Information form filtering and smoothing for Gaussian linear dynamical systems

Scott W. Linderman Matthew J. Johnson

July 18, 2016

The *information form* of the Gaussian distribution over $x \in \mathbb{R}^D$ is defined as,

$$p(x \mid J, h) = \exp\left\{-\frac{1}{2}x^\mathsf{T} J x + h^\mathsf{T} x - \log Z\right\}, \tag{1}$$

where

$$\log Z = \frac{1}{2}h^\mathsf{T} J^{-1} h - \frac{1}{2}\log|J| + \frac{D}{2}\log 2\pi. \tag{2}$$

The standard formulation is recovered by the transformations, $\Sigma = J^{-1}$, and $\mu = J^{-1}h$. When working with this natural parameterization, the parameters J and h interact linearly with the data, making it considerably easier to derive mean field variational inference algorithms.

In order to perform Kalman filtering and smoothing, we must be able to perform two operations: *conditioning* and *marginalization*.

Conditioning If,

$$p(x) = \mathcal{N}(x \mid J, h) \tag{3}$$

$$p(y \mid x) \propto \mathcal{N}(x \mid J_{\mathsf{obs}}, h_{\mathsf{obs}}) \tag{4}$$

then,

$$p(x \mid y) = \mathcal{N}(x \mid J + J_{\mathsf{obs}}, h + h_{\mathsf{obs}}). \tag{5}$$

Marginalization If,

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} J_{xx} & J_{xy} \\ J_{xy}^\mathsf{T} & J_{yy} \end{bmatrix}, \begin{bmatrix} h_x \\ h_y \end{bmatrix}\right), \tag{6}$$

then,

$$x \sim \mathcal{N}(J_{xx} - J_{xy}J_{yy}^{-1}J_{xy}^\mathsf{T},\ h_x - J_{xy}J_{yy}^{-1}h_y) \tag{7}$$

Whereas in the standard formulation, marginalization is easy (simply extract sub-blocks of the mean and covariance) and conditioning involves solving a linear system; in information form, condition is easy (simply add sufficient statistics) and marginalization requires solving a linear system.

Filtering, Sampling, and Smoothing

By interleaving these two steps we can filter, sample, and smooth the latent states in a linear dynamical system. Take the model,

$$x_1 \sim \mathcal{N}(\mu_1, Q_1) \tag{8}$$

$$x_{t+1} \sim \mathcal{N}(A_t x_t + B_t u_t, Q_t) \tag{9}$$

$$y_t \sim \mathcal{N}(C_t x_t + D_t u_t, R_t). \tag{10}$$

In information form, the initial distribution is,

$$x_1 \sim \mathcal{N}(J = Q_1^{-1}, h = Q_1^{-1}\mu_1). \tag{11}$$

The dynamics are given by,

$$p(x_{t+1} \,|\, x_t) \propto \mathcal{N}\left(\begin{bmatrix} x_t \\ x_{t+1} \end{bmatrix} \,\middle|\, \begin{bmatrix} J_{11} & J_{12} \\ J_{12}^\mathsf{T} & J_{22} \end{bmatrix}, \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \right), \tag{12}$$

with,

$$J_{11} = A_t^\mathsf{T} Q_t^{-1} A_t, \quad J_{12} = -A_t^\mathsf{T} Q_t^{-1} \quad J_{22} = Q_t^{-1} \quad h_1 = -u_t^\mathsf{T} B_t^\mathsf{T} Q_t^{-1} A_t \quad h_2 = u_t^\mathsf{T} B_t Q_t^{-1}. \tag{13}$$

Finally, the observations are given by,

$$p(y_t \,|\, x_t) \propto \mathcal{N}(x_t \,|\, J_{\mathsf{obs}}, h_{\mathsf{obs}}) \tag{14}$$

with

$$J_{\mathsf{obs}} = C_t^\mathsf{T} R_t^{-1} C_t \quad h_{\mathsf{obs}} = (y_t - D_t u_t)^\mathsf{T} R_t^{-1} C_t \tag{15}$$

Filtering

We seek the conditional distribution, $p(x_t \,|\, y_{1:t})$, which will be Gaussian. Assume, inductively, that $x_t \,|\, y_{1:t-1} \sim \mathcal{N}(J_{t|t-1}, h_{t|t-1})$. Conditioning on the t-th observation yields, Conditioned on the first observation,

$$p(x_t \,|\, y_{1:t}) = \mathcal{N}(x_t \,|\, J_{t|t}, h_{t|t}), \tag{16}$$

$$J_{t|t} = J_{t|t-1} + J_{\mathsf{obs}} \tag{17}$$

$$h_{t|t} = h_{t|t-1} + h_{\mathsf{obs}}. \tag{18}$$

Then, we predict the next latent state by writing the joint distribution of x_t and x_{t+1} and marginalizing out x_t.

$$\begin{aligned}
p(x_{t+1} \,|\, y_{1:t}) &= p(x_t \,|\, y_{1:t})\, p(x_{t+1} \,|\, x_t) && (19)\\
&= \mathcal{N}(x_t \,|\, J_{t+1|t}, h_{t+1|t}) && (20)\\
J_{t+1|t} &= J_{22} - J_{21}(J_{t|t} + J_{11})^{-1} J_{21}^{\mathsf{T}} && (21)\\
h_{t+1|t} &= h_2 - J_{21}(J_{t|t} + J_{11})^{-1}(h_{t|t} + h_1). && (22)
\end{aligned}$$

This completes one iteration and provides the input to the next. To start the recursion, we initialize,

$$J_{1|0} = \Sigma_{\mathsf{init}}^{-1}, \quad h_{1|0} = \Sigma_{\mathsf{init}}^{-1} \mu_{\mathsf{init}}. \tag{23}$$

Backward Sampling

Having computed $J_{t|t}$ and $h_{t|t}$, we the proceed backward in time to draw a joint sample of the latent states. Given $J_{t|t}$, $h_{t|t}$, and x_{t+1}, we have,

$$\begin{aligned}
p(x_t \,|\, y_{1:t}, x_{t+1}) &\propto p(x_t \,|\, y_{1:T})\, p(x_{t+1} \,|\, x_t) && (24)\\
&\propto \mathcal{N}(x_t \,|\, J_{t|t}, h_{t|t})\, \mathcal{N}(x_t \,|\, J_{11},\ h_1 - x_{t+1}^{\mathsf{T}} J_{21}) && (25)\\
&\propto \mathcal{N}(x_t \,|\, J_{t|t} + J_{11},\ h_{t|t} + h_1 - x_{t+1}^{\mathsf{T}} J_{21}) && (26)
\end{aligned}$$

We sample x_t from this conditional, then use it to sample x_{t-1}, and repeat until we reach x_1.

Rauch-Tung-Striebel Smoothing

Next we seek the conditional distribution given all the data, $p(x_t \,|\, y_{1:T})$. This will again be Gaussian, and we will call its parameters $J_{t|T}$ and $h_{t|T}$. Assume, inductively, that we have computed $J_{t+1|T}$ and $h_{t+1|T}$. We show how to compute the parameters for time t.

From the Markov properties of the model and the conditional distribution derived above, we have,

$$p(x_t \,|\, x_{t+1}, y_{1:T}) = \mathcal{N}(x_t \,|\, J_{t|t} + J_{11},\ h_{t|t} + h_1 - J_{12} x_{t+1}). \tag{27}$$

Expanding, taking care to note that x_{t+1} appears in the normalizing constant, yields,

$$\begin{aligned}
p(x_t \,|\, x_{t+1}, y_{1:T}) = \exp\Big\{ &-\frac{1}{2} x_t^{\mathsf{T}} (J_{t|t} + J_{11}) x_t + (h_{t|t} + h_1)^{\mathsf{T}} x_t - x_{t+1}^{\mathsf{T}} J_{12} x_t \\
&- \frac{1}{2} x_{t+1}^{\mathsf{T}} J_{12}^{\mathsf{T}} (J_{t|t} + J_{11})^{-1} J_{12} x_{t+1} + (h_{t|t} + h_1)^{\mathsf{T}} (J_{t|t} + J_{11})^{-1} J_{12} x_{t+1} \\
&- \frac{1}{2} (h_{t|t} + h_1)^{\mathsf{T}} (J_{t|t} + J_{11})^{-1} (h_{t|t} + h_1) \Big\}
\end{aligned} \tag{28}$$

Now consider the joint distribution of x_t and x_{t+1} given all the data,

$$p(x_t, x_{t+1} \,|\, y_{1:T}) = p(x_t \,|\, x_{t+1}, y_{1:T}) p(x_{t+1} \,|\, y_{1:T}) \tag{29}$$

$$\propto \mathcal{N}\left(\begin{bmatrix} x_t \\ x_{t+1} \end{bmatrix} \,\middle|\, \begin{bmatrix} \widetilde{J}_{11} & \widetilde{J}_{12} \\ \widetilde{J}_{12}^{\mathsf{T}} & \widetilde{J}_{22} \end{bmatrix}, \begin{bmatrix} \widetilde{h}_1 \\ \widetilde{h}_2 \end{bmatrix} \right), \tag{30}$$

with,

$$\widetilde{J}_{11} = J_{t|t} + J_{11} \tag{31}$$

$$\widetilde{J}_{12} = J_{12} \tag{32}$$

$$\widetilde{J}_{22} = J_{t+1|T} + J_{12}^{\mathsf{T}} (J_{t|t} + J_{11})^{-1} J_{12} \tag{33}$$

$$\widetilde{h}_1 = h_{t|t} + h_1 \tag{34}$$

$$\widetilde{h}_2 = h_{t+1|T} + (h_{t|t} + h_1)^{\mathsf{T}} (J_{t|t} + J_{11})^{-1} J_{12}. \tag{35}$$

Recall that,

$$J_{t+1|t} = J_{22} - J_{21} (J_{t|t} + J_{11})^{-1} J_{21}^{\mathsf{T}} \tag{36}$$

$$h_{t+1|t} = h_2 - J_{21} (J_{t|t} + J_{11})^{-1} (h_{t|t} + h_1). \tag{37}$$

Thus,

$$\widetilde{J}_{22} = J_{t+1|T} - J_{t+1|t} + J_{22} \tag{38}$$

$$\widetilde{h}_2 = h_{t+1|T} - h_{t+1|t} + h_2. \tag{39}$$

Finally, marginalize,

$$p(x_t \,|\, y_{1:T}) = \mathcal{N}(x_t \,|\, \widetilde{J}_{11} - \widetilde{J}_{12} \widetilde{J}_{22}^{-1} \widetilde{J}_{12}^{\mathsf{T}},\ \widetilde{h}_1 - \widetilde{J}_{12} \widetilde{J}_{22}^{-1} \widetilde{h}_2) \tag{40}$$

$$= \mathcal{N}(x_t \,|\, J_{t|T}, h_{t|T}). \tag{41}$$

Substituting the simplified forms above yields,

$$J_{t|T} = J_{t|t} + J_{11} - J_{12} (J_{t+1|T} - J_{t+1|t} + J_{22})^{-1} J_{12}^{\mathsf{T}} \tag{42}$$

$$h_{t|T} = h_{t|t} + h_1 - J_{12} (J_{t+1|T} - J_{t+1|t} + J_{22})^{-1} (h_{t+1|T} - h_{t+1|t} + h_2). \tag{43}$$