Linear State-Space Models

September 3, 2024

The material in these notes draws from several sources, including Chapter 6 of S&S, Chapter 11 of Chan (2010), and the online textbook Applied Time Series Analysis which was written by the authors of the `MARSS` software for `R`.

Introduction to a Simple State-Space Model

State space models give us yet another way of writing out a model for our time series. In this section, the notation we use will differ slightly. The observed time series that we intend to analyze has values $y_1, \ldots, y_n$. Let's work through a small example! For a univariate time series of length n we assume

$$\begin{aligned} y_t &= a x_t + v_t && \text{Observation Equation} \\ x_t &= \phi x_{t-1} + w_t && \text{State Equation,} \end{aligned}$$

where $v_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_v^2)$, $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$ and $x_1 = \mu$. The state equation looks like an $\mathbf{AR}(1)$ model, but values of the $\mathbf{AR}(1)$ process x_t are not observed, and it differs from an $\mathbf{AR}(1)$ model insofar as the initial state is assumed to be fixed. Rather, a linear transformation y_t of the states x_t are observed with some additional normal noise. To emphasize, we observe values of y_t, but we do *not* observe values of x_t.

Predicting, Filtering, and Smoothing the States and Forecasting the Time Series

For now, let's assume that we know σ_v^2, σ_w^2, a, ϕ, and μ. Having assumed a state-space model, we might be interested in:

- **Predicting**, i.e. estimating future values of x_t given past values of $y_{t-1}, \ldots, y_1$
  - For example, computing $\mathbb{E}[x_2|y_1]$ and $\mathbb{V}[x_2|y_1]$
- **Filtering**, i.e. estimating future values of x_t given current and past values $y_t, \ldots, y_1$
  - For example, computing $\mathbb{E}[x_2|y_1, y_2]$ and $\mathbb{V}[x_2|y_1, y_2]$
- **Smoothing**, i.e. estimating a past value of x_{t-1} given all observed past, current, and future values $y_n, \ldots, y_1$
  - For example, computing $\mathbb{E}[x_1|y_1, y_2]$ and $\mathbb{V}[x_1|y_1, y_2]$

As usual, we might also be interested in forecasting future values y_{n+k} given $y_1, \ldots, y_n$, which will require **prediction** of the states because we will need to compute $\mathbb{E}[x_{n+k}|y_1, \ldots, y_n]$. Even if we do know the values of these parameters, how do we go about predicting, filtering, and smoothing?

The first step is to realize that we can write out the joint probability distribution of $\boldsymbol{y}$ and $\boldsymbol{x}$. Because we have assumed that the errors $\boldsymbol{v}$ and $\boldsymbol{w}$ and the initial value x_1 is fixed, we know that the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{x}$ is given by:

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbb{E}[\boldsymbol{y}] \\ \mathbb{E}[\boldsymbol{x}] \end{pmatrix}, \begin{pmatrix} \mathbb{V}[\boldsymbol{y}] & \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{x}] \\ \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{x}] & \mathbb{V}[\boldsymbol{x}] \end{pmatrix} \right). \tag{1}$$

Why is this useful? Well, predicting, filtering, smoothing, and forecasting all correspond to different conditional mean and variance estimation problems, and when we have a normal

joint distribution and know its parameters, it's easy to figure out any conditional distribution we want! We can see this most easily in the case of smoothing, which corresponds to the problem of estimating the conditional means $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}]$ and variances $\mathbb{V}[\boldsymbol{x}|\boldsymbol{y}]$. Before we go further, a very useful property of the multivariate normal distribution is that:

If $\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix}, \begin{pmatrix} \boldsymbol{C} & \boldsymbol{D} \\ \boldsymbol{D}' & \boldsymbol{E} \end{pmatrix}\right)$, then

- The marginal distributions of $\boldsymbol{u}$ and $\boldsymbol{v}$ are given by

$$\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{a}, \boldsymbol{C}) \text{ and } \boldsymbol{v} \sim \mathcal{N}(\boldsymbol{b}, \boldsymbol{E})$$

- The conditional distributions of $\boldsymbol{u}|\boldsymbol{v}$ and $\boldsymbol{v}|\boldsymbol{u}$

$$\boldsymbol{u}|\boldsymbol{v} \sim \mathcal{N}\left(\boldsymbol{a} + \boldsymbol{D}'\boldsymbol{E}^{-1}(\boldsymbol{v} - \boldsymbol{b}), \boldsymbol{C} - \boldsymbol{D}'\boldsymbol{E}^{-1}\boldsymbol{D}\right)$$
$$\boldsymbol{v}|\boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{b} + \boldsymbol{D}\boldsymbol{C}^{-1}(\boldsymbol{u} - \boldsymbol{a}), \boldsymbol{E} - \boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{D}'\right)$$

We can use these facts to find the conditional distributions of the states $\boldsymbol{x}$ given the data $\boldsymbol{y}$ and accordingly, the smoothed estimates of the states, but first we'll simplify the joint distribution a bit. We can simplify $\mathbb{E}[\boldsymbol{y}]$, $\mathbb{V}[\boldsymbol{y}]$, and $\text{Cov}[\boldsymbol{x}, \boldsymbol{y}]$.

- It's pretty straightforward to see that $\mathbb{E}[\boldsymbol{y}] = a\mathbb{E}[\boldsymbol{x}]$
- A bit of algebra lets us rewrite $\mathbb{V}[\boldsymbol{y}]$

$$\begin{aligned}
\mathbb{V}[\boldsymbol{y}] &= \mathbb{E}\left[(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])'\right] \\
&= \mathbb{E}\left[(a(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]) + \boldsymbol{v})\left((a(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]) + \boldsymbol{v})'\right]\right] \\
&= a^2\mathbb{V}[\boldsymbol{x}] + \sigma_v^2\boldsymbol{I}_n.
\end{aligned}$$

- A bit more algebra lets us rewrite $\text{Cov}[\boldsymbol{x}, \boldsymbol{y}]$

$$\begin{aligned}\text{Cov}[\boldsymbol{x}, \boldsymbol{y}] &= \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])'\right] \\ &= \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(a\boldsymbol{x} - a\mathbb{E}[\boldsymbol{x}] + \boldsymbol{w})'\right] \\ &= a\mathbb{V}[\boldsymbol{x}]\end{aligned}$$

Plugging these expressions into 2 yields a nicely structured normal distribution,

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} a\mathbb{E}[\boldsymbol{x}] \\ \mathbb{E}[\boldsymbol{x}] \end{pmatrix}, \begin{pmatrix} a^2\mathbb{V}[\boldsymbol{x}] + \sigma_v^2 \boldsymbol{I}_n & a\mathbb{V}[\boldsymbol{x}] \\ a\mathbb{V}[\boldsymbol{x}] & \mathbb{V}[\boldsymbol{x}] \end{pmatrix}\right). \tag{2}$$

Using our normal distribution facts and (2), we can return to the smoothing problem

$$\begin{aligned}\boldsymbol{x}|\boldsymbol{y} \sim \mathcal{N}\big(&\mathbb{E}[\boldsymbol{x}] + a\mathbb{V}[\boldsymbol{x}]\left(a^2\mathbb{V}[\boldsymbol{x}] + \sigma_v^2\boldsymbol{I}_n\right)^{-1}(\boldsymbol{y} - a\mathbb{E}[\boldsymbol{x}]), \\ &\mathbb{V}[\boldsymbol{x}] - a^2\mathbb{V}[\boldsymbol{x}]\left(a^2\mathbb{V}[\boldsymbol{x}] + \sigma_v^2\boldsymbol{I}_n\right)^{-1}\mathbb{V}[\boldsymbol{x}]\Big).\end{aligned}$$

The smoothed values of the states are given by the conditional mean

$$\mathbb{E}[\boldsymbol{x}] + a\mathbb{V}[\boldsymbol{x}]\left(a^2\mathbb{V}[\boldsymbol{x}] + \sigma_v^2\boldsymbol{I}_n\right)^{-1}(\boldsymbol{y} - a\mathbb{E}[\boldsymbol{x}]),$$

and our uncertainty about them is quantified by the conditional variance,

$$\mathbb{V}[\boldsymbol{x}] - a^2\mathbb{V}[\boldsymbol{x}]\left(a^2\mathbb{V}[\boldsymbol{x}] + \sigma_v^2\boldsymbol{I}_n\right)^{-1}\mathbb{V}[\boldsymbol{x}].$$

Predicting and filtering can be done similarly by identifying the marginal joint distribution of a state and the corresponding observed values, and computing the corresponding conditional distribution.

- We can obtain predicted means and variances of the states x_t given the past observed values of the time series $y_1, \ldots, y_{t-1}$ by letting $u = x_t$ and $\boldsymbol{v} = (y_1, \ldots, y_{t-1})$, using our marginal and conditional normal distribution facts;

- We can obtain filtered means and variances of the states x_t given the past and current

observed values of the time series $y_1, \ldots, y_t$ using our marginal and conditional normal distribution facts, letting $u = x_t$ and $\boldsymbol{v} = (y_1, \ldots, y_t)$.

Forecasts follow straightforwardly. We have assumed $y_t = ax_t + v_t$, so it is natural to define a forecasted value of y_{n+k} as

$$\mathbb{E}\left[y_{n+k}|y_1, \ldots, y_n\right] = a\mathbb{E}\left[x_{n+k}|y_1, \ldots, y_n\right],$$

with variance

$$\mathbb{V}\left[y_{n+k}|y_1, \ldots, y_n\right] = a^2\mathbb{V}\left[x_{n+k}|y_1, \ldots, y_n\right] + \sigma_w^2.$$

In general, even though we *can* figure out any conditional distribution we want, it would be very computationally burdensome to figure out each conditional expectation one by one. Fortunately, the **Kalman filter** allows for fast, recursive estimation of:

- Predicted means and variances
- Filtered means and variances
- Smoothed means and variances
- Forecast means and variances.

The existence of a fast algorithm to compute all of these quantities is essential in practice.

- **Kalman Prediction & Filtering:** The Kalman prediction and filtering algorithms work together, iteratively computing the predictions, prediction variances, filtered values of x_t, and variances of the filtered values, starting at time $t = 1$. To aid interpretation of the algorithm, predictions and their variances are printed in blue, while filtered values and their variances are printed in red.
  - $\mathbb{E}\left[x_t|y_1, \ldots, y_{t-1}\right] = \phi\mathbb{E}\left[x_{t-1}|y_1, \ldots, y_{t-2}\right] + \left(\frac{a\phi\mathbb{V}[x_{t-1}|y_1,\ldots,y_{t-2}]}{a^2\mathbb{V}[x_{t-1}|y_1,\ldots,y_{t-2}]+\sigma_v^2}\right)\left(y_{t-1} - a\mathbb{E}\left[x_{t-1}|y_1, \ldots, y_{t-2}\right]\right)$
  - $\mathbb{V}\left[x_t|y_1, \ldots, y_{t-1}\right] = \sigma_w^2 + \phi^2\mathbb{V}\left[x_{t-1}|y_1, \ldots, y_{t-2}\right]\left(1 - \frac{a^2\mathbb{V}[x_{t-1}|y_1,\ldots,y_{t-2}]}{a^2\mathbb{V}[x_{t-1}|y_1,\ldots,y_{t-2}]+\sigma_v^2}\right)$
  - $\mathbb{E}\left[x_t|y_1, \ldots, y_t\right] = \mathbb{E}\left[x_t|y_1, \ldots, y_{t-1}\right] + \left(\frac{a\mathbb{V}[x_t|y_1,\ldots,y_{t-1}]}{a^2\mathbb{V}[x_t|y_1,\ldots,y_{t-1}]+\sigma_v^2}\right)\left(y_t - a\mathbb{E}\left[x_t|y_1, \ldots, y_{t-1}\right]\right)$

– $\mathbb{V}\left[x_t|y_1,\ldots,y_t\right] = \mathbb{V}\left[x_t|y_1,\ldots,y_{t-1}\right]\left(1-\frac{a^2\mathbb{V}[x_t|y_1,\ldots,y_{t-1}]}{a^2\mathbb{V}[x_t|y_1,\ldots,y_{t-1}]+\sigma_v^2}\right)$

- **Kalman Smoothing:** Kalman smoothing works in reverse, starting at time $t = n-1$. It uses the Kalman predictions, filtered values, and their variances. For each value of t, the smoother starts at $s = 1$ and proceeds until $s = n - t$.

  – The smoother first computes an expectation:

  $$\mathbb{E}\left[x_t|y_1,\ldots,y_{t+s}\right] = \mathbb{E}\left[x_t|y_1,\ldots,y_{t+s-1}\right] + \frac{a\mathbb{E}\left[\left(x_t-\mathbb{E}\left[x_t|y_1,\ldots,y_{t-1}\right]\right)\left(x_{t+s}-\mathbb{E}\left[x_{t+s}|y_1,\ldots,y_{t+s-1}\right]\right)\right]}{a^2\mathbb{V}\left[x_{t+s}|y_1,\ldots,y_{t+s-1}\right]+\sigma_v^2}\left(y_{t+s}-a\mathbb{E}\left[x_{t+s}|y_1,\ldots,y_{t+s-1}\right]\right)$$

  – Then some covariances as well:

  $$\mathbb{E}\left[\left(x_t-\mathbb{E}\left[x_t|y_1,\ldots,y_{t-1}\right]\right)\left(x_{t+s}-\mathbb{E}\left[x_{t+s}|y_1,\ldots,y_{t+s-1}\right]\right)\right] = \phi\mathbb{E}\left[\left(x_t-\mathbb{E}\left[x_t|y_1,\ldots,y_{t-1}\right]\right)\left(x_{t+s-1}-\phi\mathbb{E}\left[x_{t+s-1}|y_1,\ldots,y_{t+s-2}\right]\right)\right]\left(1-\frac{a^2\mathbb{V}\left[x_{t+s-1}|y_1,\ldots,y_{t+s-2}\right]}{a^2\mathbb{V}\left[x_{t+s-1}|y_1,\ldots,y_{t+s-2}\right]+\sigma_v^2}\right)$$

  – And finally a variance:

  $$\mathbb{V}\left[x_t|y_1,\ldots,y_{t+s}\right] = \mathbb{V}\left[x_t|y_1,\ldots,y_{t+s-1}\right] - a^2\left(\frac{\mathbb{E}\left[\left(x_t-\mathbb{E}\left[x_t|y_1,\ldots,y_{t-1}\right]\right)\left(x_n-\mathbb{E}\left[x_{t+s}|y_1,\ldots,y_{t+s-1}\right]\right)\right]^2}{a^2\mathbb{V}\left[x_{t+s}|y_1,\ldots,y_{t+s-1}\right]+\sigma_v^2}\right)$$

The intuition motivating these algorithms is closely related to the intuition behind the Durbin-Levinson algorithm, which we saw when we studied forecasting. The basic idea is that the algorithms let us avoid repeated computationally expensive matrix inversions by exploiting relationships between the values we want to compute.

Estimating the State-Space Model Parameters

So far, we've assumed that the parameters σ_v^2, σ_w^2, a, ϕ, and μ are known. In practice, a is often assumed to be fixed and known at $a = 1$ but we may need to estimate the rest of the parameters from the data. This requires maximizing the **marginal likelihood** of the data $\boldsymbol{y}$, having integrated the latent time series $\boldsymbol{x}$ out. This is given by:

$$p\left(\boldsymbol{y}|\sigma_v^2,\sigma_w^2,a,\phi,\mu\right) = \int p\left(\boldsymbol{y}|\boldsymbol{x},\sigma_v^2,a\right)p\left(\boldsymbol{x}|\mu,\phi,\sigma_w^2\right)d\boldsymbol{x}. \tag{3}$$

Maximizing over an integral is tricky!

Direct Maximum Marginal Likelihood

Fortunately, our normal distribution facts tell us that the marginal distribution of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}\left(a\mathbb{E}\left[\boldsymbol{x}\right], a^2\mathbb{V}\left[\boldsymbol{x}\right] + \sigma_v^2\boldsymbol{I}_n\right), \tag{4}$$

which means that the integral can be evaluated to a normal density for $\boldsymbol{y}$ equal to:

$$\frac{1}{\sqrt{2\pi}^n\sqrt{\left|a^2\mathbb{V}\left[\boldsymbol{x}\right] + \sigma_v^2\boldsymbol{I}_n\right|}}\exp\left\{-\left(\boldsymbol{y} - a\mathbb{E}\left[\boldsymbol{x}\right]\right)'\left(a^2\mathbb{V}\left[\boldsymbol{x}\right] + \sigma_v^2\boldsymbol{I}_n\right)^{-1}\left(\boldsymbol{y} - a\mathbb{E}\left[\boldsymbol{x}\right]\right)\right\},$$

where $|\boldsymbol{A}|$ refers to the determinant of the matrix $\boldsymbol{A}$. If we could maximize this over $\sigma_v^2, \sigma_w^2, a, \phi$, and μ, then we'd be all set! Unfortunately, evaluating the marginal likelihood requires inverting an $n \times n$ matrix, which can be very computationally expensive. It would be better if we could write the marginal likelihood as a function of just the data and output from the Kalman predictor, filter, or smoother. Fortunately, this is possible!

Let $r_t = y_t - a\mathbb{E}\left[x_t|y_1, \ldots, y_{t-1}\right]$ be the one-step-ahead forecast errors with variances $\mathbb{V}\left[r_t|y_1, \ldots, y_{t-1}\right] = a^2\mathbb{V}\left[x_t|y_1, \ldots, y_{t-1}\right] + \sigma_v^2$. Conveniently, it's possible to use our normal distribution facts to show that we can write the likelihood in terms of the residuals as

$$p\left(r_1\right)\prod_{t=2}^{n}p\left(r_t|y_1, \ldots, y_{t-1}\right) = \frac{1}{\sqrt{2\pi}^n}\frac{1}{\sqrt{\mathbb{V}\left[r_1\right]}}\exp\left\{-\frac{r_1^2}{2\mathbb{V}\left[r_1\right]}\right\}\prod_{t=2}^{n}\frac{1}{\sqrt{\mathbb{V}\left[r_t|y_1, \ldots, y_{t-1}\right]}}\exp\left\{-\frac{r_t^2}{2\mathbb{V}\left[r_t|y_1, \ldots, y_{t-1}\right]}\right\}. \tag{5}$$

Defining the likelihood in this way doesn't quite get us out of the woods - although we can evaluate it quickly given values of $\sigma_v^2, \sigma_w^2, a, \phi$, and μ, it is still difficult to maximize (5) over $\sigma_v^2, \sigma_w^2, a, \phi$, and μ jointly as written because they all affect the values of $\boldsymbol{r}$ as well as the variances. Fortunately, we can take a computationally simpler approach which maximizes (5) by picking starting values of the parameters $\sigma_v^2, \sigma_w^2, a, \phi$, and μ and then iteratively:

- Fixing $\sigma_v^2, \sigma_w^2, a, \phi$, and μ and computing $\boldsymbol{r}$ accordingly, using the Kalman prediction algorithm;
- Maximizing (5) over $\sigma_v^2, \sigma_w^2, a, \phi$, and μ for fixed $\boldsymbol{r}$.

This iterative process continues until the likelihood (6) or successive values of the parameters

$\sigma_v^2, \sigma_w^2, a, \phi$, and μ converge. This is a popular way to compute maximum likelihood estimates of $\sigma_v^2, \sigma_w^2, a, \phi$, and μ, but it can be difficult to work in practice because the second step of maximizing (5) over $\sigma_v^2, \sigma_w^2, a, \phi$, and μ for fixed $\boldsymbol{r}$ is still a tricky nonlinear optimization problem.

Expectation-Maximization (EM) Maximum Marginal Likelihood

One other way we can compute maximum likelihood estimates of $\sigma_v^2, \sigma_w^2, a, \phi$, and μ is to use the expectation-maximization (EM) algorithm, which is an algorithm for maximizing a function that corresponds to an integral over some latent variables, which in this case are the latent states x_t. The EM algorithm allows us to maximize the marginal likelihood in (3) by maximizing $\mathbb{E}\left[\log\left(p\left(\boldsymbol{y}|\boldsymbol{x}, \sigma_v^2, a\right) p\left(\boldsymbol{x}|\mu, \phi, \sigma_w^2\right)\right)|y_1, \ldots, y_n\right]$ using an iterative procedure. It isn't immediately obvious how this helps us, but we will get a sense by working simplifying the conditional expectation of the joint log-likelihood:

$$
\begin{aligned}
&\mathbb{E}\left[\log\left(p\left(\boldsymbol{y}|\boldsymbol{x}, \sigma_v^2, a\right) p\left(\boldsymbol{x}|\mu, \phi, \sigma_w^2\right)\right)|y_1, \ldots, y_n\right] = \\
&\mathbb{E}\left[\log\left(p\left(y_1|x_1, \sigma_v^2, a\right) p\left(x_1|\mu\right) \prod_{t=2}^{n} p\left(y_t|x_t, \sigma_v^2, a\right) p\left(x_t|x_{t-1}, \phi, \sigma_w^2\right)\right)|y_1, \ldots, y_n\right] = \\
&K - \frac{n}{2}\log\left(\sigma_v^2\right) - \frac{1}{2\sigma_v^2}\sum_{t=1}^{n}\mathbb{E}\left[\left(y_t - ax_t\right)^2|y_1, \ldots, y_n\right] - \\
&\frac{(n-1)}{2}\log\left(\sigma_w^2\right) - \frac{1}{2\sigma_w^2}\sum_{t=2}^{n}\mathbb{E}\left[\left(x_t - \phi x_{t-1}\right)^2|y_1, \ldots, y_n\right] = \\
&K - \frac{n}{2}\log\left(\sigma_v^2\right) - \frac{1}{2\sigma_v^2}\sum_{t=1}^{n}\mathbb{E}\left(y_t^2 - 2ay_t\mathbb{E}\left[x_t|y_1, \ldots, y_n\right] + a^2\mathbb{E}\left[x_t^2|y_1, \ldots, y_n\right]\right) - \\
&\frac{(n-1)}{2}\log\left(\sigma_w^2\right) - \frac{1}{2\sigma_w^2}\sum_{t=3}^{n}\left(\mathbb{E}\left[x_t^2|y_1, \ldots, y_n\right] - 2\phi\mathbb{E}\left[x_t x_{t-1}|y_1, \ldots, y_n\right] + \phi^2\mathbb{E}\left[x_{t-1}^2|y_1, \ldots, y_n\right]\right) - \\
&\frac{1}{2\sigma_w^2}\left(\mathbb{E}\left[x_2^2|y_1, \ldots, y_n\right] - 2\phi\mu\mathbb{E}\left[x_2|y_1, \ldots, y_n\right] + \phi^2\mu^2\right),
\end{aligned}
\tag{6}
$$

where $x_1 = \mu$. where K is a constant that doesn't depend on the data $\boldsymbol{y}$, or the latent states $\boldsymbol{x}$, or the state-space parameters a, ϕ, σ_v^2, σ_w^2, and μ. We can maximize the marginal likelihood in (3) by picking starting values of the parameters $\sigma_v^2, \sigma_w^2, a, \phi$, and μ and then iteratively:

- Fixing the parameter values a, ϕ, σ_v^2, σ_w^2, and μ and computing the expectations $\mathbb{E}\left[x_t|y_1, \ldots, y_n\right]$ and $\mathbb{E}\left[x_t x_{t-1}|y_1, \ldots, y_n\right]$ using output from the Kalman smoother;

- Fixing the expectations $\mathbb{E}[x_t|y_1,\ldots,y_n]$ and $\mathbb{E}[x_t x_{t-1}|y_1,\ldots,y_n]$ and maximizing (6) over a, ϕ, σ_v^2, σ_w^2, and μ.

This iterative process continues until the likelihood (6) or successive values of the parameters $\sigma_v^2, \sigma_w^2, a, \phi$, and μ converge. It's common to refer to the first step as the E-step, because it involves computing expectations, and it is common to refer to the second step as the M-step, because it involves maximizing a function. This approach is computationally much simpler than the previous one, because the parameters $\sigma_v^2, \sigma_w^2, a, \phi$, and μ enter into the M-step in much more convenient ways. In fact, there are closed form solutions to the M-step, i.e. we can write down formulas for the values of $\sigma_v^2, \sigma_w^2, a, \phi$, and μ that maximize (6) when the expectations $\mathbb{E}[x_t|y_1,\ldots,y_n]$ and $\mathbb{E}[x_t x_{t-1}|y_1,\ldots,y_n]$ are held constant! The main disadvantage to using this approach is that it can converge very slowly. In practice, we often use a hybrid of both approaches, by using the results of a (possibly unconverged) EM algorithm as starting values for a direct maximum marginal likelihood estimation procedure.

More General Linear State-Space Models

Let's make things a little more general! We can incorporate covariates $\boldsymbol{z}_t$, and allow the linear transformation from the states x_t to the observed data y_t to depend on time by letting a depend on time, denoted by a_t:

$$\begin{aligned} y_t &= a_t x_t + \boldsymbol{z}_t'\boldsymbol{\gamma} + v_t && \text{Observation Equation} \\ x_t &= \phi x_{t-1} + \boldsymbol{z}_t'\boldsymbol{\upsilon} + w_t && \text{State Equation,} \end{aligned}$$

where $v_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_v^2)$, $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$ and $x_1 \sim \mathcal{N}(\mu, \sigma_x^2)$ and a_t is either fixed and known or known up to a constant, i.e. $a_t = ac_t$ where c_t is fixed and known but a is not. The Kalman prediction, filtering, and smoothing algorithms we discussed in the simpler case can be extended to quickly compute predictions, filtered values, and smoother values for the more general model. The maximum likelihood estimation procedures can likewise be generalized to estimate σ_v^2, σ_w^2, ϕ, a, $\boldsymbol{\gamma}$, $\boldsymbol{\upsilon}$, and μ from the marginal likelihood of the observed time series.

In state-space models, **seasonal effects** are often included via the inclusion of covariates in the observed data process, $\boldsymbol{z}_t$. For example, if the data is observed daily but weekly effects are expected to be present, we might include day of the week indicators in $\boldsymbol{z}_t$. Alternatively, we could include the covariates $\boldsymbol{z}_t$ in the state process. Usually, we only include the covariates in one process or the other, i.e. we either set $\boldsymbol{\gamma} = \mathbf{0}$ and estimate $\boldsymbol{\upsilon}$ from the data or we set $\boldsymbol{\upsilon} = \mathbf{0}$ and estimate $\boldsymbol{\gamma}$ from the data. Additionally, state-space models can be used to perform **stochastic regression** with a single covariate, the values of a_t correspond to the values of a single covariate. If this is the case, we can think of x_t as a time-varying regression coefficient.

Getting Into the Weeds

It can be valuable to derive closed form expressions for $\mathbb{E}[\boldsymbol{x}]$ and $\mathbb{V}[\boldsymbol{x}]$ to delve into the implied marginal distribution of the data under the state-space models discussed here. First, we derive $\mathbb{E}[\boldsymbol{x}]$ and $\mathbb{V}[\boldsymbol{x}]$.

$$
\begin{aligned}
x_1 &= \mu + \boldsymbol{z}_1'\boldsymbol{\upsilon} \\
x_2 &= \phi(\mu + \boldsymbol{z}_1'\boldsymbol{\upsilon}) + \boldsymbol{z}_2'\boldsymbol{\upsilon} + w_2 \\
x_3 &= \phi x_2 + w_3 = \phi^2(\mu + \boldsymbol{z}_1'\boldsymbol{\upsilon}) + \phi\boldsymbol{z}_2'\boldsymbol{\upsilon} + \boldsymbol{z}_3'\boldsymbol{\upsilon} + \phi w_2 + w_3 \\
x_4 &= \phi x_3 + w_4 = \phi^3(\mu + \boldsymbol{z}_1'\boldsymbol{\upsilon}) + \phi^2\boldsymbol{z}_2'\boldsymbol{\upsilon} + \phi\boldsymbol{z}_3'\boldsymbol{\upsilon} + \boldsymbol{z}_4'\boldsymbol{\upsilon} + \phi^2 w_2 + \phi w_3 + w_4 \\
x_k &= \phi^{k-1}\mu + \left(\sum_{i=1}^{k} \phi^{k-i}\boldsymbol{z}_i'\right)\boldsymbol{\upsilon} + \left(\sum_{i=2}^{k} \phi^{k-i} w_i\right) \\
\mathbb{E}[x_t] &= \phi^{t-1}\mu + \left(\sum_{i=1}^{k} \phi^{k-i}\boldsymbol{z}_i'\right)\boldsymbol{\upsilon} \\
\mathbb{V}[x_t] &= \sigma_w^2 \sum_{i=2}^{t} \phi^{2(t-i)} \\
\mathrm{Cov}[x_t, x_s] &= \sigma_w^2 \sum_{i=2}^{\min\{s,t\}} \phi^{t+s-2i}, \quad s,t > 1 \\
\mathrm{Cov}[x_1, x_k] &= 0 \quad \text{for all k.}
\end{aligned}
$$

Combining this with what is known about the marginal distribution of $\boldsymbol{y}$ based on properties of normal distributions and letting $\boldsymbol{t} = (1, \ldots, n)$, we have:

$$\boldsymbol{y} \sim \text{normal}\left(\boldsymbol{a} \circ \left(\phi^{\boldsymbol{t}-1}\mu + \tilde{\boldsymbol{Z}}\boldsymbol{v}\right) + \boldsymbol{Z}\boldsymbol{\gamma}, \begin{pmatrix} \sigma_v^2 & \boldsymbol{0}' \\ \boldsymbol{0} & \sigma_w^2 \boldsymbol{\Omega} \circ (\boldsymbol{a}\boldsymbol{a}') + \sigma_v^2 \boldsymbol{I}_{n-1} \end{pmatrix}\right),$$

where '$\circ$' refers to the elementwise Hadamard product, $\tilde{\boldsymbol{z}}_t = \left(\sum_{i=1}^{t} \phi^{t-i}\boldsymbol{z}_i'\right)$, and $\omega_{st} = \sigma_w^2 \sum_{i=2}^{\min\{s+1,t+1\}} \phi^{t+s+2(1-i)}$. In the special case where $a_t = a$,

$$\boldsymbol{y} \sim \text{normal}\left(\phi^{\boldsymbol{t}-1}(a\mu) + \tilde{\boldsymbol{Z}}(a\boldsymbol{v}) + \boldsymbol{Z}\boldsymbol{\gamma}, \begin{pmatrix} \sigma_v^2 & \boldsymbol{0}' \\ \boldsymbol{0} & a^2\sigma_w^2 \boldsymbol{\Omega} + \sigma_v^2 \boldsymbol{I}_{n-1} \end{pmatrix}\right).$$

Note that when $\phi = 0$, $\tilde{\boldsymbol{Z}} = \boldsymbol{Z}$. This model assumes that the observations y_t are centered around a rescaled multiple of the starting value μ, a weighted sum of the past and present covariate values, and the present covariate value. Note that a, ϕ, σ_w^2, and σ_v^2 are subject to the constraints that $a^2\sigma_w^2\boldsymbol{\Omega} + \sigma_v^2\boldsymbol{I}_{n-1}$ and $\boldsymbol{\Omega}$ are positive definite, to ensure that the marginal likelihood and conditional distributions of the states are well defined.