cs.cmu.edu/~epxing/Class/10708-14/scribe_notes/scribe_note_lecture11.pdf, WHICH PROVIDES INTRODUCTION AND DETAILED DERIVATIONS OF STATE SPACE MODELS AND KALMAN FILTERING WITH ILLUSTRATIVE EXAMPLES.

1 Hidden Markov Model (HMM) v.s. Linear Gaussian State Space Models (SSM)

1.1 Forward Inference for HMM



Let the set of all possible states as $\{S_1, S_2, ..., S_N\}$.
The initial state distributions $\pi = \{\pi_i\}$, where $\pi_i = P[X_1 = S_i]$.
Let $b_i(y_t) = b_{ik}$ for $y_t = V_k$.
The forward variable is defined as $\alpha_t(i) = P(y_{1:t}, x_t = S_i)$.
For $t = 1$, trivially we have:

$$\alpha_1(i) = \pi_i b_i(y_1)$$

For $t + 1 > 1$

$$\begin{aligned}
&P(y_{1:t+1}, x_{t+1} = S_i)\\
=&P(y_{1:t}, x_{t+1} = S_i)P(y_{t+1}|x_{t+1} = S_i)\\
=&[\sum_{j=1}^{N} P(y_{1:t}, x_t = S_j)P(x_{t+1} = S_i|x_t = S_j)]P(y_{t+1}|x_{t+1} = S_i)\\
=&[\sum_{j=1}^{N} \alpha_t(j)a_{ji}]b_i(y_{t+1})
\end{aligned}$$

Therefore the induction rule is as follows(just exchange i, j):

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i)a_{ij}]b_j(y_{t+1})$$

Given α, we have:

$$P(x_t = S_j|y_{1:t}) = \frac{P(y_{1:t}, x_t = S_j)}{P(y_{1:t})} = \frac{\alpha_t(j)}{\sum_{i=1}^{N} \alpha_t(i)}$$

1.2 Kalman Filtering for SSM

State Space Models

A State Space Model (SSM) is a dynamical generalization of the Factor Analysis model. In fact, it is a collection of factor analysers connected as a chain in the time domain, with one factor analyser model per time instance. SSMs are structurally identical to Hidden Markov Models - and hence have the same independence assumptions. The only difference is that the variables in SSM follow continuous (Gaussian) instead of discrete (Multinomial) distributions as in an HMM. As we will see, despite following continuous distributions, the derivation for inference under this model does not involve complex calculus, thanks to the properties of the Gaussian distribution.

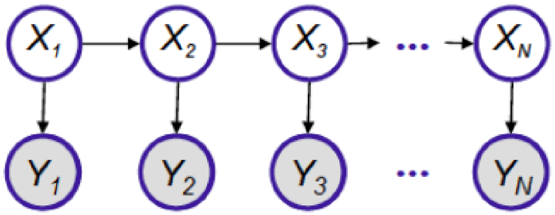


Figure 2: Graph for a State Space Model

In this model, we observe a sequence $y = (y_1, y_2, ..., y_t, ...)$ where each y_t is a continuous random variable for an instance of time, t. We assume there is a latent sequence $x = (x_1, x_2, ...x_t, ...)$ that generates this observation, where each x_t is also Gaussian. The graphical model is illustrated in Figure 2.

We introduce a transition matrix that determines the relationship between the latent variables, such that the mean of the state at time t, x_t is linear in the mean of the state at time $t - 1$.

$$x_t = Ax_{t-1} + Gw_t$$

Here, $w_t = \mathcal{N}(0, Q)$ is the Gaussian noise we have introduced into the model. Since a linear combination of Gaussians is also Gaussian, x_t is Gaussian.

To describe the output, we use the Factor Analysis model at each point. The loading matrix, say C is shared across all x_t, y_t pairs. We assume that all the data points are in the same low-dimensional space. We have

$$y_t = Cx_t + v_t$$

where $v_t = \mathcal{N}(0, R)$ is some Gaussian noise. Note that we do not make any assumptions on the Q and R matrices, these could either be full rank or low rank.

Finally, we set the starting point, $x_0 = \mathcal{N}(0, \Sigma_0)$.

Kalman Filtering

Given a sequence of observations $y_1, y_2, ..., y_t$, we have to infer the latent state at time t. This inference problem is also known as Kalman Filtering. Historically, Kalman Filtering used to be considered a stand-alone inference technique. However after graphical models gained popularity, it was clear that it is only a Gaussian analogue of the forward inference for HMMs (see Figure 4). The following equation shows this analogy:

$$p(x_t|y_{1:t}) = \alpha_t{}^i \propto p(y_t|x_t)\Sigma_{x_{t-1}} p(x_t|x_{t-1})\alpha^j{}_{t-1}$$

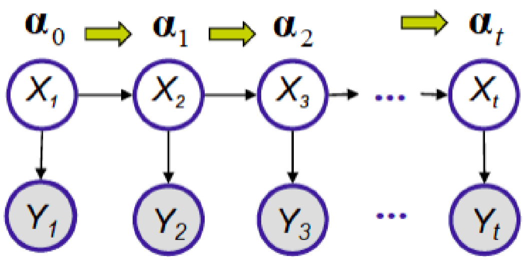


Figure 4: Kalman Filtering as Forward Inference on SSMs

Kalman Filtering for inference is widely applied in psychology. Our observation of the world through visual images, say y, is in two dimensional space even though the truth, x is in three dimensional space. This means y is a noisy version of the ground truth. However, when our brains recreate this observation, they do so in three-dimensional space. This happens dynamically in time as the brain cannot look forward in time before recreating the images. Hence, this is the Kalman Filtering estimation of $p(x_t|y_{1:t})$.

Derivation

The key observation in the SSM is that every distribution in it is Gaussian. Therefore, the distribution of interest $p(x_t|y_{1:t})$ is also a Gaussian. The task is then to estimate the mean, $\mu_{1:t} = \mathbb{E}(x_t|y_{1:t})$ and the covariance, $P_{1:t} = \mathbb{E}(x_t - \mu_{1:t})^T(x_t - \mu_{1:t})$ of this distribution.
The estimation is done in two steps to simplify computation.

- Predict step - Compute $p(x_{t+1}|y_{1:t})$ from $p(x_t|y_{1:t})$. This is equivalent to moving one step ahead of the current observation sequence. It is also called time update.

- Update step - Update the prediction in the previous step by including the new evidence in the data. The new evidence is computed according to a new observation y_{t+1} and the model parameter $p(y_{t+1}|x_{t+1})$.

The high level idea behind the estimation is the following: We are given two Gaussian vectors z_1 and z_2, which are distributed as below:

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

We shall use z_1 to generate the joint z_1, z_2 and either marginalize the joint to obtain z_2 or use the marginal to compute the conditional distribution, $z_2|z_1$.

In the prediction step, we start with $p(x_t|y_{1:t})$ and use the transition information of the model, $x_t = Ax_{t-1} + Gw_t$ to compute $p(x_{t+1}|y_{1:t})$.

In the update step, we obtain $p(y_{t+1}|x_{t+1})$ using the emission information of the model, $y_t = Cx_t + v_t$. We use this evidence to obtain a joint $p(x_{t+1}, y_{t+1}|y_{1:t})$. Finally, we invert the model to obtain $p(x_{t+1}|y_{1:t+1})$.

Predict Step

To calculate the mean, $\mu_{t+1|t}$ and the variance, $P_{t+1|t}$ of the joint distribution $p(x_{t+1}, y_{t+1}|y_{1:t})$ for the **Dynamical Model**, we proceed as below:

Mean:

$$\mathbb{E}(x_{t+1}|y_{1:t}) = \mathbb{E}(Ax_t + Gw_t) = A\mu_{1:t} + 0 = \mu_{1:t+1|t}$$

Covariance:

$$\begin{aligned} \mathbb{E}(x_{t+1} - \mu_{t+1|t})^T(x_{t+1} - \mu_{t+1|t}) &= \mathbb{E}(Ax_t + Gw_t - \mu_{t+1|t})^T(Ax_t + Gw_t - \mu_{t+1|t}) \\ &= A\mathbb{E}(x_t - \mu_{t+1|t})^T(x_t - \mu_{t+1|t})A^T + GQG^T \\ &= AP_{t+1|t}A^T + GQG^T \end{aligned}$$

To calculate the mean and the variance of the joint distribution $P(x_{t+1}, y_{t+1}|y_{1:t})$ for the **Observation Model**, we proceed as below:

Mean:

$$\mathbb{E}(y_{t+1}|y_{1:t}) = \mathbb{E}(Cx_{t+1} + v_{t+1}|y_{1:t}) = C\mu_{t+1|t}$$

Covariance:

$$\begin{aligned} \mathbb{E}[(y_{t+1} - \hat{y}_{t+1|t})(y_{t+1} - \hat{y}_{t+1|t})^T|y_{1:t}] &= CP_{t+1|t}C^T + R \\ \mathbb{E}[(y_{t+1} - \hat{y}_{t+1|t})(x_{t+1} - \mu_{t+1|t})^T|y_{1:t}] &= CP_{t+1|t} \end{aligned}$$

Update Step

From the quantities computed in the previous step, we proceed here to compute the mean and the variance of the conditional distribution, $p(X_{t+1}|Y_{t+1})$ using the formulae for conditional Gaussian distributions. $\mu_{t|t}$ and $P_{t|t}$ are the mean and covariance respectively of the distribution $p(X_t|Y_t)$.

Time Updates:

$$\mu_{t+1|t} = A\mu_{t|t}$$

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

Measurement updates:

$$\mu_{t+1|t+1} = \mu_{t+1|t} + K_{t+1}(y_{t+1} - C\mu_{t+1|t})$$

$$P_{t+1|t+1} = P_{t+1|t} - K_{t+1}CP_{t+1|t}$$

where $K_{t=1}$ is the Kalman gain. This quantity calibrates or adjusts the observed y_t such that our prediction for the next state is not biased. Kalman gain provides a trade-off between the prior and any new observation because in cases where either the prior is unreliable or the observations are noisy. The term $(y_{t+1} - C\mu_{t+1|t})$ is called the **innovation**, because it brings in additional information to the model.

Being independent of the data, the Kalman Gain can be precomputed using the following:

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

Complexity

Let $x_t \in \Re^{N_x}$ and $y_t \in \Re^{N_y}$. Computing the new variance from the old variance takes $O(N_x^2)$ time:

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

Pre-computing the Kalman Gain takes $O(N_y^3)$ time:

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

Hence, the overall time complexity is $\max(O(N_x^2), O(N_y^3))$. This makes Kalman Filtering quite an expensive inference algorithm for moderately high dimensional problems.

Kalman Filters are not popular these days due to high complexity. For instance, consider signals from 1000 aircraft coming at the same time. We need to consider all of them independent, and predicting each different trajectory is going to be highly expensive. In such cases, we should consider more complex models like Switching SSMs, which have multiple sequences of latent variables and a particular observation might depend on any combination of the latent sequences.

2 Collaborative Filtering

2.1 Cold Start Problem

When new users or items come in, their corresponding rows added to the matrix are extremely sparse, leading to very poor item-item associations and user-user associations generated using matrix factorization methods. Some methods that use SVD will fail miserably if no information from the new user is known. This will become a problem when recommending new items to old users, recommending old items to new users, or recommending new items to new users, which is a major issue when starting a whole new recommender system.

2.2 Strategies

Utilizing Meta Information

- Park, Seung-Taek, and Wei Chu. "Pairwise preference regression for cold-start recommendation." *Proceedings of the third ACM conference on Recommender systems*. ACM, 2009.

Content-driven Filtering methods can be used jointly with model based methods.

- Melville, Prem, Raymond J. Mooney, and Ramadass Nagarajan. "Content-boosted collaborative filtering for improved recommendations." *AAAI/IAAI*. 2002.
- Schein, Andrew I., et al. "Methods and metrics for cold-start recommendations."*Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002.

Customized Aspect Models and Flexible Mixture Models provide smoothed solutions with prior distributions.

- Si, Luo, and Rong Jin. "Flexible mixture model for collaborative filtering." *ICML*. Vol. 3. 2003.

Active Learning: solicit ratings from the new users

- Jin, Rong, and Luo Si. "A bayesian approach toward active learning for collaborative filtering." *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004.