PMR Discrete Latent State Dynamical Models

Probabilistic Modelling and Reasoning

Amos Storkey

School of Informatics, University of Edinburgh

Based on Slides of David Barber that accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from `www.cs.ucl.ac.uk/staff/D.Barber/brml`

Outline

Outline

Outline

Observed Linear Dynamical Systems

The OLDS defines the temporal evolution of a vector $\mathbf{v}_t$ according to the discrete-time update equation

$$\mathbf{v}_t = \mathbf{A}_t \mathbf{v}_{t-1}$$

where $\mathbf{A}_t$ is the transition matrix at time t.

Uses

A motivation for studying OLDSs is that many equations that describe the physical world can be written as an OLDS. OLDSs are interesting since they may be used as simple prediction models: if $\mathbf{v}_t$ describes the state of the environment at time t, then $\mathbf{A}\mathbf{v}_t$ predicts the environment at time $t+1$. As such, these models, have widespread application in many branches of science, from engineering and physics to economics.

Eigenanalysis

For the deterministic OLDS if we specify $\mathbf{v}_1$, all future values $\mathbf{v}_2, \mathbf{v}_3, \ldots,$ are defined.

$$\mathbf{v}_t = \mathbf{A}^{t-1}\mathbf{v}_1 = \mathbf{P}\boldsymbol{\Lambda}^{t-1}\mathbf{P}^{-1}\mathbf{v}_1$$

where $\boldsymbol{\Lambda} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_V)$, is the diagonal eigenvalue matrix, and $\mathbf{P}$ is the corresponding eigenvector matrix of $\mathbf{A}$.

Stability criteria

If $\lambda_i > 1$ then for large t, $\mathbf{v}_t$ will explode. On the other hand, if $\lambda_i < 1$, then λ_i^{t-1} will tend to zero. For stable systems we require therefore no eigenvalues of magnitude greater than 1 and only unit eigenvalues will contribute in long term.

Observed Linear Dynamical System

More generally, we consider a system with additive Gaussian noise:

$$\mathbf{v}_t = \mathbf{A}_t \mathbf{v}_{t-1} + \boldsymbol{\eta}_t$$

where $\boldsymbol{\eta}_t$ is a noise vector sampled from a Gaussian distribution,

$$\mathcal{N}\left(\boldsymbol{\eta}_t \middle| \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\right)$$

This is equivalent to a first order Markov model

$$p(\mathbf{v}_t|\mathbf{v}_{t-1}) = \mathcal{N}\left(\mathbf{v}_t \middle| \mathbf{A}_t \mathbf{v}_{t-1} + \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\right)$$

At $t = 1$ we have an initial distribution $p(\mathbf{v}_1) = \mathcal{N}\left(\mathbf{v}_1 \middle| \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$. For $t > 1$ if the parameters are time-independent, $\boldsymbol{\mu}_t \equiv \boldsymbol{\mu}$, $\mathbf{A}_t \equiv \mathbf{A}$, $\boldsymbol{\Sigma}_t \equiv \boldsymbol{\Sigma}$, the process is called time-invariant.

Stationary distribution

Consider the one-dimensional linear system with independent additive noise

$$v_t = av_{t-1} + \eta_t, \qquad \eta_t \sim \mathcal{N}\left(\eta_t \middle| 0, \sigma_v^2\right)$$

Assuming $v_{t-1} \sim \mathcal{N}\left(v_{t-1} \middle| \mu_{t-1}, \sigma_{t-1}^2\right)$, then using $\langle \eta_t \rangle = 0$ we have

$$\begin{aligned}
&\langle v_t \rangle = a \langle v_{t-1} \rangle + \langle \eta_t \rangle \Rightarrow \mu_t = a\mu_{t-1} \\
&\left\langle v_t^2 \right\rangle = \langle av_{t-1} + \eta_t \rangle^2 = a^2 \left\langle v_{t-1}^2 \right\rangle + 2a \langle v_{t-1} \rangle \langle \eta_t \rangle + \left\langle \eta_t^2 \right\rangle \\
\Rightarrow\ &\sigma_t^2 = a^2 \sigma_{t-1}^2 + \sigma_v^2
\end{aligned}$$

so that $v_t \sim \mathcal{N}\left(v_t \middle| a\mu_{t-1}, a^2\sigma_{t-1}^2 + \sigma_v^2\right)$. The stationary distribution satisfies

$$\sigma_\infty^2 = a^2 \sigma_\infty^2 + \sigma_v^2 \quad \Rightarrow \sigma_\infty^2 = \frac{\sigma_v^2}{1 - a^2}, \quad \mu_\infty = a^\infty \mu_1$$

If $a \geq 1$ the variance increases indefinitely. For $a < 1$, the noise remains steady in the long run.

Auto-Regressive Models

A scalar time-invariant auto-regressive model is defined by

$$v_t = \sum_{l=1}^{L} a_l v_{t-l} + \eta_t, \qquad \eta_t \sim \mathcal{N}\left(\eta_t \middle| \mu, \sigma^2\right)$$

where $\mathbf{a} = (a_1, \ldots, a_L)^{\mathsf{T}}$ are called the AR coefficients and σ^2 is called the innovation noise. The model predicts the future based on a linear combination of the previous L observations. This is an L^{th} order Markov model:

$$p(v_{1:T}) = \prod_{t=1}^{T} p(v_t | v_{t-1}, \ldots, v_{t-L}), \qquad \text{with } v_i = \emptyset \text{ for } i \leq 0 \text{ and with}$$

$$p(v_t | v_{t-1}, \ldots, v_{t-L}) = \mathcal{N}\left(v_t \middle| \sum_{l=1}^{L} a_l v_{t-l}, \sigma^2\right)$$

Introducing the vector of the L previous observations $\hat{\mathbf{v}}_{t-1} \equiv [v_{t-1}, v_{t-2}, \ldots, v_{t-L}]^{\mathsf{T}}$ we can write more compactly

$$p(v_t | v_{t-1}, \ldots, v_{t-L}) = \mathcal{N}\left(v_t \middle| \mathbf{a}^{\mathsf{T}} \hat{\mathbf{v}}_{t-1}, \sigma^2\right)$$

Fitting a trend

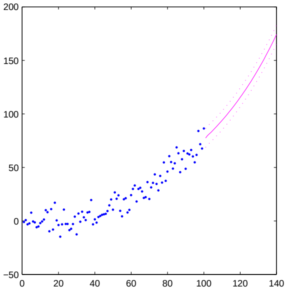


Figure : Fitting an order 3 AR model to the training points. The x axis represents time, and the y axis the value of the timeseries. The solid line is the mean prediction and the dashed lines $\pm$ one standard deviation.

AR models are heavily used in financial time-series prediction, being able to capture simple trends in the data. Another application area is speech processing whereby for a one-dimensional speech signal partitioned into windows of length T, the AR coefficients best able to describe the signal in each window are found. These AR coefficients then form a compressed representation of the signal and are subsequently transmitted for each window. Such a representation is used for example in telephones and known as a linear predictive vocoder.

Training an AR model

Maximum Likelihood training of the AR coefficients is straightforward based on

$$\log p(v_{1:T}) = \sum_{t=1}^{T} \log p(v_t|\hat{\mathbf{v}}_{t-1}) = -\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(v_t - \hat{\mathbf{v}}_{t-1}^{\mathsf{T}}\mathbf{a}\right)^2 - \frac{T}{2}\log(2\pi\sigma^2)$$

Differentiating *w.r.t.* $\mathbf{a}$ and equating to zero we arrive at

$$\sum_t \left(v_t - \hat{\mathbf{v}}_{t-1}^{\mathsf{T}}\mathbf{a}\right)\hat{\mathbf{v}}_{t-1} = 0$$

so that optimally

$$\mathbf{a} = \left(\sum_t \hat{\mathbf{v}}_{t-1}\hat{\mathbf{v}}_{t-1}^{\mathsf{T}}\right)^{-1}\sum_t v_t\hat{\mathbf{v}}_{t-1}$$

These equations can be solved by Gaussian elimination. Similarly, optimally,

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}\left(v_t - \hat{\mathbf{v}}_{t-1}^{\mathsf{T}}\mathbf{a}\right)^2$$

Above we assume that ‘negative’ timesteps are available in order to keep the notation simple.

AR model as an OLDS

We can write an OLDS using

$$\begin{pmatrix} v_t \\ v_{t-1} \\ \vdots \\ v_{t-L+1} \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & \dots & a_L \\ 1 & 0 & \dots & 0 \\ \vdots & 1 & \dots & 0 \\ 0 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} v_{t-1} \\ v_{t-2} \\ \vdots \\ v_{t-L} \end{pmatrix} + \begin{pmatrix} \eta_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

which can be written as

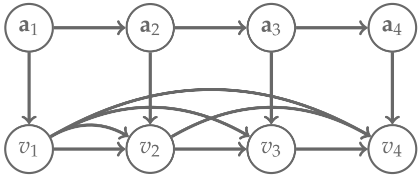
$$\hat{\mathbf{v}}_t = \mathbf{A}\hat{\mathbf{v}}_{t-1} + \eta_t, \qquad \eta_t \sim \mathcal{N}\left(\eta_t | \mathbf{0}, \mathbf{\Sigma}\right)$$

where we define the block matrices

$$\mathbf{A} = \left(\begin{array}{c|c} a_{1:L-1} & a_L \\ \hline \mathbf{I} & \mathbf{0} \end{array}\right), \qquad \mathbf{\Sigma} = \left(\begin{array}{c|c} \sigma^2 & 0_{1,1:L-1} \\ \hline 0_{1:L-1,1} & 0_{1:L-1,1:L-1} \end{array}\right)$$

In this representation, the first component of the vector is updated according to the standard AR model, with the remaining components being copies of the previous values.

Time-varying AR model



If $\mathbf{a}_t$ are the latent AR coefficients, the term

$$v_t = \hat{\mathbf{v}}_{t-1}^{\mathsf{T}} \mathbf{a}_t + \eta_t, \qquad \eta_t \sim \mathcal{N}\left(\eta_t \middle| 0, \sigma^2\right)$$

can be viewed as the emission distribution of a latent LDS in which the hidden variable is $\mathbf{a}_t$ and the time-dependent emission matrix is given by $\hat{\mathbf{v}}_{t-1}^{\mathsf{T}}$. By placing a simple latent transition

$$\mathbf{a}_t = \mathbf{a}_{t-1} + \boldsymbol{\eta}_t^a, \qquad \boldsymbol{\eta}_t^a \sim \mathcal{N}\left(\boldsymbol{\eta}_t^a \middle| 0, \sigma_a^2 \mathbf{I}\right)$$

we encourage the AR coefficients to change slowly with time. This defines a model

$$p(v_{1:T}, \mathbf{a}_{1:T}) = \prod_t p(v_t | \mathbf{a}_t, \hat{\mathbf{v}}_{t-1}) p(\mathbf{a}_t | \mathbf{a}_{t-1})$$

Standard smoothing algorithms can then be applied to yield the time-varying AR coefficients.

Discrete Fourier Transform

For a sequence $x_{0:N-1}$ the DFT $f_{0:N-1}$ is defined as

$$f_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn}, \qquad k = 0, \dots, N-1$$

f_k is a (complex) representation as to how much frequency k is present in the sequence $x_{0:N-1}$. The power of component k is defined as the absolute length of the complex f_k.

Spectrogram

Given a timeseries $x_{1:T}$ the spectrogram at time t is a representation of the frequencies present in a window localised around t. For each window one computes the Discrete Fourier Transform, from which we obtain a vector of log power in each frequency. The window is then moved (usually) one step forward and the DFT recomputed. Note that by taking the logarithm, small values in the original signal can translate to visibly appreciable values in the spectrogram.

Nightingale

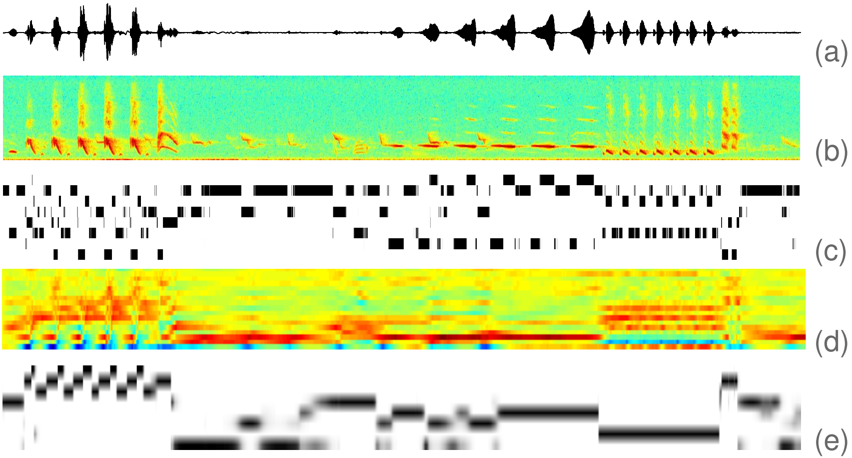


Figure : **(a)**: The raw recording of 5 seconds of a nightingale song (with additional background birdsong). **(b)**: Spectrogram of (a) up to 20,000 Hz. **(c)**: Clustering of the results in panel (b) using an 8 component Gaussian mixture model. The index (from 1 to 8) of the component most probably responsible for the observation is indicated vertically in black. **(d)**: The 20 AR coefficients learned using $\sigma_v^2 = 0.001$, $\sigma_h^2 = 0.001$. **(e)**: Clustering the results in panel (d) using a Gaussian mixture model with 8 components. The AR components group roughly according to the different song regimes.

Latent Linear Dynamical Systems

The Latent LDS defines a stochastic linear dynamical system in a latent (or 'hidden') space on a sequence of vectors $\mathbf{h}_{1:T}$:

$$
\begin{array}{lll}
\mathbf{h}_t = \mathbf{A}_t \mathbf{h}_{t-1} + \eta_t^h & \eta_t^h \sim \mathcal{N}\left(\eta_t^h \middle| \bar{\mathbf{h}}_t, \mathbf{\Sigma}_t^H\right) & \text{transition model} \\
\mathbf{v}_t = \mathbf{B}_t \mathbf{h}_t + \eta_t^v & \eta_t^v \sim \mathcal{N}\left(\eta_t^v \middle| \bar{\mathbf{v}}_t, \mathbf{\Sigma}_t^V\right) & \text{emission model}
\end{array}
$$

where η_t^h and η_t^v are noise vectors. $\mathbf{A}_t$ is called the transition matrix and $\mathbf{B}_t$ the emission matrix. The terms $\bar{\mathbf{h}}_t$ and $\bar{\mathbf{v}}_t$ are the hidden and output bias respectively.

Kalman Filter

Another term for the (latent) LDS is Kalman Filter, particularly in the engineering literature. A 'filter' is typically some operation on a signal. We prefer here to focus on the model viewpoint from which various operations of inference will be applied.

Probabilistic model description

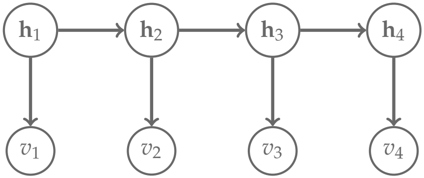


Figure : A (latent) LDS. Both hidden and visible variables are Gaussian distributed.

The transition and emission models define a first order Markov model

$$p(\mathbf{h}_{1:T}, \mathbf{v}_{1:T}) = p(\mathbf{h}_1)p(\mathbf{v}_1|\mathbf{h}_1)\prod_{t=2}^{T} p(\mathbf{h}_t|\mathbf{h}_{t-1})p(\mathbf{v}_t|\mathbf{h}_t)$$

with the transitions and emissions given by Gaussian distributions

$$p(\mathbf{h}_t|\mathbf{h}_{t-1}) = \mathcal{N}\left(\mathbf{h}_t \middle| \mathbf{A}_t\mathbf{h}_{t-1} + \bar{\mathbf{h}}_t, \boldsymbol{\Sigma}_t^H\right), \qquad p(\mathbf{h}_1) = \mathcal{N}\left(\mathbf{h}_1 \middle| \boldsymbol{\mu}_\pi, \boldsymbol{\Sigma}_\pi\right)$$
$$p(\mathbf{v}_t|\mathbf{h}_t) = \mathcal{N}\left(\mathbf{v}_t \middle| \mathbf{B}_t\mathbf{h}_t + \bar{\mathbf{v}}_t, \boldsymbol{\Sigma}_t^V\right)$$

One may also include an external input $\mathbf{o}_t$ at each time, which will add $\mathbf{C}\mathbf{o}_t$ to the mean of the hidden variable and $\mathbf{D}\mathbf{o}_t$ to the mean of the observation.

The emission and transition distributions

Explicit expressions for the transition and emission distributions are given below for the time-invariant case with $\bar{\mathbf{v}}_t = \mathbf{0}, \bar{\mathbf{h}}_t = \mathbf{0}$. Each hidden variable is a multidimensional Gaussian distributed vector $\mathbf{h}_t$, with

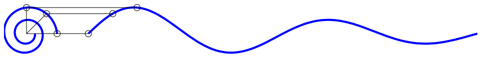
$$p(\mathbf{h}_t|\mathbf{h}_{t-1}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_H|}} \exp\left(-\frac{1}{2}\left(\mathbf{h}_t - \mathbf{A}\mathbf{h}_{t-1}\right)^\mathsf{T} \boldsymbol{\Sigma}_H^{-1} \left(\mathbf{h}_t - \mathbf{A}\mathbf{h}_{t-1}\right)\right)$$

which states that $\mathbf{h}_{t+1}$ has a mean equal to $\mathbf{A}\mathbf{h}_t$ with Gaussian fluctuations described by the covariance matrix $\boldsymbol{\Sigma}_H$. Similarly,

$$p(\mathbf{v}_t|\mathbf{h}_t) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_V|}} \exp\left(-\frac{1}{2}\left(\mathbf{v}_t - \mathbf{B}\mathbf{h}_t\right)^\mathsf{T} \boldsymbol{\Sigma}_V^{-1} \left(\mathbf{v}_t - \mathbf{B}\mathbf{h}_t\right)\right)$$

describes an output $\mathbf{v}_t$ with mean $\mathbf{B}\mathbf{h}_t$ and covariance $\boldsymbol{\Sigma}_V$.

Example: Phasor



Consider a dynamical system defined on two dimensional vectors $\mathbf{h}_t$:

$$\mathbf{h}_{t+1} = \gamma \mathbf{R}_\theta \mathbf{h}_t, \quad 0 < \gamma < 1 \qquad \text{with} \quad \mathbf{R}_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

$\mathbf{R}_\theta$ rotates the vector $\mathbf{h}_t$ through angle θ in one timestep. Under this LDS $\mathbf{h}$ will trace out points on a circle through time. By taking a scalar projection of $\mathbf{h}_t$, for example,

$$v_t = [\mathbf{h}_t]_1 = [1 \quad 0]^\mathsf{T} \mathbf{h}_t,$$

the elements v_t, $t = 1, \ldots, T$ describe a sinusoid through time. By using a block diagonal $\mathbf{R} = \text{blkdiag}\left(\mathbf{R}_{\theta_1}, \ldots, \mathbf{R}_{\theta_m}\right)$ and taking a scalar projection of the extended $m \times 2$ dimensional $\mathbf{h}$ vector, one can construct a representation of a signal in terms of m sinusoidal components.

Inference

Given an observation sequence $\mathbf{v}_{1:T}$ we wish to consider filtering and smoothing, as we did for the HMM. Since the LDS has the same independence structure as the HMM, we can use the same independence assumptions in deriving the updates for the LDS.

Dealing with continuous messages

The filtering recursion becomes

$$p(\mathbf{h}_t|\mathbf{v}_{1:t}) \propto \int_{\mathbf{h}_{t-1}} p(\mathbf{v}_t|\mathbf{h}_t)p(\mathbf{h}_t|\mathbf{h}_{t-1})p(\mathbf{h}_{t-1}|\mathbf{v}_{1:t-1}), \qquad t > 1$$

Since the product of two Gaussians is another Gaussian, and the integral of a Gaussian is another Gaussian, the resulting $p(\mathbf{h}_t|\mathbf{v}_{1:t})$ is also Gaussian. This closure property of Gaussians means that we may represent $p(\mathbf{h}_{t-1}|\mathbf{v}_{1:t-1}) = \mathcal{N}(\mathbf{h}_{t-1}|\mathbf{f}_{t-1}, \mathbf{F}_{t-1})$ with mean $\mathbf{f}_{t-1}$ and covariance $\mathbf{F}_{t-1}$. The effect of a message update is equivalent to updating the mean $\mathbf{f}_{t-1}$ and covariance $\mathbf{F}_{t-1}$ into a mean $\mathbf{f}_t$ and covariance $\mathbf{F}_t$ for $p(\mathbf{h}_t|\mathbf{v}_{1:t})$.

Filtering

We represent the filtered distribution as a Gaussian with mean $\mathbf{f}_t$ and covariance $\mathbf{F}_t$,

$$p(\mathbf{h}_t|\mathbf{v}_{1:t}) \sim \mathcal{N}(\mathbf{h}_t|\mathbf{f}_t, \mathbf{F}_t)$$

Our task is then to find a recursion for $\mathbf{f}_t, \mathbf{F}_t$ in terms of $\mathbf{f}_{t-1}$, $\mathbf{F}_{t-1}$.

The big picture

A convenient approach is to first find the joint distribution $p(\mathbf{h}_t, \mathbf{v}_t|\mathbf{v}_{1:t-1})$ and then condition on $\mathbf{v}_t$ to find the distribution $p(\mathbf{h}_t|\mathbf{v}_{1:t})$.

Filtering

$p(\mathbf{h}_t, \mathbf{v}_t | \mathbf{v}_{1:t-1})$ is a Gaussian whose statistics can be found from

$$\mathbf{v}_t = \mathbf{B}\mathbf{h}_t + \eta_t^v, \qquad \mathbf{h}_t = \mathbf{A}\mathbf{h}_{t-1} + \eta_t^h$$

Using the above, and assuming time-invariance and zero biases, we readily find

$$\left\langle \Delta\mathbf{h}_t \Delta\mathbf{h}_t^\top | \mathbf{v}_{1:t-1} \right\rangle = \mathbf{A} \left\langle \Delta\mathbf{h}_{t-1} \Delta\mathbf{h}_{t-1}^\top | \mathbf{v}_{1:t-1} \right\rangle \mathbf{A}^\top + \mathbf{\Sigma}_H$$

$$\left\langle \Delta\mathbf{v}_t \Delta\mathbf{h}_t^\top | \mathbf{v}_{1:t-1} \right\rangle = \mathbf{B} \left\langle \Delta\mathbf{h}_t \Delta\mathbf{h}_t^\top | \mathbf{v}_{1:t-1} \right\rangle$$

$$\left\langle \Delta\mathbf{v}_t \Delta\mathbf{v}_t^\top | \mathbf{v}_{1:t-1} \right\rangle = \mathbf{B} \left\langle \Delta\mathbf{h}_t \Delta\mathbf{h}_t^\top | \mathbf{v}_{1:t-1} \right\rangle \mathbf{B}^\top + \mathbf{\Sigma}_V$$

$$\langle \mathbf{v}_t | \mathbf{v}_{1:t-1} \rangle = \mathbf{B}\mathbf{A} \langle \mathbf{h}_{t-1} | \mathbf{v}_{1:t-1} \rangle, \qquad \langle \mathbf{h}_t | \mathbf{v}_{1:t-1} \rangle = \mathbf{A} \langle \mathbf{h}_{t-1} | \mathbf{v}_{1:t-1} \rangle$$

In the above, using our moment representation of the forward messages

$$\langle \mathbf{h}_{t-1} | \mathbf{v}_{1:t-1} \rangle \equiv \mathbf{f}_{t-1}, \qquad \left\langle \Delta\mathbf{h}_{t-1} \Delta\mathbf{h}_{t-1}^\top | \mathbf{v}_{1:t-1} \right\rangle \equiv \mathbf{F}_{t-1}$$

Filtering update

Then, using conditioning

$$\mathbf{f}_t = \mathbf{A}\mathbf{f}_{t-1} + \mathbf{P}\mathbf{B}^\top \left(\mathbf{B}\mathbf{P}\mathbf{B}^\top + \Sigma_V\right)^{-1} (\mathbf{v}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1})$$

and covariance

$$\mathbf{F}_t = \mathbf{P} + \Sigma_H - \mathbf{P}\mathbf{B}^\top \left(\mathbf{B}\mathbf{P}\mathbf{B}^\top + \Sigma_V\right)^{-1} \mathbf{B}\mathbf{P}$$

where

$$\mathbf{P} \equiv \mathbf{A}\mathbf{F}_{t-1}\mathbf{A}^\top + \Sigma_H$$

One can write the covariance update as

$$\mathbf{F}_t = (\mathbf{I} - \mathbf{K}\mathbf{B})\,\mathbf{P}$$

where we define the Kalman gain matrix

$$\mathbf{K} = \mathbf{P}\mathbf{B}^\top \left(\Sigma_V + \mathbf{B}\mathbf{P}\mathbf{B}^\top\right)^{-1}$$

Smoothing

By representing the posterior as a Gaussian with mean $\mathbf{g}_t$ and covariance $\mathbf{G}_t$,

$$p(\mathbf{h}_t|\mathbf{v}_{1:T}) \sim \mathcal{N}(\mathbf{h}_t|\mathbf{g}_t, \mathbf{G}_t)$$

we can form a recursion for $\mathbf{g}_t$ and $\mathbf{G}_t$ as follows:

$$\begin{aligned} p(\mathbf{h}_t|\mathbf{v}_{1:T}) &= \int_{\mathbf{h}_{t+1}} p(\mathbf{h}_t, \mathbf{h}_{t+1}|\mathbf{v}_{1:T}) \\ &= \int_{\mathbf{h}_{t+1}} p(\mathbf{h}_t|\mathbf{v}_{1:T}, \mathbf{h}_{t+1}) p(\mathbf{h}_{t+1}|\mathbf{v}_{1:T}) \\ &= \int_{\mathbf{h}_{t+1}} p(\mathbf{h}_t|\mathbf{v}_{1:t}, \mathbf{h}_{t+1}) p(\mathbf{h}_{t+1}|\mathbf{v}_{1:T}) \end{aligned}$$

The term $p(\mathbf{h}_t|\mathbf{v}_{1:t}, \mathbf{h}_{t+1})$ can be found by conditioning the joint distribution

$$p(\mathbf{h}_t, \mathbf{h}_{t+1}|\mathbf{v}_{1:t}) = p(\mathbf{h}_{t+1}|\mathbf{h}_t, \cancel{\mathbf{v}_{1:t}}) p(\mathbf{h}_t|\mathbf{v}_{1:t})$$

This procedure is the Rauch-Tung-Striebel Kalman smoother. This is called a 'correction' method since it takes the filtered estimate $p(\mathbf{h}_t|\mathbf{v}_{1:t})$ and 'corrects' it to form a smoothed estimate $p(\mathbf{h}_t|\mathbf{v}_{1:T})$.

Newtonian Trajectory Analysis

A toy rocket with unknown mass and initial velocity is launched in the air. In addition, the constant accelerations from the rocket's propulsion system are unknown. Based on noisy measurements of $x(t)$ and $y(t)$, our task is to infer the position of the rocket at each time. Although this is perhaps most appropriately considered from the using continuous time dynamics, we will translate this into a discrete time approximation.

Newton's laws

$$\frac{d^2}{dt^2}x = \frac{f_x(t)}{m}, \qquad \frac{d^2}{dt^2}y = \frac{f_y(t)}{m}$$

where m is the mass of the object and $f_x(t), f_y(t)$ are the horizontal and vertical forces respectively.

Discretising time

A naive approach is to reparameterise time to use the variable $\tilde{t}$ such that $t \equiv \tilde{t}\Delta$, where $\tilde{t}$ is integer and Δ is a unit of time. The dynamics is then

$$x((\tilde{t}+1)\Delta) = x(\tilde{t}\Delta) + \Delta x'(\tilde{t}\Delta)$$

$$y((\tilde{t}+1)\Delta) = y(\tilde{t}\Delta) + \Delta y'(\tilde{t}\Delta)$$

where $y'(t) \equiv \frac{dy}{dt}$. We can write an update equation for the x' and y' as

$$x'((\tilde{t}+1)\Delta) = x'(\tilde{t}\Delta) + f_x\Delta/m, \qquad y'((\tilde{t}+1)\Delta) = y'(\tilde{t}\Delta) + f_y\Delta/m$$

The instrument which measures $x(t)$ and $y(t)$ is not completely accurate. For simplicity, we relabel $a_x(t) = f_x(t)/m(t)$, $a_y(t) = f_y(t)/m(t)$ – these accelerations will be assumed to be roughly constant, but unknown :

$$a_x((\tilde{t}+1)\Delta) = a_x(\tilde{t}\Delta) + \eta_x, \quad a_y((\tilde{t}+1)\Delta) = a_y(\tilde{t}\Delta) + \eta_y,$$

where η_x and η_y are small noise terms.

Description as an LDS

We describe the above model by considering $x'(t)$, $x(t)$, $y'(t)$, $y(t)$, $a_x(t)$, $a_y(t)$ as hidden variables, giving rise to a $H = 6$ dimensional LDS with transition and emission matrices as below:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & \Delta & 0 \\ \Delta & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta \\ 0 & 0 & \Delta & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

We place a large variance on their initial values, and attempt to infer the unknown trajectory.

Example

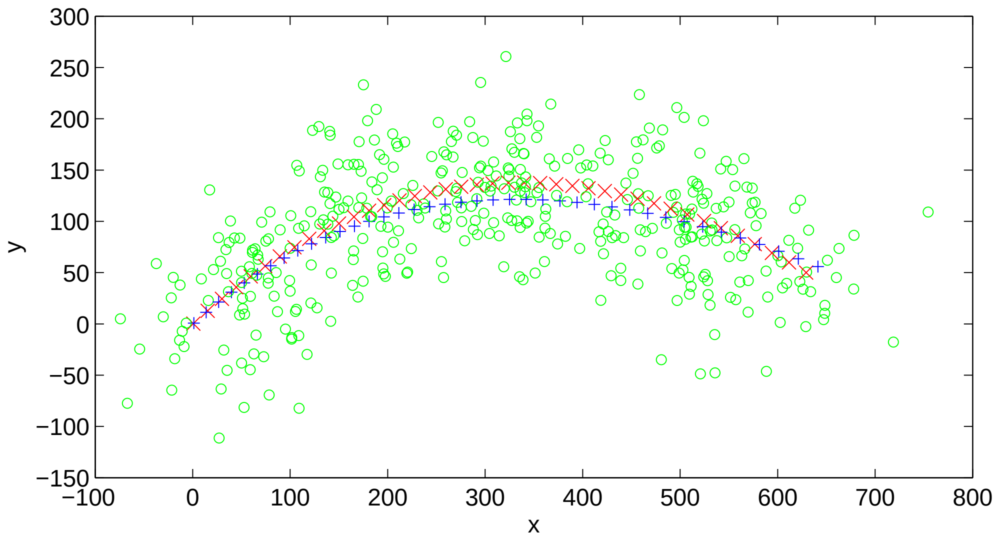


Figure : Estimate of the trajectory of a Newtonian ballistic object based on noisy observations (small circles). All time labels are known but omitted in the plot. The 'x' points are the true positions of the object, and the crosses '+' are the estimated smoothed mean positions $\langle x_t, y_t | \mathbf{v}_{1:T} \rangle$ of the object plotted every several time steps.