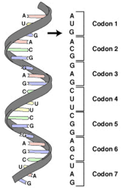
Sequence Modelling

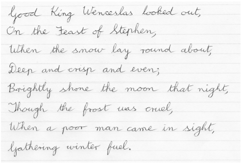
Rich Turner and José Miguel Hernández-Lobato

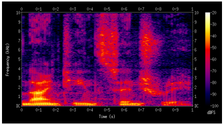
Sequence data

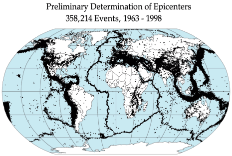


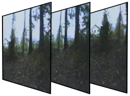
RNA

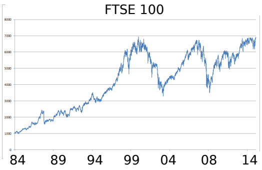
Ribonucleic acid

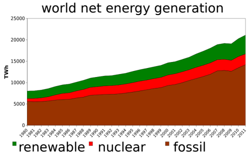












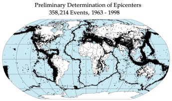
I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. A. Turing

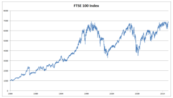
Some images taken from wikipedia

Goals of sequence modelling

Predict future items in sequence

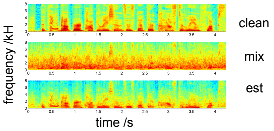
$$p(y_t | y_1, \ldots, y_{t-1})$$





Remove noise from a sequence

$$p(y'_1, \ldots, y'_t | y_1, \ldots, y_t)$$



Predict one sequence from another

$$p(y'_1, \ldots, y'_t | y_1, \ldots, y_t)$$

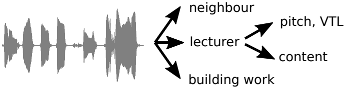
I believe that →

↓

私はでそれを信じて

Discover underlying latent variables

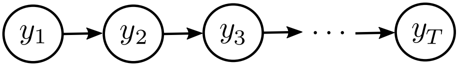
$$p(x_1, \ldots, x_t | y_1, \ldots, y_t)$$



Markov models

First order Markov

$$p(y_1, y_2, y_3, \ldots, y_T) = p(y_1)p(y_2|y_1)p(y_3|y_2)\ldots p(y_T|y_{T-1})$$



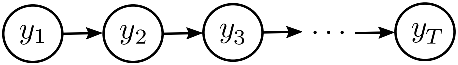
Markov models

parameters tied

∞ number of variables
finite number of
parameters

First order Markov

$$p(y_1, y_2, y_3, \ldots, y_T) = p(y_1)p(y_2|y_1)p(y_3|y_2)\ldots p(y_T|y_{T-1})$$



Markov models

parameters tied

∞ number of variables
finite number of parameters

First order Markov

$$p(y_1, y_2, y_3, \ldots, y_T) = p(y_1)p(y_2|y_1)p(y_3|y_2)\ldots p(y_T|y_{T-1})$$

$$y_1 \to y_2 \to y_3 \to \cdots \to y_T$$

Matlab notation
$1{:}t{-}1 = 1, 2, 3, \ldots, t-1$

Markov model = conditional independence relationship + product rule

independent of past

future $\to$ $y_{t+1} \perp y_{1:t-1} | y_t$ $\leftarrow$ given present

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1})$$

$T=4$

$$p(y_1, y_2, y_3, y_4) = p(y_1)\,p(y_2|y_1)\,p(y_3|y_2, \cancel{y_1})\,p(y_4|y_3, \cancel{y_2, y_1})$$

$$p(y_3|y_2, \cancel{y_1}) = p(y_3|y_2)$$

$$p(y_4|y_3, \cancel{y_2, y_1}) = p(y_4|y_3)$$

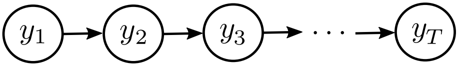
Markov models

parameters tied

∞ number of variables
finite number of parameters

First order Markov

$$p(y_1, y_2, y_3, \ldots, y_T) = p(y_1)p(y_2|y_1)p(y_3|y_2)\ldots p(y_T|y_{T-1})$$



Markov model = conditional independence relationship + product rule

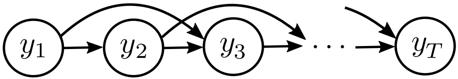
independent of past

future $\;\; y_{t+1} \perp y_{1:t-1} | y_t \;\;$ given present

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1})$$

Second order Markov

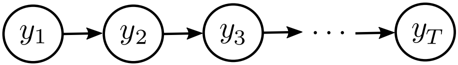
$$p(y_1, y_2, y_3, \ldots, y_T) = p(y_1)p(y_2|y_1)p(y_3|y_2, y_1)\ldots p(y_T|y_{T-1}, y_{T-2})$$



Markov models for discrete data: n-gram models

First order Markov (bi-gram)

$$p(y_1, y_2, y_3, \ldots, y_T) = p(y_1)p(y_2|y_1)p(y_3|y_2)\ldots p(y_T|y_{T-1})$$



$T_{l,k}$

$y_t \in \{1, ..., K\}$ — discrete states

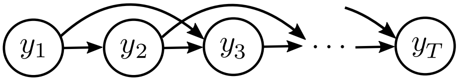
$p(y_1 = k) = \pi_k^0$ — initial state probabilities

$\boxed{p(y_t = k|y_{t-1} = l) = T_{k,l}}$ — transition probabilities (stochastic matrix)

$\sum_{k=1}^{K} T_{k,l} = 1$

Second order Markov (tri-gram)

$$p(y_1, y_2, y_3, \ldots, y_T) = p(y_1)p(y_2|y_1)p(y_3|y_2, y_1)\ldots p(y_T|y_{T-1}, y_{T-2})$$



$p(y_t = k|y_{t-1} = l, y_{t-2} = m) = T_{k,l,m}$

n-grams require large multidimensional arrays

Some questions about n-gram models

First order Markov (bi-gram)

$$y_t \in \{1, ..., K\} \quad p(y_1 = k) = \pi_k^0 \qquad p(y_t = k | y_{t-1} = l) = T_{k,l}$$

Q1. How can we compute the marginal distribution over the second state?

$$p(y_2 = l) = \sum_k \underline{p(y_2 = l \mid y_1 = k)}\, p(y_1 = k) = \sum_k T_{lk} \pi_k^0$$

$$p(y_2) = \underline{\underline{T}}\, \underline{\pi}^0$$

$$(\underline{\pi}^0)^T \underline{\underline{T}}^T$$

Some questions about n-gram models

First order Markov (bi-gram)

$$y_t \in \{1, ..., K\} \quad p(y_1 = k) = \pi_k^0 \qquad p(y_t = k | y_{t-1} = l) = T_{k,l}$$

Q1. How can we compute the marginal distribution over the second state?

$$p(y_2 = k) = \sum_{l=1}^{K} p(y_2 = k | y_1 = l) p(y_1 = l) = \sum_{l=1}^{K} T_{k,l} \pi_l^0$$

Some questions about n-gram models

First order Markov (bi-gram)

$$y_t \in \{1, ..., K\} \quad p(y_1 = k) = \pi_k^0 \qquad p(y_t = k|y_{t-1} = l) = T_{k,l}$$

Q1. How can we compute the marginal distribution over the second state?

$$p(y_2 = k) = \sum_{l=1}^{K} p(y_2 = k|y_1 = l)p(y_1 = l) = \sum_{l=1}^{K} T_{k,l}\pi_l^0$$

Q2. How can we compute the stationary distribution for the Markov chain?

$\equiv$

invariant distribution

eigenvalues of transition matrix

$$p(y_\infty = k) = \sum_l p(y_\infty = k | y_{\infty-1} = l)\, p(y_{\infty-1} = l)$$

equate

$$\underline{p(y_\infty = k)} = \sum_l T_{kl}\, p(\underline{y_{\infty-1}} = l)$$

$\lambda_\mu = 1$

$$1 \times \underline{p}_\infty = \underline{\underline{T}}\, \underline{p}_\infty \qquad \underline{\underline{T}}\, \underline{e}_\mu = \lambda_\mu \underline{e}_\mu$$

Some questions about n-gram models

First order Markov (bi-gram)

$$y_t \in \{1, ..., K\} \quad p(y_1 = k) = \pi_k^0 \qquad p(y_t = k | y_{t-1} = l) = T_{k,l}$$

Q1. How can we compute the marginal distribution over the second state?

$$p(y_2 = k) = \sum_{l=1}^{K} p(y_2 = k | y_1 = l) p(y_1 = l) = \sum_{l=1}^{K} T_{k,l} \pi_l^0$$

Q2. How can we compute the stationary distribution for the Markov chain?

$$p(y_t = k) = \sum_{l=1}^{K} p(y_t = k | y_{t-1} = l) p(y_{t-1} = l)$$

Some questions about n-gram models

First order Markov (bi-gram)

$$y_t \in \{1, ..., K\} \quad p(y_1 = k) = \pi_k^0 \qquad p(y_t = k | y_{t-1} = l) = T_{k,l}$$

Q1. How can we compute the marginal distribution over the second state?

$$p(y_2 = k) = \sum_{l=1}^{K} p(y_2 = k | y_1 = l) p(y_1 = l) = \sum_{l=1}^{K} T_{k,l} \pi_l^0$$

Q2. How can we compute the stationary distribution for the Markov chain?

$$p(y_t = k) = \sum_{l=1}^{K} p(y_t = k | y_{t-1} = l) p(y_{t-1} = l)$$

$$\pi_k^\infty = \sum_{l=1}^{K} T_{k,l} \pi_l^\infty$$

eigenvectors of
transition matrix
with eigenvalue = 1

Some questions about n-gram models

First order Markov (bi-gram)

$$y_t \in \{1, ..., K\} \quad p(y_1 = k) = \pi_k^0 \qquad p(y_t = k | y_{t-1} = l) = T_{k,l}$$

Q1. How can we compute the marginal distribution over the second state?

$$p(y_2 = k) = \sum_{l=1}^K p(y_2 = k | y_1 = l) p(y_1 = l) = \sum_{l=1}^K T_{k,l} \pi_l^0$$

Q2. How can we compute the stationary distribution for the Markov chain?

$$p(y_t = k) = \sum_{l=1}^K p(y_t = k | y_{t-1} = l) p(y_{t-1} = l)$$

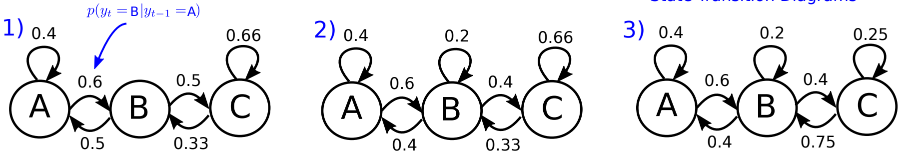
$$\pi_k^\infty = \sum_{l=1}^K T_{k,l} \pi_l^\infty$$

eigenvectors of transition matrix with eigenvalue = 1

Q3. Which transition matrix is most compatible with the following sequence?

ABAAABBABCCCBC

'State Transition Diagrams'



Some questions about n-gram models

to state → / from state	A	B	C	total
A	2/5	3/5	0	5
B	2/5	1/5	2/5	5
C	0	1/3	2/3	3

First order Markov (bi-gram)

$$y_t \in \{1, ..., K\} \quad p(y_1 = k) = \pi_k^0 \qquad p(y_t = k | y_{t-1} = l) = T_{k,l}$$

Q1. How can we compute the marginal distribution over the second state?

$$p(y_2 = k) = \sum_{l=1}^{K} p(y_2 = k | y_1 = l) p(y_1 = l) = \sum_{l=1}^{K} T_{k,l} \pi_l^0$$

Q2. How can we compute the stationary distribution for the Markov chain?

$$p(y_t = k) = \sum_{l=1}^{K} p(y_t = k | y_{t-1} = l) p(y_{t-1} = l)$$

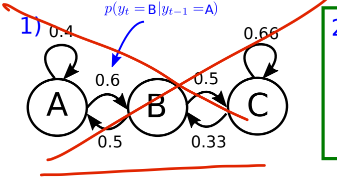
$$\pi_k^\infty = \sum_{l=1}^{K} T_{k,l} \pi_l^\infty$$

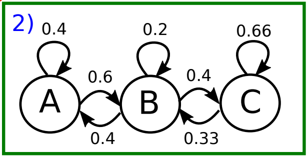
eigenvectors of transition matrix with eigenvalue = 1

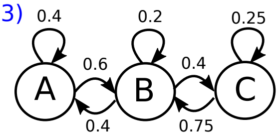
Q3. Which transition matrix is most compatible with the following sequence?

ABAAABBABCCCBC

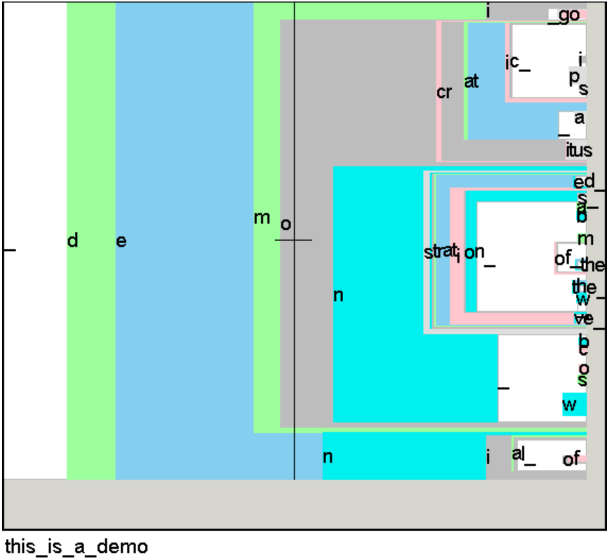
'State Transition Diagrams'







Example application of n-grams: text modelling for dasher

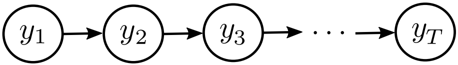


http://www.inference.phy.cam.ac.uk/dasher/ https://www.youtube.com/watch?v=nr3s4613DX8

Markov models for discrete data: n-gram models

First order Markov (bi-gram)

$$p(y_1, y_2, y_3, \ldots, y_T) = p(y_1)p(y_2|y_1)p(y_3|y_2)\ldots p(y_T|y_{T-1})$$



$y_t \in \{1, ..., K\}$ — discrete states

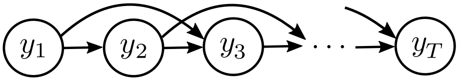
$p(y_1 = k) = \pi_k^0$ — initial state probabilities

$p(y_t = k|y_{t-1} = l) = T_{k,l}$ — transition probabilities (stochastic matrix)

$$\sum_{k=1}^{K} T_{k,l} = 1$$

Second order Markov (tri-gram)

$$p(y_1, y_2, y_3, \ldots, y_T) = p(y_1)p(y_2|y_1)p(y_3|y_2, y_1)\ldots p(y_T|y_{T-1}, y_{T-2})$$



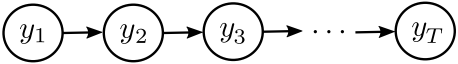
$p(y_t = k|y_{t-1} = l, y_{t-2} = m) = T_{k,l,m}$

n-grams require large multidimensional arrays

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$p(y_1, y_2, y_3, \ldots, y_T) = p(y_1)p(y_2|y_1)p(y_3|y_2)\ldots p(y_T|y_{T-1})$$



$y_t \in \mathbb{R}^D$ — continuous vector states

$p(y_1) = \mathcal{G}(y_1; \mu_0, \Sigma_0)$ — initial state density

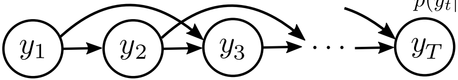
$p(y_t|y_{t-1}) = \mathcal{G}(y_t; \Lambda y_{t-1}, \Sigma)$ — transition density

$$\mathcal{G}(y; \mu, \Sigma) = \frac{1}{(2\pi)^{D/2}\det(\Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right\}$$

Second order Markov (AR(2))

$$p(y_1, y_2, y_3, \ldots, y_T) = p(y_1)p(y_2|y_1)p(y_3|y_2, y_1)\ldots p(y_T|y_{T-1}, y_{T-2})$$

$$p(y_t|y_{t-1}, y_{t-2}) = \mathcal{G}(y_t; \Lambda_1 y_{t-1} + \Lambda_2 y_{t-2}, \Sigma)$$

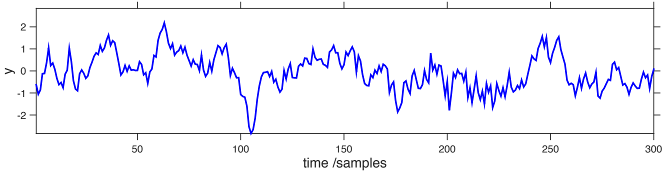


joint distribution over all variables is always multivariate Gaussian

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

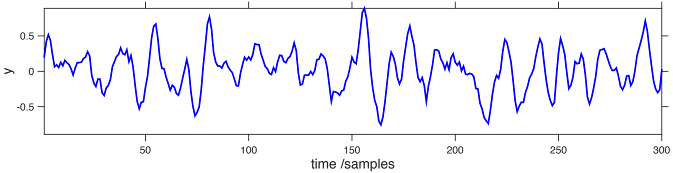
$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



Second order Markov (AR(2))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}, y_{t-2}) = \mathcal{G}(y_t; \lambda_1 y_{t-1} + \lambda_2 y_{t-2}, \sigma^2)$$

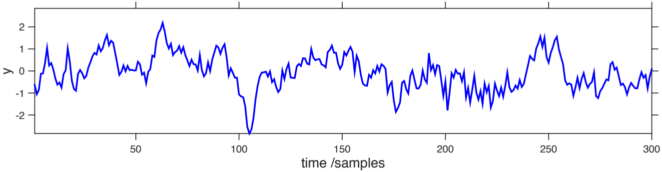
$$[\lambda_1, \lambda_2] = [1.57, -0.78] \qquad \sigma^2 = 0.01$$



Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$

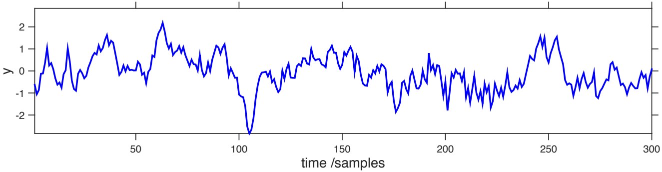


What is the stationary distribution of this process? $p(y_\infty) = ?$

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



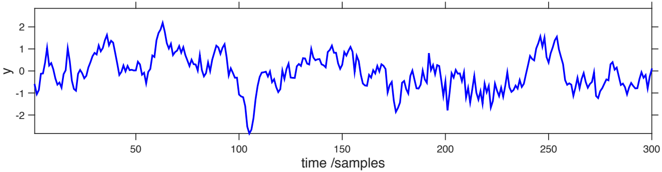
What is the stationary distribution of this process? $p(y_\infty) = \, ?$

Everything is linear Gaussian => must be Gaussian $p(y_\infty) = \mathcal{G}(y_\infty; \mu_\infty, \sigma^2_\infty)$

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



What is the stationary distribution of this process? $p(y_\infty) = ?$

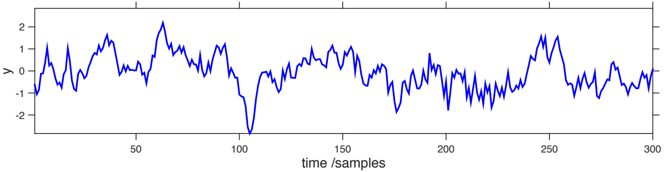
Everything is linear Gaussian => must be Gaussian $p(y_\infty) = \mathcal{G}(y_\infty; \mu_\infty, \sigma^2_\infty)$

$$y_t = \lambda y_{t-1} + \sigma \epsilon_t \;\; \epsilon_t \sim \mathcal{G}(0, 1)$$

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



What is the stationary distribution of this process? $p(y_\infty) = \,?$

Everything is linear Gaussian => must be Gaussian $p(y_\infty) = \mathcal{G}(y_\infty; \mu_\infty, \sigma^2_\infty)$

$$y_t = \lambda y_{t-1} + \sigma\epsilon_t \;\; \epsilon_t \sim \mathcal{G}(0, 1)$$

Mean: $\langle y_t \rangle = \langle \lambda y_{t-1} + \sigma \epsilon_t \rangle = \lambda \langle y_{t-1} \rangle + \sigma \langle \epsilon_t \rangle$

$= \lambda \langle y_{t-1} \rangle$ $\qquad 0$

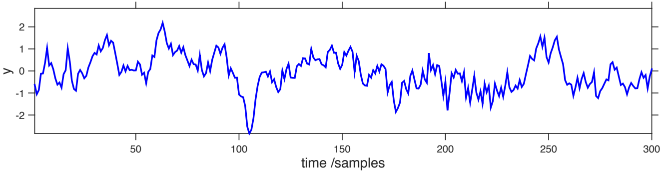
$\mathbb{E}(y_t)$

$\mu_\infty = \lambda \mu_\infty \Rightarrow \mu_\infty = 0$

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



What is the stationary distribution of this process? $p(y_\infty) = \,?$

Everything is linear Gaussian => must be Gaussian $p(y_\infty) = \mathcal{G}(y_\infty; \mu_\infty, \sigma^2_\infty)$

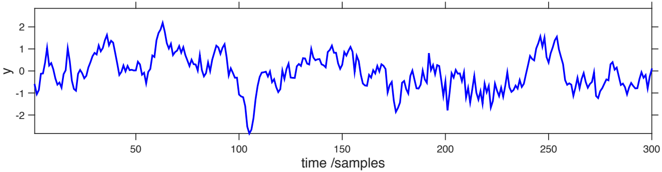
$$y_t = \lambda y_{t-1} + \sigma \epsilon_t \;\; \epsilon_t \sim \mathcal{G}(0,1)$$

Mean: $\langle y_t \rangle = \lambda \langle y_{t-1} \rangle + \sigma \langle \epsilon_t \rangle = 0$

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



What is the stationary distribution of this process? $p(y_\infty) = ?$

Everything is linear Gaussian => must be Gaussian $p(y_\infty) = \mathcal{G}(y_\infty; \mu_\infty, \sigma^2_\infty)$

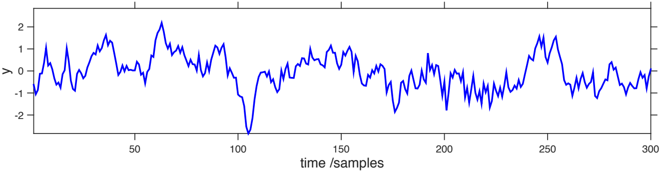
$$y_t = \lambda y_{t-1} + \sigma \epsilon_t \;\; \epsilon_t \sim \mathcal{G}(0, 1)$$

Mean: $\langle y_t \rangle = \lambda \langle y_{t-1} \rangle + \sigma \langle \epsilon_t \rangle = 0 \qquad \mu_\infty = 0$

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



What is the stationary distribution of this process? $p(y_\infty) = \, ?$

Everything is linear Gaussian => must be Gaussian $p(y_\infty) = \mathcal{G}(y_\infty; \mu_\infty, \sigma^2_\infty)$

$$y_t = \lambda y_{t-1} + \sigma \epsilon_t \;\; \epsilon_t \sim \mathcal{G}(0,1)$$

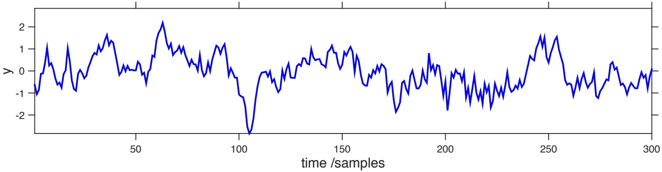
Mean: $\langle y_t \rangle = \lambda \langle y_{t-1} \rangle + \sigma \langle \epsilon_t \rangle = 0$ $\qquad \mu_\infty = 0$

Variance: $\langle y_t^2 \rangle$

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



What is the stationary distribution of this process? $p(y_\infty) = \, ?$

Everything is linear Gaussian => must be Gaussian $p(y_\infty) = \mathcal{G}(y_\infty; \mu_\infty, \sigma^2_\infty)$

$$y_t = \lambda y_{t-1} + \sigma \epsilon_t \;\; \epsilon_t \sim \mathcal{G}(0,1)$$

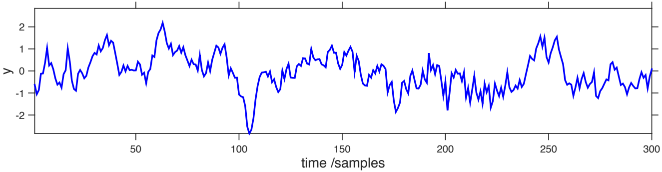
Mean: $\langle y_t \rangle = \lambda \langle y_{t-1} \rangle + \sigma \langle \epsilon_t \rangle = 0$ $\qquad \mu_\infty = 0$

Variance: $\langle y_t^2 \rangle = \langle (\lambda y_{t-1} + \sigma \epsilon_t)^2 \rangle$

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



What is the stationary distribution of this process? $p(y_\infty) = \,?$

Everything is linear Gaussian => must be Gaussian $p(y_\infty) = \mathcal{G}(y_\infty; \mu_\infty, \sigma^2_\infty)$

$$y_t = \lambda y_{t-1} + \sigma \epsilon_t \;\; \epsilon_t \sim \mathcal{G}(0,1)$$

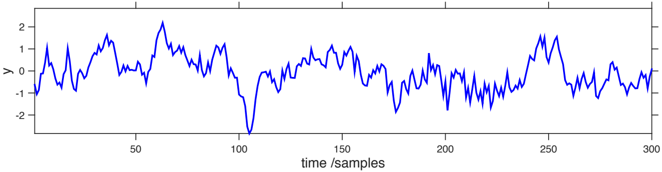
Mean: $\langle y_t \rangle = \lambda \langle y_{t-1} \rangle + \sigma \langle \epsilon_t \rangle = 0 \qquad \mu_\infty = 0$

Variance: $\langle y_t^2 \rangle = \langle (\lambda y_{t-1} + \sigma \epsilon_t)^2 \rangle = \lambda^2 \langle y_{t-1}^2 \rangle + 2\lambda\, \sigma \langle y_{t-1} \epsilon_t \rangle + \sigma^2 \langle \epsilon_t^2 \rangle$

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



What is the stationary distribution of this process? $p(y_\infty) = \, ?$

Everything is linear Gaussian => must be Gaussian $p(y_\infty) = \mathcal{G}(y_\infty; \mu_\infty, \sigma^2_\infty)$

$$y_t = \lambda y_{t-1} + \sigma \epsilon_t \;\; \epsilon_t \sim \mathcal{G}(0,1)$$

Mean: $\langle y_t \rangle = \lambda \langle y_{t-1} \rangle + \sigma \langle \epsilon_t \rangle = 0 \qquad \mu_\infty = 0$

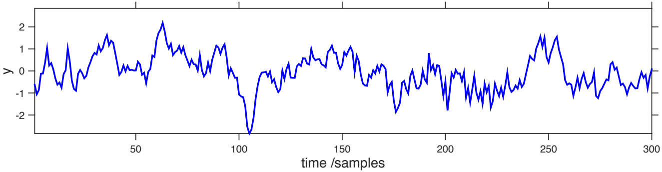
Variance: $\langle y_t^2 \rangle = \langle (\lambda y_{t-1} + \sigma \epsilon_t)^2 \rangle = \lambda^2 \langle y_{t-1}^2 \rangle + 2\lambda\,\sigma \langle y_{t-1} \epsilon_t \rangle + \sigma^2 \langle \epsilon_t^2 \rangle$

$$\langle y_t^2 \rangle = \lambda^2 \langle y_{t-1}^2 \rangle + \sigma^2$$

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



What is the stationary distribution of this process? $p(y_\infty) = \,?$

Everything is linear Gaussian => must be Gaussian $p(y_\infty) = \mathcal{G}(y_\infty; \mu_\infty, \sigma^2_\infty)$

$$y_t = \lambda y_{t-1} + \sigma \epsilon_t \;\; \epsilon_t \sim \mathcal{G}(0,1)$$

Mean: $\langle y_t \rangle = \lambda \langle y_{t-1} \rangle + \sigma \langle \epsilon_t \rangle = 0$ $\qquad \mu_\infty = 0$

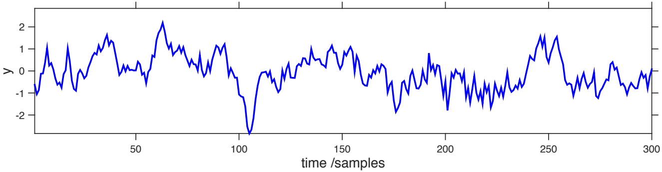
Variance: $\langle y_t^2 \rangle = \langle (\lambda y_{t-1} + \sigma \epsilon_t)^2 \rangle = \lambda^2 \langle y_{t-1}^2 \rangle + 2\lambda\,\sigma \langle y_{t-1} \epsilon_t \rangle + \sigma^2 \langle \epsilon_t^2 \rangle$

$$\langle y_t^2 \rangle = \lambda^2 \langle y_{t-1}^2 \rangle + \sigma^2 \quad \sigma^2_\infty = \lambda^2 \sigma^2_\infty + \sigma^2$$

Markov models for continuous data: Auto-Regressive (AR) Gaussian models

First order Markov (AR(1))

$$y_t \in \mathbb{R}^1 \qquad p(y_t|y_{t-1}) = \mathcal{G}(y_t; \lambda y_{t-1}, \sigma^2) \qquad \lambda = 0.9 \quad \sigma^2 = 0.01$$



What is the stationary distribution of this process? $p(y_\infty) = \, ?$

Everything is linear Gaussian => must be Gaussian $p(y_\infty) = \mathcal{G}(y_\infty; \mu_\infty, \sigma^2_\infty)$

$$y_t = \lambda y_{t-1} + \sigma \epsilon_t \;\; \epsilon_t \sim \mathcal{G}(0,1)$$

Mean: $\langle y_t \rangle = \lambda \langle y_{t-1} \rangle + \sigma \langle \epsilon_t \rangle = 0 \qquad \mu_\infty = 0$

Variance: $\langle y_t^2 \rangle = \langle (\lambda y_{t-1} + \sigma \epsilon_t)^2 \rangle = \lambda^2 \langle y_{t-1}^2 \rangle + 2\lambda \, \sigma \langle y_{t-1} \epsilon_t \rangle + \sigma^2 \langle \epsilon_t^2 \rangle$

$$\langle y_t^2 \rangle = \lambda^2 \langle y_{t-1}^2 \rangle + \sigma^2 \quad \sigma^2_\infty = \lambda^2 \sigma^2_\infty + \sigma^2 \qquad \sigma^2_\infty = \frac{\sigma^2}{1-\lambda^2}$$

3F8 Q&A Wed 9.30 – 11.00
CBL Seminar Room
See Moodle for details

Markov Models

1st Order

$$p(y_{1:T}) = p(y_1)p(y_2|y_1)\cdots p(y_T|y_{T-1})$$

discrete y ⇒ bigram models

$$p(y_1 = k) = \pi_k^0$$

$$p(y_t = k | y_{t-1} = l) = T_{kl}$$

$$\Downarrow$$

Stationary / invariant distribution

$$p(y_\infty = k) = \pi_k^\infty = \sum_k T_{kl} \pi_k^\infty$$

continuous y ⇒ autoregressive

$$p(\underline{y}_1) = \mathcal{G}(\underline{y}_1 ; \underline{\mu}_0, \underline{\underline{\Sigma}}_0)$$

$$\dim(\underline{y}_1) = D$$

D×D matrices

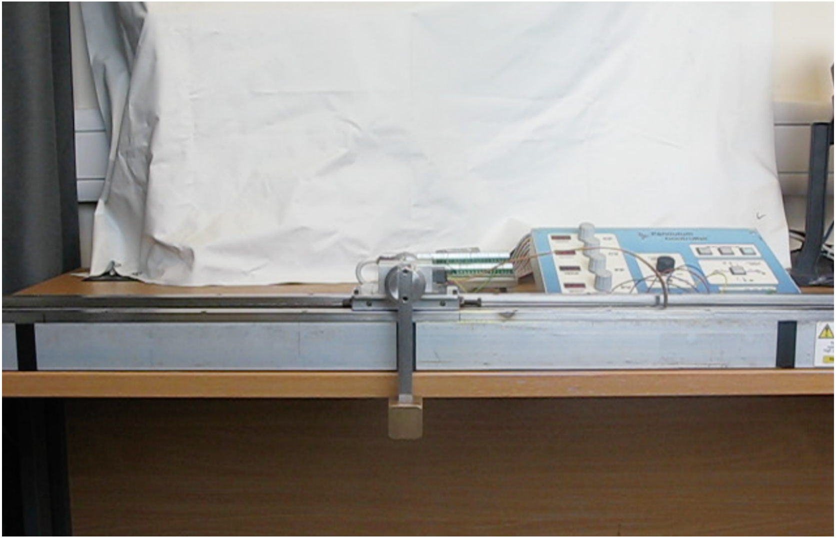
$$p(\underline{y}_t | \underline{y}_{t-1}) = \mathcal{G}(\underline{y}_t ; \underline{\underline{\Lambda}}\,\underline{y}_{t-1}, \underline{\underline{\Sigma}})$$

$$\Downarrow$$

Stationary distribution

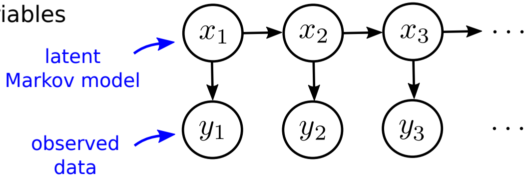
$$p(y_\infty) = \mathcal{G}\left(y_\infty ; \mu_\infty = 0, \sigma_\infty^2 = \frac{\sigma^2}{1-\lambda^2}\right)$$

Example application of Markov Models: pendulum swing up control problem



Hidden Markov models

Real data depend on latent variables



ASR

x phonemes/words

y waveform/feature

Computer Vision

x objects, pose, lighting

y image pixel intensities

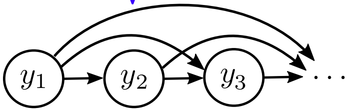
Natural Language Processing

x topics

y words

$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

marginalise latents



$p(y_{1:T})$

fully connected

Two prevelant Examples:

Hidden Markov Models (discrete x)
Linear Gaussian State Space Models (Gaussian x and y)

Hidden Markov models: discrete hidden state

Discrete Hidden State

$$x_t \in \{1, ..., K\}$$

$$p(x_t = k | x_{t-1} = l) = T_{k,l}$$

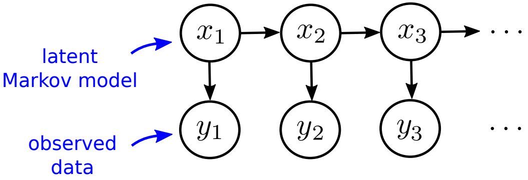
E.g. in examples below $K = 2$

$$T = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$$

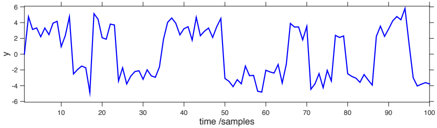
Continuous Observed State

$$p(y_t | x_t = k) = \mathcal{G}(y_t; \mu_k, \Sigma_k)$$

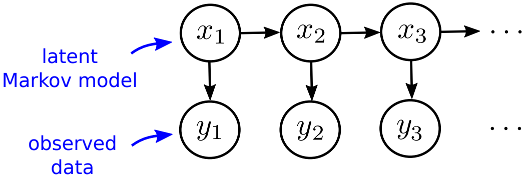
$\mu_1 = 3 \qquad \mu_2 = -3 \qquad \sigma_1^2 = \sigma_2^2 = 1$



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t | x_{t-1}) p(y_t | x_t)$$



Hidden Markov models: discrete hidden state



Discrete Hidden State

$$x_t \in \{1, ..., K\}$$

$$p(x_t = k | x_{t-1} = l) = T_{k,l}$$

E.g. in examples below $K = 2$

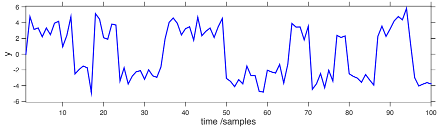
$$T = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$$

$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t | x_{t-1}) p(y_t | x_t)$$

Continuous Observed State

$$p(y_t | x_t = k) = \mathcal{G}(y_t; \mu_k, \Sigma_k)$$

$\mu_1 = 3$ $\mu_2 = -3$ $\sigma_1^2 = \sigma_2^2 = 1$



Discrete Observed State

$$p(y_t = l | x_t = k) = S_{l,k}$$

$$S = \begin{bmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{bmatrix}$$

ABBBBAAABAAACCCCBBBBBBCCCCCCCCCC
AAABBBBAABAAABBCCCCCCCCCCCCCCCBBA
AACCCCCCBABCCCCCCCAABBAABABCCCCC

Hidden Markov models: discrete hidden state

Discrete Hidden State, Continuous Observed State

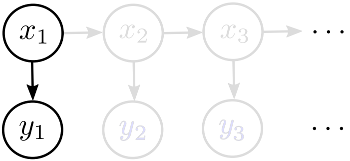
$x_t \in \{1, ..., K\}$

$p(x_1 = k) = \pi_k^0$

$p(x_t = k | x_{t-1} = l) = T_{k,l}$

$p(y_t | x_t = k) = \mathcal{G}(y_t; \mu_k, \Sigma_k)$

$t=1$



Consider T = 1

Q1: What type of distribution is $p(y_1)$?

sum rule

product rule

$$p(y_1) = \sum_{k=1}^{K} p(y_1, x_1 = k) = \sum_{k=1}^{K} p(x_1 = k)\, p(y_1 | x_1 = k)$$

$$\boxed{p(y_1) = \sum_{k=1}^{K} \pi_k^0 \, \mathcal{G}(y_1; \mu_k, \Sigma_k)}$$

Hidden Markov models: discrete hidden state

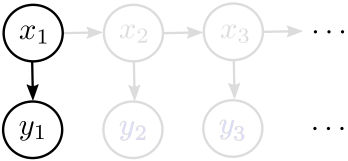
Discrete Hidden State, Continuous Observed State

$$x_t \in \{1, ..., K\}$$

$$p(x_1 = k) = \pi_k^0$$

$$p(x_t = k | x_{t-1} = l) = T_{k,l}$$

$$p(y_t | x_t = k) = \mathcal{G}(y_t; \mu_k, \Sigma_k)$$



Consider T = 1

Q1: What type of distribution is $p(y_1)$?

$$p(y_1) = \sum_k p(y_1 | x_1 = k) p(x_1 = k)$$

Hidden Markov models: discrete hidden state

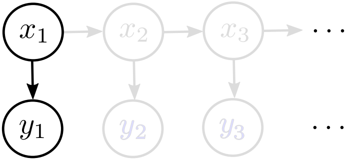
Discrete Hidden State, Continuous Observed State

$x_t \in \{1, ..., K\}$

$p(x_1 = k) = \pi_k^0$

$p(x_t = k | x_{t-1} = l) = T_{k,l}$

$p(y_t | x_t = k) = \mathcal{G}(y_t; \mu_k, \Sigma_k)$



Consider T = 1

Q1: What type of distribution is $p(y_1)$?

$$p(y_1) = \sum_k p(y_1 | x_1 = k) p(x_1 = k) = \sum_k \pi_k^0 \mathcal{G}(y_1; \mu_k, \Sigma_k)$$

Hidden Markov models: discrete hidden state

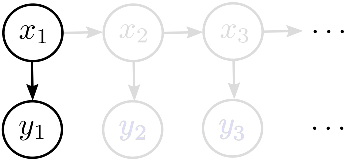
Discrete Hidden State, Continuous Observed State

$$x_t \in \{1, ..., K\}$$

$$p(x_1 = k) = \pi^0_k$$

$$p(x_t = k | x_{t-1} = l) = T_{k,l}$$

$$p(y_t | x_t = k) = \mathcal{G}(y_t; \mu_k, \Sigma_k)$$



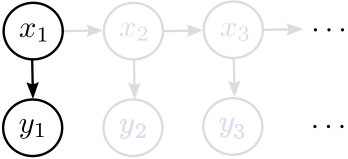
Consider T = 1

Q1: What type of distribution is $p(y_1)$?

$$p(y_1) = \sum_k p(y_1 | x_1 = k) p(x_1 = k) = \sum_k \pi^0_k \mathcal{G}(y_1; \mu_k, \Sigma_k)$$

Q2: What distribution does $p(y_t)$ converge to after a long time?

Hidden Markov models: discrete hidden state



Consider T = 1

Discrete Hidden State, Continuous Observed State

$$x_t \in \{1, ..., K\}$$
$$p(x_1 = k) = \pi_k^0$$
$$p(x_t = k | x_{t-1} = l) = T_{k,l}$$
$$p(y_t | x_t = k) = \mathcal{G}(y_t; \mu_k, \Sigma_k)$$

Q1: What type of distribution is $p(y_1)$?

$$p(y_1) = \sum_k p(y_1 | x_1 = k) p(x_1 = k) = \sum_k \pi_k^0 \mathcal{G}(y_1; \mu_k, \Sigma_k)$$

Q2: What distribution does $p(y_t)$ converge to after a long time?

stationary distribution of Markov chain satifies $\pi_k^\infty = \sum_{l=1}^K T_{k,l} \pi_l^\infty$

Hidden Markov models: discrete hidden state

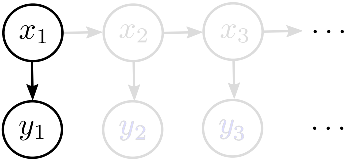
Discrete Hidden State, Continuous Observed State

$$x_t \in \{1, ..., K\}$$

$$p(x_1 = k) = \pi_k^0$$

$$p(x_t = k | x_{t-1} = l) = T_{k,l}$$

$$p(y_t | x_t = k) = \mathcal{G}(y_t; \mu_k, \Sigma_k)$$



Consider T = 1

Q1: What type of distribution is $p(y_1)$?

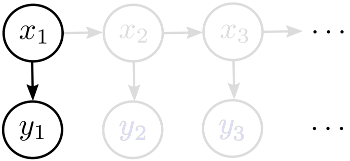
$$p(y_1) = \sum_k p(y_1 | x_1 = k) p(x_1 = k) = \sum_k \pi_k^0 \mathcal{G}(y_1; \mu_k, \Sigma_k)$$

Q2: What distribution does $p(y_t)$ converge to after a long time?

stationary distribution of Markov chain satifies $\pi_k^\infty = \sum_{l=1}^K T_{k,l} \pi_l^\infty$

$$p(y_t) = \sum_k p(y_t | x_t = k) p(x_t = k)$$

Hidden Markov models: discrete hidden state



Consider T = 1

Discrete Hidden State, Continuous Observed State

$$x_t \in \{1, ..., K\}$$

$$p(x_1 = k) = \pi_k^0$$

$$p(x_t = k | x_{t-1} = l) = T_{k,l}$$

$$p(y_t | x_t = k) = \mathcal{G}(y_t; \mu_k, \Sigma_k)$$

Q1: What type of distribution is $p(y_1)$?

$$p(y_1) = \sum_k p(y_1 | x_1 = k) p(x_1 = k) = \sum_k \pi_k^0 \mathcal{G}(y_1; \mu_k, \Sigma_k)$$

Q2: What distribution does $p(y_t)$ converge to after a long time?

stationary distribution of Markov chain satifies $\pi_k^\infty = \sum_{l=1}^K T_{k,l} \pi_l^\infty$

$$p(y_t) = \sum_k p(y_t | x_t = k) p(x_t = k) \rightarrow \sum_k \pi_k^\infty \mathcal{G}(y_t; \mu_k, \Sigma_k)$$

Hidden Markov models: discrete hidden state

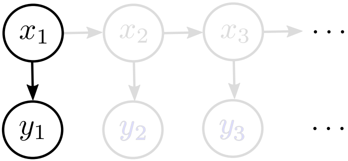
Discrete Hidden State, Continuous Observed State

$$x_t \in \{1, ..., K\}$$

$$p(x_1 = k) = \pi_k^0$$

$$p(x_t = k | x_{t-1} = l) = T_{k,l}$$

$$p(y_t | x_t = k) = \mathcal{G}(y_t; \mu_k, \Sigma_k)$$



Consider T = 1

Q1: What type of distribution is $p(y_1)$?

$$p(y_1) = \sum_k p(y_1 | x_1 = k) p(x_1 = k) = \sum_k \pi_k^0 \mathcal{G}(y_1; \mu_k, \Sigma_k)$$

Q2: What distribution does $p(y_t)$ converge to after a long time?

stationary distribution of Markov chain satifies $\pi_k^\infty = \sum_{l=1}^K T_{k,l} \pi_l^\infty$

$$p(y_t) = \sum_k p(y_t | x_t = k) p(x_t = k) \rightarrow \sum_k \pi_k^\infty \mathcal{G}(y_t; \mu_k, \Sigma_k)$$

this HMM = Mixture of Gaussian Models with dynamic cluster assignments

Hidden Markov models: continuous hidden state (LGSSMs)

Continuous Hidden State

$$x_t \in \mathbb{R}^K$$

$$p(x_t|x_{t-1}) = \mathcal{G}(x_t; Ax_{t-1}, Q)$$

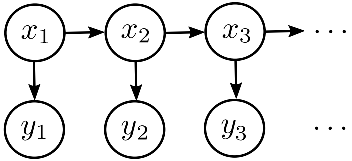
Continuous Observed State

$$y_t \in \mathbb{R}^D$$

$$p(y_t|x_t) = \mathcal{G}(y_t; Cx_t, R)$$

latent
Markov model

observed
data



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

E.g. simple example $\quad K = 2 \quad D = 1$

$$A = \lambda \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$$

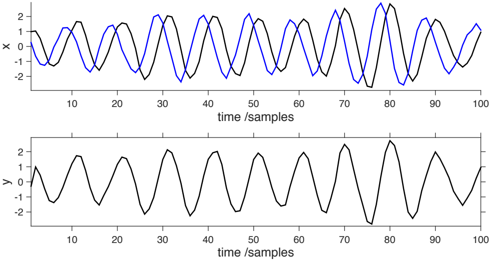
$$\lambda = 0.99 \qquad \theta = 2\pi/10$$

$$Q = (1 - \lambda^2) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

dynamics model

$$C = [1, 0] \qquad R = 0.01$$

obs. model



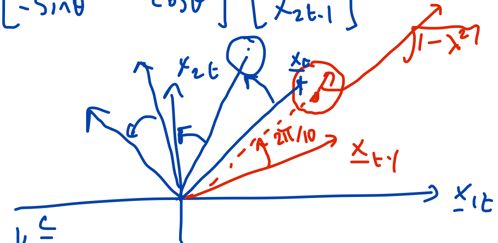
$$\hat{Z}\, N(\theta; \mu, \sigma^2) \approx p^*(\theta) = Z\, p(\theta) \qquad Z = \int p^*(\theta)\, d\theta \approx \boxed{p(\{y\} \mid \{x\}, \sigma, \ell)}$$

$$\underline{x}_t = \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \underline{\underline{A}}\, \underline{x}_{t-1} + \underline{\underline{Q}}^{1/2} \underline{\epsilon}_t \qquad \underline{\epsilon}_t \sim \mathcal{G}(\underline{0}, \underline{\underline{I}})$$

matrix square root

0.99 $\qquad$ $2\pi/10$

$$\begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \lambda \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_{1t-1} \\ x_{2t-1} \end{bmatrix} + (1-\lambda^2)^{1/2} \underline{\epsilon}_t$$



$$y_t = \overset{\underline{c}}{[1, 0]} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} + \text{noise} = x_{1t} + \text{noise}$$

Summary Sequence Modelling Lecture II

$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t \mid x_{t-1})\, p(y_t \mid x_t)$$

observed ($y_{1:T}$) latent ($x_{1:T}$)

Discrete Hidden State $x_t \in \{1 \dots K\}$ (also called HMMs!)

$p(x_t = k \mid x_{t-1} = l) = T_{kl}$

$p(y_t \mid x_t = k) = \mathcal{G}(y_t ; \mu_k, \Sigma_k) \qquad y_t \in \mathbb{R}^D$

emission

$p(y_t = l \mid x_t = k) = \underline{S_{lk}} \qquad y_t \in \{\underline{1 \dots} D\}$

Continuous Hidden State $\underline{x_t \in \mathbb{R}^{K}}$ (linear Gaussian state space models)

$\underline{p(\underline{x}_t \mid \underline{x}_{t-1}) = \mathcal{G}(\underline{x}_t ; \underline{\underline{A}}\, \underline{x}_{t-1}, \underline{\underline{Q}})} \iff \underline{x}_t = \underline{\underline{A}}\, \underline{x}_{t-1} + \underline{\underline{Q}}^{1/2} \underline{\epsilon}_t$

$p(\underline{y}_t \mid \underline{x}_t) = \mathcal{G}(\underline{y}_t ; \underline{\underline{C}}\, \underline{x}_t, \underline{\underline{R}}) \qquad \underline{y}_t \in \mathbb{R}^D$

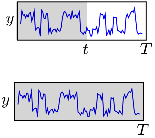
Today: Inference & Learning

Varieties of Inference

Distributional estimates

infer single state or sequence?

future data available?



	marginal	joint
filter	$p(x_t \mid y_{1:t})$	$p(x_{1:t} \mid y_{1:t})$
smoother	$p(x_t \mid y_{1:T})$	$p(x_{1:T} \mid y_{1:T})$

1. LGSSM $p(x_{1:T} \mid y_{1:T}) = \mathcal{G}(x_{1:T}; \mu_{1:T}, \Sigma_{1:T})$

$\Rightarrow x'_{1:T} = \mu_{1:T}$ $\quad \int dx_{\neq t}$

$p(x_t \mid y_{1:T}) = \mathcal{G}(x_t; \mu_t, \Sigma_{tt})$

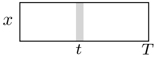
$x_t^* = \mu_t$

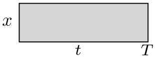
$\Rightarrow x^*_{1:T} = x'_{1:T}$

Varieties of Inference

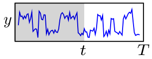
Distributional estimates

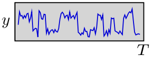
infer single state or sequence?





future data available?





	marginal	joint
filter	$p(x_t \| y_{1:t})$	$p(x_{1:t} \| y_{1:t})$
smoother	$p(x_t \| y_{1:T})$	$p(x_{1:T} \| y_{1:T})$

Point estimates

most probable state @ t

$$x_t^* = \underset{x_t}{\arg\max}\; p(x_t | y_{1:T})$$

most probable sequence

$$x'_{1:T} = \underset{x_{1:T}}{\arg\max}\; p(x_{1:T} | y_{1:T})$$

2. Discrete Hidden State HMM

T=2

x_1	x_2	$p(x_1, x_2 \| y_1, y_2)$
0	0	0.3
0	1	0.4
1	0	0.3
1	1	0

$x^*_{1:2} = [0, 0]$ $\not=$

$x'_{1:2} = [0, 1]$

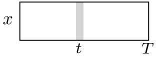
$p(x_1 | y_1, y_2) = [0.7, 0.3]$ $\Rightarrow$ $x_1^* = 0$

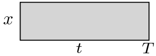
$p(x_2 | y_1, y_2) = [0.6, 0.4]$ $\Rightarrow$ $x_2^* = 0$

Varieties of Inference

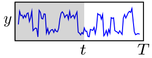
Distributional estimates

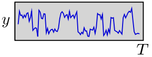
infer single state or sequence?





future data available?





	marginal	joint
filter	$p(x_t \mid y_{1:t})$	$p(x_{1:t} \mid y_{1:t})$
smoother	$p(x_t \mid y_{1:T})$	$p(x_{1:T} \mid y_{1:T})$

Point estimates

most probable state @ t

$$x_t^* = \arg\max_{x_t} p(x_t \mid y_{1:T})$$

most probable sequence

$$x'_{1:T} = \arg\max_{x_{1:T}} p(x_{1:T} \mid y_{1:T})$$

Question: are these estimates the same $x^*_{1:T} \stackrel{?}{=} x'_{1:T}$ for

1. Linear Gaussian State Spate Models?

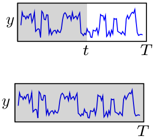
2. Discrete Hidden State HMMs?

Varieties of Inference

Distributional estimates

infer single state or sequence?

future data available?



	marginal	joint
filter	$p(x_t \vert y_{1:t})$	$p(x_{1:t} \vert y_{1:t})$
smoother	$p(x_t \vert y_{1:T})$	$p(x_{1:T} \vert y_{1:T})$

Point estimates

most probable state @ t

most probable sequence

$$x_t^* = \arg\max_{x_t} p(x_t|y_{1:T}) \qquad x'_{1:T} = \arg\max_{x_{1:T}} p(x_{1:T}|y_{1:T})$$

Question: are these estimates the same $x^*_{1:T} \stackrel{?}{=} x'_{1:T}$ for

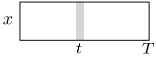
1. Linear Gaussian State Spate Models? $x^*_{1:T} = x'_{1:T}$ (Gaussian)

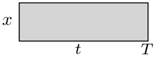
2. Discrete Hidden State HMMs?

Varieties of Inference

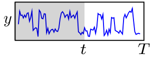
Distributional estimates

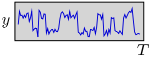
infer single state or sequence?





future data available?





	marginal	joint
filter	$p(x_t \vert y_{1:t})$	$p(x_{1:t} \vert y_{1:t})$
smoother	$p(x_t \vert y_{1:T})$	$p(x_{1:T} \vert y_{1:T})$

Point estimates

most probable state @ t

most probable sequence

$$x_t^* = \arg\max_{x_t} p(x_t|y_{1:T}) \qquad x'_{1:T} = \arg\max_{x_{1:T}} p(x_{1:T}|y_{1:T})$$

Question: are these estimates the same $x^*_{1:T} \stackrel{?}{=} x'_{1:T}$ for

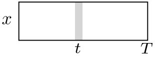
1. Linear Gaussian State Spate Models? $x^*_{1:T} = x'_{1:T}$ (Gaussian)

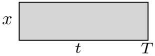
2. Discrete Hidden State HMMs? $x^*_{1:T} \neq x'_{1:T}$

Varieties of Inference

Distributional estimates

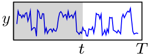
infer single state or sequence?

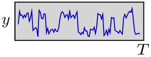




look at this next

future data available?





	marginal	joint
filter	$\boxed{p(x_t \mid y_{1:t})}$	$p(x_{1:t} \mid y_{1:t})$
smoother	$p(x_t \mid y_{1:T})$	$p(x_{1:T} \mid y_{1:T})$

Point estimates

most probable state @ t

most probable sequence

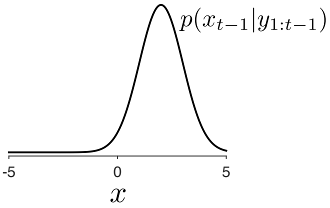
$$x_t^* = \arg\max_{x_t} p(x_t \mid y_{1:T}) \qquad x'_{1:T} = \arg\max_{x_{1:T}} p(x_{1:T} \mid y_{1:T})$$

Question: are these estimates the same $x^*_{1:T} \stackrel{?}{=} x'_{1:T}$ for

1. Linear Gaussian State Spate Models? $x^*_{1:T} = x'_{1:T}$ (Gaussian)

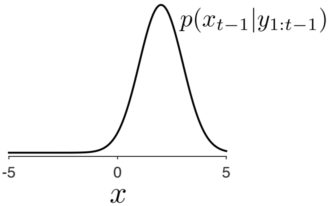
2. Discrete Hidden State HMMs? $x^*_{1:T} \neq x'_{1:T}$

Inference: Kalman Filter



$$p(x_{t-1}|y_{1:t-1})$$

Inference: Kalman Filter



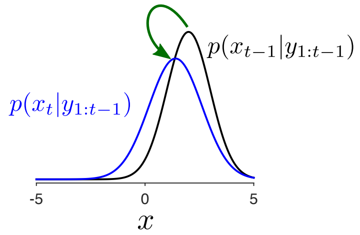
diffuse via dynamics

sum for discrete hidden state

$$p(x_{t-1}|y_{1:t-1})$$

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}$$

Inference: Kalman Filter



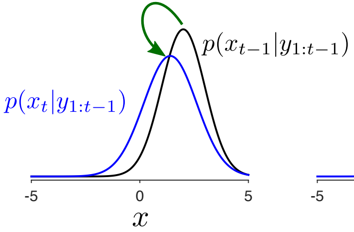
diffuse via dynamics

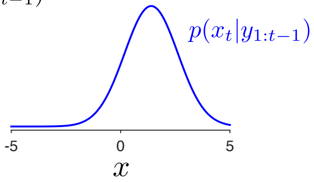
$$p(x_{t-1}|y_{1:t-1})$$

sum for discrete hidden state

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}$$

Inference: Kalman Filter





$$p(x_{t-1}|y_{1:t-1})$$

sum for discrete hidden state

diffuse via dynamics

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}$$

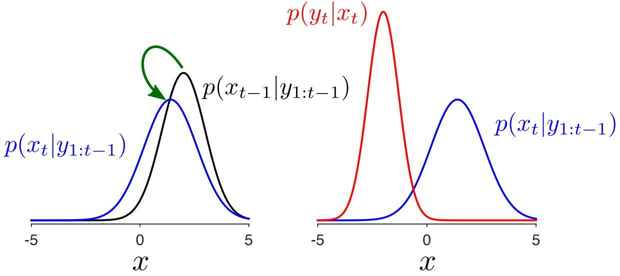
combine with likelihood

$$p(x_t|y_{1:t}) \propto p(x_t|y_{1:t-1})p(y_t|x_t)$$

Bayes' Rule

prior likelihood

Inference: Kalman Filter



diffuse via dynamics

$$p(x_{t-1}|y_{1:t-1})$$

sum for discrete hidden state

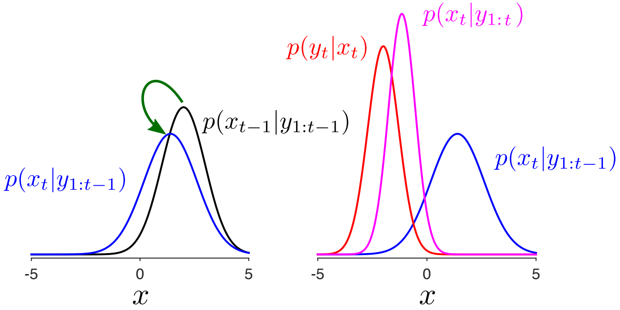
$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}$$

combine with likelihood

$$p(x_t|y_{1:t}) \propto \underbrace{p(x_t|y_{1:t-1})}_{\text{prior}}\underbrace{p(y_t|x_t)}_{\text{likelihood}}$$

Bayes' Rule

Inference: Kalman Filter



diffuse via dynamics

$$p(x_{t-1}|y_{1:t-1})$$

sum for discrete hidden state

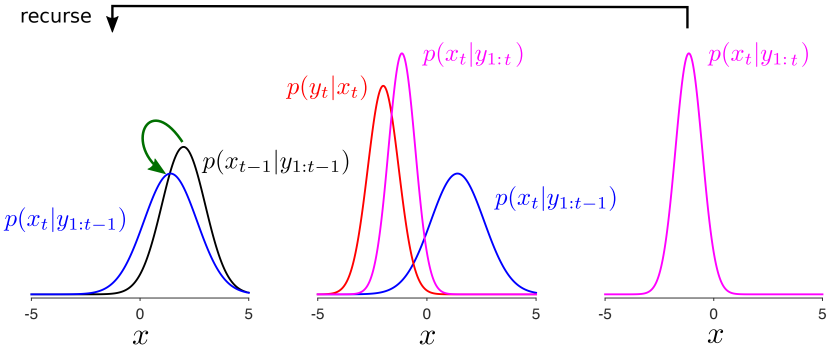
$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}$$

combine with likelihood

$$p(x_t|y_{1:t}) \propto \underbrace{p(x_t|y_{1:t-1})}_{\text{prior}}\underbrace{p(y_t|x_t)}_{\text{likelihood}}$$

Bayes' Rule

Inference: Kalman Filter



$$p(x_{t-1}|y_{1:t-1})$$

diffuse via dynamics

sum for discrete hidden state

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}$$

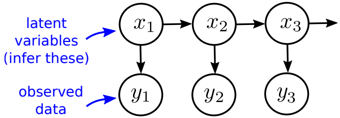
combine with likelihood

$$p(x_t|y_{1:t}) \propto \underbrace{p(x_t|y_{1:t-1})}_{\text{prior}}\underbrace{p(y_t|x_t)}_{\text{likelihood}}$$

Bayes' Rule

Inference: Derivation of General Filtering Equations

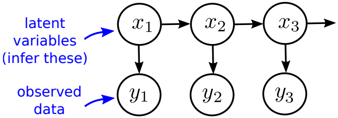
Model



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

Inference: Derivation of General Filtering Equations

Model



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

\+

Rules of probability

product rule

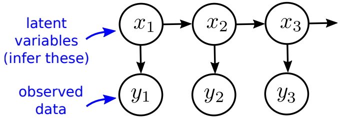
$$p(A|B,C) = \frac{1}{p(B|C)}\, p(B|A,C)p(A|C)$$

sum rule

$$p(A|C) = \sum_B p(A,B|C)$$

Inference: Derivation of General Filtering Equations

Model



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1}) p(y_t|x_t)$$

+

Rules of probability

product rule

$$p(A|B,C) = \frac{1}{p(B|C)} p(B|A,C)p(A|C)$$

sum rule

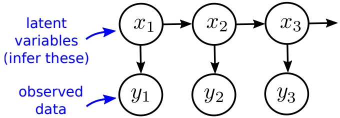
$$p(A|C) = \sum_B p(A,B|C)$$

= ?

Inference

Inference: Derivation of General Filtering Equations

Model



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

+

Rules of probability

product rule

$$p(A|B,C) = \frac{1}{p(B|C)}\, p(B|A,C)p(A|C)$$

sum rule

$$p(A|C) = \sum_B p(A,B|C)$$

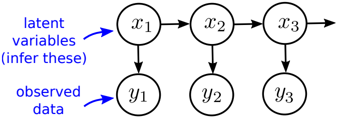
= ?

Inference

$$p(x_t|y_{1:t})$$

Inference: Derivation of General Filtering Equations

Model



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1}) p(y_t|x_t)$$

+

Rules of probability

product rule

$$p(A|B,C) = \frac{1}{p(B|C)} p(B|A,C)p(A|C)$$

sum rule

$$p(A|C) = \sum_B p(A,B|C)$$

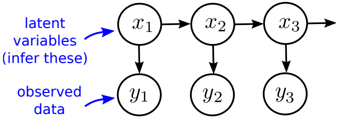
= ?

Inference

$$p(x_t|y_{1:t}) = p(x_t|y_t, y_{1:t-1})$$

Inference: Derivation of General Filtering Equations

Model



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

+

Rules of probability

product rule

$$p(A|B,C) = \frac{1}{p(B|C)}\, p(B|A,C)p(A|C)$$

sum rule

$$p(A|C) = \sum_B p(A,B|C)$$

= ?

Inference

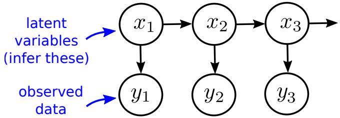
$$\begin{aligned} p(x_t|y_{1:t}) &= p(x_t|y_t, y_{1:t-1}) \\ &= \frac{1}{p(y_t|y_{1:t-1})}\, p(y_t|x_t, y_{1:t-1})p(x_t|y_{1:t-1}) \end{aligned}$$

product rule

$A = x_t \;\; B = y_t \;\; C = y_{1:t-1}$

Inference: Derivation of General Filtering Equations

Model



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

+

Rules of probability

product rule

$$p(A|B,C) = \frac{1}{p(B|C)}\, p(B|A,C)p(A|C)$$

sum rule

$$p(A|C) = \sum_B p(A,B|C)$$

= ?

Inference

$$\begin{aligned}
p(x_t|y_{1:t}) &= p(x_t|y_t, y_{1:t-1}) \\
&= \frac{1}{p(y_t|y_{1:t-1})}\, p(y_t|x_t, y_{1:t-1})p(x_t|y_{1:t-1}) \\
&= \frac{1}{p(y_t|y_{1:t-1})}\, p(y_t|x_t)p(x_t|y_{1:t-1})
\end{aligned}$$

product rule
$A = x_t \;\; B = y_t \;\; C = y_{1:t-1}$

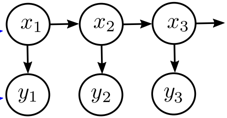
conditional independence from model
$y_t \perp y_{1:t-1}|x_t$

Inference: Derivation of General Filtering Equations

Model

latent variables (infer these)

observed data



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1}) p(y_t|x_t)$$

$+$

Rules of probability

product rule

$$p(A|B,C) = \frac{1}{p(B|C)} p(B|A,C) p(A|C)$$

sum rule

$$p(A|C) = \sum_B p(A,B|C)$$

$=$ **?**

Inference

$$\begin{aligned}
p(x_t|y_{1:t}) &= p(x_t|y_t, y_{1:t-1}) \\
&= \frac{1}{p(y_t|y_{1:t-1})} p(y_t|x_t, y_{1:t-1}) p(x_t|y_{1:t-1}) \\
&= \frac{1}{p(y_t|y_{1:t-1})} p(y_t|x_t) p(x_t|y_{1:t-1}) \\
&\propto p(y_t|x_t) p(x_t|y_{1:t-1})
\end{aligned}$$

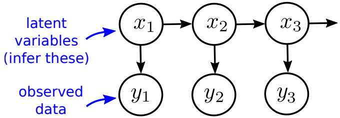
product rule $A = x_t$ $B = y_t$ $C = y_{1:t-1}$

conditional independence from model $y_t \perp y_{1:t-1} | x_t$

constant of proportionality $p(y_t|y_{1:t-1})$ (see learning)

Inference: Derivation of General Filtering Equations

Model



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

$+$

Rules of probability

product rule

$$p(A|B,C) = \frac{1}{p(B|C)}\, p(B|A,C)p(A|C)$$

sum rule

$$p(A|C) = \sum_B p(A,B|C)$$

$= \ ?$

Inference

$$\begin{aligned}
p(x_t|y_{1:t}) &= p(x_t|y_t, y_{1:t-1}) \\
&= \frac{1}{p(y_t|y_{1:t-1})}\, p(y_t|x_t, y_{1:t-1})p(x_t|y_{1:t-1}) \\
&= \frac{1}{p(y_t|y_{1:t-1})}\, p(y_t|x_t)p(x_t|y_{1:t-1}) \\
&\propto p(y_t|x_t)p(x_t|y_{1:t-1})
\end{aligned}$$

product rule $A = x_t \ B = y_t \ C = y_{1:t-1}$

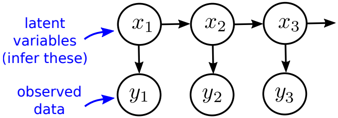
conditional independence from model $y_t \perp y_{1:t-1}|x_t$

constant of proportionality $p(y_t|y_{1:t-1})$ (see learning)

$$p(x_t|y_{1:t-1})$$

Inference: Derivation of General Filtering Equations

Model



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1}) p(y_t|x_t)$$

+

Rules of probability

product rule

$$p(A|B,C) = \frac{1}{p(B|C)} p(B|A,C) p(A|C)$$

sum rule

$$p(A|C) = \textstyle\sum_B p(A,B|C)$$

= ?

Inference

$$
\begin{aligned}
p(x_t|y_{1:t}) &= p(x_t|y_t, y_{1:t-1}) \\
&= \frac{1}{p(y_t|y_{1:t-1})} p(y_t|x_t, y_{1:t-1}) p(x_t|y_{1:t-1}) \\
&= \frac{1}{p(y_t|y_{1:t-1})} p(y_t|x_t) p(x_t|y_{1:t-1}) \\
&\propto p(y_t|x_t) p(x_t|y_{1:t-1})
\end{aligned}
$$

product rule $A = x_t \;\; B = y_t \;\; C = y_{1:t-1}$

conditional independence from model $y_t \perp y_{1:t-1}|x_t$

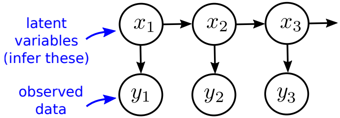
constant of proportionality $p(y_t|y_{1:t-1})$ (see learning)

$$p(x_t|y_{1:t-1}) = \int p(x_t, x_{t-1}|y_{1:t-1}) \mathrm{d}x_{t-1}$$

sum rule $A = x_t \;\; B = x_{t-1} \;\; C = y_{1:t-1}$

Inference: Derivation of General Filtering Equations

Model



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

+

Rules of probability

product rule

$$p(A|B,C) = \frac{1}{p(B|C)}\, p(B|A,C)p(A|C)$$

sum rule

$$p(A|C) = \textstyle\sum_B p(A,B|C)$$

= ?

Inference

$$p(x_t|y_{1:t}) = p(x_t|y_t, y_{1:t-1})$$

$$= \frac{1}{p(y_t|y_{1:t-1})}\, p(y_t|x_t, y_{1:t-1})p(x_t|y_{1:t-1})$$

product rule $A = x_t \;\; B = y_t \;\; C = y_{1:t-1}$

$$= \frac{1}{p(y_t|y_{1:t-1})}\, p(y_t|x_t)p(x_t|y_{1:t-1})$$

conditional independence from model $y_t \perp y_{1:t-1}|x_t$

$$\propto p(y_t|x_t)p(x_t|y_{1:t-1})$$

constant of proportionality $p(y_t|y_{1:t-1})$ (see learning)

$$p(x_t|y_{1:t-1}) = \int p(x_t, x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}$$

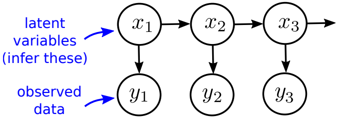
sum rule $A = x_t \;\; B = x_{t-1} \;\; C = y_{1:t-1}$

$$= \int p(x_t|x_{t-1}, y_{1:t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}$$

product rule

Inference: Derivation of General Filtering Equations

Model



$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

+

Rules of probability

product rule

$$p(A|B,C) = \frac{1}{p(B|C)}\, p(B|A,C)p(A|C)$$

sum rule

$$p(A|C) = \textstyle\sum_B p(A,B|C)$$

= ?

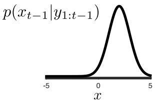
Inference

$$
\begin{aligned}
p(x_t|y_{1:t}) &= p(x_t|y_t, y_{1:t-1}) \\
&= \frac{1}{p(y_t|y_{1:t-1})}\, p(y_t|x_t, y_{1:t-1})p(x_t|y_{1:t-1}) && \text{product rule } A = x_t \;\; B = y_t \;\; C = y_{1:t-1} \\
&= \frac{1}{p(y_t|y_{1:t-1})}\, p(y_t|x_t)p(x_t|y_{1:t-1}) && \text{conditional independence from model } y_t \perp y_{1:t-1}|x_t \\
&\propto p(y_t|x_t)p(x_t|y_{1:t-1}) && \text{constant of proportionality } p(y_t|y_{1:t-1}) \text{ (see learning)}
\end{aligned}
$$

$$
\begin{aligned}
p(x_t|y_{1:t-1}) &= \int p(x_t, x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1} && \text{sum rule } A = x_t \;\; B = x_{t-1} \;\; C = y_{1:t-1} \\
&= \int p(x_t|x_{t-1}, y_{1:t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1} && \text{product rule} \\
&= \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1} && \text{conditional independence from model}
\end{aligned}
$$

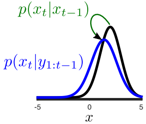
Inference: Kalman Filter

$$p(x_{t-1}|y_{1:t-1})$$



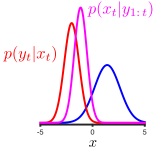
diffuse via dynamics

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}$$



combine with likelihood

$$p(x_t|y_{1:t}) \propto \underbrace{p(x_t|y_{1:t-1})}_{\text{prior}}\underbrace{p(y_t|x_t)}_{\text{likelihood}}$$

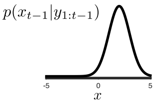


Inference: Kalman Filter

$$p(x_{t-1}|y_{1:t-1}) = \mathcal{G}(x_{t-1}; \mu_{t-1}^{t-1}, V_{t-1}^{t-1})$$

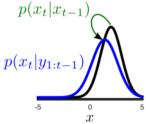
most recent data used in prediction

variable being predicted



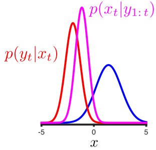
diffuse via dynamics

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) \mathrm{d}x_{t-1}$$



combine with likelihood

$$p(x_t|y_{1:t}) \propto \underbrace{p(x_t|y_{1:t-1})}_{\text{prior}} \underbrace{p(y_t|x_t)}_{\text{likelihood}}$$

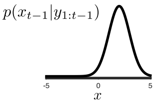


Inference: Kalman Filter

$$p(x_{t-1}|y_{1:t-1}) = \mathcal{G}(x_{t-1}; \mu_{t-1}^{t-1}, V_{t-1}^{t-1})$$

most recent data used in prediction

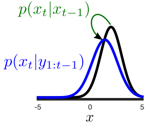
variable being predicted



diffuse via dynamics

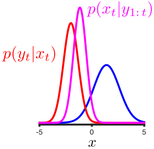
$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) \mathrm{d}x_{t-1}$$

$$p(x_t|y_{1:t-1}) = \mathcal{G}(x_t; \mu_t^{t-1}, V_t^{t-1})$$



combine with likelihood

$$p(x_t|y_{1:t}) \propto \underbrace{p(x_t|y_{1:t-1})}_{\text{prior}} \underbrace{p(y_t|x_t)}_{\text{likelihood}}$$

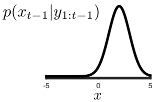


Inference: Kalman Filter

$$p(x_{t-1}|y_{1:t-1}) = \mathcal{G}(x_{t-1}; \mu_{t-1}^{t-1}, V_{t-1}^{t-1})$$

most recent data used in prediction

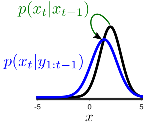
variable being predicted



diffuse via dynamics

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) \mathrm{d}x_{t-1}$$

diffuses toward 0

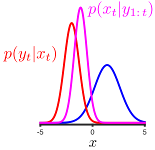


$$p(x_t|y_{1:t-1}) = \mathcal{G}(x_t; \mu_t^{t-1}, V_t^{t-1}) \qquad \mu_t^{t-1} = A\mu_{t-1}^{t-1}$$

$$V_t^{t-1} = AV_{t-1}^{t-1}A^\top + Q$$

variance inflates

combine with likelihood



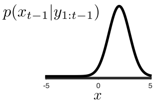
$$p(x_t|y_{1:t}) \propto \underbrace{p(x_t|y_{1:t-1})}_{\text{prior}} \underbrace{p(y_t|x_t)}_{\text{likelihood}}$$

Inference: Kalman Filter

$$p(x_{t-1}|y_{1:t-1}) = \mathcal{G}(x_{t-1}; \mu_{t-1}^{t-1}, V_{t-1}^{t-1})$$

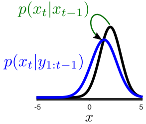
most recent data used in prediction

variable being predicted



diffuse via dynamics

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) \mathrm{d}x_{t-1}$$



diffuses toward 0

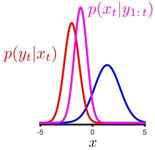
$$p(x_t|y_{1:t-1}) = \mathcal{G}(x_t; \mu_t^{t-1}, V_t^{t-1})$$

$$\mu_t^{t-1} = A\mu_{t-1}^{t-1}$$

$$V_t^{t-1} = AV_{t-1}^{t-1}A^\top + Q$$

variance inflates

combine with likelihood



$$p(x_t|y_{1:t}) \propto \underbrace{p(x_t|y_{1:t-1})}_{\text{prior}} \underbrace{p(y_t|x_t)}_{\text{likelihood}}$$

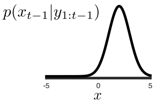
$$p(x_t|y_{1:t}) = \mathcal{G}(x_t; \mu_t^t, V_t^t)$$

Inference: Kalman Filter

$$p(x_{t-1}|y_{1:t-1}) = \mathcal{G}(x_{t-1}; \mu_{t-1}^{t-1}, V_{t-1}^{t-1})$$

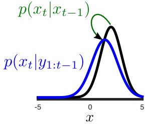
most recent data used in prediction

variable being predicted



diffuse via dynamics

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) \mathrm{d}x_{t-1}$$



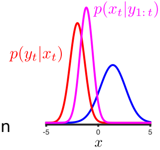
diffuses toward 0

$$p(x_t|y_{1:t-1}) = \mathcal{G}(x_t; \mu_t^{t-1}, V_t^{t-1}) \qquad \mu_t^{t-1} = A\mu_{t-1}^{t-1}$$

$$V_t^{t-1} = AV_{t-1}^{t-1}A^\top + Q$$

variance inflates

combine with likelihood



$$p(x_t|y_{1:t}) \propto \underbrace{p(x_t|y_{1:t-1})}_{\text{prior}} \underbrace{p(y_t|x_t)}_{\text{likelihood}}$$

prediction

correction

$$p(x_t|y_{1:t}) = \mathcal{G}(x_t; \mu_t^t, V_t^t) \qquad \mu_t^t = \mu_t^{t-1} + K_t(y_t - C\mu_t^{t-1})$$

$$V_t^t = V_t^{t-1} - K_t C V_t^{t-1}$$

Kalman gain

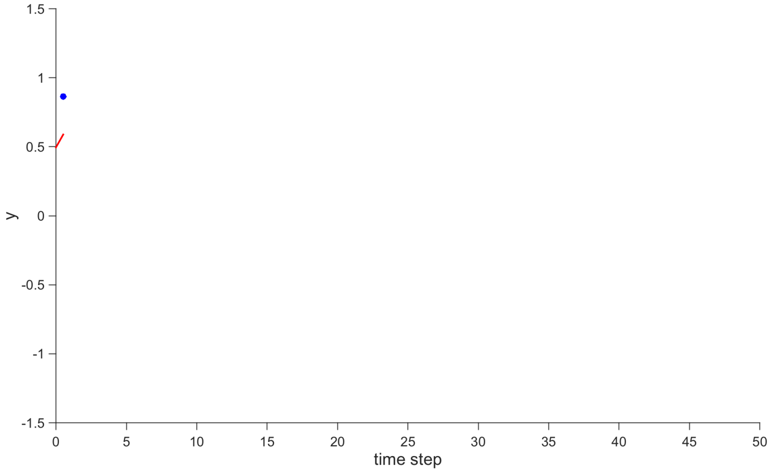
$$K_t = V_t^{t-1}C^\top(CV_t^{t-1}C^\top + R)^{-1}$$

Kalman Filter Demo

- data: $y_t = \sin(\omega t) + \sigma_y \epsilon_t$ where $\sigma_y^2 = 0.1$
- model: $x_t = \lambda x_{t-1} + \sigma \eta$ and $y_t = x_t + \sigma_y \eta'_t$
  where $\lambda = 0.99$ and $\sigma^2 = 1 - \lambda^2$
- demo shows how the Kalman filter processes the data to form estimates of the hidden state at each time point $p(x_t|y_{1:t})$

Kalman Filter Demo

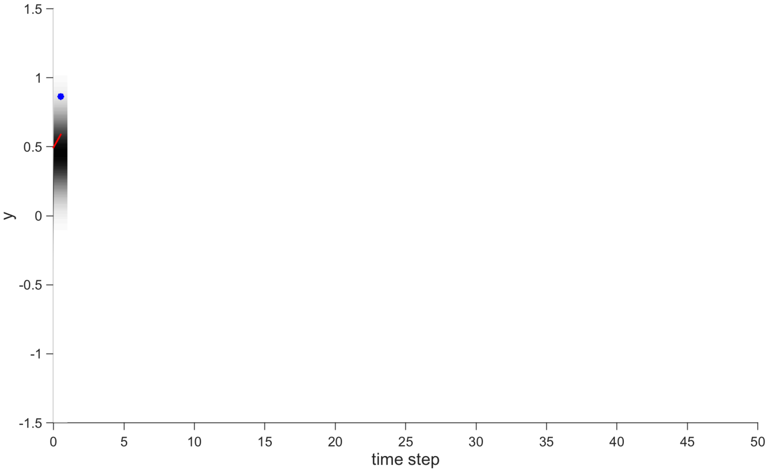
observed noisy data y_t, ground truth sinusoid



observe first data point y_1

Kalman Filter Demo

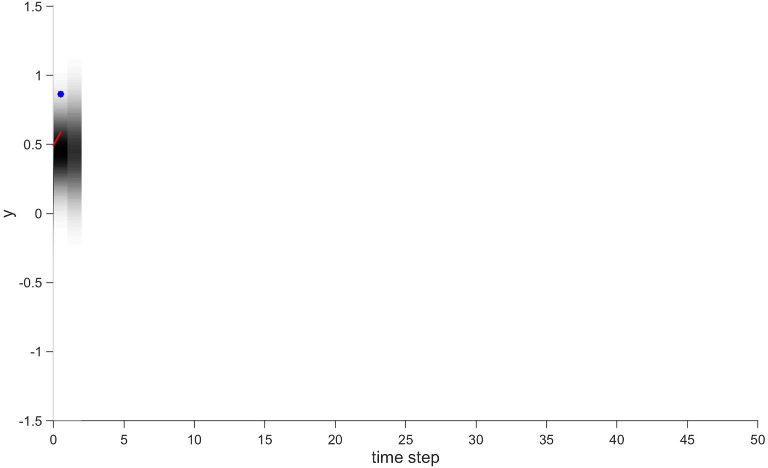
observed noisy data y_t, ground truth sinusoid



posterior over first latent variable $p(x_1|y_1)$

Kalman Filter Demo

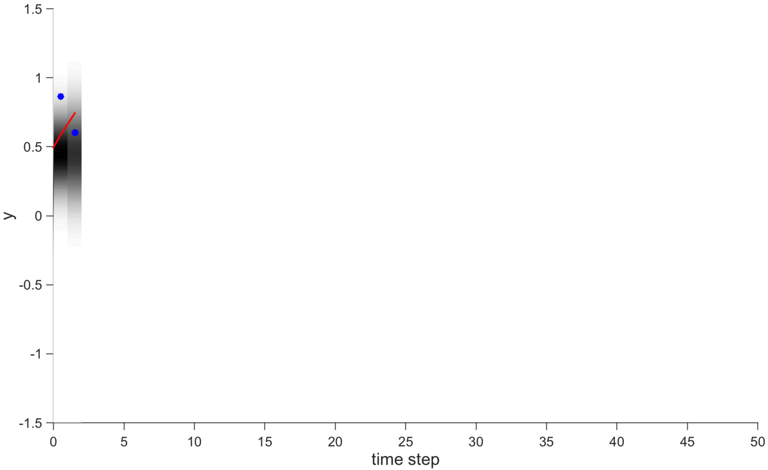
observed noisy data y_t, ground truth sinusoid



prediction for second latent variable $p(x_2|y_1)$

Kalman Filter Demo

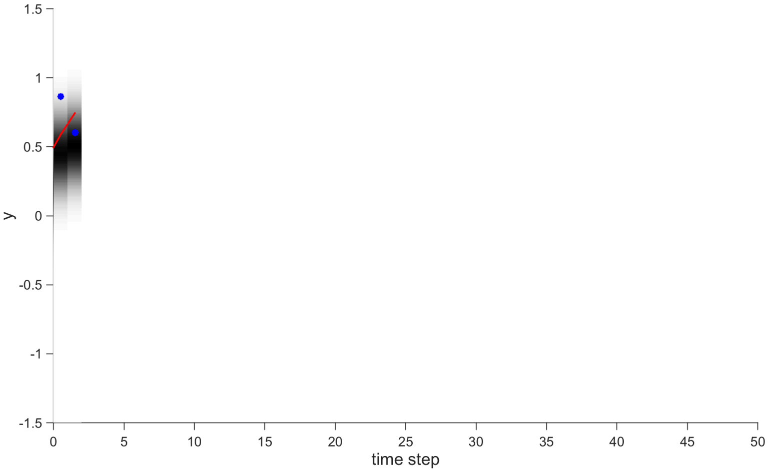
observed noisy data y_t, ground truth sinusoid



observe next data point y_2

Kalman Filter Demo

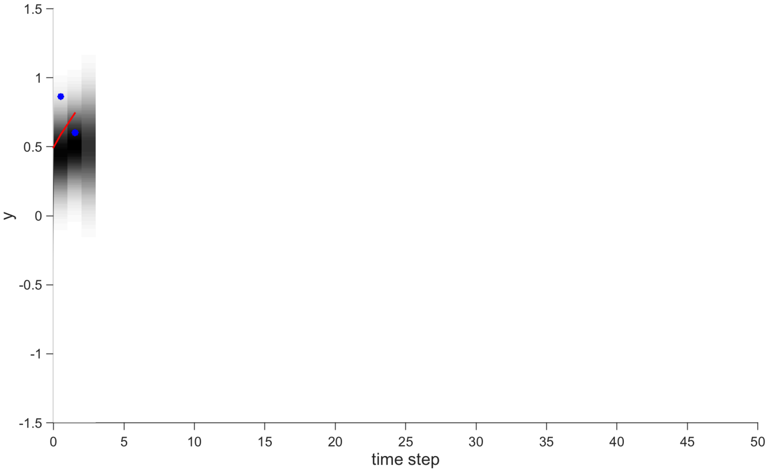
observed noisy data y_t, ground truth sinusoid



form posterior over second latent variable $p(x_2|y_1, y_2)$

Kalman Filter Demo

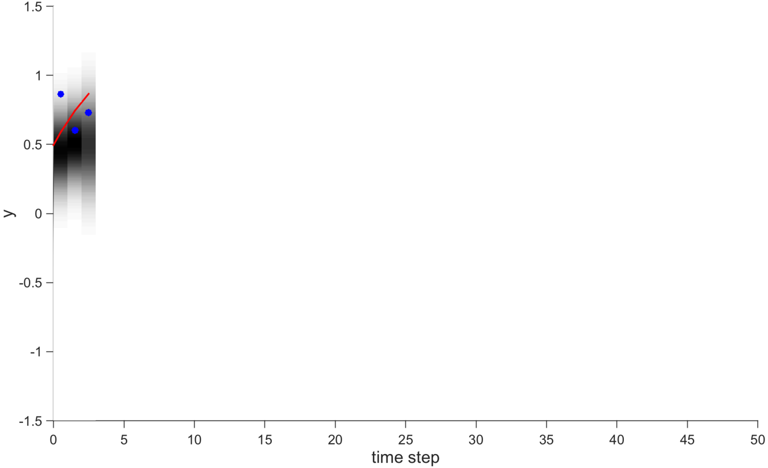
observed noisy data y_t, ground truth sinusoid



prediction for third latent variable $p(x_3|y_1, y_2)$

Kalman Filter Demo

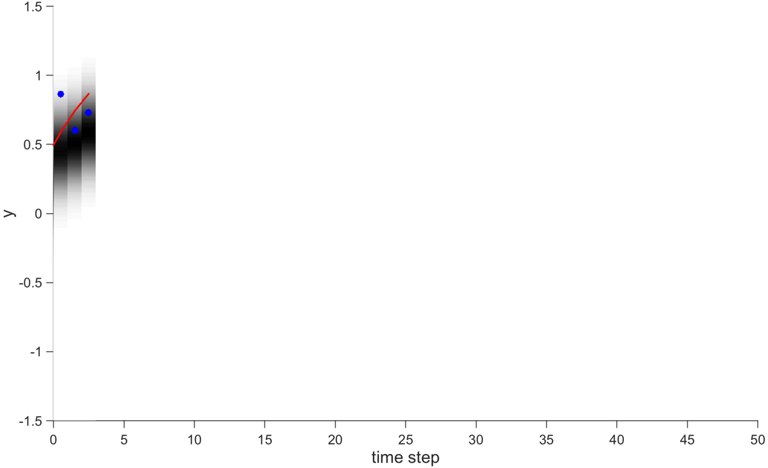
observed noisy data y_t, ground truth sinusoid



observe next data point y_3

Kalman Filter Demo

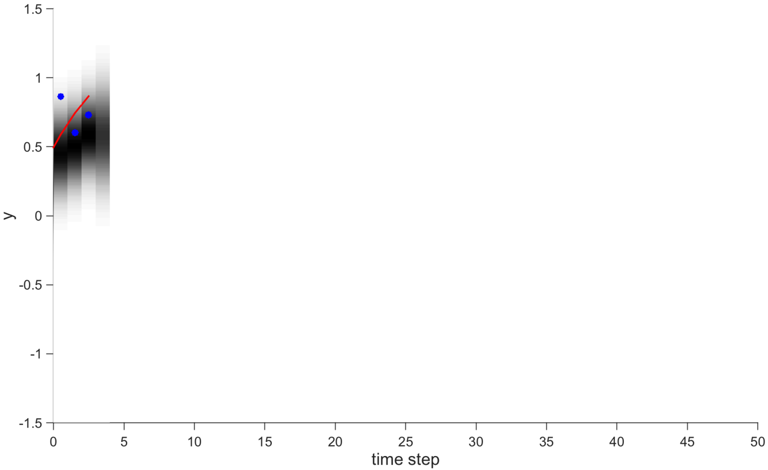
observed noisy data y_t, ground truth sinusoid



form posterior over third latent variable $p(x_3|y_1, y_2, y_3)$

Kalman Filter Demo

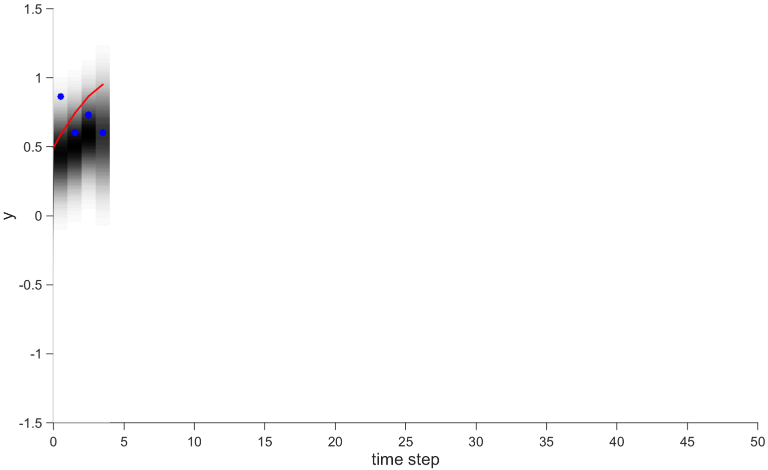
observed noisy data y_t, ground truth sinusoid



prediction for fourth latent variable $p(x_4|y_{1:3})$

Kalman Filter Demo

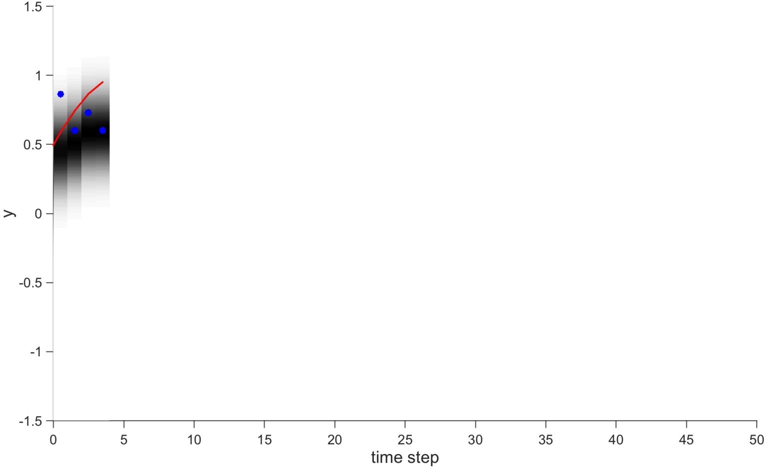
observed noisy data y_t, ground truth sinusoid



observe next data point y_4

Kalman Filter Demo

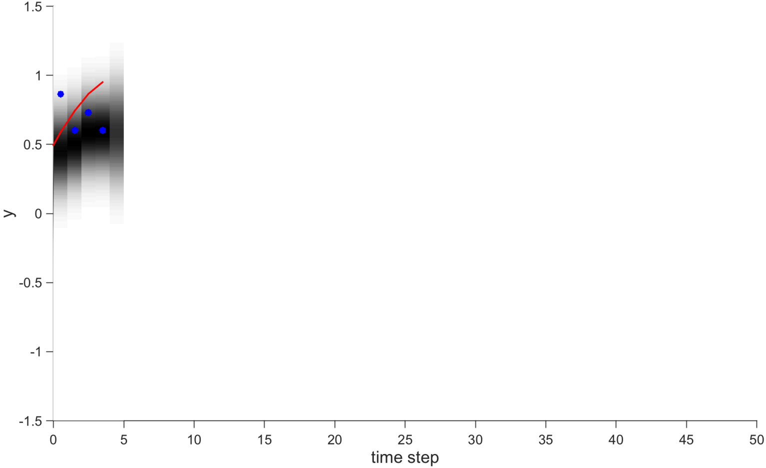
observed noisy data y_t, ground truth sinusoid



form posterior over fourth latent variable $p(x_4|y_{1:4})$

Kalman Filter Demo

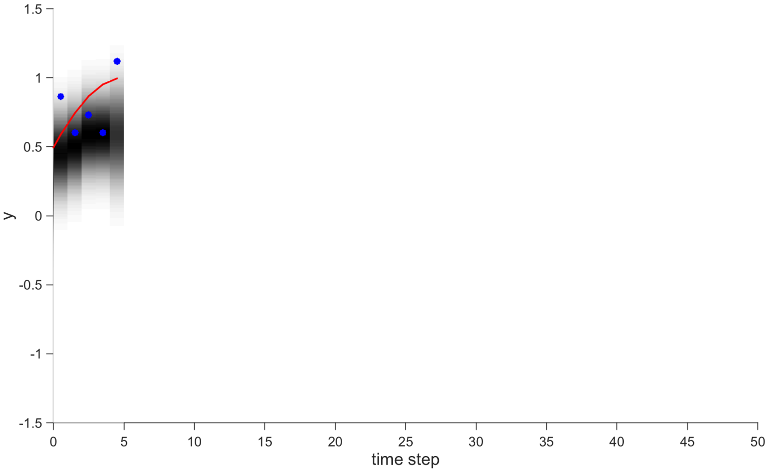
observed noisy data y_t, ground truth sinusoid



prediction for fifth latent variable $p(x_5|y_{1:4})$

Kalman Filter Demo

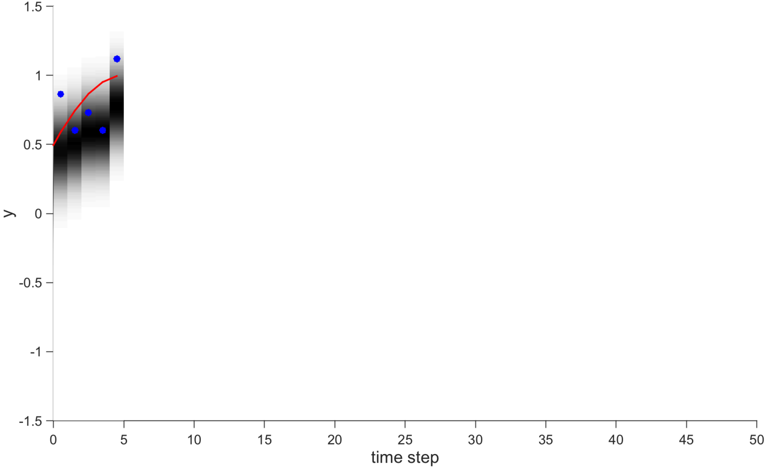
observed noisy data y_t, ground truth sinusoid



observe next data point y_5

Kalman Filter Demo

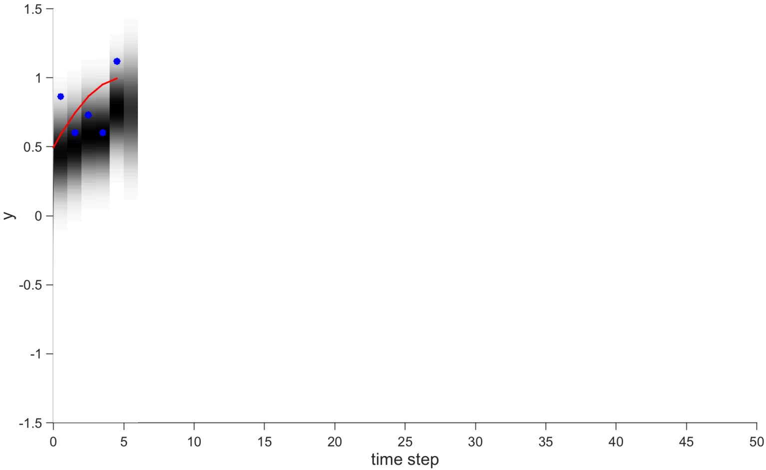
observed noisy data y_t, ground truth sinusoid



form posterior over fifth latent variable $p(x_5|y_{1:5})$

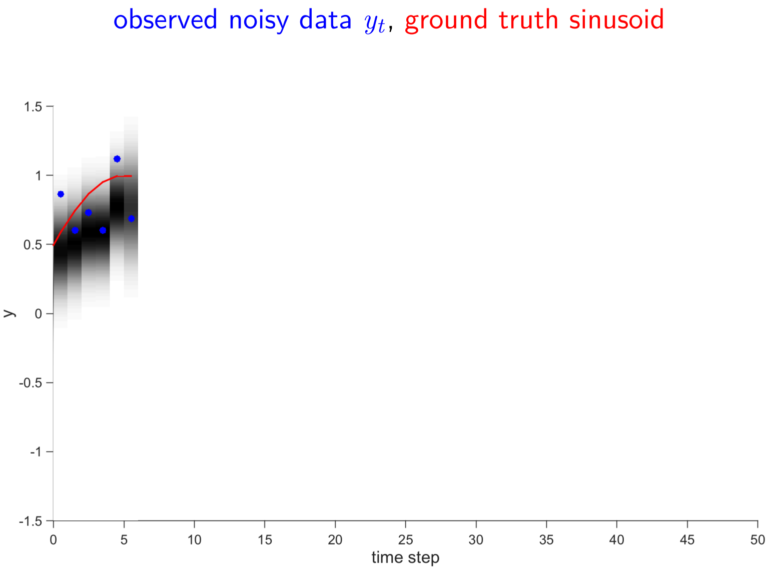
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



prediction for sixth latent variable $p(x_6|y_{1:5})$

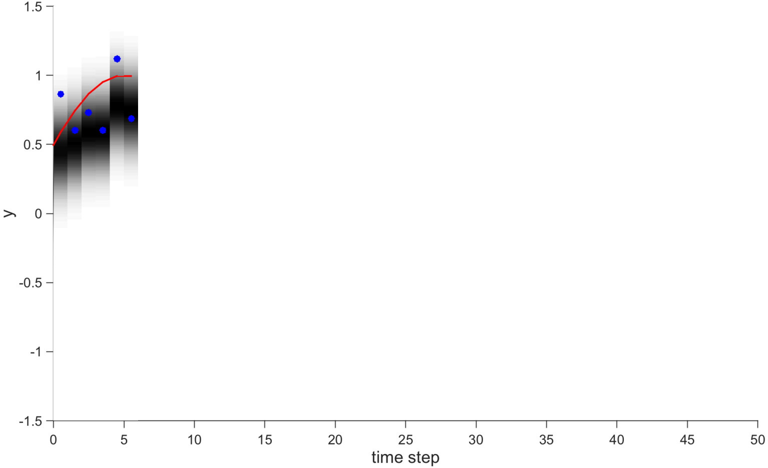
Kalman Filter Demo



observe next data point y_6

Kalman Filter Demo

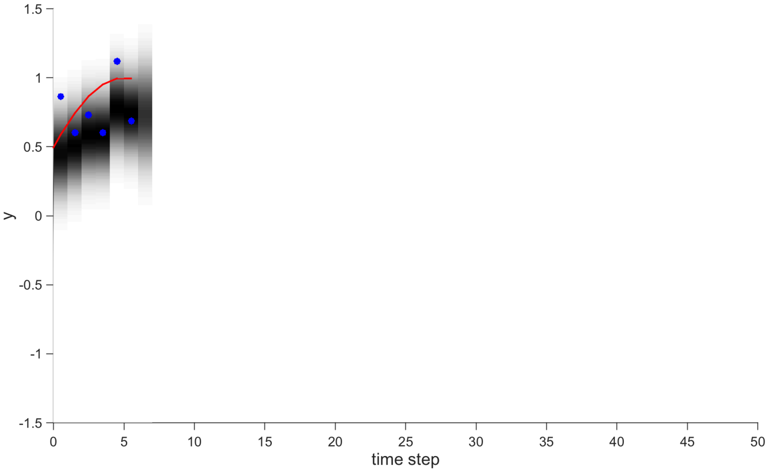
observed noisy data y_t, ground truth sinusoid



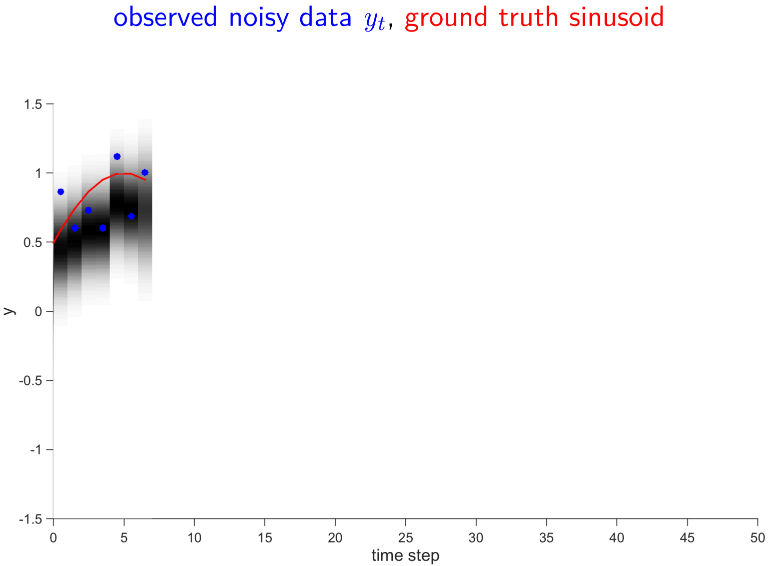
form posterior over sixth latent variable $p(x_6|y_{1:6})$

Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

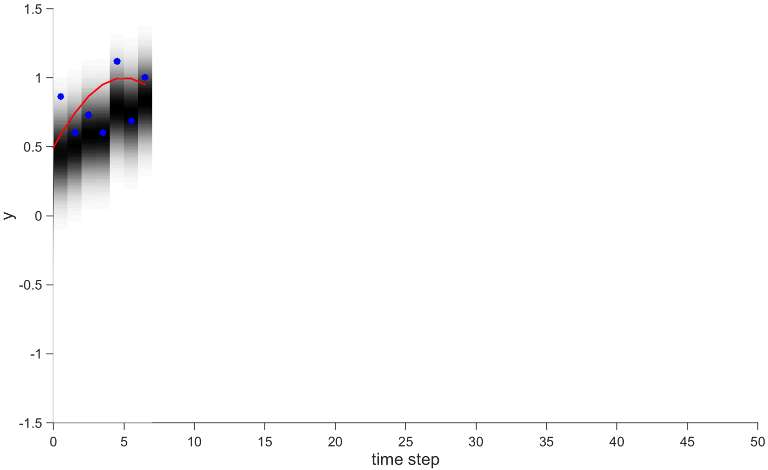


Kalman Filter Demo



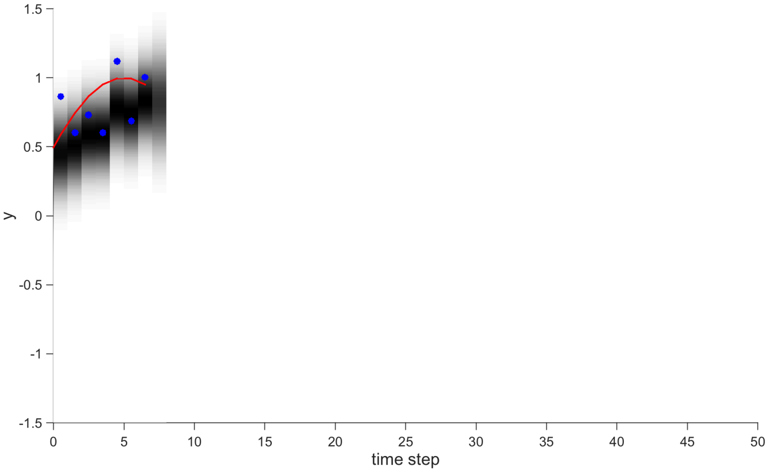
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



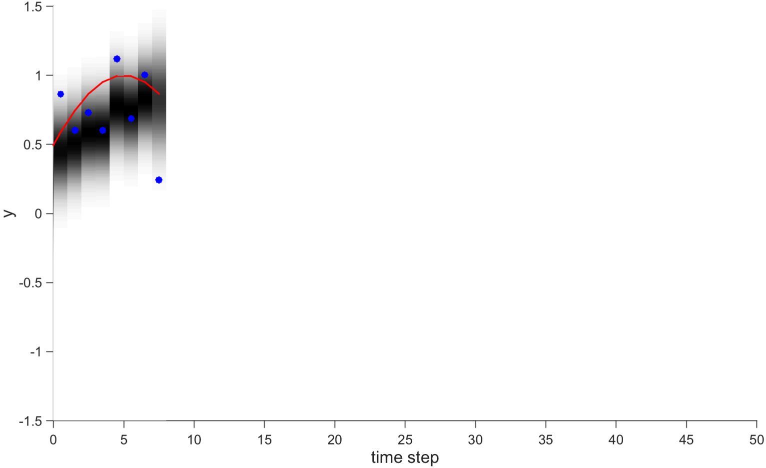
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



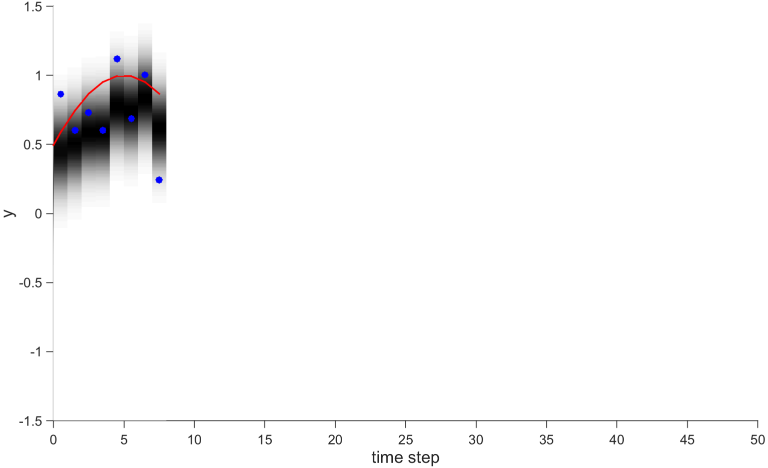
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

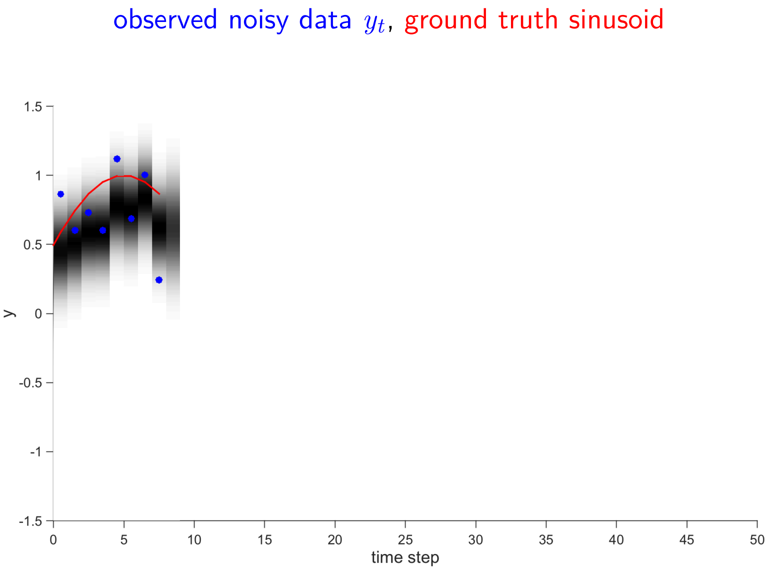


Kalman Filter Demo

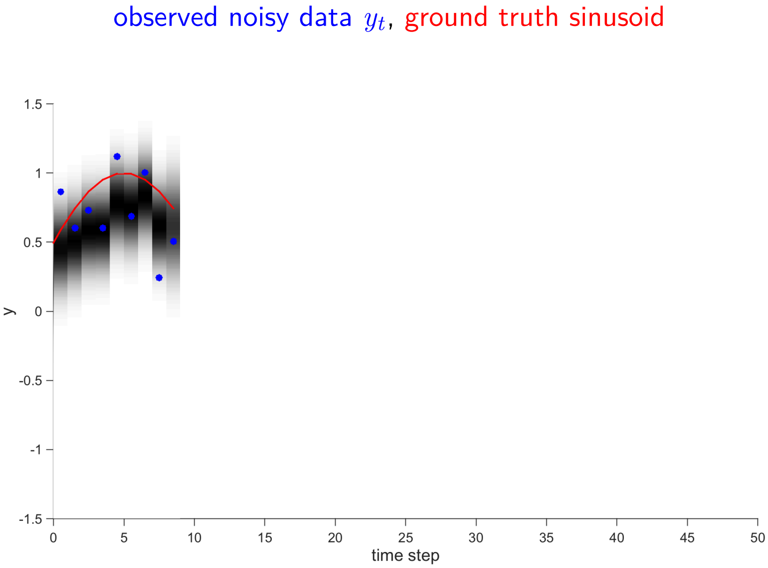
observed noisy data y_t, ground truth sinusoid



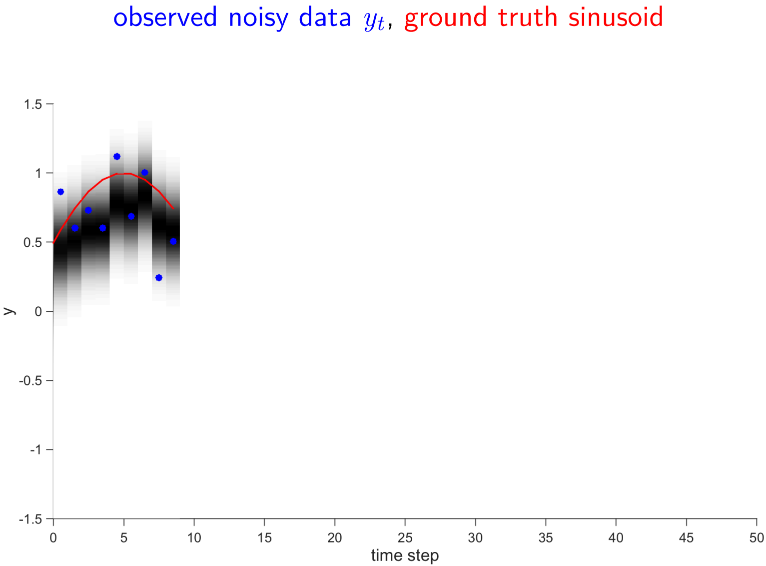
Kalman Filter Demo



Kalman Filter Demo

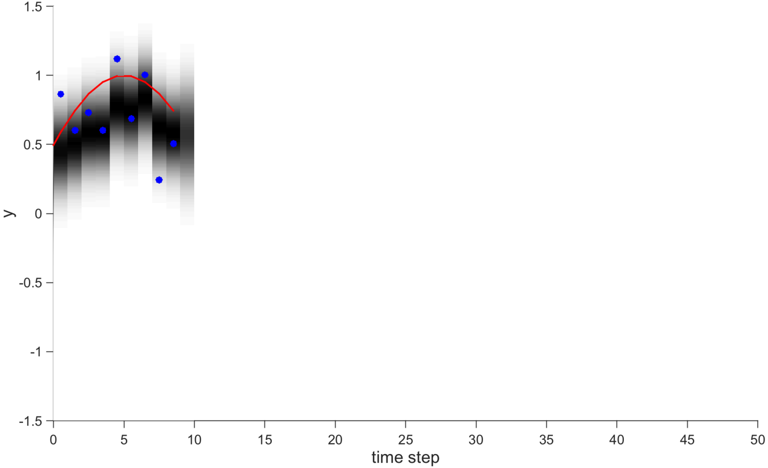


Kalman Filter Demo



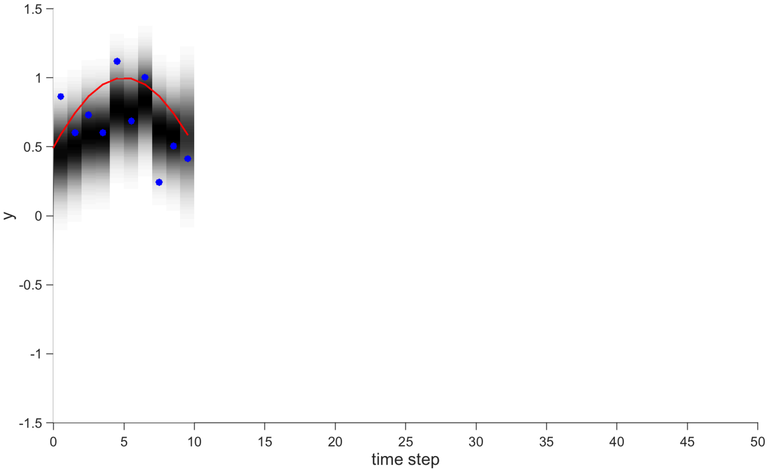
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

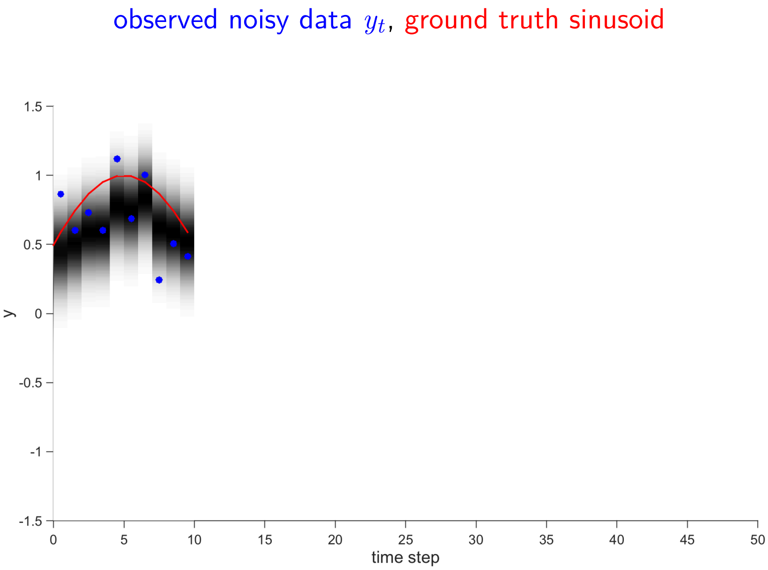


Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

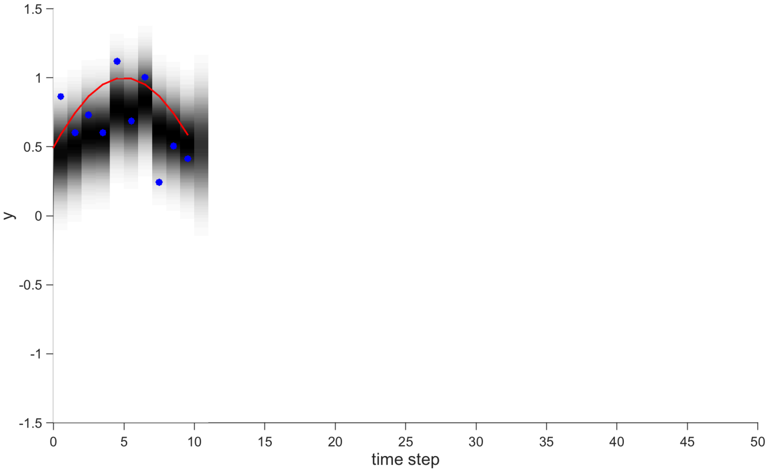


Kalman Filter Demo



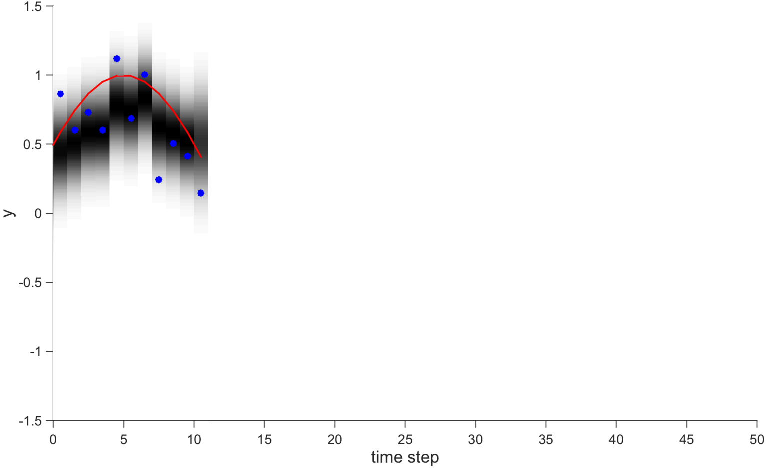
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



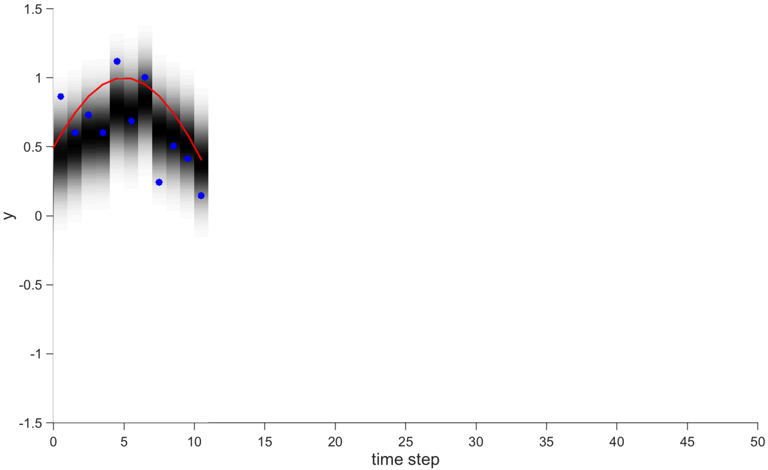
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



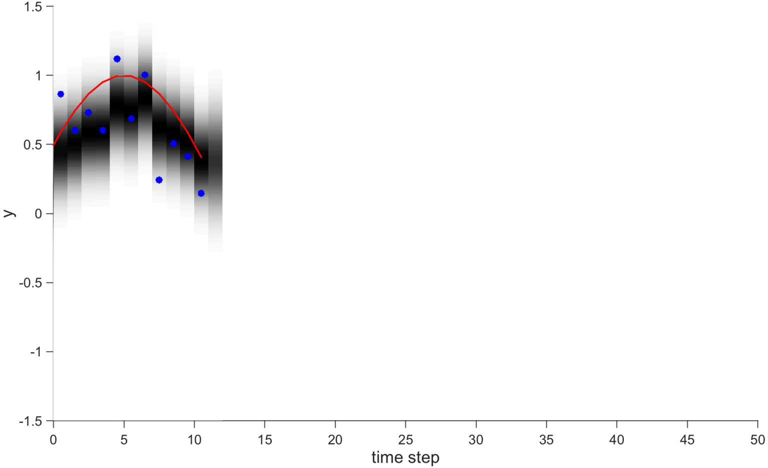
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



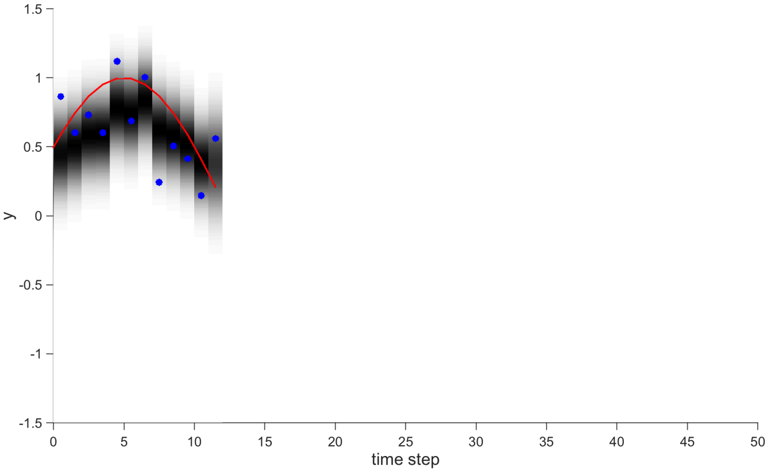
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



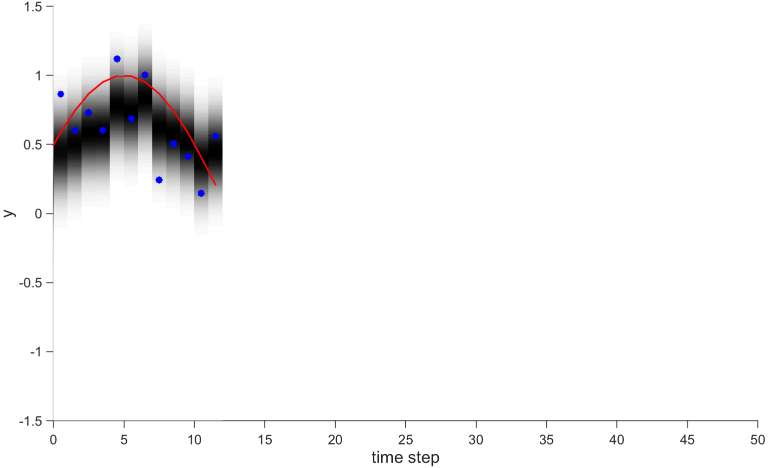
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



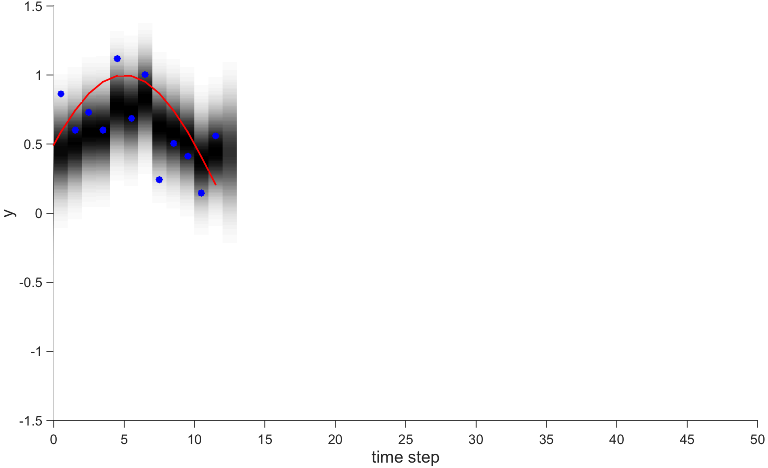
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



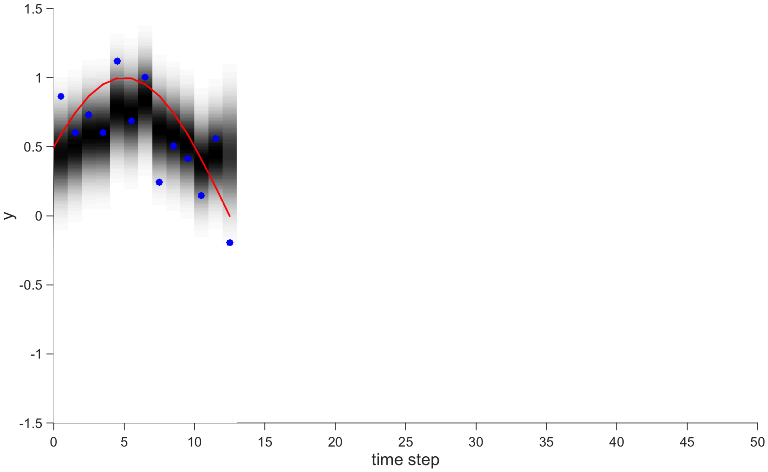
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



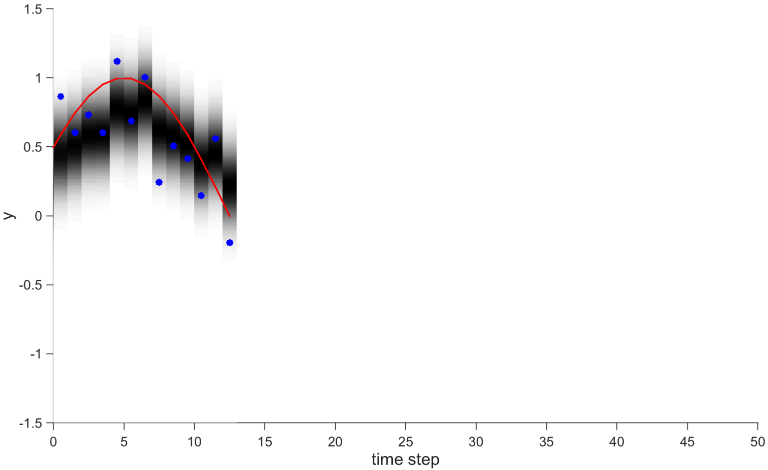
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



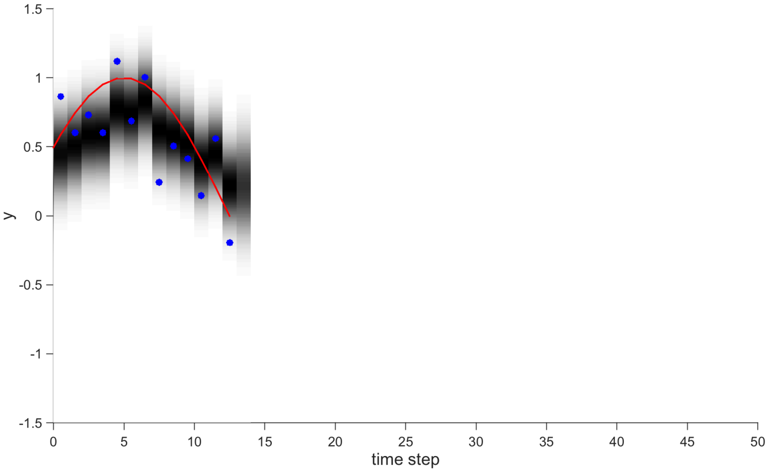
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



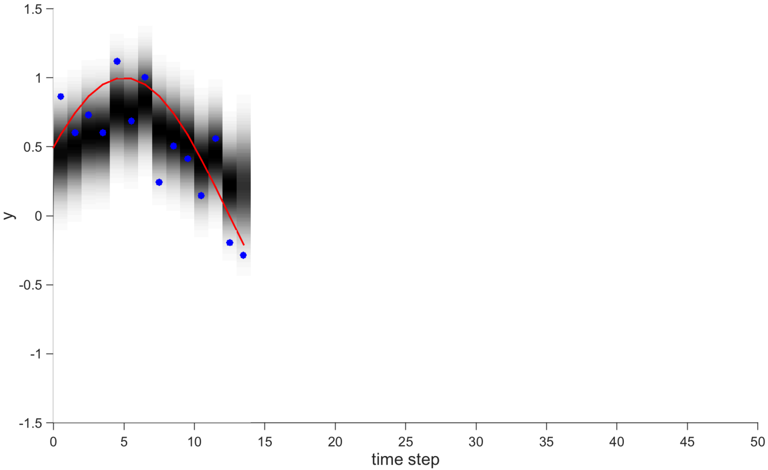
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



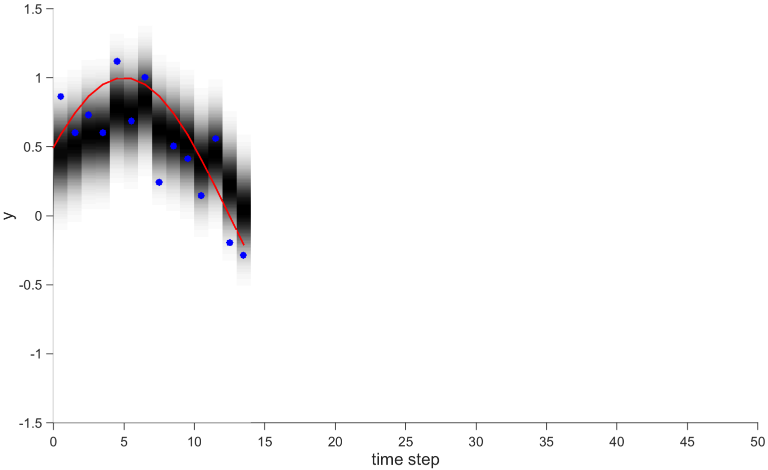
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



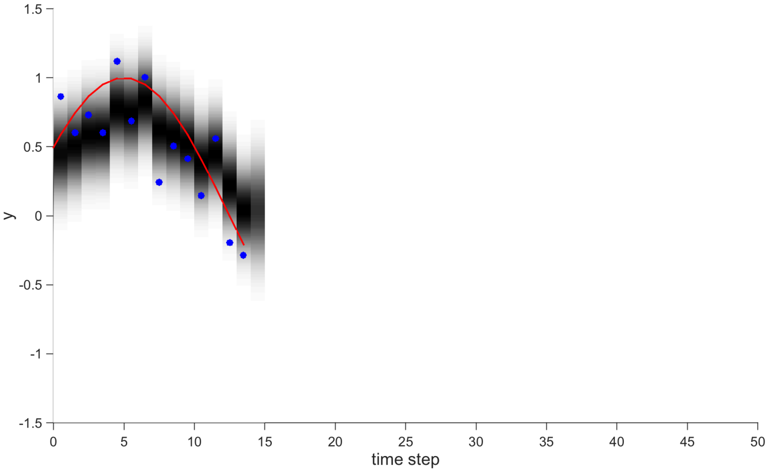
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



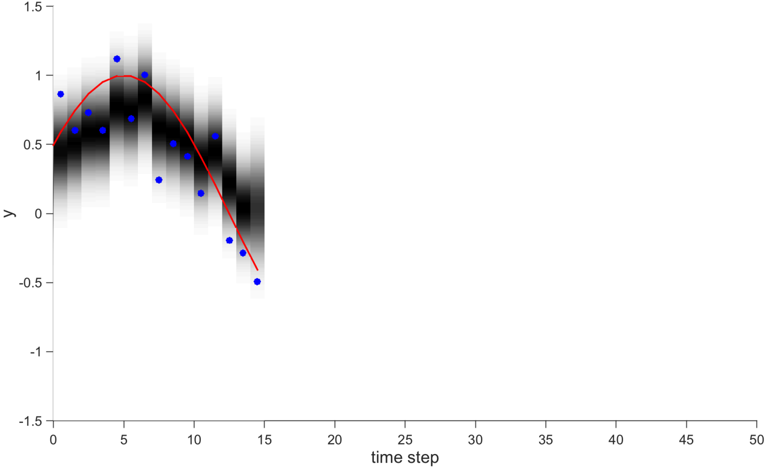
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



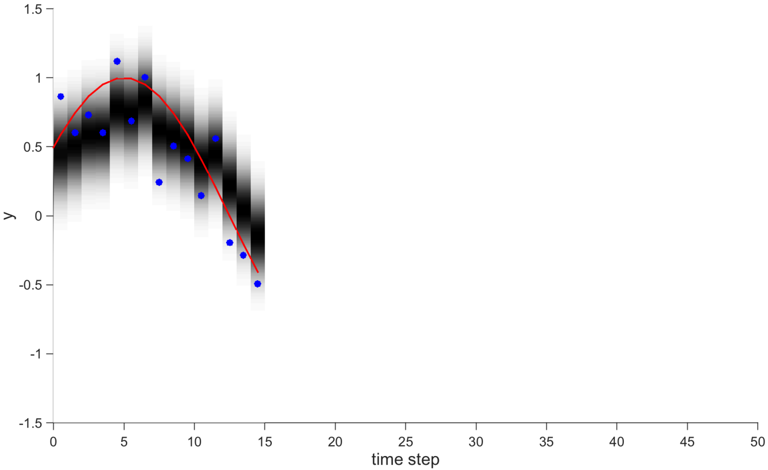
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



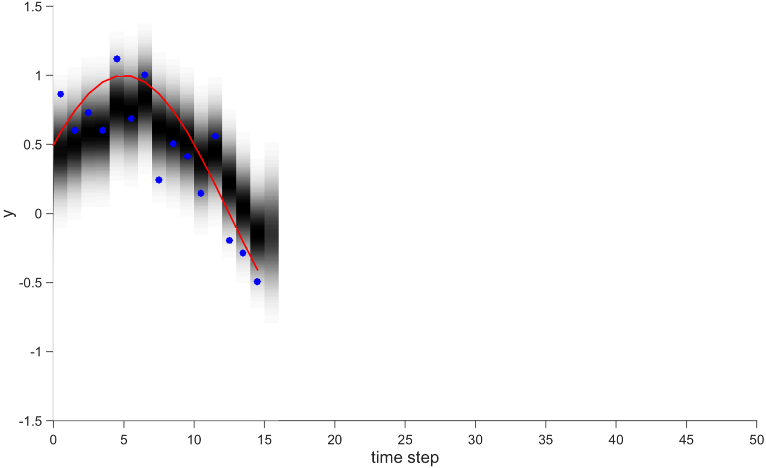
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



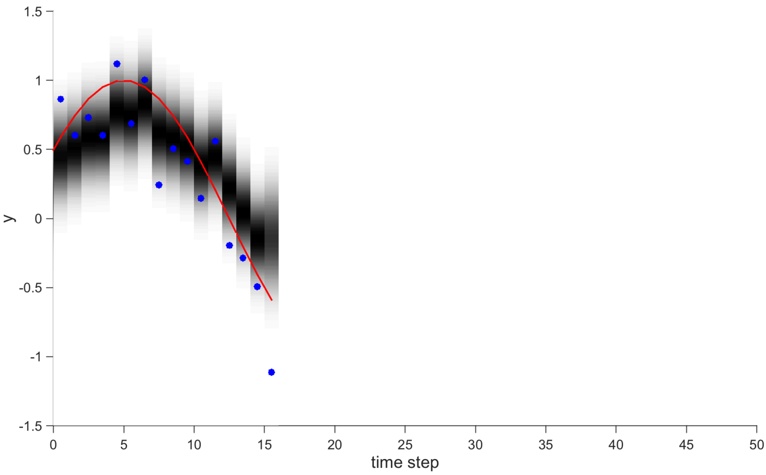
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



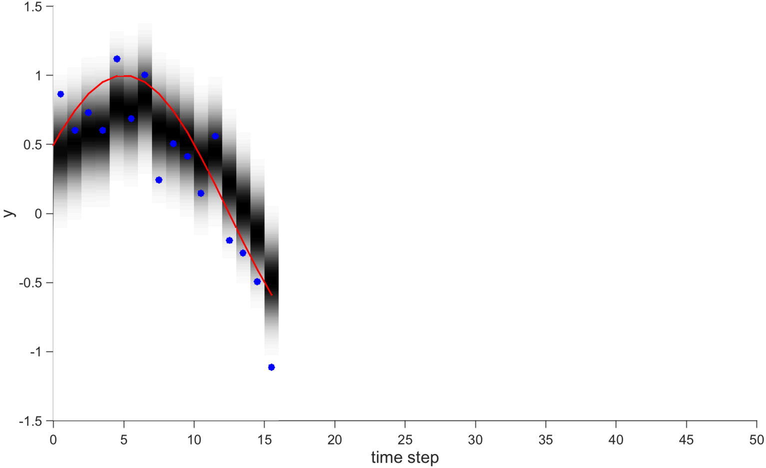
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



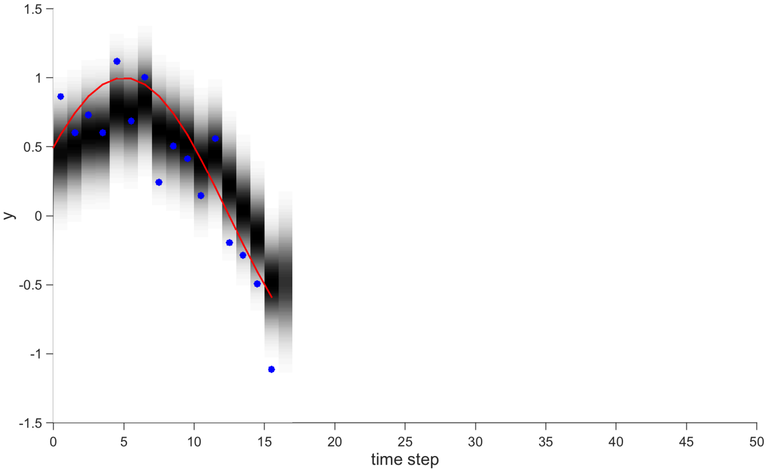
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



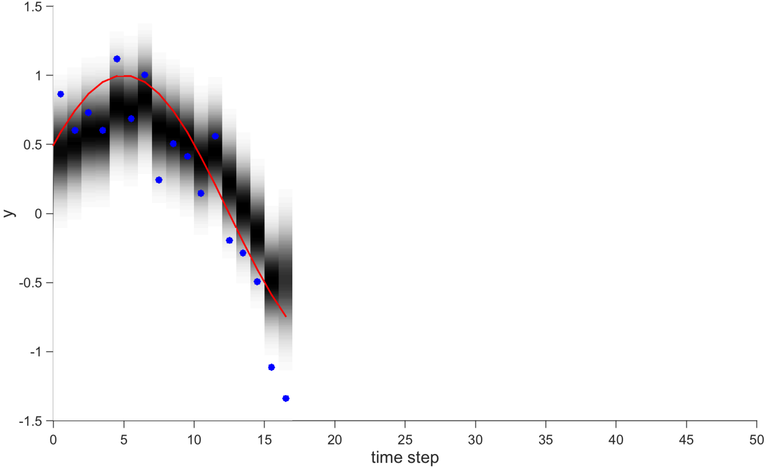
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



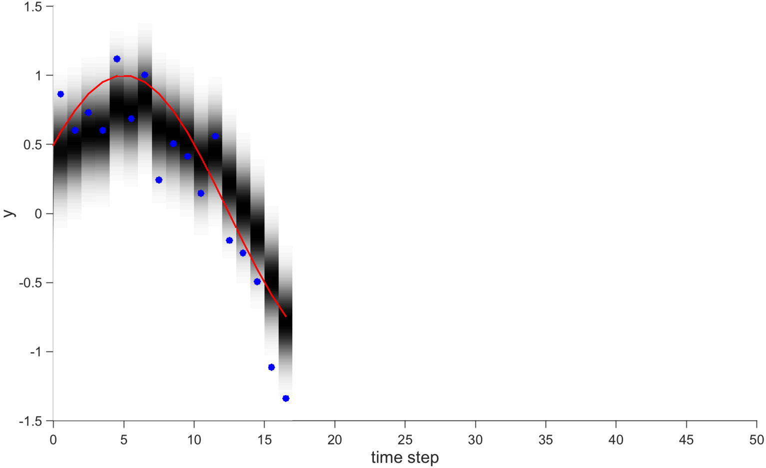
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



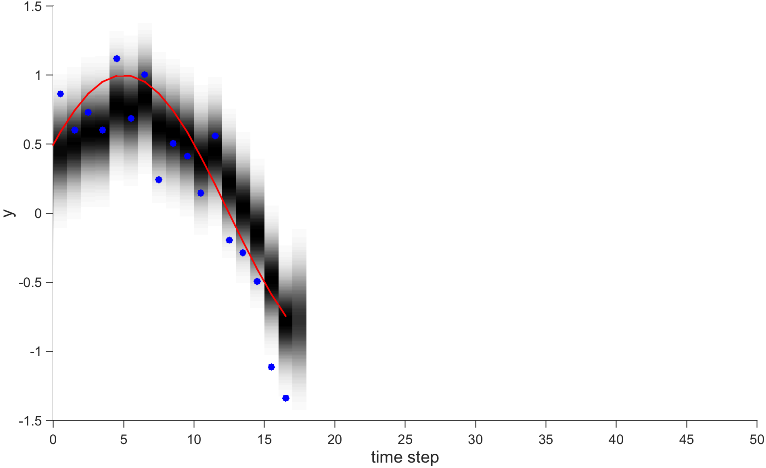
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



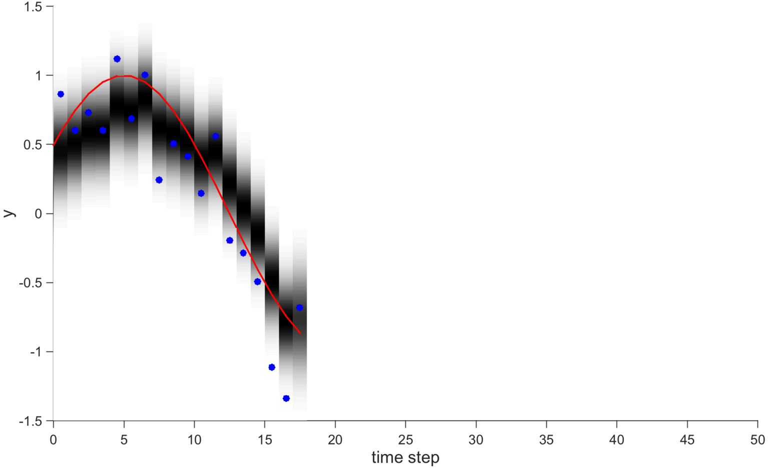
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



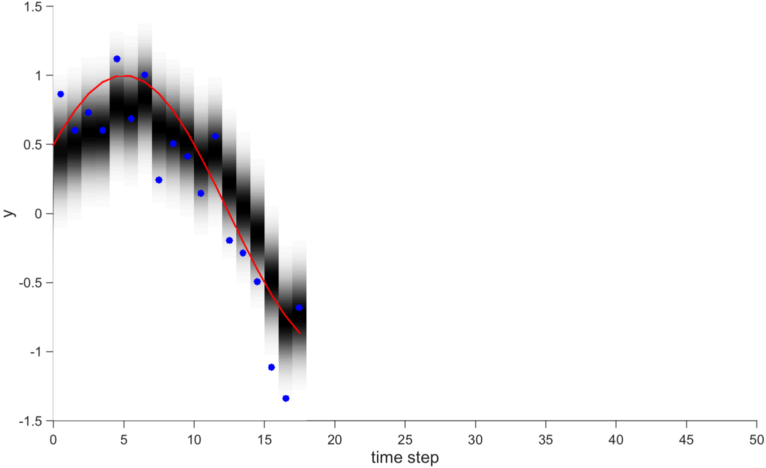
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



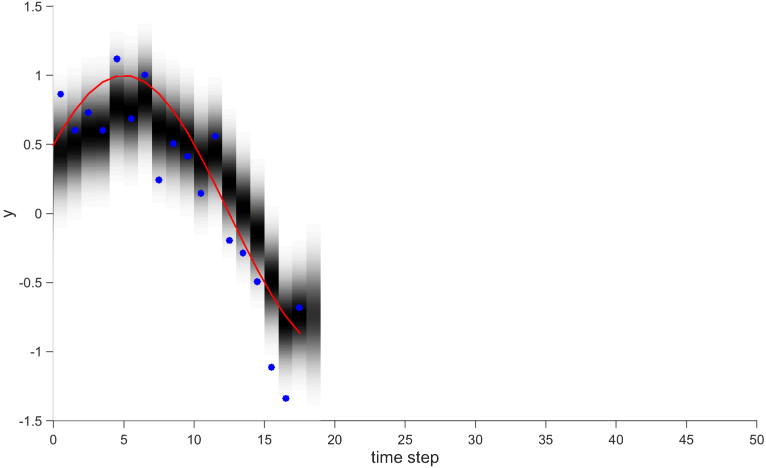
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



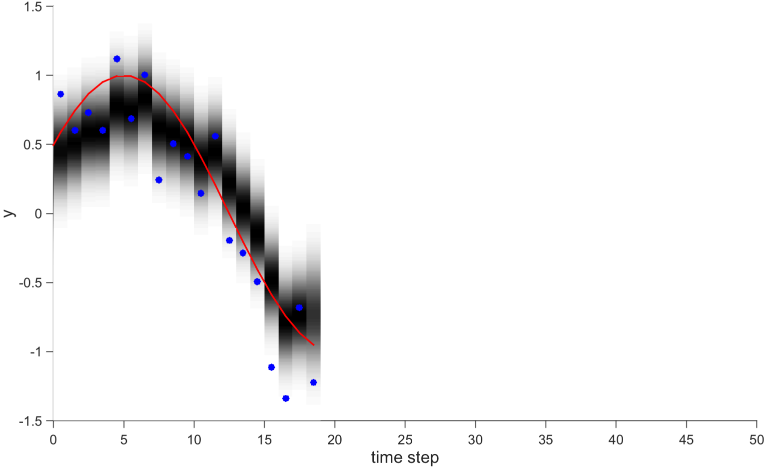
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



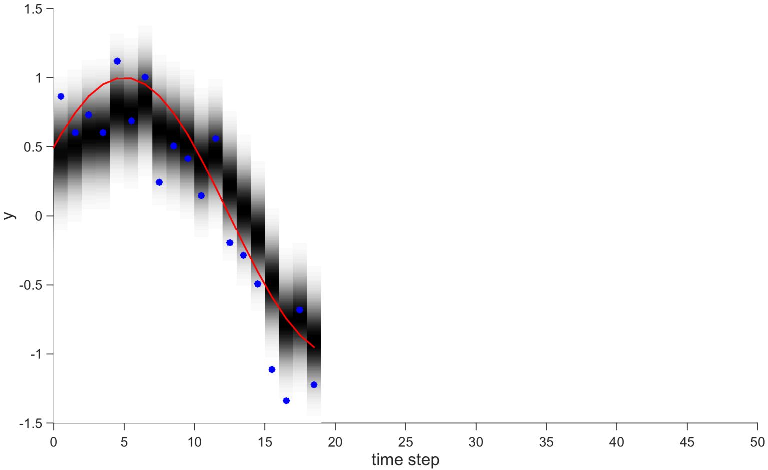
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



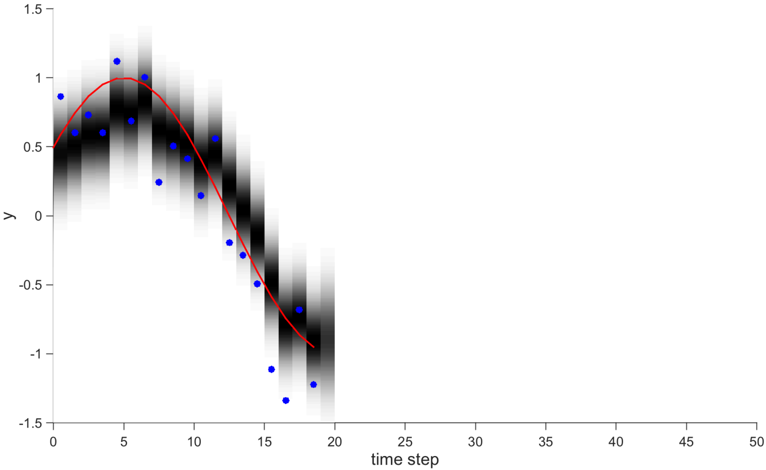
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



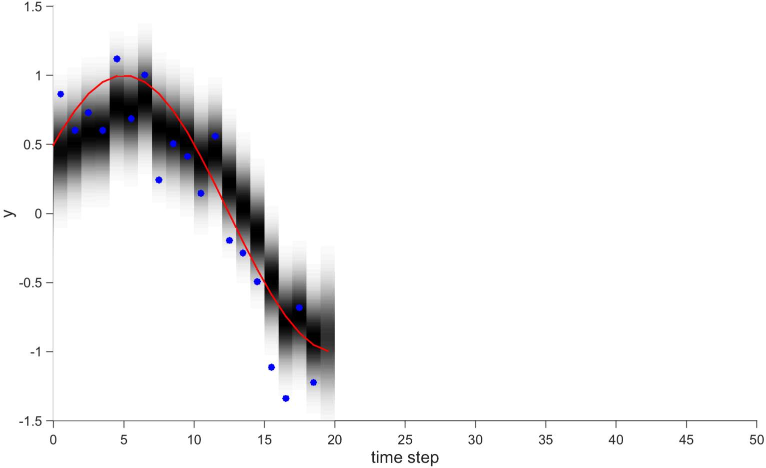
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

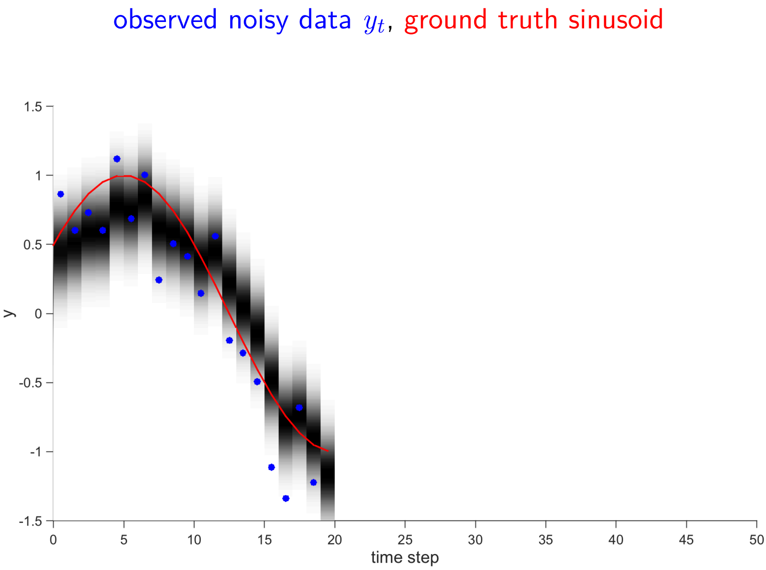


Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

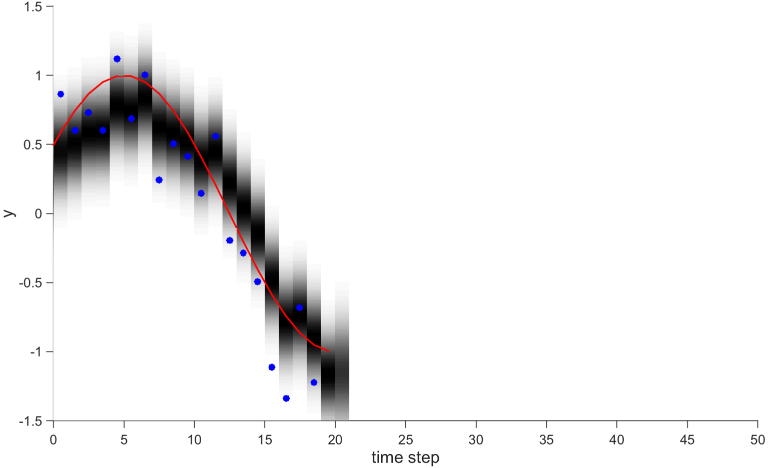


Kalman Filter Demo



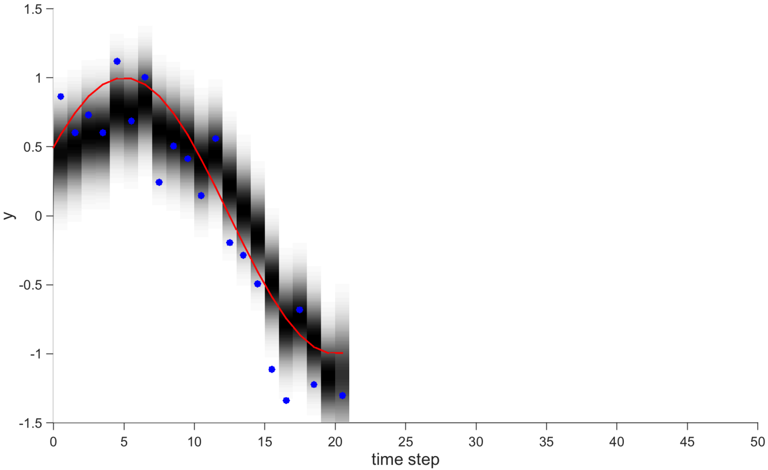
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



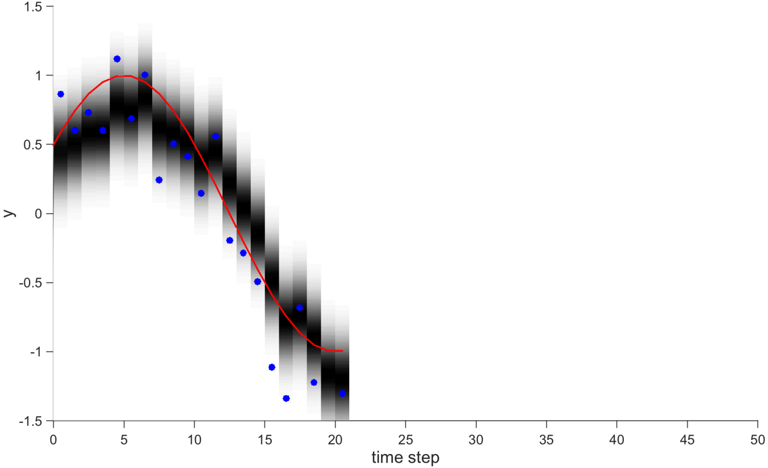
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



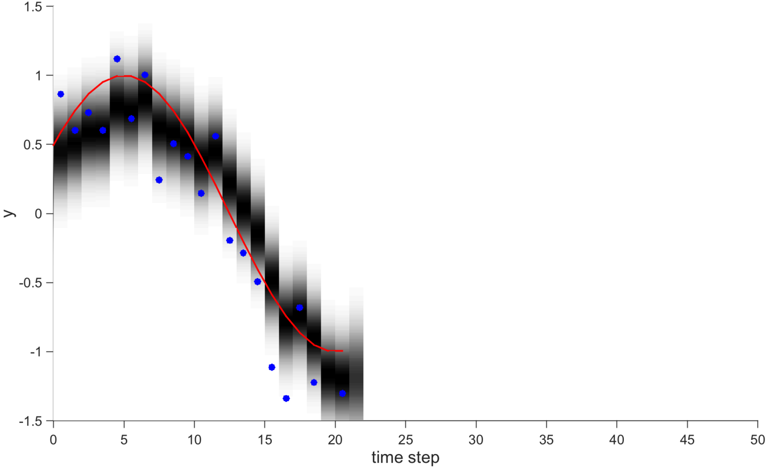
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



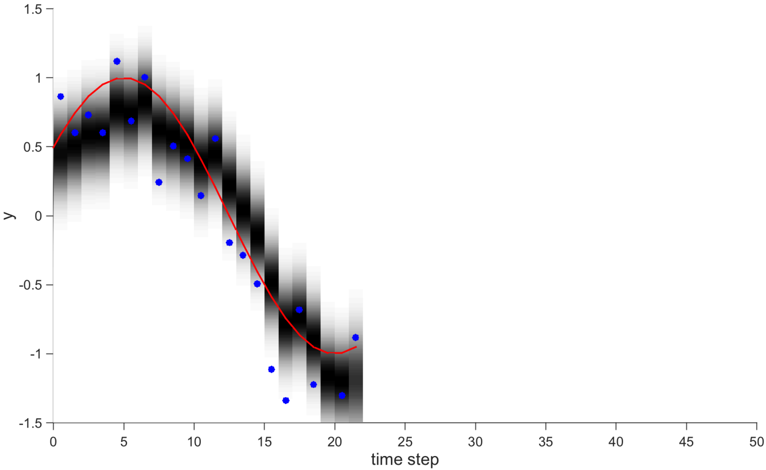
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



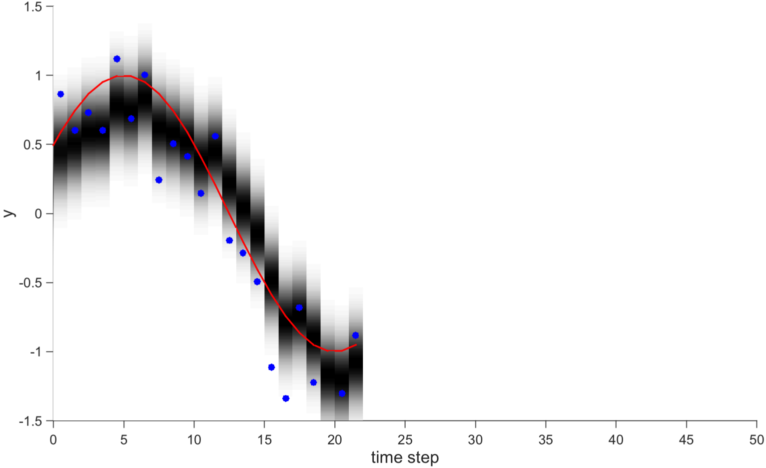
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



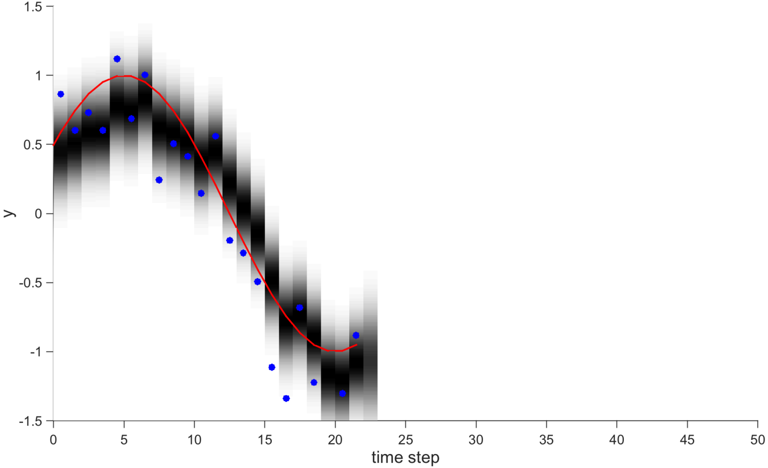
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



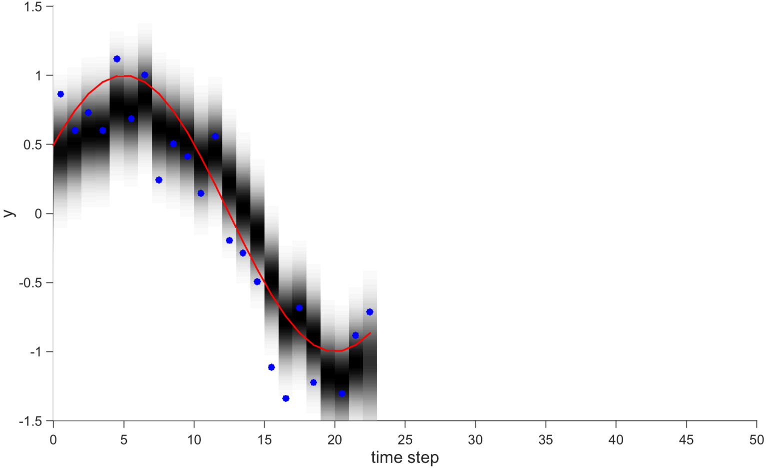
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



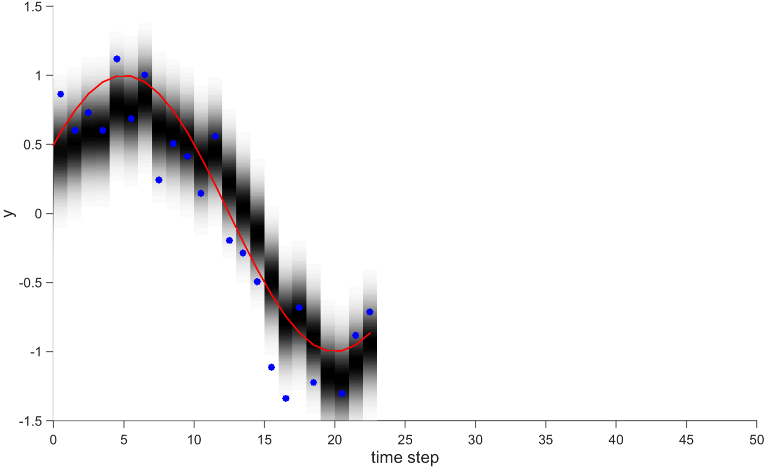
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



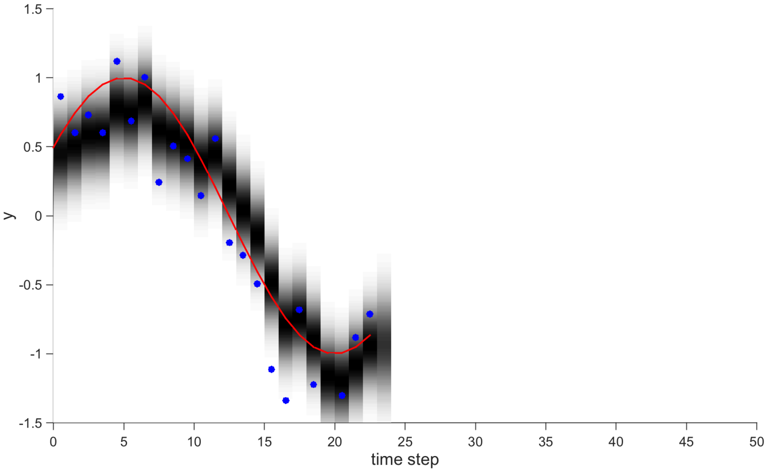
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



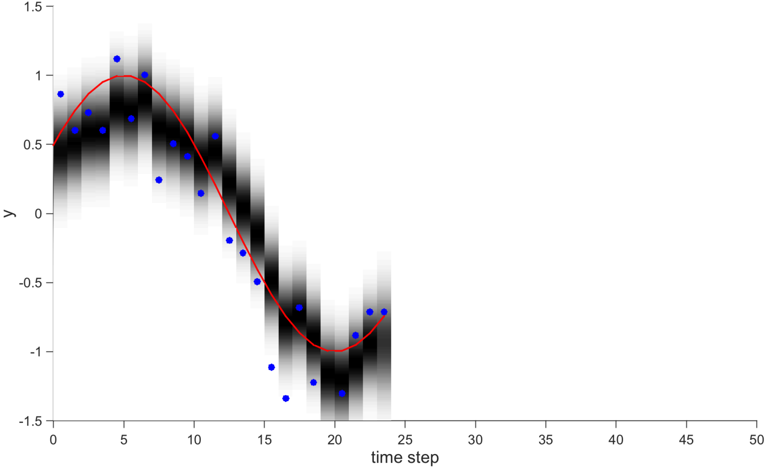
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



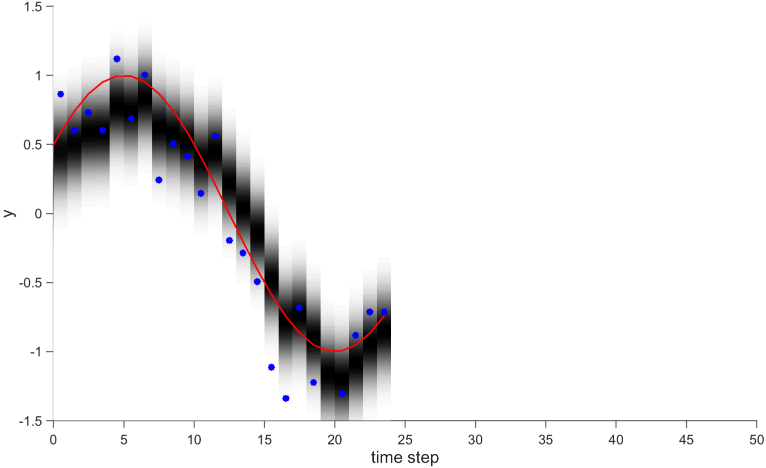
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



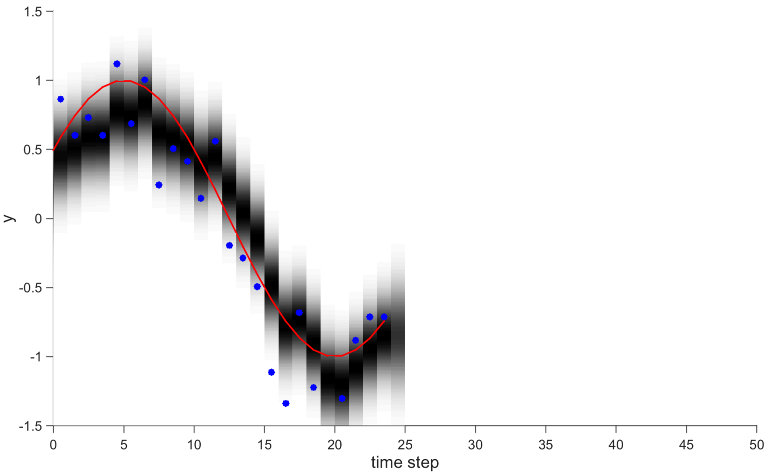
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



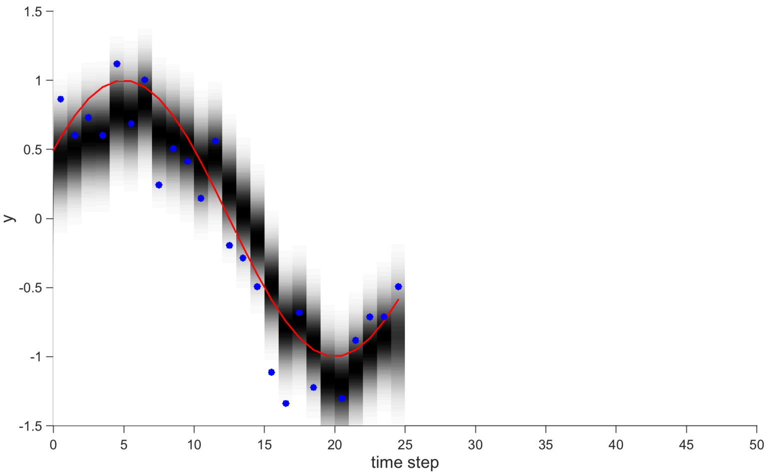
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



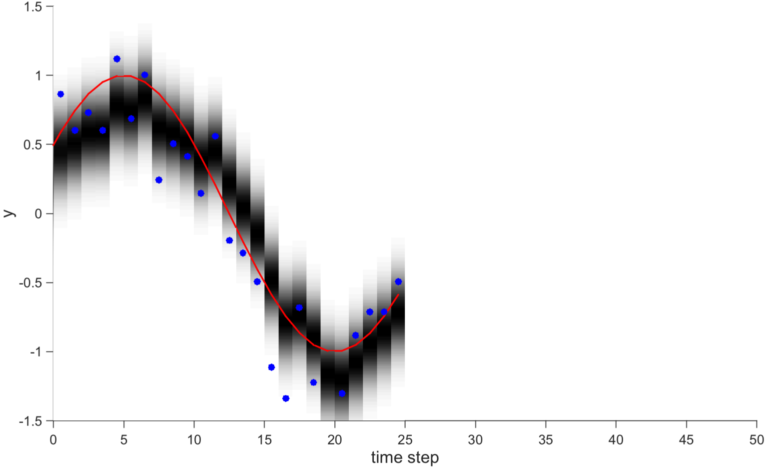
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

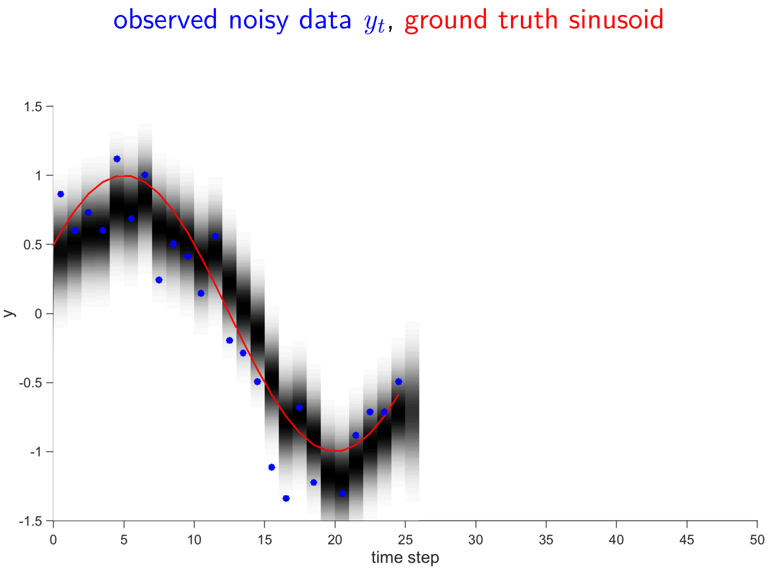


Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

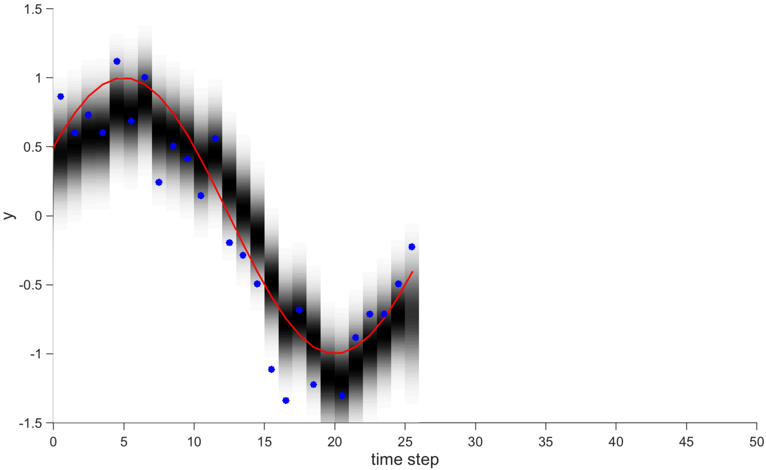


Kalman Filter Demo



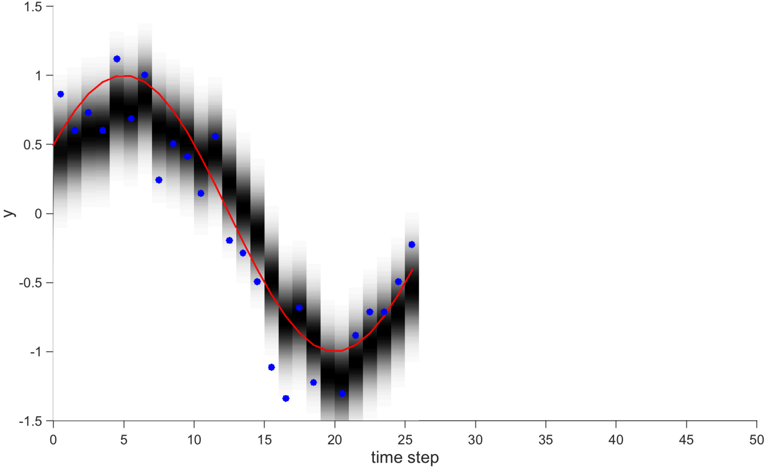
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



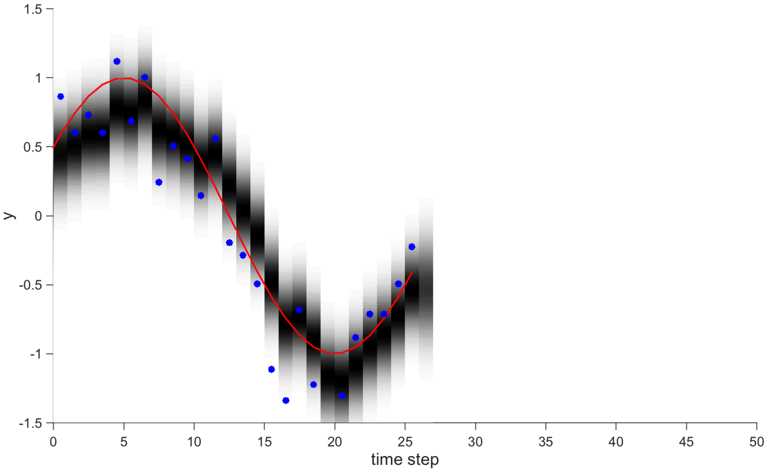
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



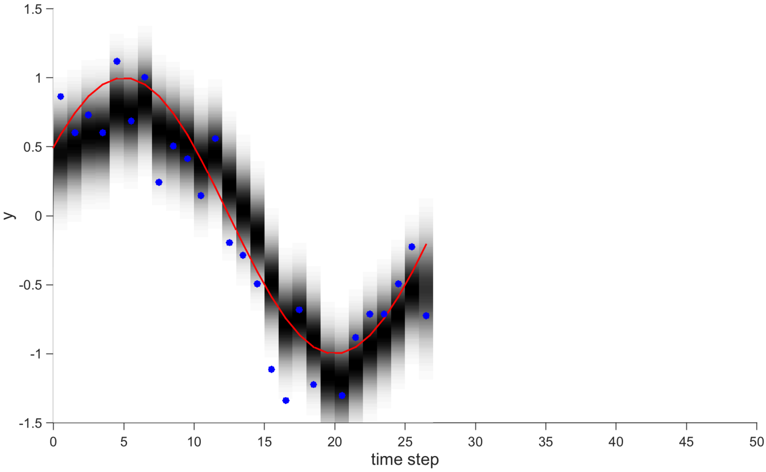
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



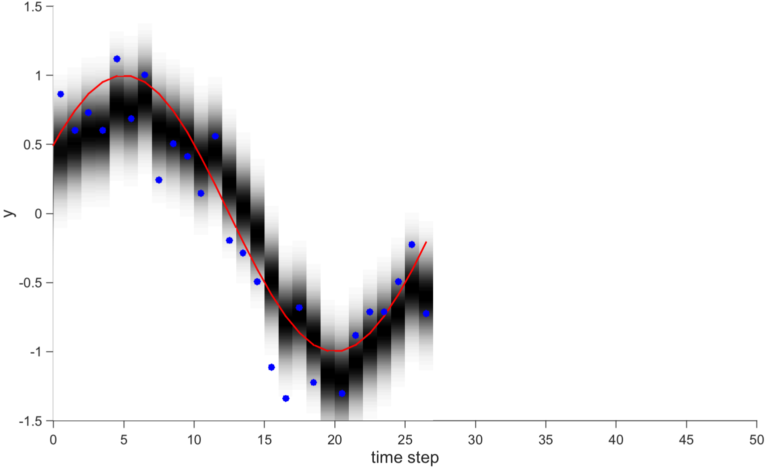
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



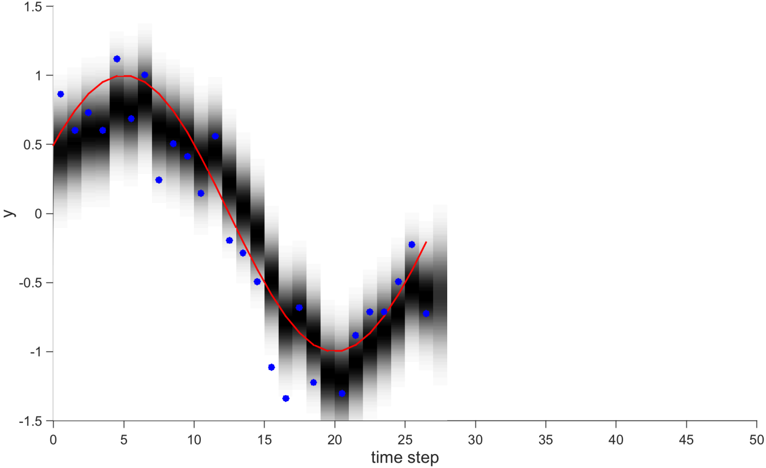
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



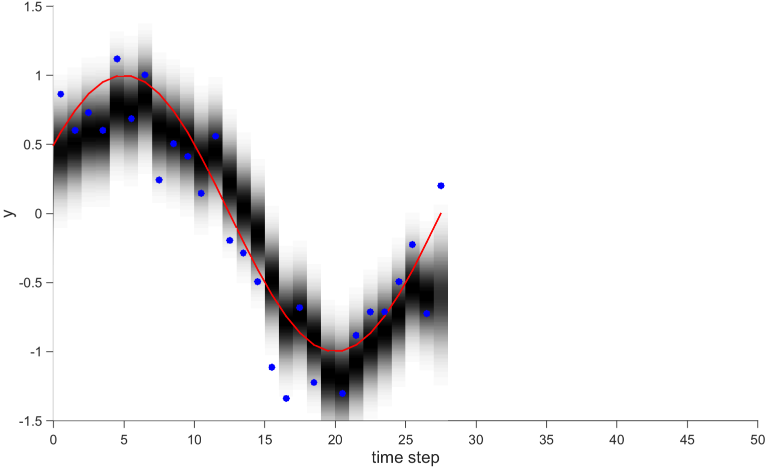
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



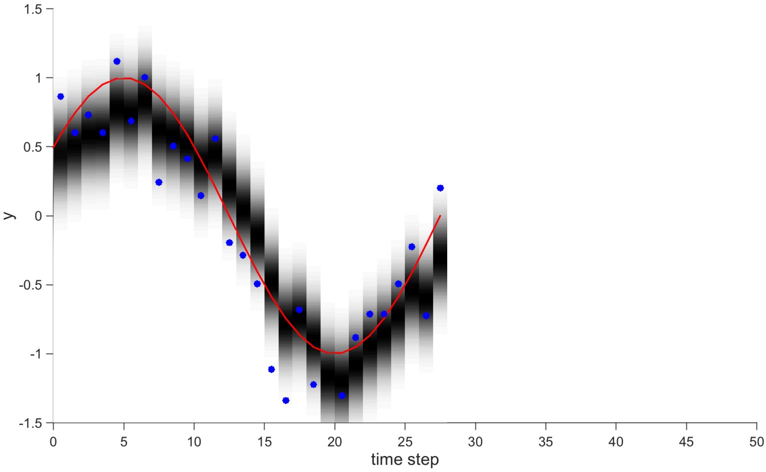
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



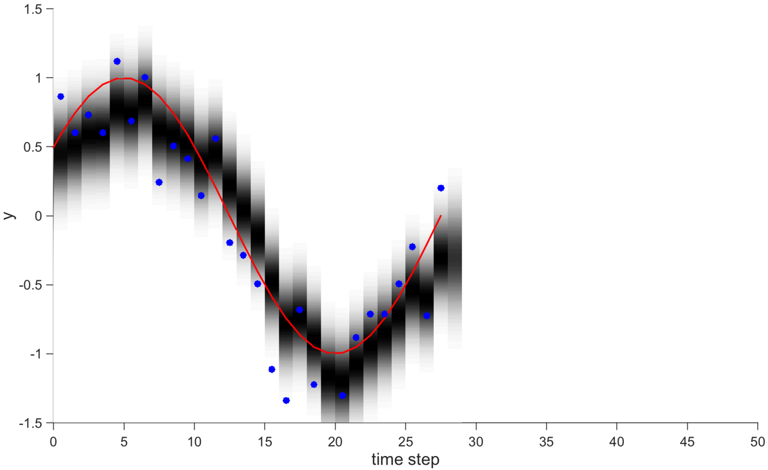
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



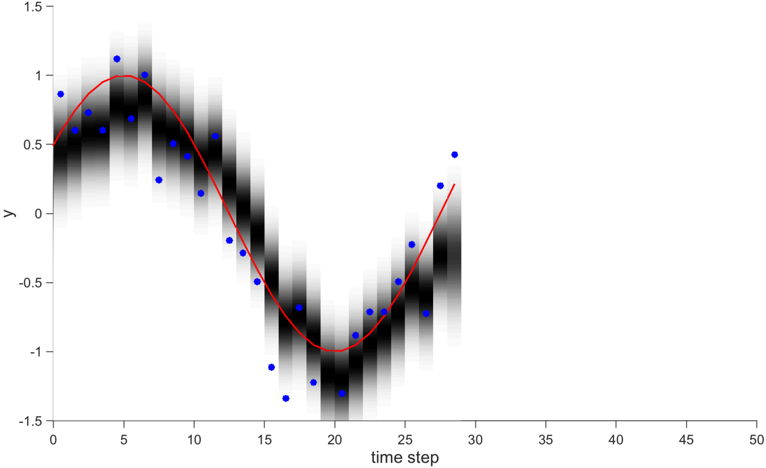
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



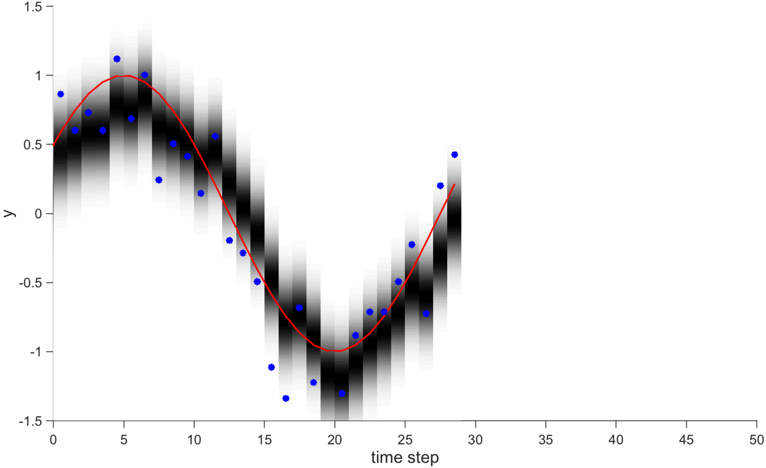
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



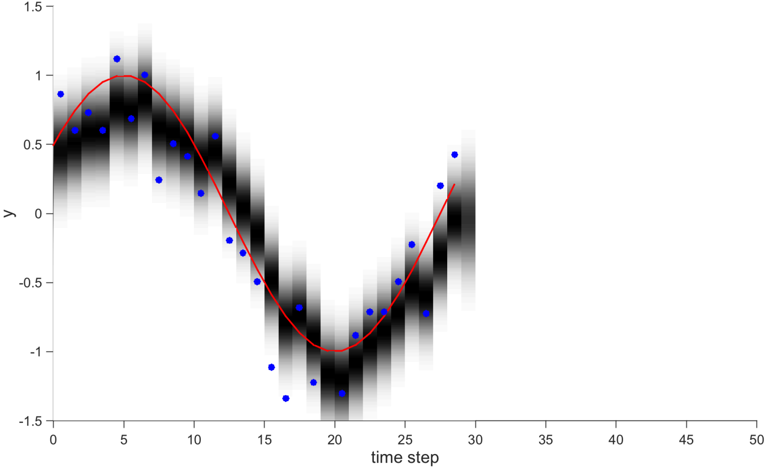
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



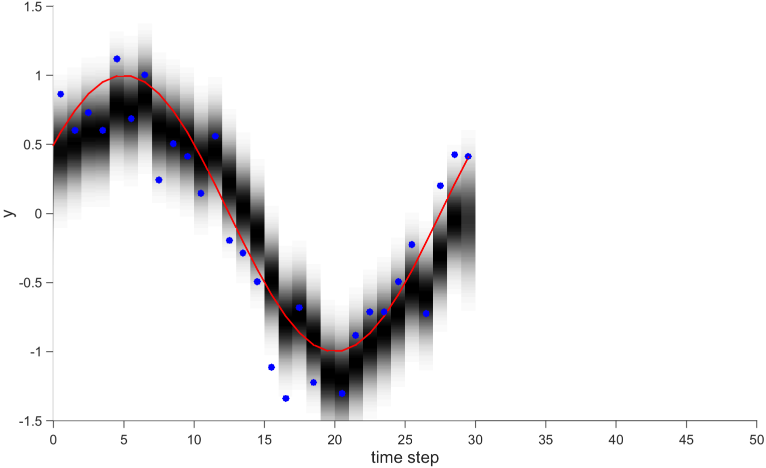
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



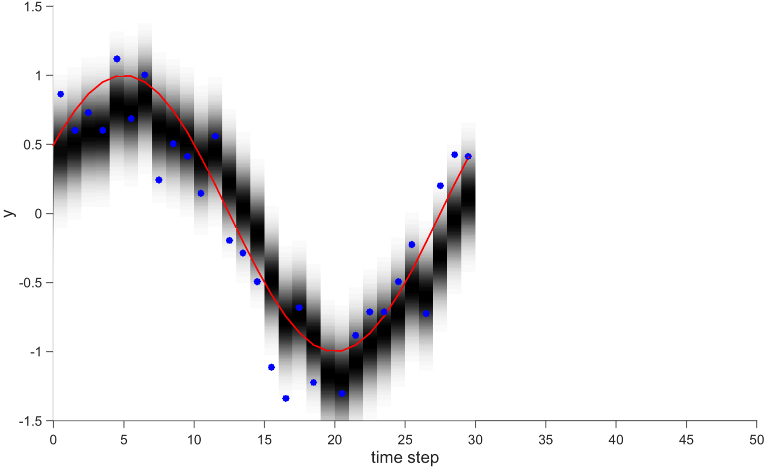
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



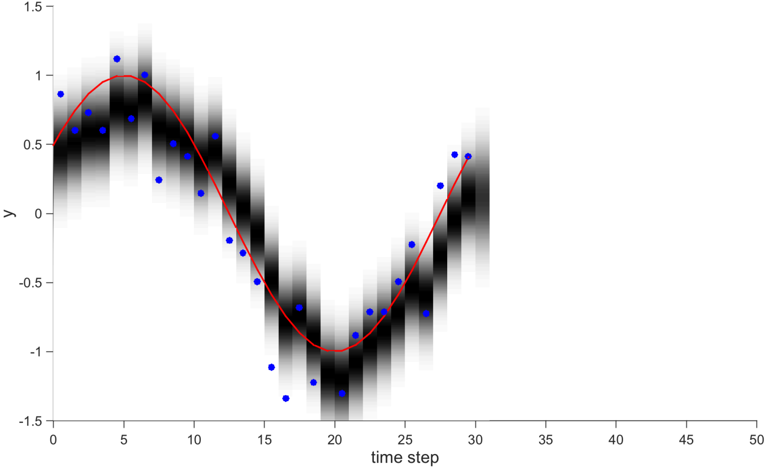
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



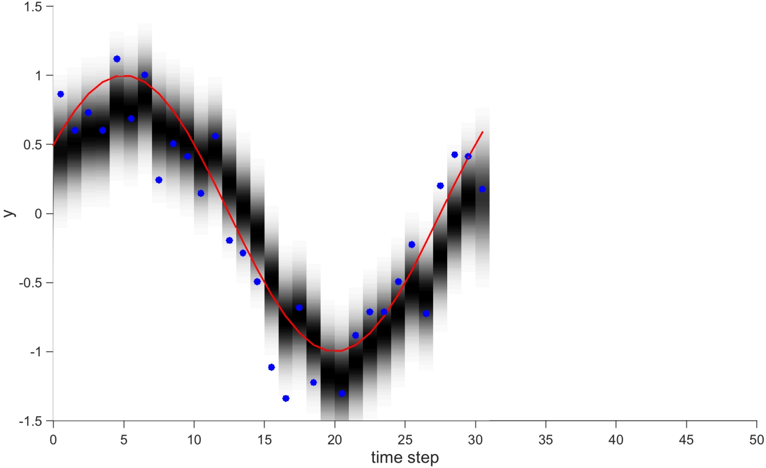
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



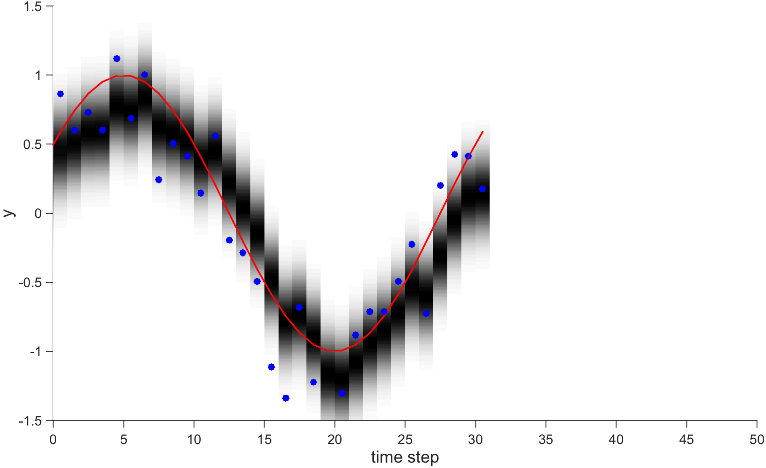
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



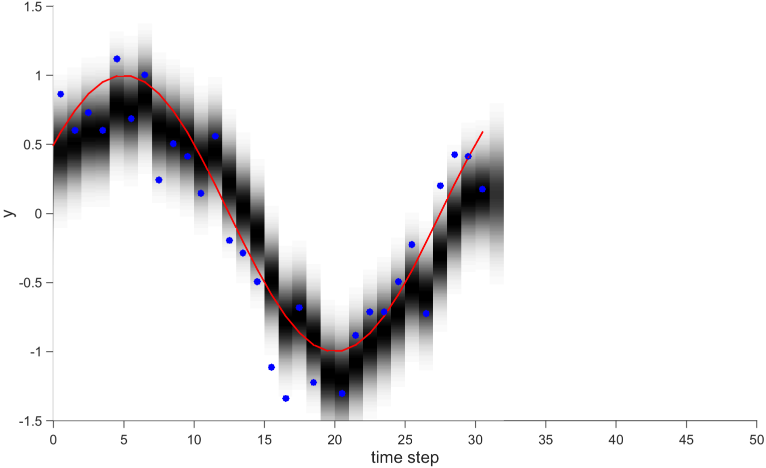
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



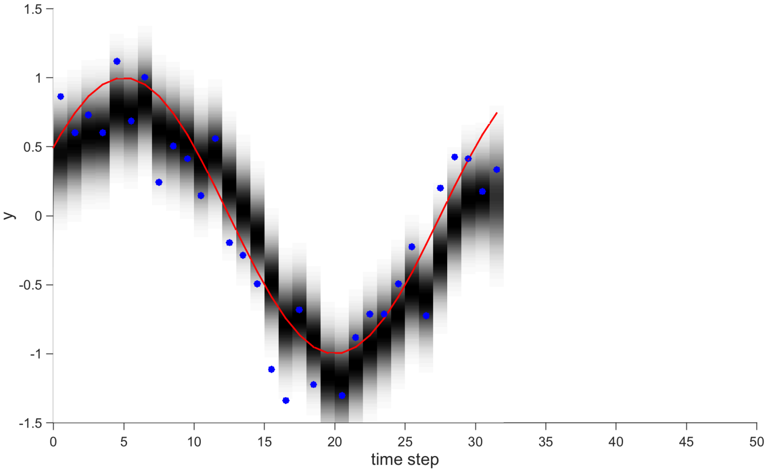
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



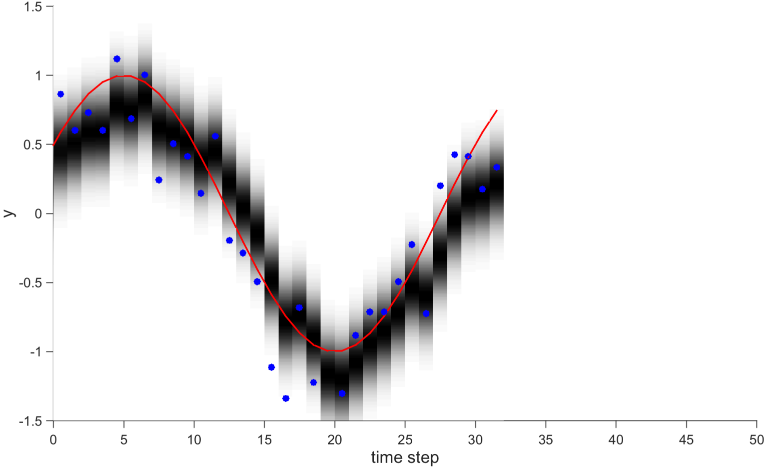
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



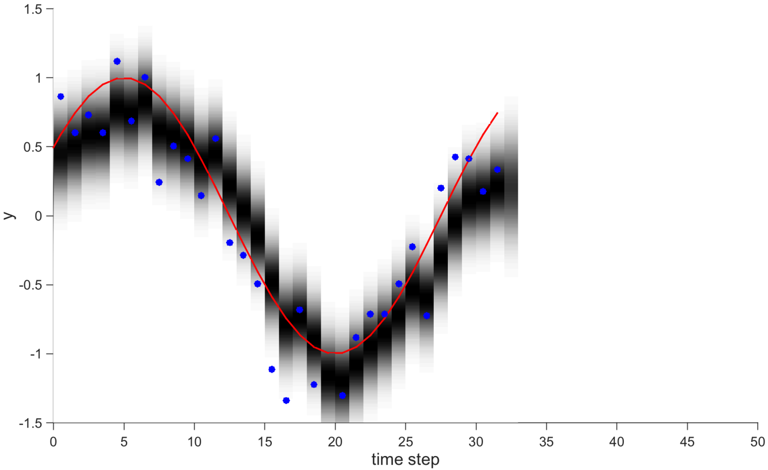
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



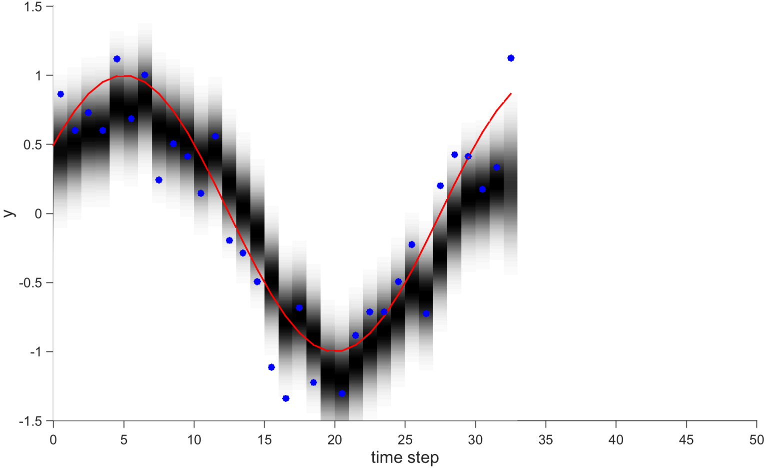
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



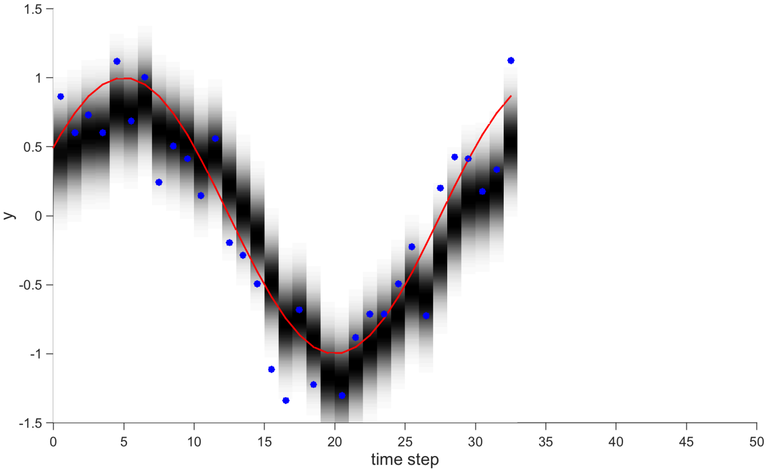
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



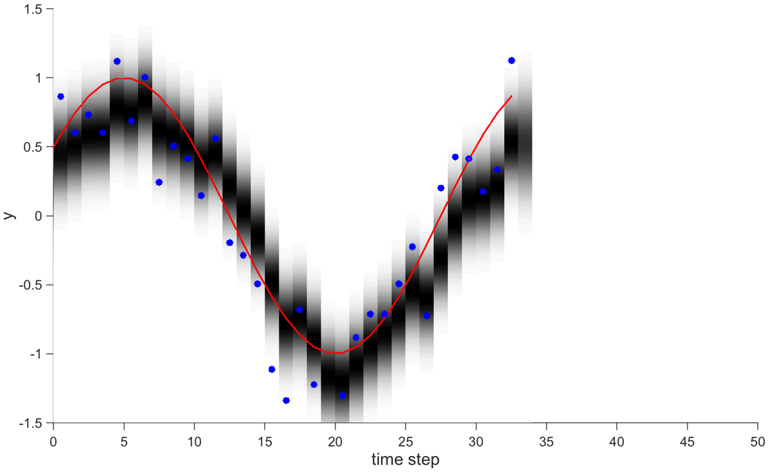
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



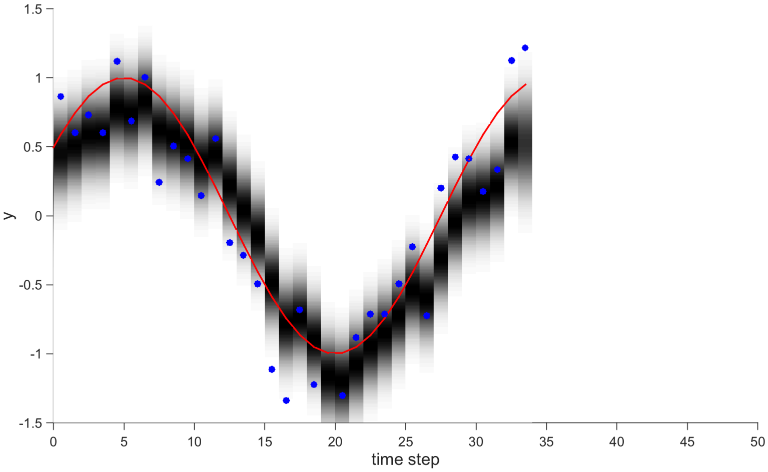
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



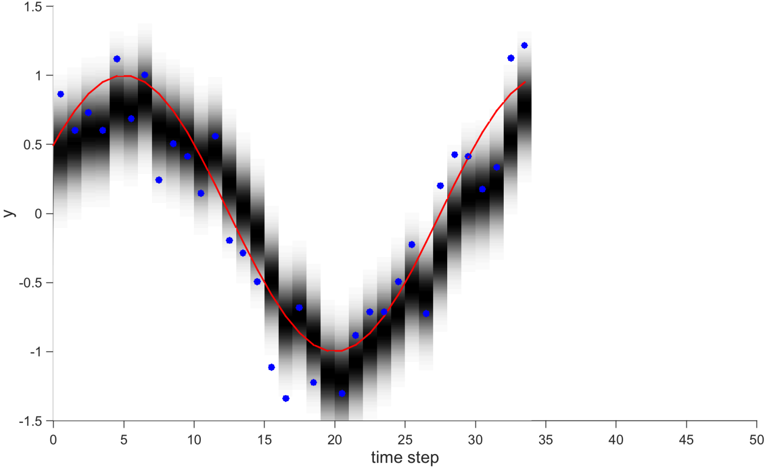
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



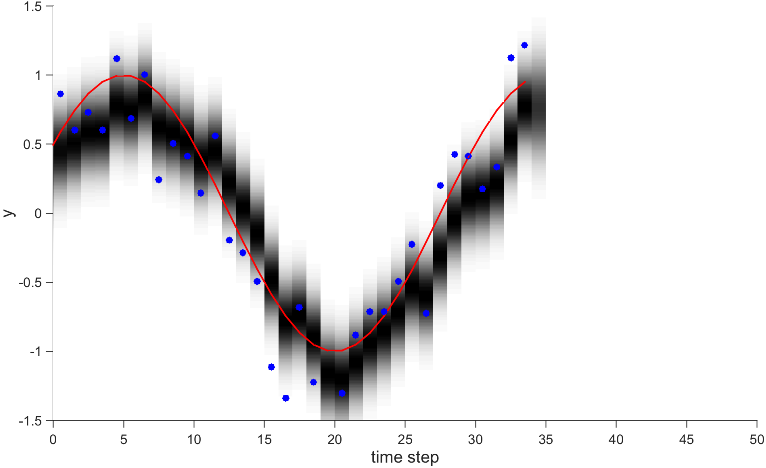
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



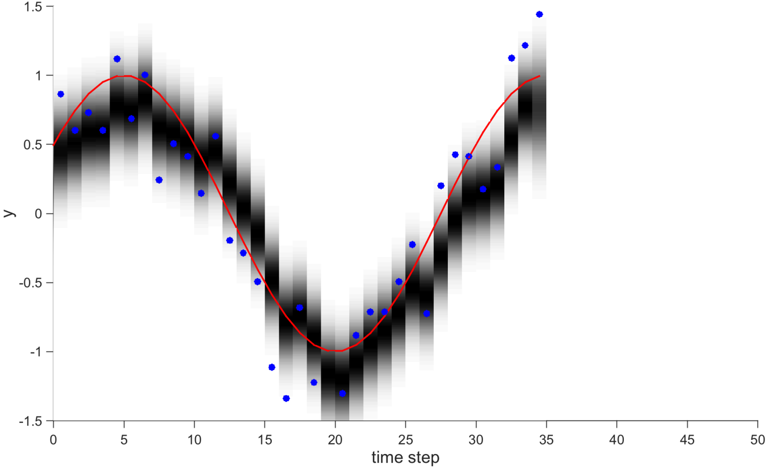
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



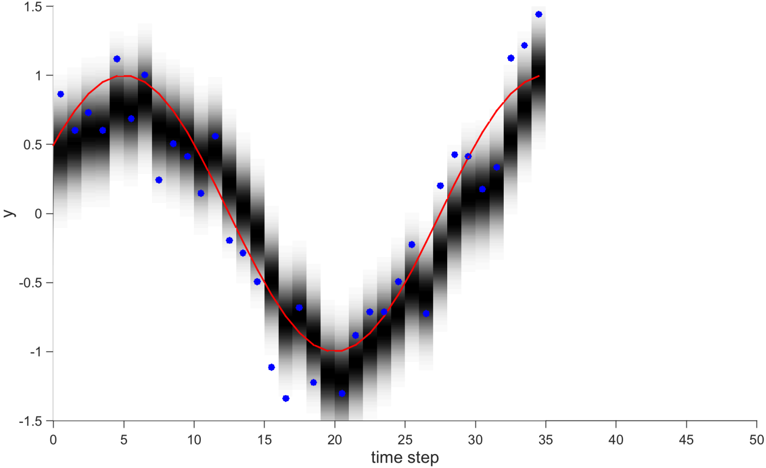
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

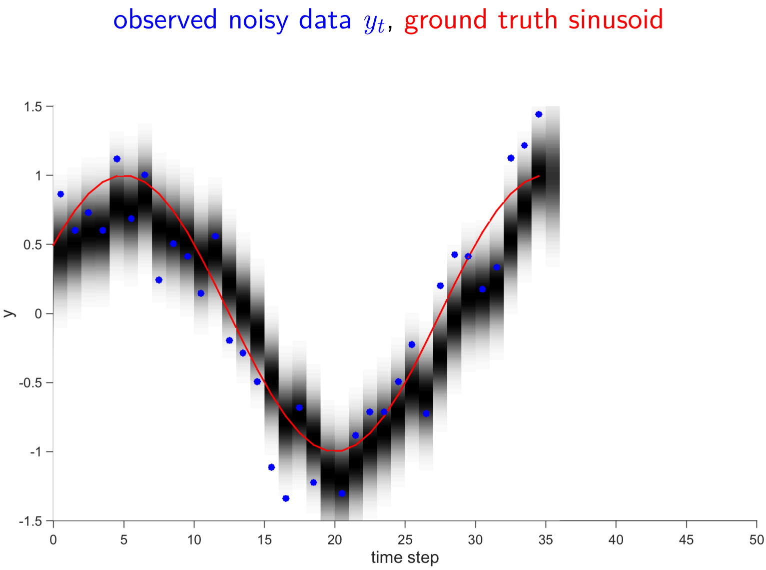


Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

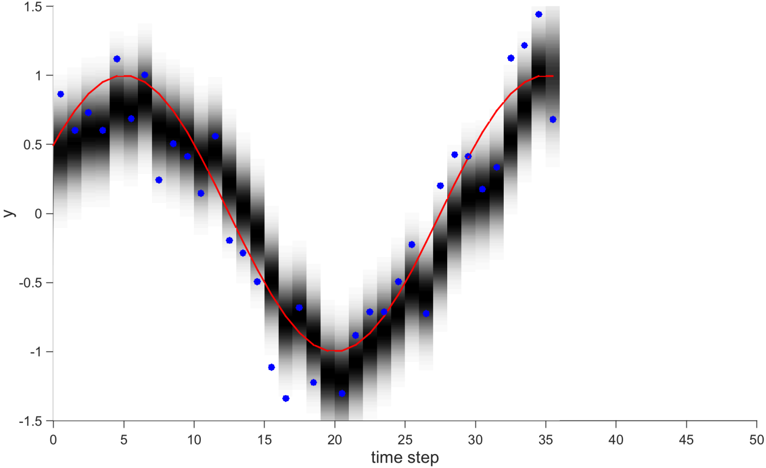


Kalman Filter Demo



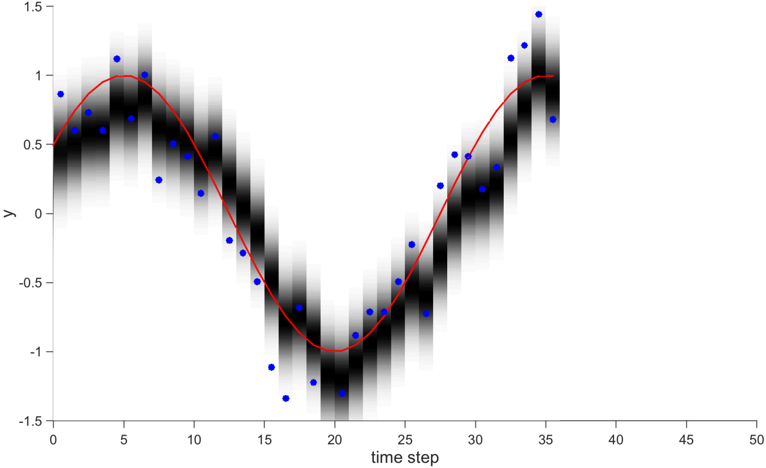
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



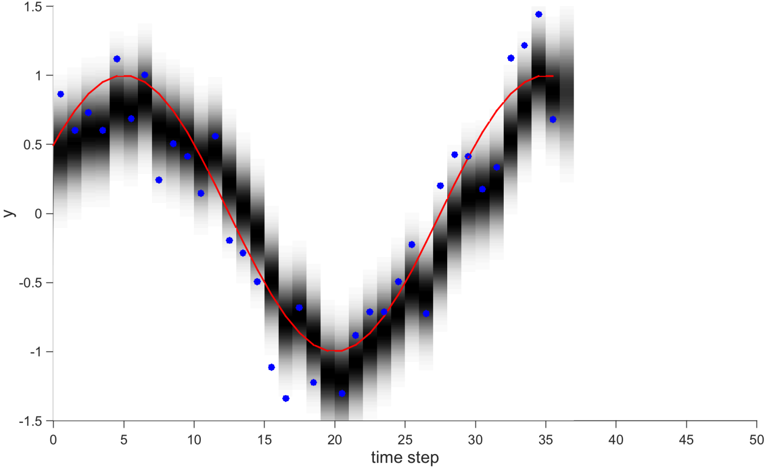
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



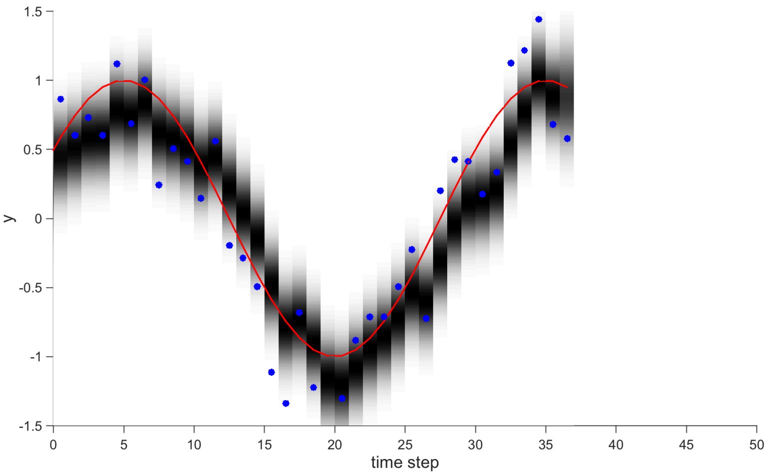
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



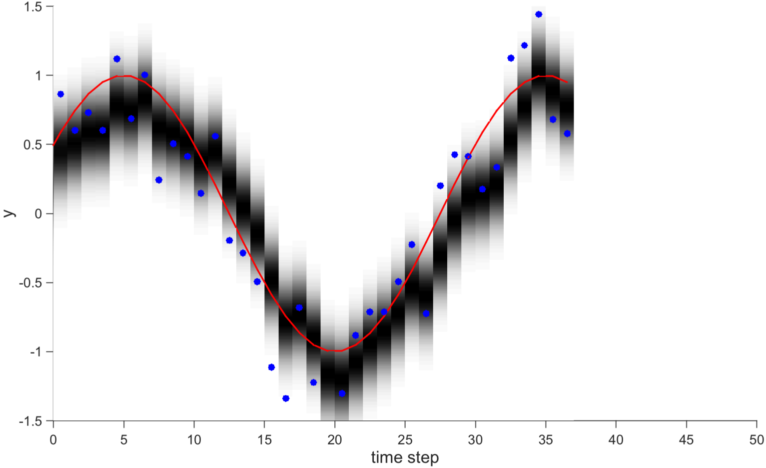
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



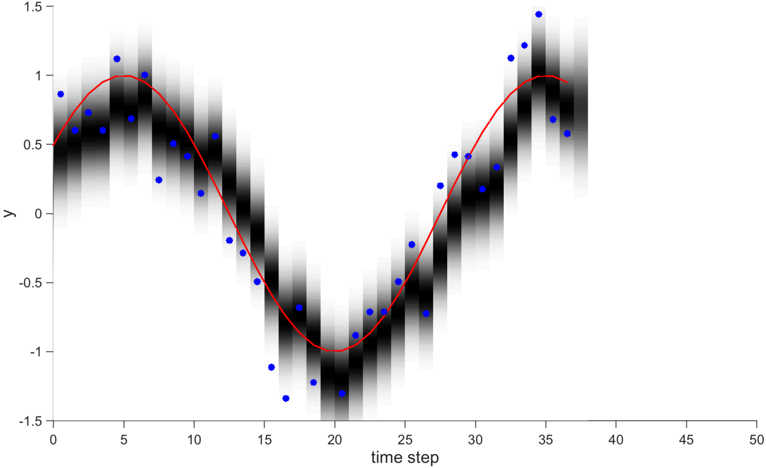
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



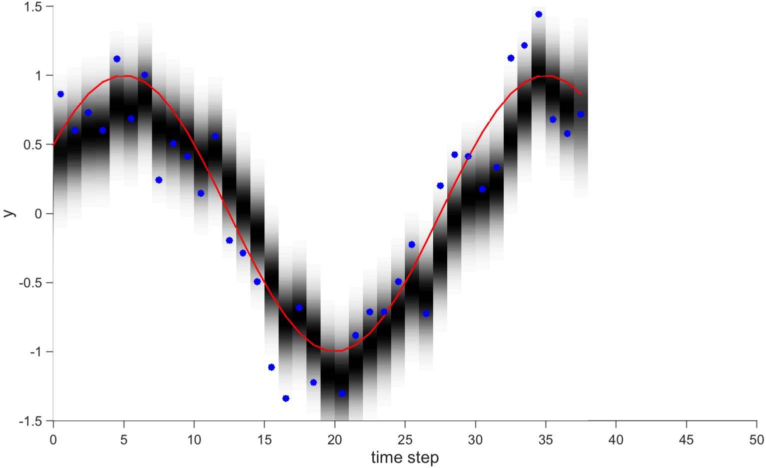
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



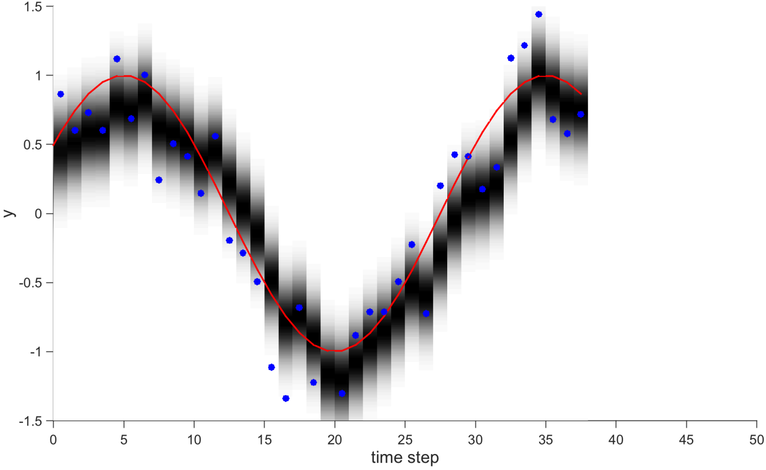
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



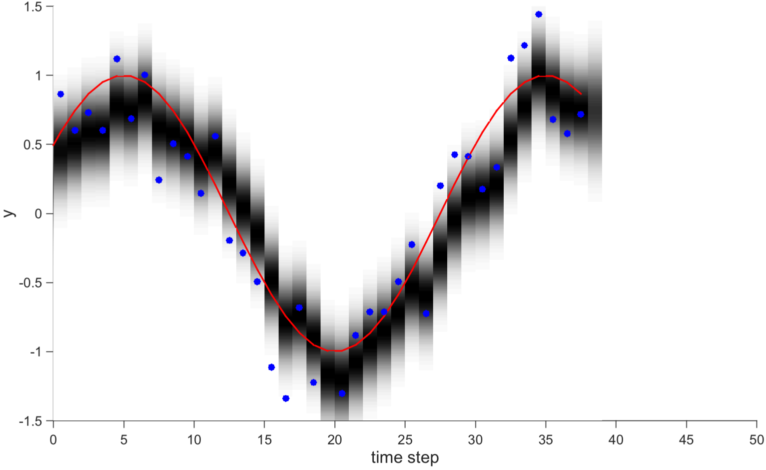
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



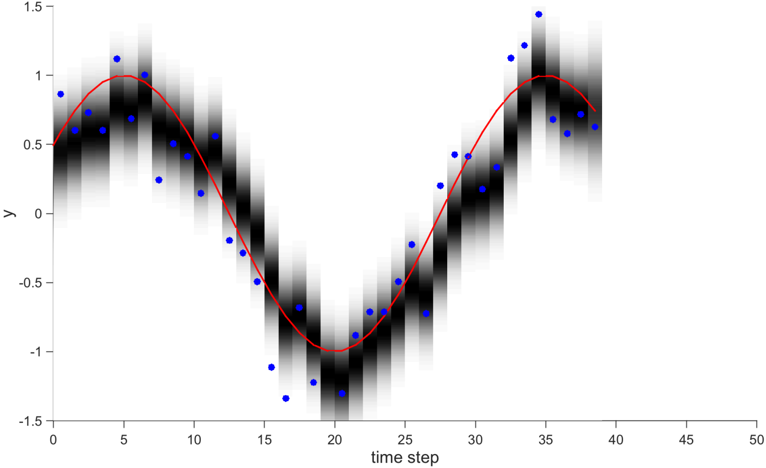
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

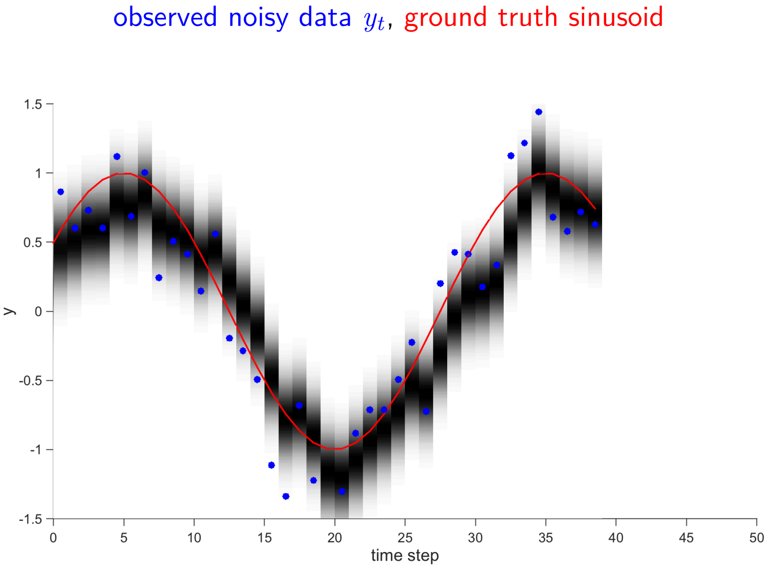


Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

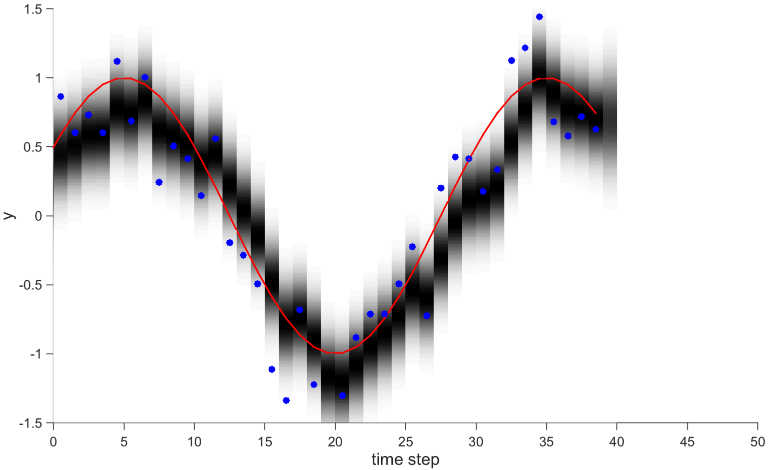


Kalman Filter Demo



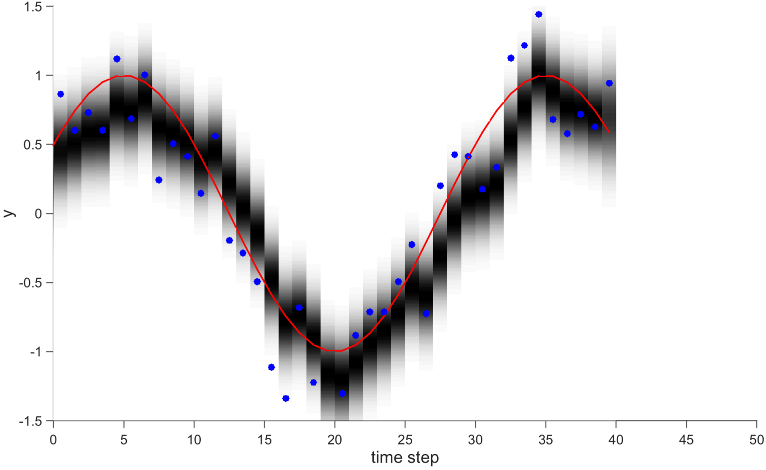
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



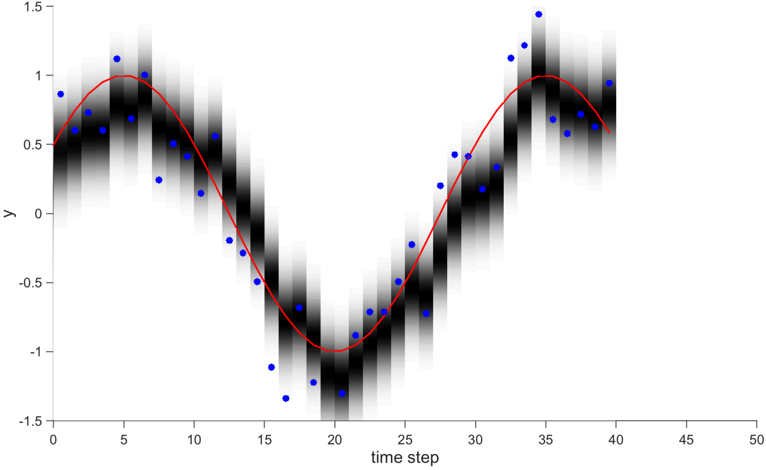
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



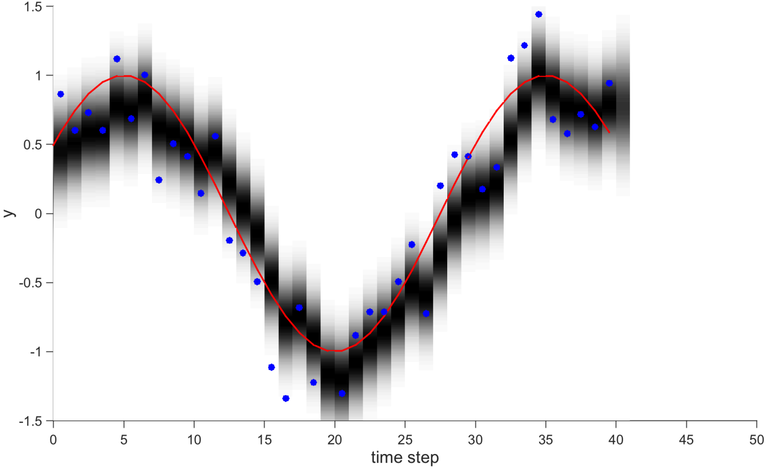
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



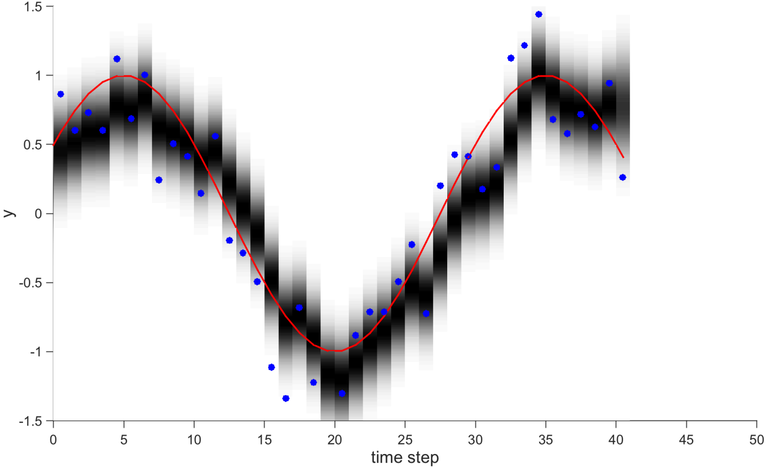
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



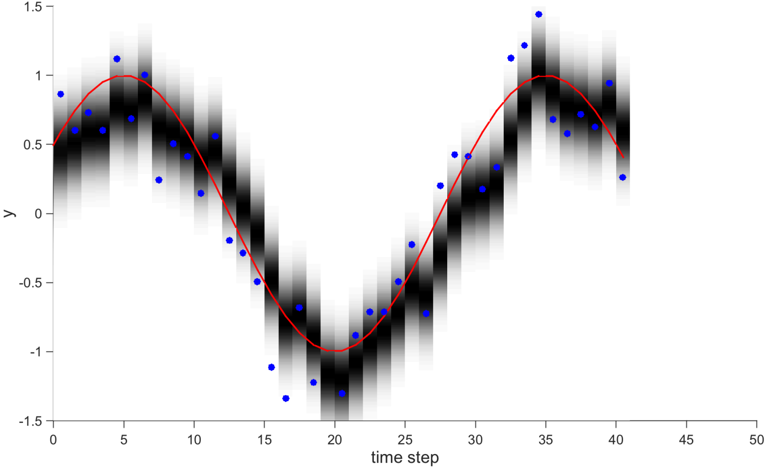
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



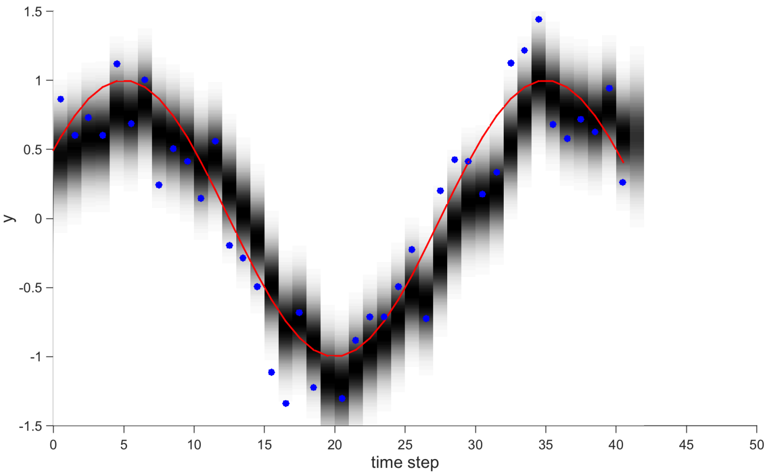
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



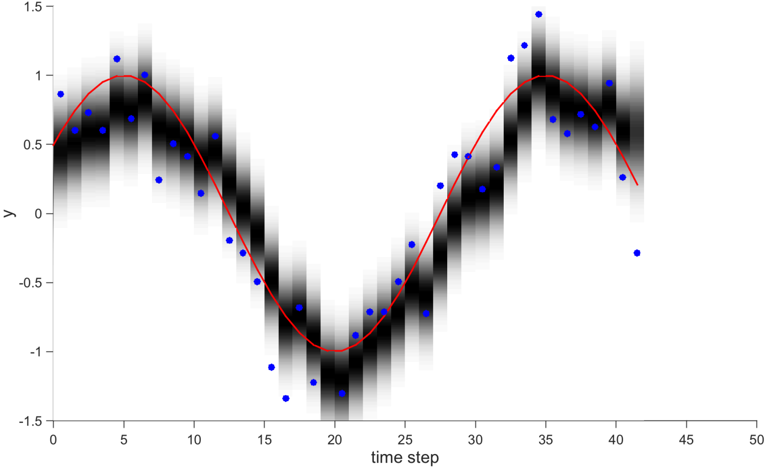
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



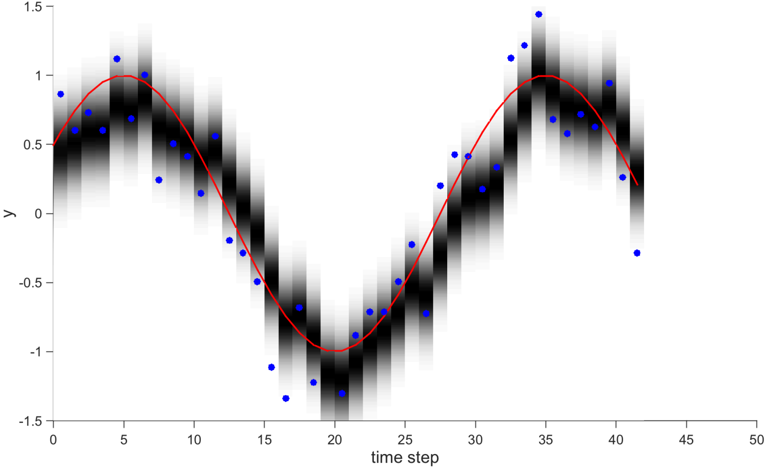
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



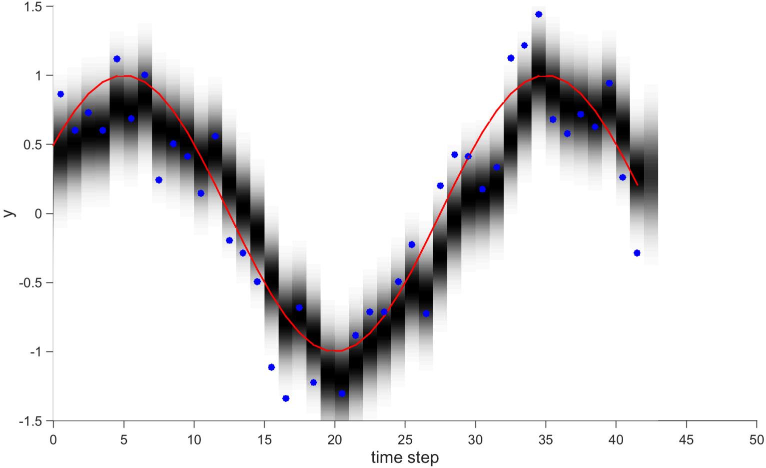
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



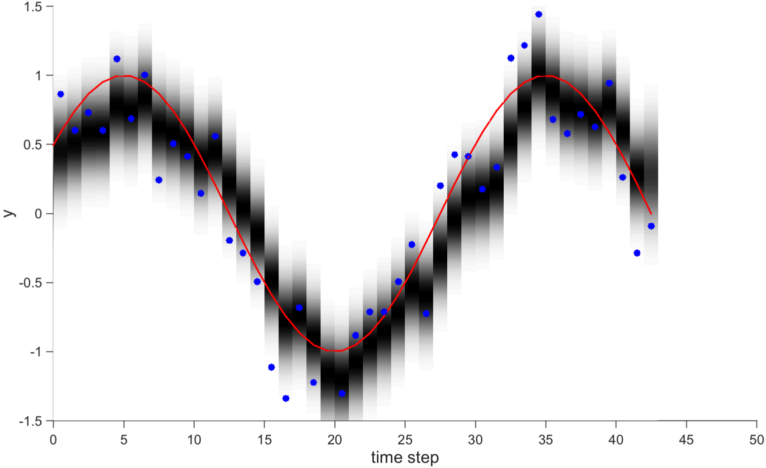
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



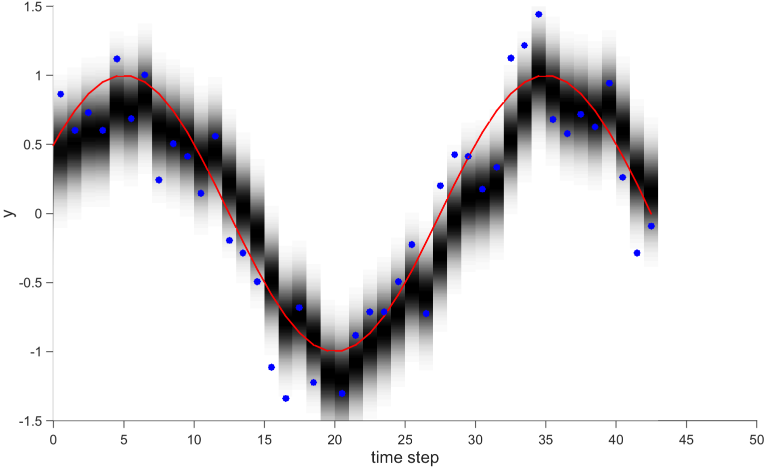
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



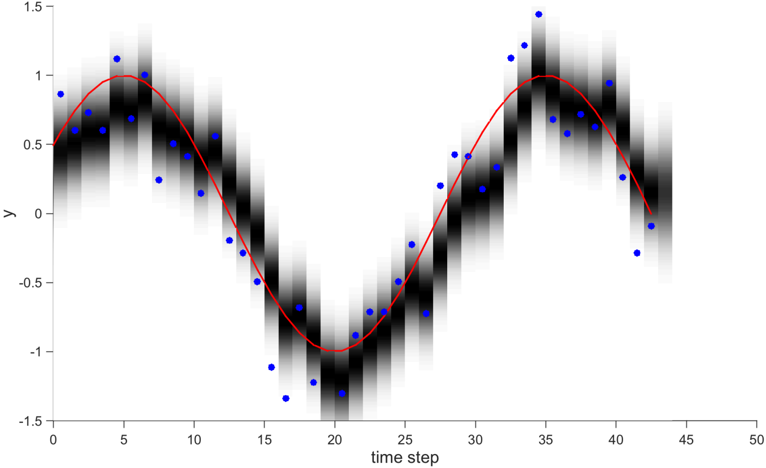
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



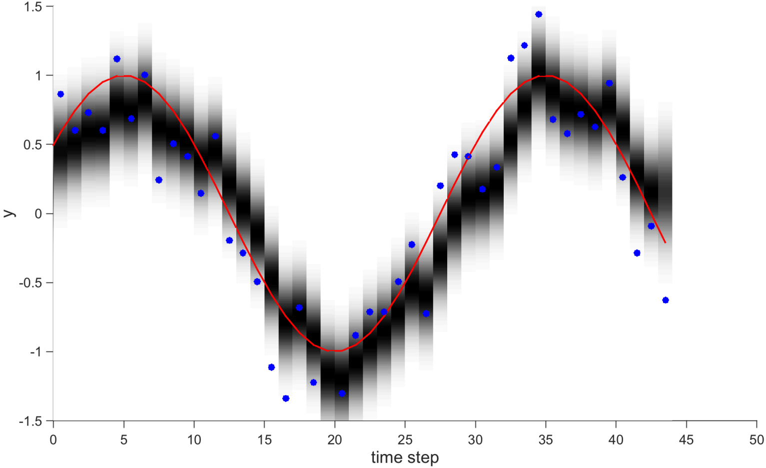
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



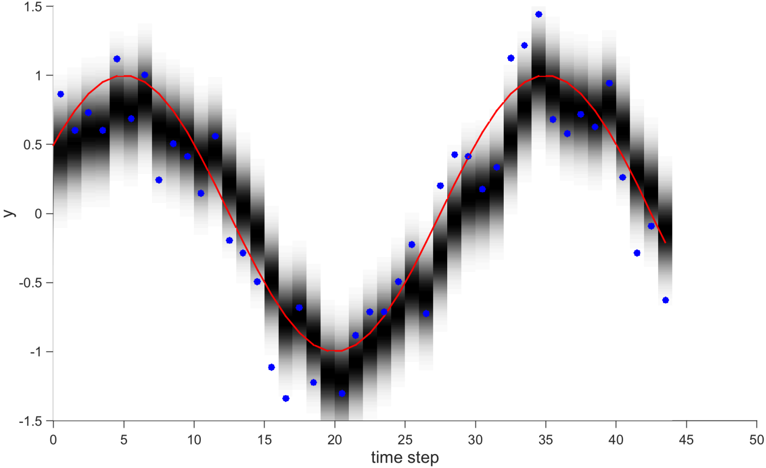
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



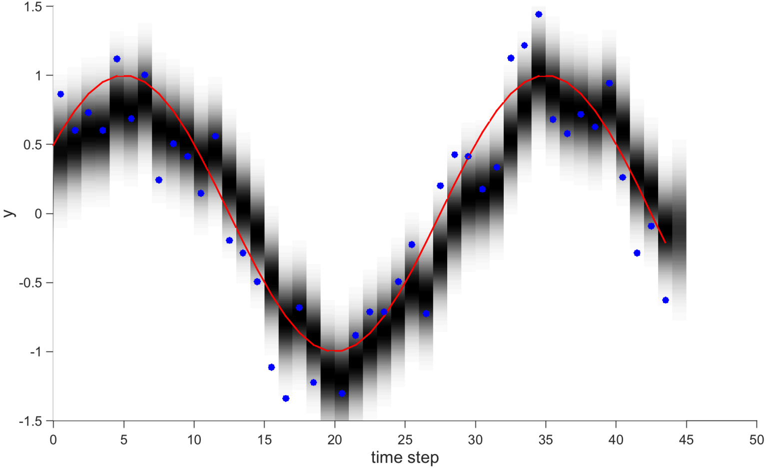
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



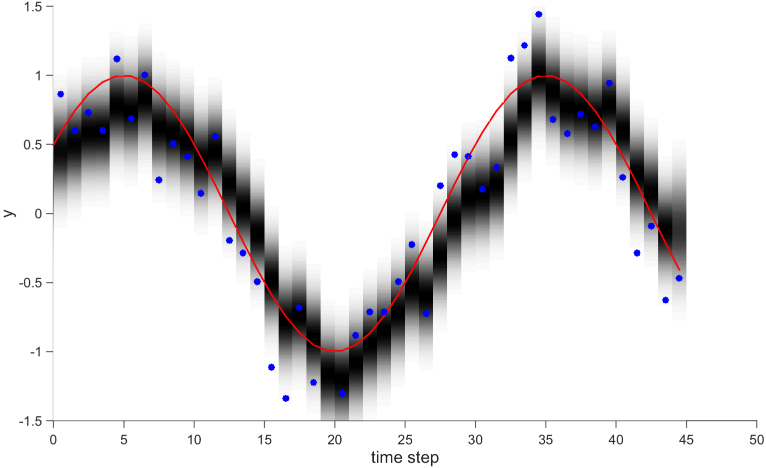
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



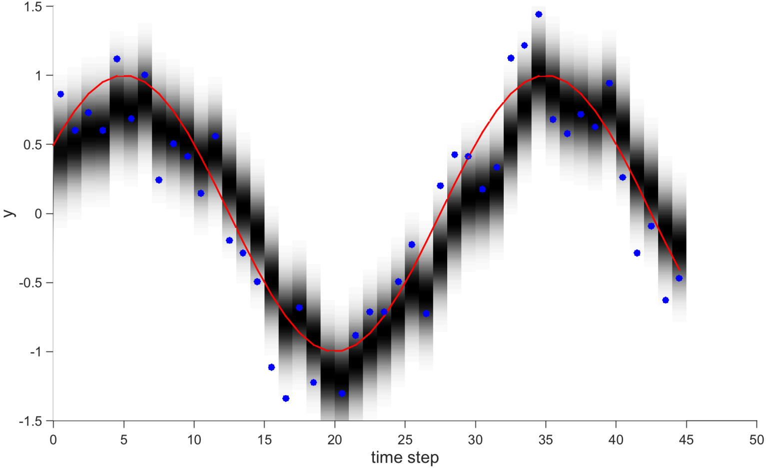
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



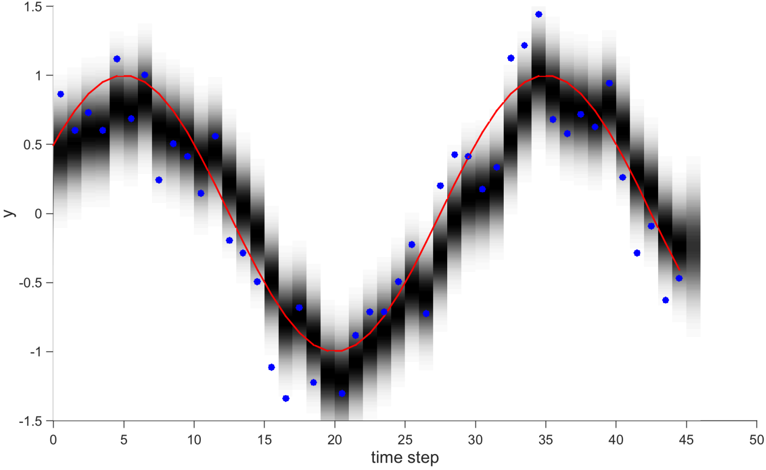
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



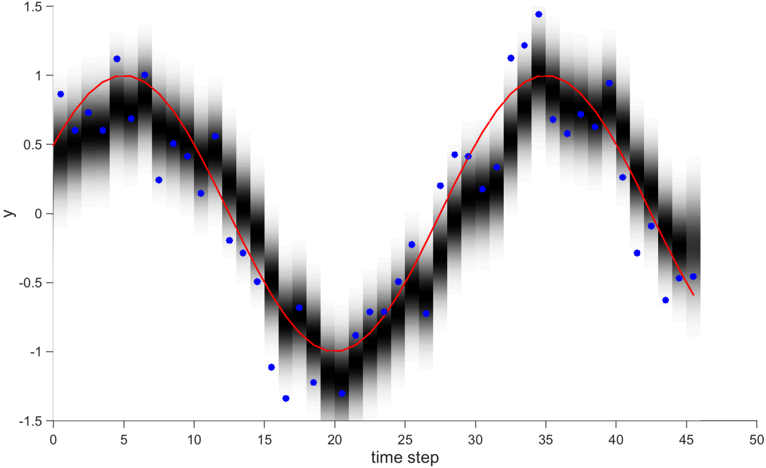
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



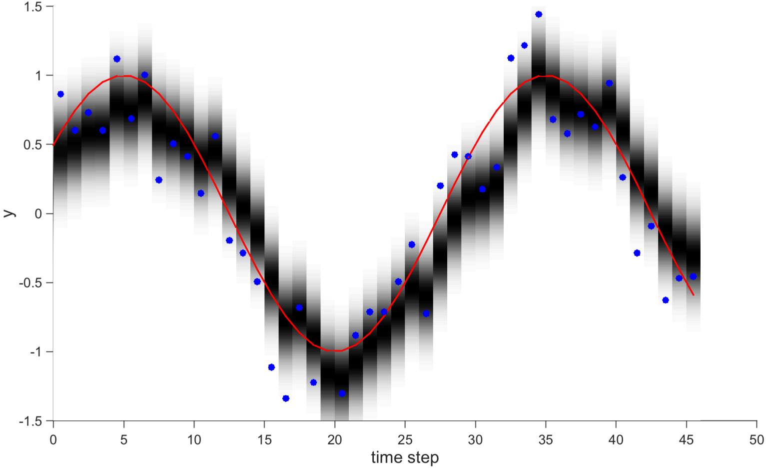
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



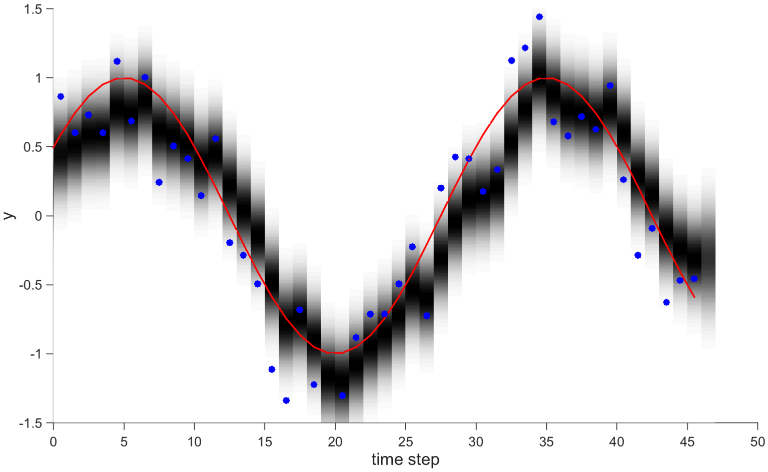
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



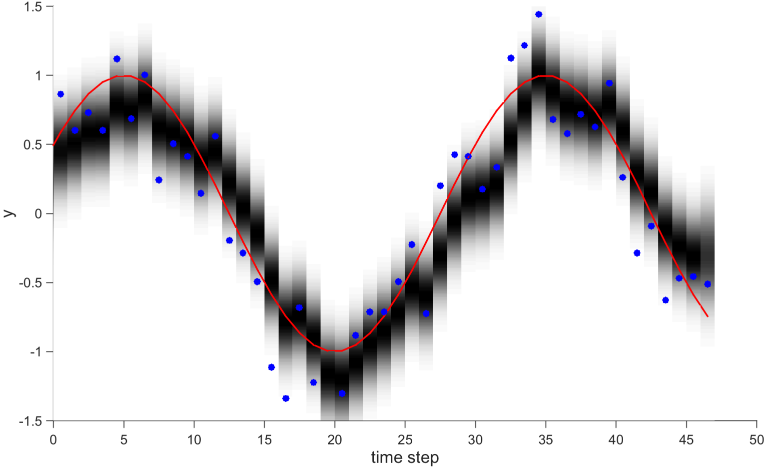
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



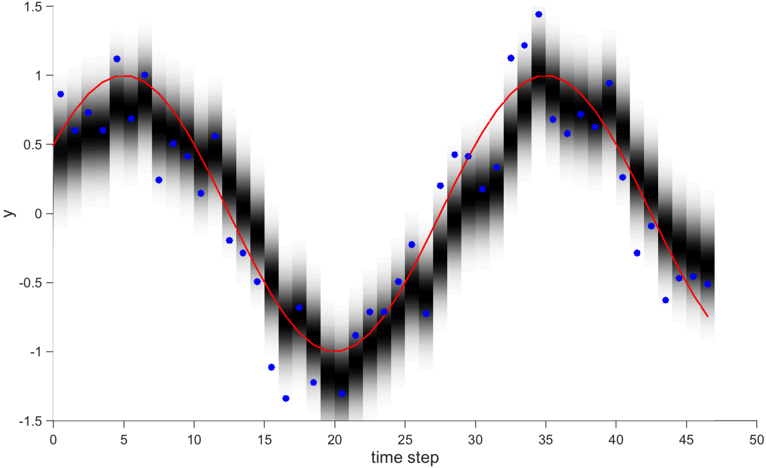
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



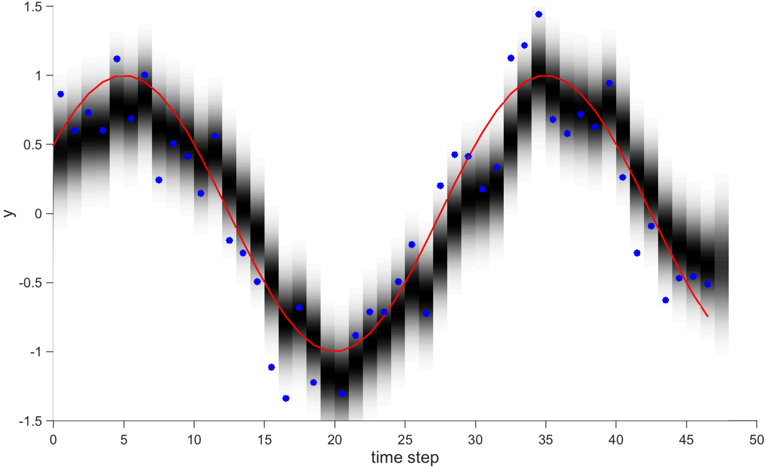
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



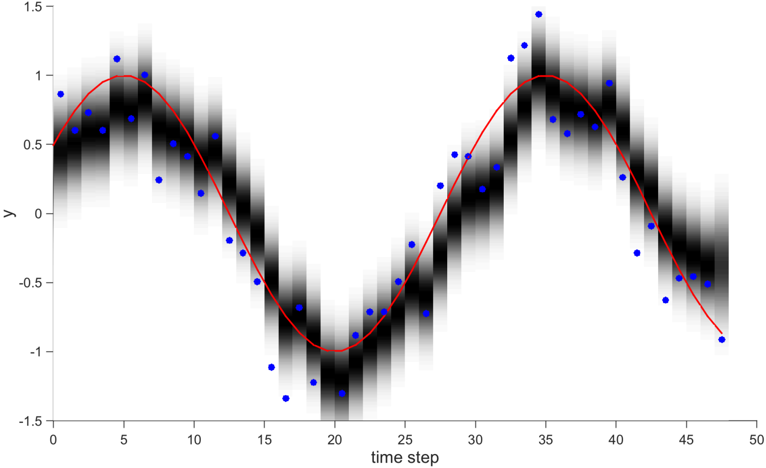
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



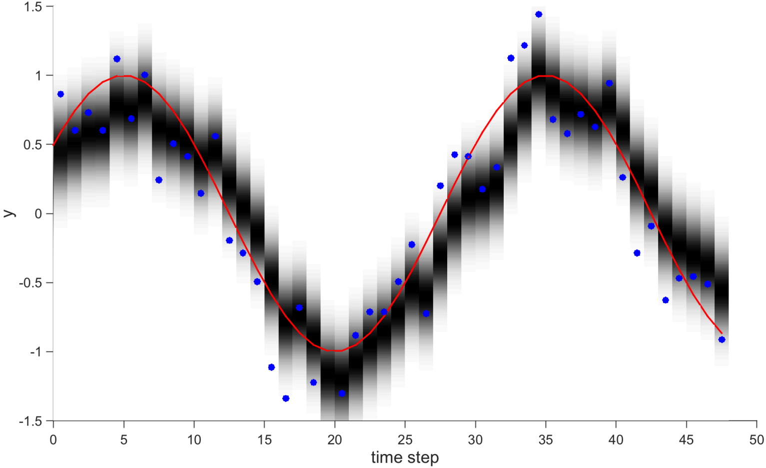
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



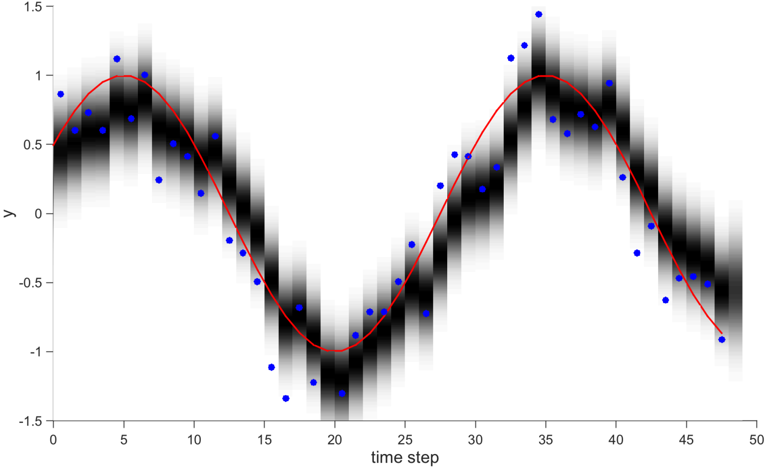
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



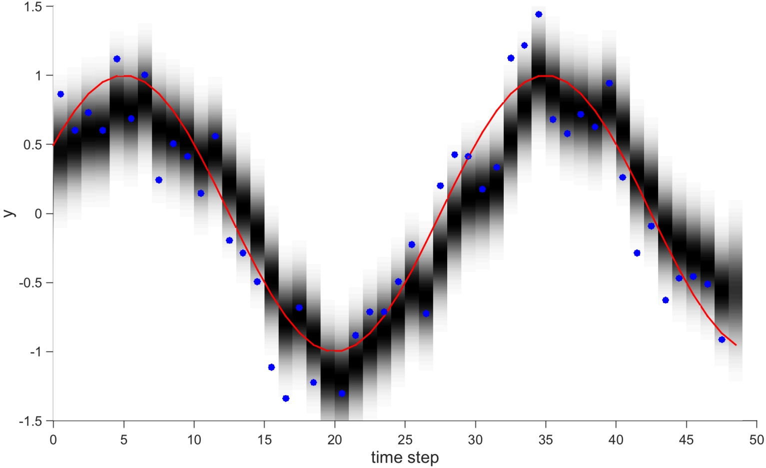
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



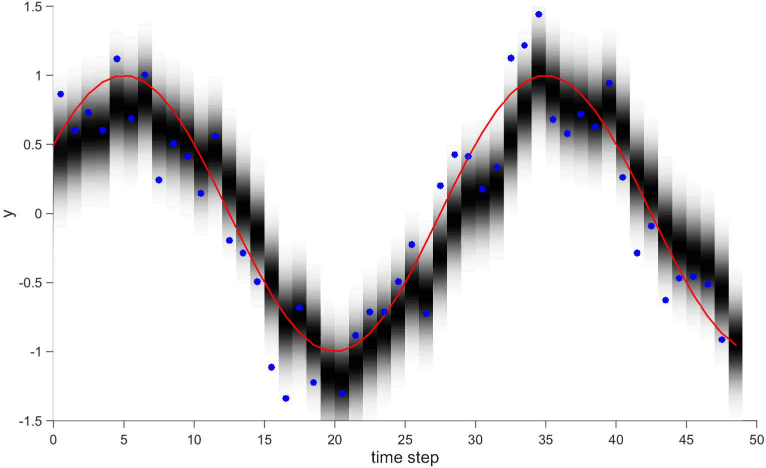
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



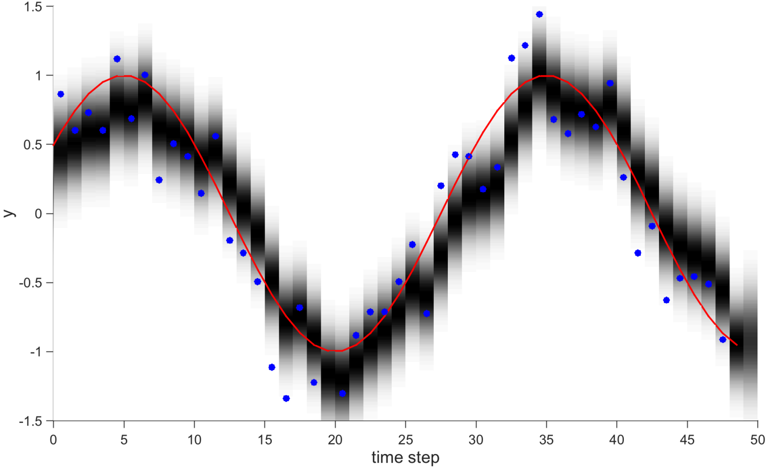
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



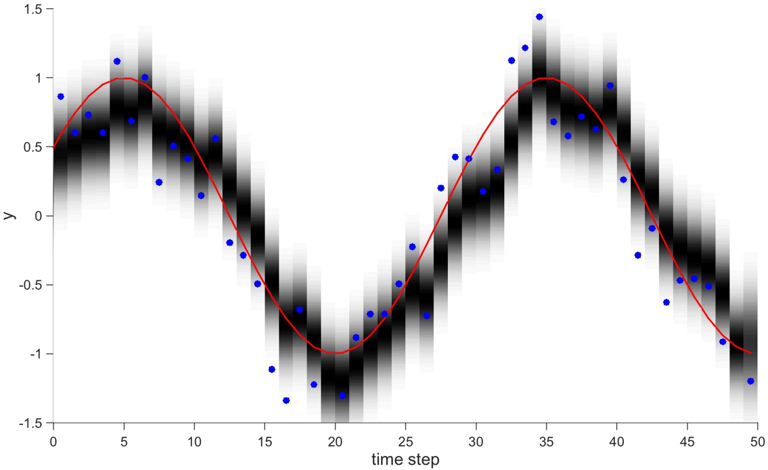
Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid



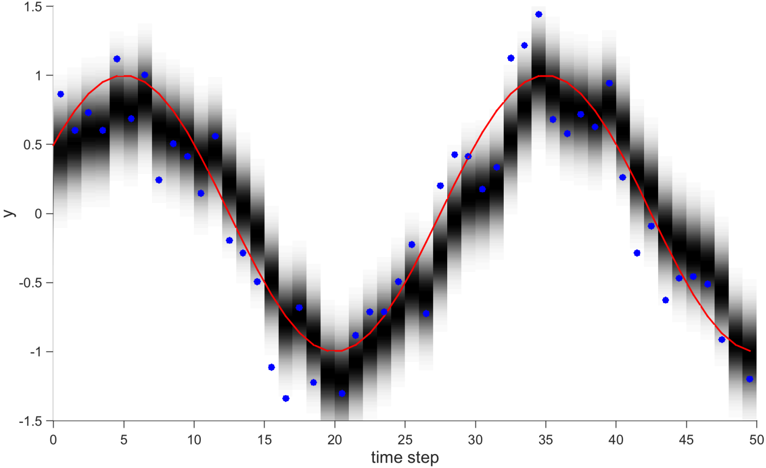
Kalman Filter Demo

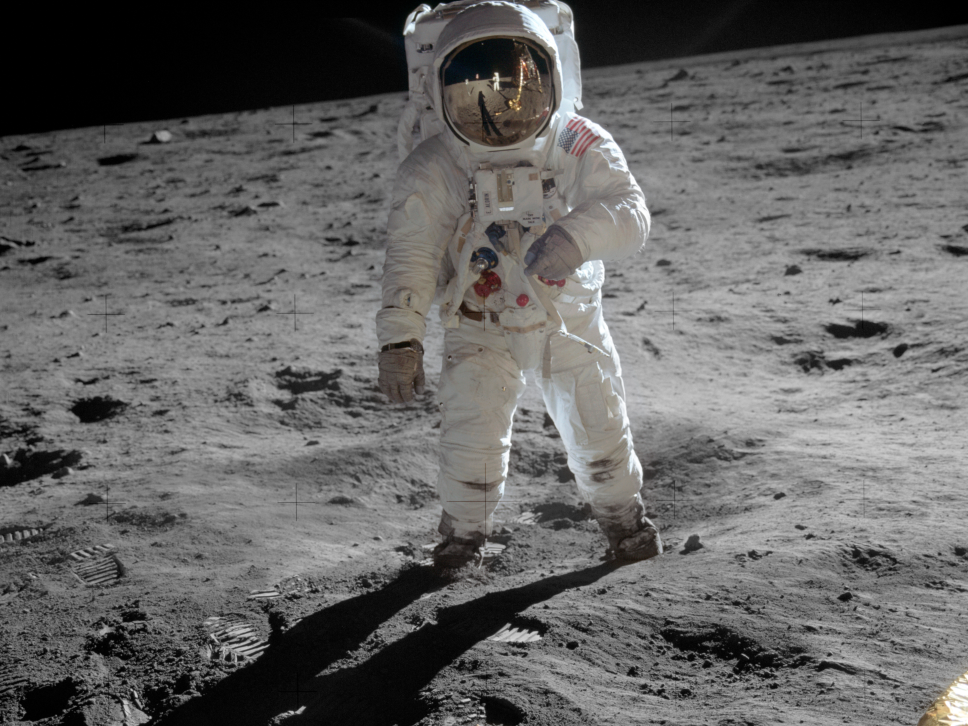
observed noisy data y_t, ground truth sinusoid



Kalman Filter Demo

observed noisy data y_t, ground truth sinusoid

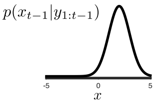




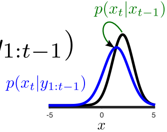
Inference: Forward Algorithm

$$p(x_{t-1} = k | y_{1:t-1})$$

diffuse via
dynamics

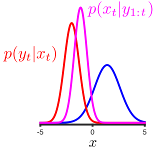


$$p(x_t = k|y_{1:t-1}) = \sum_{l=1}^{K} p(x_t = k|x_{t-1} = l)p(x_{t-1} = l|y_{1:t-1})$$



combine
with
likelihood

$$p(x_t = k|y_{1:t}) \propto \underbrace{p(x_t = k|y_{1:t-1})}_{\text{prior}}\underbrace{p(y_t|x_t = k)}_{\text{likelihood}}$$

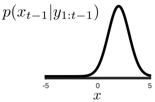


Inference: Forward Algorithm

$$p(x_{t-1} = k | y_{1:t-1}) = \rho_{t-1}^{t-1}(k)$$

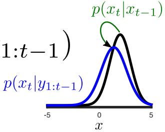
most recent data used in prediction

variable being predicted



diffuse via dynamics

$$p(x_t = k|y_{1:t-1}) = \sum_{l=1}^{K} p(x_t = k|x_{t-1} = l)p(x_{t-1} = l|y_{1:t-1})$$

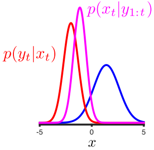


combine with likelihood

$$p(x_t = k|y_{1:t}) \propto p(x_t = k|y_{1:t-1})p(y_t|x_t = k)$$

prior

likelihood

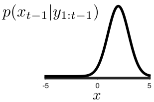


Inference: Forward Algorithm

most recent data used in prediction

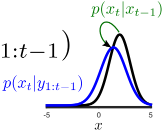
$$p(x_{t-1} = k | y_{1:t-1}) = \rho_{t-1}^{t-1}(k)$$

variable being predicted



diffuse via dynamics

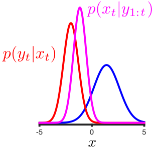
$$p(x_t = k | y_{1:t-1}) = \sum_{l=1}^{K} p(x_t = k | x_{t-1} = l) p(x_{t-1} = l | y_{1:t-1})$$



$$\rho_t^{t-1}(k) = \sum_{l=1}^{K} T(k,l) \rho_{t-1}^{t-1}(l)$$

combine with likelihood

$$p(x_t = k | y_{1:t}) \propto \underbrace{p(x_t = k | y_{1:t-1})}_{\text{prior}} \underbrace{p(y_t | x_t = k)}_{\text{likelihood}}$$

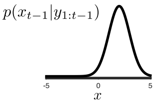


Inference: Forward Algorithm

$$p(x_{t-1}=k|y_{1:t-1}) = \rho_{t-1}^{t-1}(k)$$

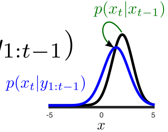
most recent data used in prediction

variable being predicted



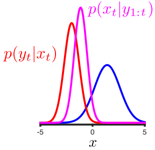
diffuse via dynamics

$$p(x_t=k|y_{1:t-1}) = \sum_{l=1}^{K} p(x_t=k|x_{t-1}=l)p(x_{t-1}=l|y_{1:t-1})$$



$$\rho_t^{t-1}(k) = \sum_{l=1}^{K} T(k,l)\rho_{t-1}^{t-1}(l)$$

combine with likelihood



$$p(x_t=k|y_{1:t}) \propto \underbrace{p(x_t=k|y_{1:t-1})}_{\text{prior}}\underbrace{p(y_t|x_t=k)}_{\text{likelihood}}$$

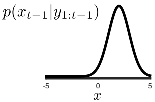
$$\rho_t^t(k) \propto \rho_t^{t-1}(k)p(y_t|x_t=k)$$

Inference: Forward Algorithm

$$p(x_{t-1}=k|y_{1:t-1}) = \rho_{t-1}^{t-1}(k)$$

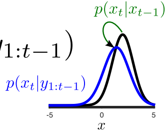
most recent data used in prediction

variable being predicted



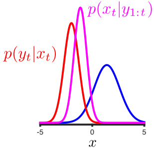
diffuse via dynamics

$$p(x_t=k|y_{1:t-1}) = \sum_{l=1}^{K} p(x_t=k|x_{t-1}=l)p(x_{t-1}=l|y_{1:t-1})$$



$$\rho_t^{t-1}(k) = \sum_{l=1}^{K} T(k,l)\rho_{t-1}^{t-1}(l)$$

combine with likelihood



$$p(x_t=k|y_{1:t}) \propto \underbrace{p(x_t=k|y_{1:t-1})}_{\text{prior}}\underbrace{p(y_t|x_t=k)}_{\text{likelihood}}$$

$$\rho_t^t(k) \propto \rho_t^{t-1}(k)p(y_t|x_t=k)$$

When implementing, take care with numerical underflow/overflow.

Computing the likelihood

How can we compute the likelihood efficiently?

Computing the likelihood

How can we compute the likelihood efficiently?

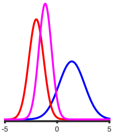
$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1})$$

Computing the likelihood

How can we compute the likelihood efficiently?

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1})$$

already retuned by Kalman Filter/Forward algorithm



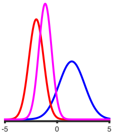
$$\begin{aligned} p(x_t|y_{1:t}) &= \frac{1}{p(y_t|y_{1:t-1})} \; p(y_t|x_t)p(x_t|y_{1:t-1}) \\ &\propto p(y_t|x_t)p(x_t|y_{1:t-1}) \end{aligned}$$

Computing the likelihood

How can we compute the likelihood efficiently?

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1})$$

already retuned by Kalman Filter/Forward algorithm



$$p(x_t|y_{1:t}) = \frac{1}{p(y_t|y_{1:t-1})} p(y_t|x_t)p(x_t|y_{1:t-1})$$

$$\propto p(y_t|x_t)p(x_t|y_{1:t-1})$$

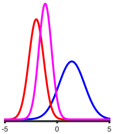
$p(y_t|y_{1:t-1})$ is normaliser of filter/forward algorithm update

Computing the likelihood

How can we compute the likelihood efficiently?

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1})$$

already retuned by Kalman Filter/Forward algorithm



$$p(x_t|y_{1:t}) = \frac{1}{p(y_t|y_{1:t-1})} p(y_t|x_t)p(x_t|y_{1:t-1})$$

$$\propto p(y_t|x_t)p(x_t|y_{1:t-1})$$

$p(y_t|y_{1:t-1})$ is normaliser of filter/forward algorithm update

How can we compute the smoothing estimate?

$$p(x_t|y_{1:T})$$

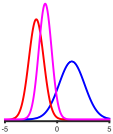
LGSSM: Kalman Smoother
HMM: Forward-Backward= Algorithm

Computing the likelihood

How can we compute the likelihood efficiently?

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1})$$

already retuned by Kalman Filter/Forward algorithm



$$p(x_t|y_{1:t}) = \frac{1}{p(y_t|y_{1:t-1})} p(y_t|x_t)p(x_t|y_{1:t-1})$$

$$\propto p(y_t|x_t)p(x_t|y_{1:t-1})$$

$p(y_t|y_{1:t-1})$ is normaliser of filter/forward algorithm update

How can we compute the smoothing estimate?

$$p(x_t|y_{1:T})$$

LGSSM: Kalman Smoother
HMM: Forward-Backward= Algorithm

How can we compute the most probable sequence?

$$x'_{1:T} = \arg\max_{x_{1:T}} p(x_{1:T}|y_{1:T})$$

LGSSM: Kalman Smoother
HMM: Viterbi Decoding

The magic of the Forward Algorithm: Dynamic Programming

What's going on here?

In discrete case, likelihood involves sum over all sequences: $x_{1:T}^{(k)}$

$$p(y_{1:T}) = \sum_{\text{all sequences } k} p(y_{1:T}, x_{1:T}^{(k)})$$

The magic of the Forward Algorithm: Dynamic Programming

What's going on here?

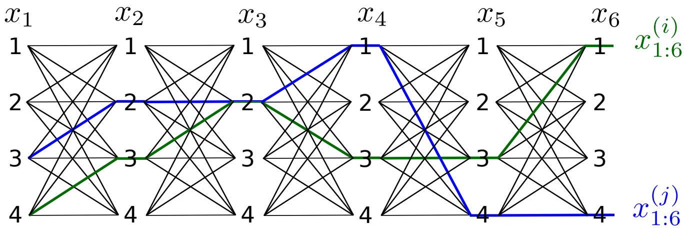
In discrete case, likelihood involves sum over all sequences: $x_{1:T}^{(k)}$

$$p(y_{1:T}) = \sum_{\text{all sequences } k} p(y_{1:T}, x_{1:T}^{(k)})$$

Trellis diagram represents possible sequences:

$K = 4$

$T = 6$



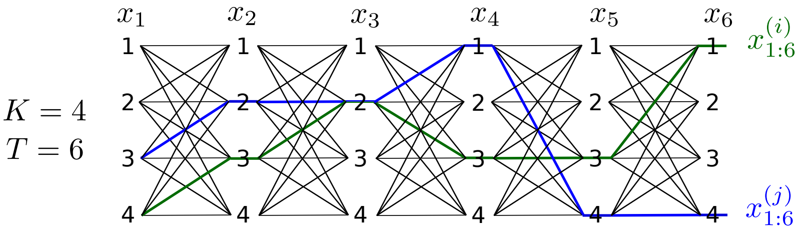
The magic of the Forward Algorithm: Dynamic Programming

What's going on here?

In discrete case, likelihood involves sum over all sequences: $x_{1:T}^{(k)}$

$$p(y_{1:T}) = \sum_{\text{all sequences } k} p(y_{1:T}, x_{1:T}^{(k)})$$

Trellis diagram represents possible sequences:



Exponential number of sequences: K^T

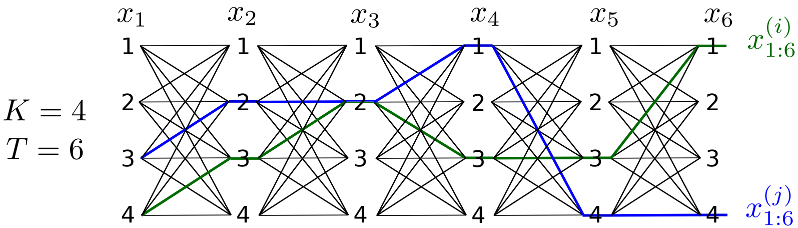
The magic of the Forward Algorithm: Dynamic Programming

What's going on here?

In discrete case, likelihood involves sum over all sequences: $x_{1:T}^{(k)}$

$$p(y_{1:T}) = \sum_{\text{all sequences } k} p(y_{1:T}, x_{1:T}^{(k)})$$

Trellis diagram represents possible sequences:



Exponential number of sequences: K^T

But Forward algorithm had linear complexity in time (loop over t)

The magic of the Forward Algorithm: Dynamic Programming

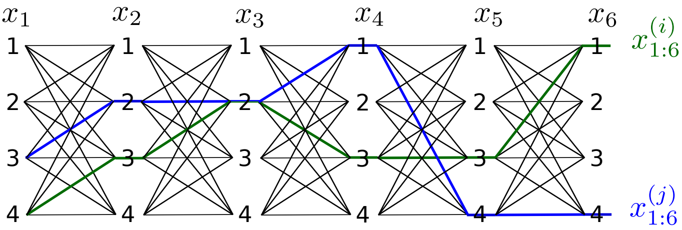
What's going on here?

In discrete case, likelihood involves sum over all sequences: $x_{1:T}^{(k)}$

$$p(y_{1:T}) = \sum_{\text{all sequences } k} p(y_{1:T}, x_{1:T}^{(k)})$$

Trellis diagram represents possible sequences:

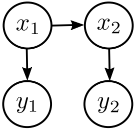
$K = 4$
$T = 6$



Exponential number of sequences: K^T

But Forward algorithm had linear complexity in time (loop over t)

Markov property means we can forget history of previous states: just remember last one (dynamic programming/belief propagation)



Maximum Likelihood Learning of HMMs: simple once inference is solved

log-likelihood: $$\log p(y_{1:T}|\theta) = \log \int p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

Maximum Likelihood Learning of HMMs: simple once inference is solved

log-likelihood: $$\log p(y_{1:T}|\theta) = \log \int p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

gradient of log-likelihood: $$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

Maximum Likelihood Learning of HMMs: simple once inference is solved

log-likelihood:

$$\log p(y_{1:T}|\theta) = \log \int p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

gradient of log-likelihood:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

show gradient depends on simple moments of posterior:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta)$$

Maximum Likelihood Learning of HMMs: simple once inference is solved

log-likelihood: $$\log p(y_{1:T}|\theta) = \log \int p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

gradient of log-likelihood: $$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

show gradient depends on simple moments of posterior:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} \exp(\log p(y_{1:T}, x_{1:T}|\theta)) dx_{1:T}$$

Maximum Likelihood Learning of HMMs: simple once inference is solved

log-likelihood: $$\log p(y_{1:T}|\theta) = \log \int p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

gradient of log-likelihood: $$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

show gradient depends on simple moments of posterior:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} \exp(\overbrace{\log p(y_{1:T}, x_{1:T}|\theta)}^{E(\theta; x_{1:T}, y_{1:T})}) dx_{1:T}$$

Maximum Likelihood Learning of HMMs: simple once inference is solved

log-likelihood: $$\log p(y_{1:T}|\theta) = \log \int p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

gradient of log-likelihood: $$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

show gradient depends on simple moments of posterior:

$$E(\theta; x_{1:T}, y_{1:T}) = \sum_t \left[\log p(y_t|x_t, \theta) + \log p(x_t|x_{t-1}, \theta)\right]$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} \exp(\overbrace{\log p(y_{1:T}, x_{1:T}|\theta)}^{E(\theta; x_{1:T}, y_{1:T})}) dx_{1:T}$$

Maximum Likelihood Learning of HMMs: simple once inference is solved

log-likelihood: $$\log p(y_{1:T}|\theta) = \log \int p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

gradient of log-likelihood: $$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

show gradient depends on simple moments of posterior:

simple form: e.g. quadratic in x for LGSSMs

$E(\theta; x_{1:T}, y_{1:T}) = \sum_t \left[\log p(y_t|x_t, \theta) + \log p(x_t|x_{t-1}, \theta)\right]$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} \exp(\overbrace{\log p(y_{1:T}, x_{1:T}|\theta)}^{E(\theta; x_{1:T}, y_{1:T})}) dx_{1:T}$$

Maximum Likelihood Learning of HMMs: simple once inference is solved

log-likelihood: $$\log p(y_{1:T}|\theta) = \log \int p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

gradient of log-likelihood: $$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

simple form: e.g. quadratic in x for LGSSMs

$E(\theta; x_{1:T}, y_{1:T}) = \sum_t \left[\log p(y_t|x_t, \theta) + \log p(x_t|x_{t-1}, \theta)\right]$

show gradient depends on simple moments of posterior:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} \exp(\overbrace{\log p(y_{1:T}, x_{1:T}|\theta)}^{E(\theta; x_{1:T}, y_{1:T})}) dx_{1:T}$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int p(y_{1:T}, x_{1:T}|\theta) \frac{\mathrm{d}}{\mathrm{d}\theta} E(\theta; x_{1:T}, y_{1:T}) dx_{1:T}$$

Maximum Likelihood Learning of HMMs: simple once inference is solved

log-likelihood:

$$\log p(y_{1:T}|\theta) = \log \int p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

gradient of log-likelihood:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

simple form: e.g. quadratic in x for LGSSMs

$E(\theta; x_{1:T}, y_{1:T}) = \sum_t \left[\log p(y_t|x_t,\theta) + \log p(x_t|x_{t-1},\theta)\right]$

show gradient depends on simple moments of posterior:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} \exp(\overbrace{\log p(y_{1:T}, x_{1:T}|\theta)}^{E(\theta; x_{1:T}, y_{1:T})}) dx_{1:T}$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int p(y_{1:T}, x_{1:T}|\theta) \frac{\mathrm{d}}{\mathrm{d}\theta} E(\theta; x_{1:T}, y_{1:T}) dx_{1:T}$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \int p(x_{1:T}|y_{1:T},\theta) \frac{\mathrm{d}}{\mathrm{d}\theta} E(\theta; x_{1:T}, y_{1:T}) dx_{1:T}$$

Maximum Likelihood Learning of HMMs: simple once inference is solved

log-likelihood:

$$\log p(y_{1:T}|\theta) = \log \int p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

gradient of log-likelihood:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} p(y_{1:T}, x_{1:T}|\theta) dx_{1:T}$$

simple form: e.g. quadratic in x for LGSSMs

$$E(\theta; x_{1:T}, y_{1:T}) = \sum_t \left[\log p(y_t|x_t, \theta) + \log p(x_t|x_{t-1}, \theta)\right]$$

show gradient depends on simple moments of posterior:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int \frac{\mathrm{d}}{\mathrm{d}\theta} \exp(\overbrace{\log p(y_{1:T}, x_{1:T}|\theta)}^{E(\theta; x_{1:T}, y_{1:T})}) dx_{1:T}$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \frac{1}{p(y_{1:T}|\theta)} \int p(y_{1:T}, x_{1:T}|\theta) \frac{\mathrm{d}}{\mathrm{d}\theta} E(\theta; x_{1:T}, y_{1:T}) dx_{1:T}$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \int p(x_{1:T}|y_{1:T}, \theta) \frac{\mathrm{d}}{\mathrm{d}\theta} E(\theta; x_{1:T}, y_{1:T}) dx_{1:T}$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(y_{1:T}|\theta) = \left\langle \frac{\mathrm{d}}{\mathrm{d}\theta} E(\theta; x_{1:T}, y_{1:T}) \right\rangle_{p(x_{1:T}|y_{1:T}, \theta)}$$

requires posterior moments: marginals and pairwise marginals

Course Survey: please complete this!