State space models

Introduction to dynamical systems #11

Hiroaki Sakamoto

§ Contents

1 Auto regression models

1.1 Random walk

- **Model**
  - Consider the following dynamical system:

$$x_t = x_{t-1} + v_t, \quad v_t \sim \mathcal{N}(0, V), \quad t = 0, 1, 2, \ldots, \tag{1}$$

  - Suppose that
    - the value of V is unknown to us
    - we observed a sample path $X_n := (x_0, x_1, x_2, \ldots, x_n)$
  - We want to obtain an estimate (i.e., the best guess) of V based on X_n

- **Maximum likelihood estimation**
  - What is the value of V that 'justifies' the observed data X_n?
    1. for each possible value of V, derive the probability of observing X_n (density $p(X_n)$)
    2. the maximum likelihood estimator, $\hat{V}$, is the value of V that maximizes the probability of observing what was actually observed, X_n
  - The density $p(X_n)$ of $X_n = (x_0, x_1, x_2, \ldots, x_n)$ may be decomposed as

$$p(X_n) = p(x_n|X_{n-1})p(X_{n-1}) = p(x_n|X_{n-1})p(x_{n-1}|X_{n-2})p(X_{n-2}) = \left(\prod_{t=1}^{n} p(x_t|X_{t-1})\right) p(x_0),$$

where (1) implies $x_t|X_{t-1} \sim \mathcal{N}(x_{t-1}, V)$ and thus

$$p(x_t|X_{t-1}) = \frac{1}{(2\pi)^{\frac{1}{2}} V^{\frac{1}{2}}} e^{-\frac{1}{2}\frac{(x_t - x_{t-1})^2}{V}} \quad \forall t \geq 1$$

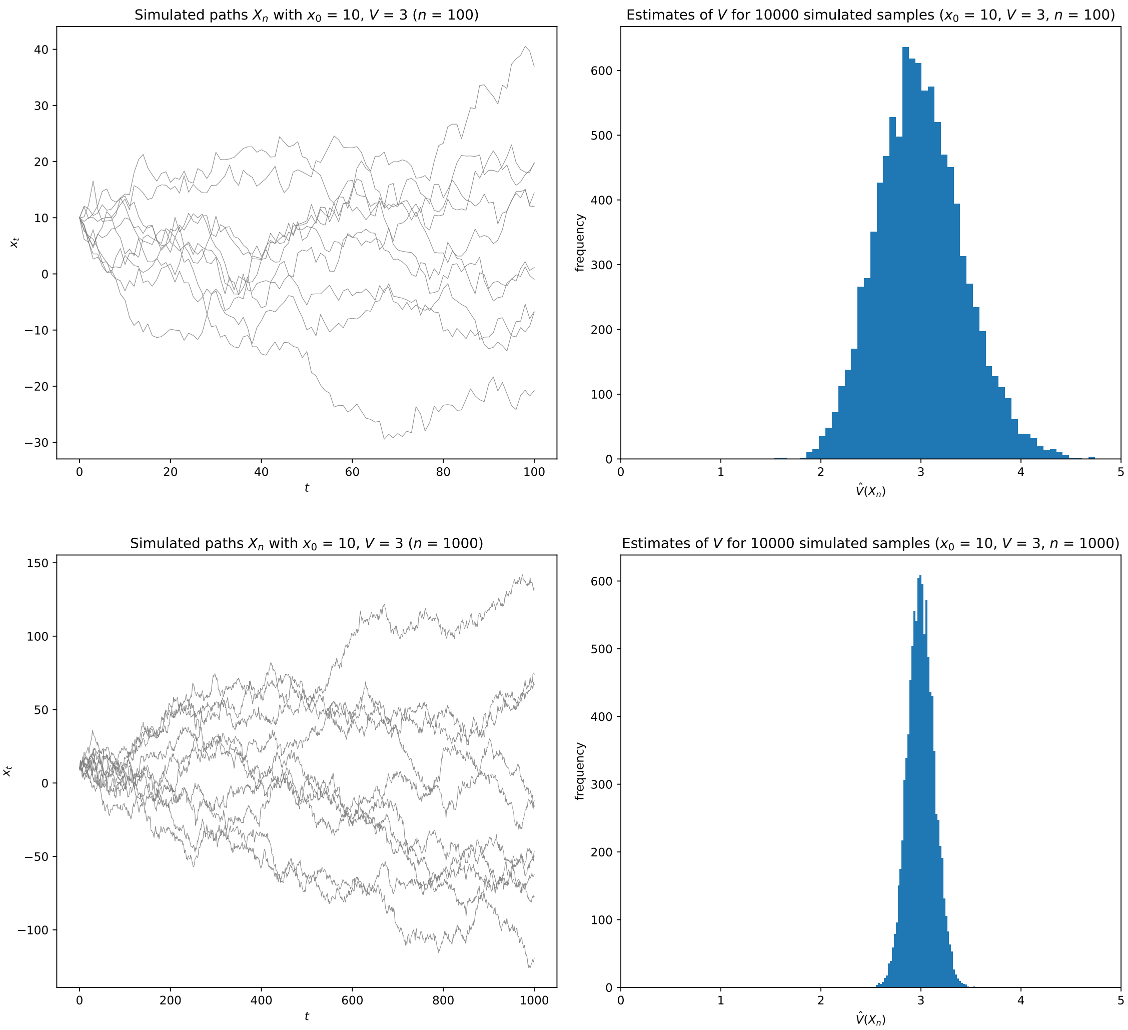

Figure 1: Sample paths X_n generated from (1) where $V = 3$ (left) and the maximum likelihood estimator $\hat{V}(X_n)$ computed as (2) (right).

- Assuming $p(x_0) = 1$, the *likelihood function* (density seen as a function of parameter) is

$$L(V; X_n) = \prod_{t=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}} V^{\frac{1}{2}}} e^{-\frac{1}{2}\frac{(x_t - x_{t-1})^2}{V}} = \left(\frac{1}{(2\pi)^{\frac{1}{2}} V^{\frac{1}{2}}} \right)^n e^{-\frac{1}{2V}\sum_{t=1}^{n}(x_t - x_{t-1})^2}$$

- The *maximum likelihood estimator* (MLE) of V is the one that maximizes $L(V; X_n)$, which for this particular example, is given as

$$\frac{dL(\hat{V}; X_n)}{dV} = 0 \iff \hat{V} = \frac{1}{n}\sum_{t=1}^{n}(x_t - x_{t-1})^2 \tag{2}$$

- Remarks:
  - $\hat{V}(X_n)$ is a function of stochastically generated data (different draw of X_n yields a different estimate $\hat{V}$)
  - If you are unlucky, you may observe X_n that rarely occurs (without knowing that it is a rare event), in which case $\hat{V}(X_n)$ may significantly deviate from the true value
  - In theory, however, MLE gives you a fairly 'good' estimate of V, ensuring $\mathbb{E}[\hat{V}(X_n)] = V$ and $\lim_{n\to\infty} \hat{V}(X_n) = V$; See Figure 1 for an illustration

1.2 AR1 model

- **Model**
  - Consider the following dynamical system
$$x_t = ax_{t-1} + b + \nu_t, \quad \nu_t \sim \mathcal{N}(0, V), \tag{3}$$
    which is often called the *autoregressive model* of order 1 (or AR1 model)
  - Suppose that
    - a, b, V are all unknown to us
    - we observed a sample path $X_n := (x_0, x_1, x_2, \dots, x_n)$
  - We want to obtain an estimate of unknown parameters $\boldsymbol{\theta} := (a, b, V)$ based on X_n
- **Likelihood function**
  - Model (3) implies that the probability density of observing X_n is
$$p(X_n) = \left(\prod_{t=1}^{n} p(x_t|X_{t-1})\right) p(x_0) = \left(\frac{1}{(2\pi)^{\frac{n}{2}} V^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{t=1}^{n} \frac{(x_t - ax_{t-1} - b)^2}{V}}\right) p(x_0),$$
    which is a function of unknown parameters, $\boldsymbol{\theta} = (a, b, V)$
  - Two alternative ways to specify $p(x_0)$:
    a) $p(x_0) = 1$ (assuming x_0 is fixed or improper/uniform prior)
    b) If we can reasonably assume $|a| < 1$, we solve the difference equation (3) for x_0 as
$$x_0 = ax_{-1} + b + \nu_0 = a(ax_{-2} + b + \nu_{-1}) + b + \nu_0 = \frac{1}{1-a}b + \sum_{k=0}^{\infty} a^k \nu_{-k} + \underbrace{\lim_{k\to\infty} a^k x_{-k}}_{=0},$$
    which implies $x_0 \sim \mathcal{N}(\mathbb{E}[x_0], \mathbb{V}[x_0])$ with
$$\mathbb{E}[x_0] = \mathbb{E}\left[\frac{1}{1-a}b + \sum_{k=0}^{\infty} a^k \nu_{-k}\right] = \frac{1}{1-a}b + \sum_{k=0}^{\infty} a^k \mathbb{E}\left[\nu_{-k}\right] = \frac{1}{1-a}b,$$
$$\mathbb{V}[x_0] = \mathbb{V}\left[\frac{1}{1-a}b + \sum_{k=0}^{\infty} a^k \nu_{-k}\right] = \sum_{k=0}^{\infty} a^{2k} \mathbb{V}\left[\nu_{-k}\right] = \frac{1}{1-a^2}V,$$
    and therefore
$$p(x_0) = \frac{1}{(2\pi)^{\frac{1}{2}} \left(\frac{1}{1-a^2}V\right)^{\frac{1}{2}}} e^{-\frac{1}{2}\frac{\left(x_0 - \frac{1}{1-a}b\right)^2}{\frac{1}{1-a^2}V}}$$
  - The likelihood function is
$$L(\boldsymbol{\theta}; X_n) = \begin{cases} \frac{1}{(2\pi)^{\frac{n}{2}} V^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{t=1}^{n} \frac{(x_t - ax_{t-1} - b)^2}{V}} & \text{if we can assume } p(x_0) = 1 \\ \frac{(1-a^2)^{\frac{1}{2}}}{(2\pi)^{\frac{n+1}{2}} V^{\frac{n+1}{2}}} e^{-\frac{1}{2}\sum_{t=1}^{n} \frac{(x_t - ax_{t-1} - b)^2}{V} - \frac{1-a^2}{2}\frac{\left(x_0 - \frac{1}{1-a}b\right)^2}{V}} & \text{otherwise} \end{cases}$$

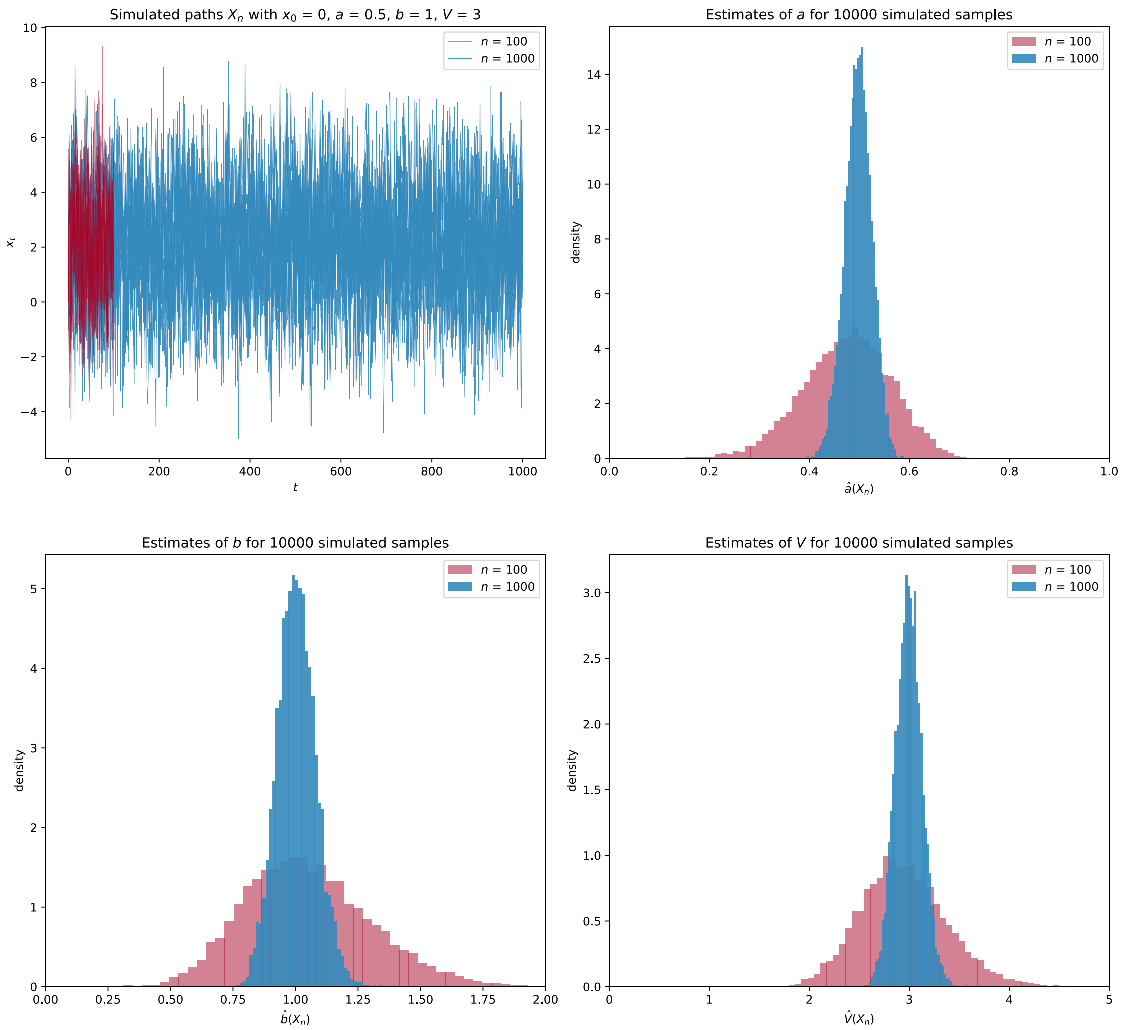

Figure 2: Sample paths X_n generated from (3) where $x_0 = 0$, $a = 0.5$, $b = 1$, $V = 3$ (top left) and the maximum likelihood estimator $\hat{\theta}(X_n) = (\hat{a}(X_n), \hat{b}(X_n), \hat{V}(X_n))$ computed as (5).

- **Maximum likelihood estimator**
  - The maximum likelihood estimator, $\hat{\boldsymbol{\theta}} = (\hat{a}, \hat{b}, \hat{V})$, must satisfy the first-order condition

$$\frac{\partial L(\hat{\boldsymbol{\theta}}; X_n)}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{4}$$

  - In case of $p(x_0) = 1$, the first-order condition (4) yields

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{n} x_{t-1}^2 & \sum_{t=1}^{n} x_{t-1} \\ \sum_{t=1}^{n} x_{t-1} & n \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{n} x_t x_{t-1} \\ \sum_{t=1}^{n} x_t \end{bmatrix}, \quad \text{and} \quad \hat{V} = \frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{a} x_{t-1} - \hat{b})^2 \tag{5}$$

  - The estimator $\hat{\boldsymbol{\theta}}(X_n)$ is a function of data:
    - it typically involves estimation errors but gives the true parameter values on average
    - the estimation errors become smaller as the sample size n increases
    - See Figure 2 for an illustration
  - In case of $p(x_0) \neq 1$, no closed-form expression is available for $\hat{\boldsymbol{\theta}}$ and we resort to numerically solving the maximization problem

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}; X_n)$$

2 Random walk with measurement noise

2.1 Model

- **Description**
  - Suppose that we cannot directly observe $X_n = (x_0, x_1, \ldots, x_n)$ due, for example, to:
    - measurement noise
    - limited data availability
  - The simplest possible case is

$$\begin{aligned} x_t &= x_{t-1} + v_t, \quad v_t = V^{\frac{1}{2}} z_{v,t} \sim \mathcal{N}(0, V) \\ y_t &= x_t + \omega_t, \quad \omega_t = W^{\frac{1}{2}} z_{\omega,t} \sim \mathcal{N}(0, W) \end{aligned} \qquad \forall t = 1, 2, \ldots, n \tag{6}$$

    where
    - x_t is a state variable, which is NOT directly observable (i.e., latent variable)
    - y_t is an observable variable (i.e., measurement), from which we indirectly infer x_t
    - v_t is state disturbance (random component outside the model)
    - ω_t is observation disturbance (measurement noise)
  - For example:
    - you may be remotely monitoring the location of your cat using a GPS device
    - x_t is the actual location of your cat that is randomly walking around
    - y_t is a (noisy) signal sent from the GPS device attached to the cat

- **Our task**
  - Suppose that
    - the values of V, W are unknown to us
    - we observed $Y_n := (y_1, y_2, \ldots, y_n)$
    - the state trajectory $X_n := (x_0, x_1, \ldots, x_n)$ is NOT observable
  - We want to obtain an estimate of
    - the value of parameter V, W
    - the state trajectory $X_n := (x_0, x_1, \ldots, x_n)$

    both based on the measurement data Y_n
  - For maximum likelihood estimation, we need to compute the probability density

$$p(Y_n) = p(y_n | Y_{n-1}) p(Y_{n-1}) = \prod_{t=1}^{n} p(y_t | Y_{t-1}), \tag{7}$$

    which in turn requires us to compute $p(y_t | Y_{t-1})$ for each t (but how?)

2.2 Kalman filter

- **The idea**
  - We sequentially compute the distribution of $y_t | Y_{t-1}$ as follows:

$$x_0 | Y_0 \overset{(6)}{\Longrightarrow} (x_1, y_1) | Y_0 \overset{y_1}{\Longrightarrow} x_1 | Y_1 \overset{(6)}{\Longrightarrow} (x_2, y_2) | Y_1 \overset{y_2}{\Longrightarrow} x_2 | Y_2 \overset{(6)}{\Longrightarrow} (x_3, y_3) | Y_2 \overset{y_3}{\Longrightarrow} \cdots$$

  - This sequential process is called the *Kalman filtering*

- **Details**
  - STEP 0: Initial distribution $x_0|Y_0$
    - Assume the distribution of initial state x_0 as

$$x_0 = x_{0|0} + P_{0|0}^{\frac{1}{2}} z_0 \sim \mathcal{N}(x_{0|0}, P_{0|0}) \tag{8}$$

for a Gaussian white noise $z_0 \sim \mathcal{N}(0,1)$ and some **known** constants $x_{0|0}$ and $P_{0|0}$ (but see below for the case where these constants are unknown)

  - STEP 1: Prior $x_1|Y_0$, forecast $y_1|Y_0$, and posterior $x_1|Y_1$
    - Using model (6) and initial distribution (8), we have

$$\begin{bmatrix} x_1|Y_0 \\ y_1|Y_0 \end{bmatrix} = \begin{bmatrix} x_0|Y_0 + \nu_1 \\ x_1|Y_0 + \omega_1 \end{bmatrix} = \begin{bmatrix} x_0|Y_0 + \nu_1 \\ x_0|Y_0 + \nu_1 + \omega_1 \end{bmatrix} = \begin{bmatrix} x_{0|0} + P_{0|0}^{\frac{1}{2}} z_0 + V^{\frac{1}{2}} z_{\nu,1} \\ x_{0|0} + P_{0|0}^{\frac{1}{2}} z_0 + V^{\frac{1}{2}} z_{\nu,1} + W^{\frac{1}{2}} z_{\omega,1} \end{bmatrix}$$
$$= \begin{bmatrix} x_{0|0} \\ x_{0|0} \end{bmatrix} + \begin{bmatrix} P_{0|0}^{\frac{1}{2}} & V^{\frac{1}{2}} & 0 \\ P_{0|0}^{\frac{1}{2}} & V^{\frac{1}{2}} & W^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} z_0 \\ z_{\nu,1} \\ z_{\omega,1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} x_{0|0} \\ x_{0|0} \end{bmatrix}, \begin{bmatrix} P_{0|0} + V & P_{0|0} + V \\ P_{0|0} + V & P_{0|0} + V + W \end{bmatrix} \right), \tag{9}$$

from which we can compute the marginal distributions as

$$\underbrace{x_1|Y_0}_{\text{prior on } x_1} \sim \mathcal{N}(\underbrace{x_{0|0}}_{=:\hat{x}_1}, \underbrace{P_{0|0} + V}_{=:\hat{P}_1}), \qquad \underbrace{y_1|Y_0}_{\text{forecast on } y_1} \sim \mathcal{N}(\underbrace{\hat{x}_1}_{=:\hat{y}_1}, \underbrace{\hat{P}_1 + W}_{=:\hat{Q}_1}) \tag{10}$$

    - Once y_1 is observed, combine it with (9) to obtain the conditional distribution

$$x_1|Y_1 \sim \mathcal{N}\left(x_{0|0} + \frac{P_{0|0} + V}{P_{0|0} + V + W}(y_1 - x_{0|0}), (P_{0|0} + V) - \frac{P_{0|0} + V}{P_{0|0} + V + W}(P_{0|0} + V) \right)$$
$$= \mathcal{N}\left(\underbrace{\hat{x}_1 + \frac{\hat{P}_1}{\hat{Q}_1}(y_1 - \hat{y}_1)}_{=:x_{1|1}}, \underbrace{\hat{P}_1 - \frac{\hat{P}_1}{\hat{Q}_1}\hat{Q}_1\frac{\hat{P}_1}{\hat{Q}_1}}_{=:P_{1|1}} \right) \tag{11}$$

    - Note (11) may be written as

$$x_1|Y_1 = x_{1|1} + P_{1|1}^{\frac{1}{2}} z_1, \tag{12}$$

where $z_1 \sim \mathcal{N}(0,1)$ is independent of (ν_2, ω_2) because it comes from (z_0, ν_1, ω_1)

  - STEP 2: Prior $x_2|Y_1$, forecast $y_2|Y_1$, and posterior $x_2|Y_2$
    - Using $x_1|Y_1$ defined as (12) and model (6), we have

$$\begin{bmatrix} x_2|Y_1 \\ y_2|Y_1 \end{bmatrix} = \begin{bmatrix} x_{1|1} \\ x_{1|1} \end{bmatrix} + \begin{bmatrix} P_{1|1}^{\frac{1}{2}} & V^{\frac{1}{2}} & 0 \\ P_{1|1}^{\frac{1}{2}} & V^{\frac{1}{2}} & W^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} z_1 \\ z_{\nu,2} \\ z_{\omega,2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} x_{1|1} \\ x_{1|1} \end{bmatrix}, \begin{bmatrix} P_{1|1} + V & P_{1|1} + V \\ P_{1|1} + V & P_{1|1} + V + W \end{bmatrix} \right), \tag{13}$$

from which we can compute the marginal distributions as

$$x_2|Y_1 \sim \mathcal{N}(\underbrace{x_{1|1}}_{=:\hat{x}_2}, \underbrace{P_{1|1} + V}_{=:\hat{P}_2}), \quad y_2|Y_1 \sim \mathcal{N}(\underbrace{\hat{x}_2}_{=:\hat{y}_2}, \underbrace{\hat{P}_2 + W}_{=:\hat{Q}_2}) \tag{14}$$

- Once y_2 is observed, combine it with (13) to obtain the conditional distribution

$$x_2|Y_2 \sim \mathcal{N}\Bigg(\underbrace{\hat{x}_2 + \frac{\hat{P}_2}{\hat{Q}_2}(y_2 - \hat{y}_2)}_{=:x_{2|2}}, \underbrace{\hat{P}_2 - \frac{\hat{P}_2}{\hat{Q}_2}\hat{Q}_2\frac{\hat{P}_2}{\hat{Q}_2}}_{=:P_{2|2}}\Bigg) \tag{15}$$

- Note (15) may be written as

$$x_2|Y_2 = x_{2|2} + P_{2|2}^{\frac{1}{2}} z_2, \tag{16}$$

where $z_2 \sim \mathcal{N}(0,1)$ is independent of (ν_3, ω_3)

◦ STEP t: Prior $x_t|Y_{t-1}$, forecast $y_t|Y_{t-1}$, and posterior $x_t|Y_t$

- Using $x_{t-1}|Y_{t-1}$ (from the preceding step) and model (6), we have

$$\begin{bmatrix} x_t|Y_{t-1} \\ y_t|Y_{t-1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} x_{t-1|t-1} \\ x_{t-1|t-1} \end{bmatrix}, \begin{bmatrix} P_{t-1|t-1} + V & P_{t-1|t-1} + V \\ P_{t-1|t-1} + V & P_{t-1|t-1} + V + W \end{bmatrix}\right), \tag{17}$$

from which the marginal distributions follow as

$$x_t|Y_{t-1} \sim \mathcal{N}(\underbrace{x_{t-1|t-1}}_{=:\hat{x}_t}, \underbrace{P_{t-1|t-1} + V}_{=:\hat{P}_t}), \quad y_t|Y_{t-1} \sim \mathcal{N}(\underbrace{\hat{x}_t}_{=:\hat{y}_t}, \underbrace{\hat{P}_t + W}_{=:\hat{Q}_t}) \tag{18}$$

- Once y_t is observed, combine it with (17) to obtain the conditional distribution

$$x_t|Y_t \sim \mathcal{N}\Bigg(\underbrace{\hat{x}_t + \frac{\hat{P}_t}{\hat{Q}_t}(y_t - \hat{y}_t)}_{=:x_{t|t}}, \underbrace{\hat{P}_t - \frac{\hat{P}_t}{\hat{Q}_t}\hat{Q}_t\frac{\hat{P}_t}{\hat{Q}_t}}_{=:P_{t|t}}\Bigg)$$

- **Kalman filter equations**

◦ The incremental updating process described above is summarized as follows:

1. Given the posterior $x_{t-1}|Y_{t-1} \sim \mathcal{N}(x_{t-1|t-1}, P_{t-1|t-1})$ from the previous period, compute the prior:

$$x_t|Y_{t-1} \sim \mathcal{N}(\hat{x}_t, \hat{P}_t) \quad \text{where} \quad \begin{aligned} \hat{x}_t &= x_{t-1|t-1} \\ \hat{P}_t &= P_{t-1|t-1} + V \end{aligned} \tag{19}$$

2. Given the prior $x_t|Y_{t-1} \sim \mathcal{N}(\hat{x}_t, \hat{P}_t)$, compute the forecast:

$$y_t|Y_{t-1} \sim \mathcal{N}(\hat{y}_t, \hat{Q}_t) \quad \text{where} \quad \begin{aligned} \hat{y}_t &= \hat{x}_t \\ \hat{Q}_t &= \hat{P}_t + W \end{aligned} \tag{20}$$

3. Compute

$$K_t = \frac{\hat{P}_t}{\hat{Q}_t}, \tag{21}$$

which is called the *Kalman gain*

4. Once y_t is observed, compute the forecast error $\hat{q}_t$ as

$$\hat{q}_t = y_t - \hat{y}_t \tag{22}$$

and derive the posterior distribution:

$$x_t|Y_t \sim \mathcal{N}(x_{t|t}, P_{t|t}) \quad \text{where} \quad \begin{aligned} x_{t|t} &= \hat{x}_t + K_t\hat{q}_t \\ P_{t|t} &= \hat{P}_t - K_t\hat{Q}_tK_t \end{aligned} \tag{23}$$

◦ Equations (19)–(23) are called the Kalman filter equations

2.3 Parameter estimation

- **Maximum likelihood estimator**
  - The probability density of Y_n is then

$$p(Y_n) = \prod_{t=1}^{n} p(y_t|Y_{t-1}) = \prod_{t=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}} \hat{Q}_t^{\frac{1}{2}}} e^{-\frac{1}{2}\frac{(y_t-\hat{y}_t)^2}{\hat{Q}_t}} \tag{24}$$

  - MLE of $\boldsymbol{\theta} = (V, W)$ is the one that maximizes the log likelihood

$$\begin{aligned}
\ln L(\boldsymbol{\theta}; Y_n) &:= \ln\left(p(Y_n)\right) = \sum_{t=1}^{n} \ln\left(p(y_t|Y_{t-1})\right) \\
&= \sum_{t=1}^{n} \left(-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\hat{Q}_t) - \frac{1}{2}\frac{(y_t-\hat{y}_t)^2}{\hat{Q}_t}\right) \\
&= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\left(\ln(\hat{Q}_t) + \frac{(y_t-\hat{y}_t)^2}{\hat{Q}_t}\right)
\end{aligned} \tag{25}$$

- **Uninformative case**
  - We have assumed that the initial state distribution (8) is known, which is reasonable when we have prior knowledge about the state
  - In practice, however, the initial distribution may be unknown, in which case we use the so-called uninformative (or diffuse) prior
  - To be more precise, we put $P_{0|0} = \kappa$ and take the limit of $\kappa \to \infty$ in (11) to obtain

$$\lim_{\kappa\to\infty} x_1|Y_1 \sim \mathcal{N}(\underbrace{y_1}_{x_{1|1}}, \underbrace{W}_{P_{1|1}}), \tag{26}$$

    from which we recursively compute $(x_t, y_t)|Y_{t-1}$ and $x_t|Y_t$ for all $t = 2, 3, \ldots$

- **Maximum likelihood estimator: uninformative case**
  - One can normalize the log-likelihood function by adding a constant $\frac{1}{2}\ln(P_{0|0})$

$$\begin{aligned}
&\ln \bar{L}(\boldsymbol{\theta}; Y_n) := \ln\left(p(Y_n)\right) + \frac{1}{2}\ln(P_{0|0}) \\
&= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=2}^{n}\left(\ln(\hat{Q}_t) + \frac{(y_t-\hat{y}_t)^2}{\hat{Q}_t}\right) - \frac{1}{2}\underbrace{\ln\left(\frac{P_{0|0}+V+W}{P_{0|0}}\right)}_{\to 0\ (P_{0|0}\to\infty)} - \frac{1}{2}\underbrace{\frac{(y_1-\hat{y}_1)^2}{P_{0|0}+V+W}}_{\to 0\ (P_{0|0}\to\infty)}
\end{aligned}$$

  - Clearly, $\boldsymbol{\theta}$ maximizes $\ln \bar{L}(\boldsymbol{\theta}; Y_n)$ if and only if it maximizes $\ln L(\boldsymbol{\theta}; Y_n)$
  - Putting $P_{0|0} = \kappa$ and taking the limit of $\kappa \to \infty$, we obtain the diffuse log-likelihood

$$\begin{aligned}
\ln L_d(\boldsymbol{\theta}; Y_n) &:= \lim_{\kappa\to\infty} \ln \bar{L}(\boldsymbol{\theta}; Y_n) \\
&= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=2}^{n}\left(\ln\left(\hat{Q}_t\right) + \frac{(y_t-\hat{y}_t)^2}{\hat{Q}_t}\right),
\end{aligned} \tag{27}$$

    where $\hat{Q}_t$ and $\hat{y}_t$ are all computed based on the initialization given by (26)
  - In the uninformative case, MLE of $\boldsymbol{\theta}$ is the one that maximizes (27)
  - See Figure 3 for illustration

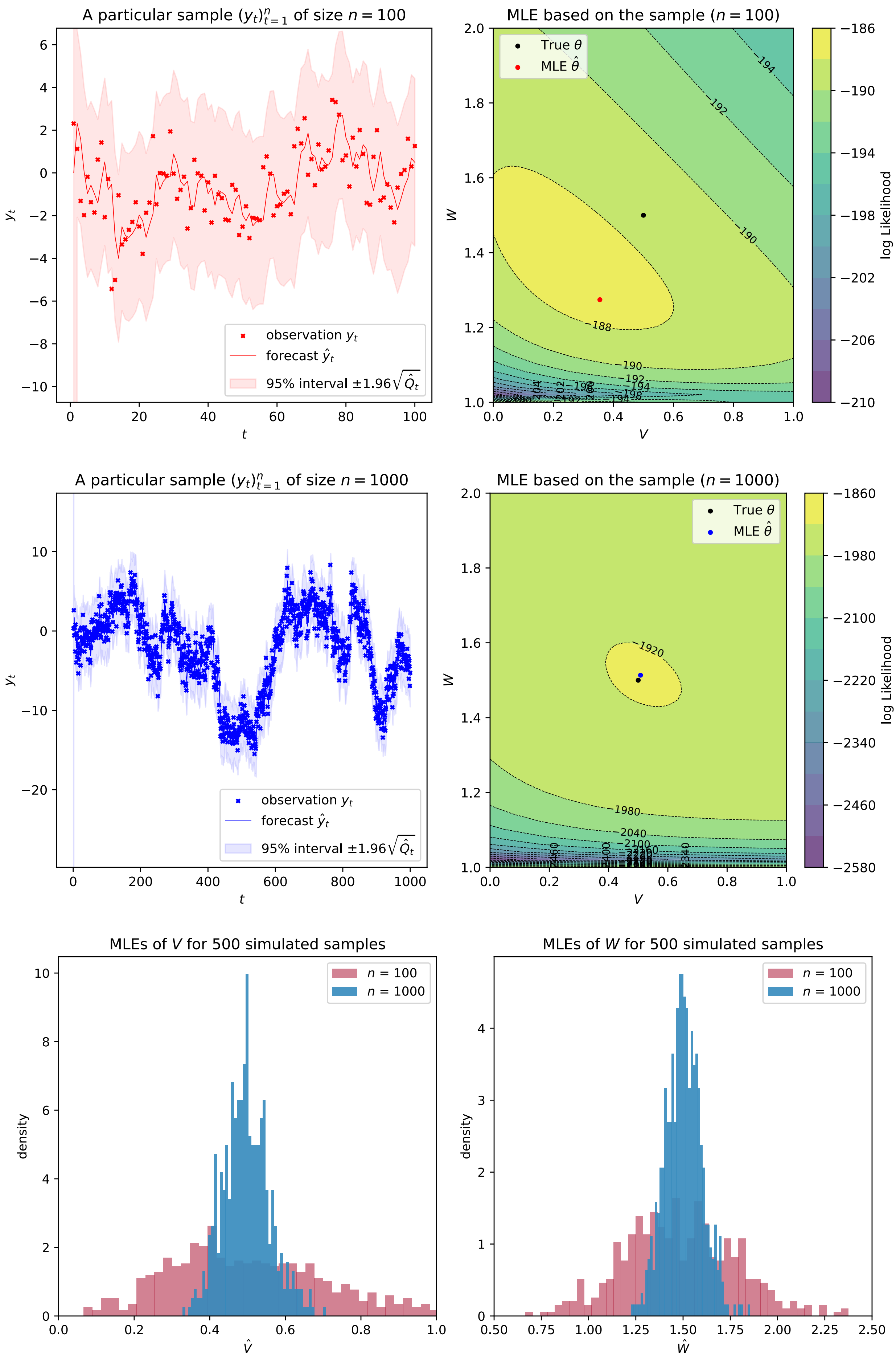

Figure 3: Sample path $Y_n = (y_1, \ldots, y_n)$ generated from (6) where $V = 0.5, W = 1.5$ for different sample size, $n = 100$ (top) or $n = 1000$ (middle). The distribution of maximum likelihood estimator $\hat{V}, \hat{W}$ (bottom).

3 General linear state-space model

3.1 Model

- **Linear Gaussian state space model**
  - Consider a time series $(y_t)_{t=1}^n$ from the following data generating process:

$$\begin{aligned} x_t &= A_t x_{t-1} + v_t + \nu_t, \quad \nu_t = V_t^{\frac{1}{2}} z_{\nu,t} \sim \mathcal{N}(\mathbf{0}, V_t) \\ y_t &= C_t x_t + w_t + \omega_t, \quad \omega_t = W_t^{\frac{1}{2}} z_{\omega,t} \sim \mathcal{N}(\mathbf{0}, W_t) \end{aligned} \qquad \forall t = 1, 2, \ldots, n, \tag{28}$$

    where $A_t \in \mathbb{R}^{m\times m}$, $C_t \in \mathbb{R}^{p\times m}$, $v_t \in \mathbb{R}^m$, $w_t \in \mathbb{R}^p$, $V_t \in \mathbb{R}^{m\times m}$, $W_t \in \mathbb{R}^{p\times p}$, and
    - $x_t \in \mathbb{R}^m$: potentially unobservable (latent) state vector at time t
    - $y_t \in \mathbb{R}^p$: observed vector at time t
    - $\nu_t \in \mathbb{R}^m$: state disturbance at time t (white noise)
    - $\omega_t \in \mathbb{R}^p$: observation disturbance at time t (white noise)

3.2 Kalman filter

- **Filtering process**
  - STEP 0: Initial state distribution
    - Assume there exists a standard (multivariate) Gaussian $z_0 \sim \mathcal{N}(0, I)$ and

$$x_0 = x_{0|0} + P_{0|0}^{\frac{1}{2}} z_0 \sim \mathcal{N}(x_{0|0}, P_{0|0}) \tag{29}$$

      for some known vector $x_{0|0} \in \mathbb{R}^m$ and positive definite matrix $P_{0|0} \in \mathbb{R}^{m\times m}$
  - STEP 1: Prior $x_1|Y_0$, forecast $x_1|Y_0$, and posterior $x_1|Y_1$
    - Given (28) and (29), the joint distribution is

$$\begin{aligned} \begin{bmatrix} x_1|Y_0 \\ y_1|Y_0 \end{bmatrix} &= \begin{bmatrix} A_1 x_0|Y_0 + v_1 + \nu_1 \\ C_1 x_1|Y_0 + w_1 + \omega_1 \end{bmatrix} \\ &= \begin{bmatrix} A_1 x_{0|0} + v_1 \\ C_1 A_1 x_{0|0} + C_1 v_1 + w_1 \end{bmatrix} + \begin{bmatrix} A_1 P_{0|0}^{\frac{1}{2}} & V_1^{\frac{1}{2}} & O \\ C_1 A_1 P_{0|0}^{\frac{1}{2}} & C_1 V_1^{\frac{1}{2}} & W_1^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} z_0 \\ z_{\nu,1} \\ z_{\omega,1} \end{bmatrix} \\ &\sim \mathcal{N}\left(\begin{bmatrix} A_1 x_{0|0} + v_1 \\ C_1 A_1 x_{0|0} + C_1 v_1 + w_1 \end{bmatrix}, \Sigma \right), \end{aligned} \tag{30}$$

      where

$$\begin{aligned} \Sigma &= \begin{bmatrix} A_1 P_{0|0}^{\frac{1}{2}} & V_1^{\frac{1}{2}} & O \\ C_1 A_1 P_{0|0}^{\frac{1}{2}} & C_1 V_1^{\frac{1}{2}} & W_1^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} (A_1 P_{0|0}^{\frac{1}{2}})^\top & (C_1 A_1 P_{0|0}^{\frac{1}{2}})^\top \\ (V_1^{\frac{1}{2}})^\top & (C_1 V_1^{\frac{1}{2}})^\top \\ O & (W_1^{\frac{1}{2}})^\top \end{bmatrix} \\ &= \begin{bmatrix} A_1 P_{0|0} A_1^\top + V_1 & (A_1 P_{0|0} A_1^\top + V_1) C_1^\top \\ C_1 (A_1 P_{0|0} A_1^\top + V_1)^\top & C_1 (A_1 P_{0|0} A_1^\top + V_1) C_1^\top + W_1 \end{bmatrix} \end{aligned} \tag{31}$$

    - The prior on x_1 and the forecast on y_1 are therefore

$$x_1|Y_0 \sim \mathcal{N}\Big(\underbrace{A_1 x_{0|0} + v_1}_{=:\hat{x}_1}, \underbrace{A_1 P_{0|0} A_1^\top + V_1}_{=:\hat{P}_1} \Big), \quad y_1|Y_0 \sim \mathcal{N}\Big(\underbrace{C_1 \hat{x}_1 + w_1}_{=:\hat{y}_1}, \underbrace{C_1 \hat{P}_1 C_1^\top + W_1}_{=:\hat{Q}_1} \Big) \tag{32}$$

– Once y_1 is observed, it follows from (30) that the posterior $x_1|Y_1$ is[1]

$$
\begin{aligned}
x_1|Y_1 \sim \mathcal{N}\Bigg(& A_1 x_{0|0} + v_1 + \frac{(A_1 P_{0|0} A_1^\top + V_1) C_1^\top}{C_1 (A_1 P_{0|0} A_1^\top + V_1) C_1^\top + W_1} (y_1 - C_1 A_1 x_{0|0} - C_1 v_1 - w_1), \\
& (A_1 P_{0|0} A_1^\top + V_1) - \frac{(A_1 P_{0|0} A_1^\top + V_1) C_1^\top}{C_1 (A_1 P_{0|0} A_1^\top + V_1) C_1^\top + W_1} C_1 (A_1 P_{0|0} A_1^\top + V_1) \Bigg) \\
= \mathcal{N}\Bigg(& \underbrace{\hat{x}_1 + \frac{\hat{P}_1 C_1^\top}{\hat{Q}_1} (y_1 - \hat{y}_1)}_{=:x_{1|1}}, \underbrace{\hat{P}_1 - \frac{\hat{P}_1 C_1^\top}{\hat{Q}_1} \hat{Q}_1 \left(\frac{\hat{P}_1 C_1^\top}{\hat{Q}_1} \right)^\top}_{=:P_{1|1}} \Bigg)
\end{aligned} \tag{33}
$$

– Note (33) may be written as

$$
x_1|Y_1 = x_{1|1} + P_{1|1}^{\frac{1}{2}} z_1, \tag{34}
$$

where $z_1 \sim \mathcal{N}(\mathbf{0}, I)$ is independent of (ν_2, ω_2) because it comes from (z_0, ν_1, ω_1)

◦ STEP 2: Prior $x_2|Y_1$, forecast $x_2|Y_1$, and posterior $x_2|Y_2$

– Using $x_1|Y_1$ defined as (34) and the model (28), we have

$$
\begin{aligned}
\begin{bmatrix} x_2|Y_1 \\ y_2|Y_1 \end{bmatrix} &= \begin{bmatrix} A_2 x_1|Y_1 + v_2 + \nu_2 \\ C_2 x_2|Y_1 + w_2 + \omega_2 \end{bmatrix} \\
&= \begin{bmatrix} A_2 x_{1|1} + v_2 \\ C_2 A_2 x_{1|1} + C_2 v_2 + w_2 \end{bmatrix} + \begin{bmatrix} A_2 P_{1|1}^{\frac{1}{2}} & V_2^{\frac{1}{2}} & O \\ C_2 A_2 P_{1|1}^{\frac{1}{2}} & C_2 V_2^{\frac{1}{2}} & W_2^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} z_1 \\ z_{\nu,2} \\ z_{\omega,2} \end{bmatrix} \\
&\sim \mathcal{N}\left(\begin{bmatrix} A_2 x_{1|1} + v_2 \\ C_2 A_2 x_{1|1} + C_2 v_2 + w_2 \end{bmatrix}, \Sigma \right),
\end{aligned} \tag{35}
$$

where

$$
\Sigma = \begin{bmatrix} A_2 P_{1|1} A_2^\top + V_2 & (A_2 P_{1|1} A_2^\top + V_2) C_2^\top \\ C_2 (A_2 P_{1|1} A_2^\top + V_2)^\top & C_2 (A_2 P_{1|1} A_2^\top + V_2) C_2^\top + W_2 \end{bmatrix},
$$

from which we can compute the marginal distributions as

$$
x_2|Y_1 \sim \mathcal{N}\Big(\underbrace{A_2 x_{1|1} + v_2}_{=:\hat{x}_2}, \underbrace{A_2 P_{1|1} A_2^\top + V_2}_{=:\hat{P}_2} \Big), \quad y_2|Y_1 \sim \mathcal{N}\Big(\underbrace{C_2 \hat{x}_2 + w_2}_{=:\hat{y}_2}, \underbrace{C_2 \hat{P}_2 C_2^\top + W_2}_{=:\hat{Q}_2} \Big)
$$

– Once y_2 is observed, it follows from (35) that the posterior $x_2|Y_2$ is

$$
x_2|Y_2 \sim \mathcal{N}\Bigg(\underbrace{\hat{x}_2 + \frac{\hat{P}_2 C_2^\top}{\hat{Q}_2} (y_2 - \hat{y}_2)}_{=:x_{2|2}}, \underbrace{\hat{P}_2 - \frac{\hat{P}_2 C_2^\top}{\hat{Q}_2} \hat{Q}_2 \left(\frac{\hat{P}_2 C_2^\top}{\hat{Q}_2} \right)^\top}_{=:P_{2|2}} \Bigg) \tag{36}
$$

– Note (36) may be written as

$$
x_2|Y_2 = x_{2|2} + P_{2|2}^{\frac{1}{2}} z_2, \tag{37}
$$

where $z_2 \sim \mathcal{N}(\mathbf{0}, I)$ is independent of (ν_3, ω_3)

[1] Here, just to make the expression easier to read, we introduce 'division' of matrices as $\frac{X_1}{X_2} := X_1 X_2^{-1}$.

- STEP t: Prior $\boldsymbol{x}_t|Y_{t-1}$, forecast $\boldsymbol{x}_t|Y_{t-1}$, and posterior $\boldsymbol{x}_t|Y_t$
  - Using $\boldsymbol{x}_{t-1}|Y_{t-1}$ (from the preceding step) and the model (28), we have

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{x}_t|Y_{t-1} \\ \boldsymbol{y}_t|Y_{t-1} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{A}_t\boldsymbol{x}_{t-1}|Y_{t-1} + \boldsymbol{v}_t + \boldsymbol{\nu}_t \\ \boldsymbol{C}_t\boldsymbol{x}_t|Y_{t-1} + \boldsymbol{w}_t + \boldsymbol{\omega}_t \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{A}_t\boldsymbol{x}_{t-1|t-1} + \boldsymbol{v}_t \\ \boldsymbol{C}_t\boldsymbol{A}_t\boldsymbol{x}_{t-1|t-1} + \boldsymbol{C}_t\boldsymbol{v}_t + \boldsymbol{w}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{A}_t\boldsymbol{P}^{\frac{1}{2}}_{t-1|t-1} & \boldsymbol{V}_t^{\frac{1}{2}} & \boldsymbol{O} \\ \boldsymbol{C}_t\boldsymbol{A}_t\boldsymbol{P}^{\frac{1}{2}}_{t-1|t-1} & \boldsymbol{C}_t\boldsymbol{V}_t^{\frac{1}{2}} & \boldsymbol{W}_t^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_{t-1} \\ \boldsymbol{z}_{\nu,t} \\ \boldsymbol{z}_{\omega,t} \end{bmatrix} \\
&\sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{A}_t\boldsymbol{x}_{t-1|t-1} + \boldsymbol{v}_t \\ \boldsymbol{C}_t\boldsymbol{A}_t\boldsymbol{x}_{t-1|t-1} + \boldsymbol{C}_t\boldsymbol{v}_t + \boldsymbol{w}_t \end{bmatrix}, \boldsymbol{\Sigma} \right),
\end{aligned} \tag{38}$$

    where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{A}_t\boldsymbol{P}_{t-1|t-1}\boldsymbol{A}_t^\top + \boldsymbol{V}_t & (\boldsymbol{A}_t\boldsymbol{P}_{t-1|t-1}\boldsymbol{A}_t^\top + \boldsymbol{V}_t)\boldsymbol{C}_t^\top \\ \boldsymbol{C}_t(\boldsymbol{A}_t\boldsymbol{P}_{t-1|t-1}\boldsymbol{A}_t^\top + \boldsymbol{V}_t)^\top & \boldsymbol{C}_t(\boldsymbol{A}_t\boldsymbol{P}_{t-1|t-1}\boldsymbol{A}_t^\top + \boldsymbol{V}_t)\boldsymbol{C}_t^\top + \boldsymbol{W}_t \end{bmatrix},$$

    from which we can compute the marginal distributions as

$$\boldsymbol{x}_t|Y_{t-1} \sim \mathcal{N}\Big(\underbrace{\boldsymbol{A}_t\boldsymbol{x}_{t-1|t-1} + \boldsymbol{v}_t}_{=:\hat{\boldsymbol{x}}_t}, \underbrace{\boldsymbol{A}_t\boldsymbol{P}_{t-1|t-1}\boldsymbol{A}_t^\top + \boldsymbol{V}_t}_{=:\hat{\boldsymbol{P}}_t} \Big), \quad \boldsymbol{y}_t|Y_{t-1} \sim \mathcal{N}\Big(\underbrace{\boldsymbol{C}_t\hat{\boldsymbol{x}}_t + \boldsymbol{w}_t}_{=:\hat{\boldsymbol{y}}_t}, \underbrace{\boldsymbol{C}_t\hat{\boldsymbol{P}}_t\boldsymbol{C}_t^\top + \boldsymbol{W}_t}_{=:\hat{\boldsymbol{Q}}_t} \Big)$$

  - Once $\boldsymbol{y}_t$ is observed, it follows from (38) that the posterior $\boldsymbol{x}_t|Y_t$ is

$$\boldsymbol{x}_t|Y_t \sim \mathcal{N}\left(\underbrace{\hat{\boldsymbol{x}}_t + \frac{\hat{\boldsymbol{P}}_t\boldsymbol{C}_t^\top}{\hat{\boldsymbol{Q}}_t}(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)}_{=:\boldsymbol{x}_{t|t}}, \underbrace{\hat{\boldsymbol{P}}_t - \frac{\hat{\boldsymbol{P}}_t\boldsymbol{C}_t^\top}{\hat{\boldsymbol{Q}}_t}\hat{\boldsymbol{Q}}_t\left(\frac{\hat{\boldsymbol{P}}_t\boldsymbol{C}_t^\top}{\hat{\boldsymbol{Q}}_t}\right)^\top}_{=:\boldsymbol{P}_{t|t}} \right) \tag{39}$$

- **Kalman filter equations**
  - Priors and posteriors can be incrementally computed in the following sequential manner:
    1. Given the posterior $\boldsymbol{x}_{t-1}|Y_{t-1} \sim \mathcal{N}(\boldsymbol{x}_{t-1|t-1}, \boldsymbol{P}_{t-1|t-1})$ from the previous period, compute the prior:

$$\boldsymbol{x}_t|Y_{t-1} \sim \mathcal{N}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{P}}_t) \quad \text{where} \quad \begin{aligned} \hat{\boldsymbol{x}}_t &= \boldsymbol{A}_t\boldsymbol{x}_{t-1|t-1} + \boldsymbol{v}_t \\ \hat{\boldsymbol{P}}_t &= \boldsymbol{A}_t\boldsymbol{P}_{t-1|t-1}\boldsymbol{A}_t^\top + \boldsymbol{V}_t \end{aligned} \tag{40}$$

    2. Given the prior $\boldsymbol{x}_t|Y_{t-1} \sim \mathcal{N}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{P}}_t)$, compute the forecast:

$$\boldsymbol{y}_t|Y_{t-1} \sim \mathcal{N}(\hat{\boldsymbol{y}}_t, \hat{\boldsymbol{Q}}_t) \quad \text{where} \quad \begin{aligned} \hat{\boldsymbol{y}}_t &= \boldsymbol{C}_t\hat{\boldsymbol{x}}_t + \boldsymbol{w}_t \\ \hat{\boldsymbol{Q}}_t &= \boldsymbol{C}_t\hat{\boldsymbol{P}}_t\boldsymbol{C}_t^\top + \boldsymbol{W}_t \end{aligned} \tag{41}$$

    3. Compute the Kalman gain

$$\boldsymbol{K}_t = \hat{\boldsymbol{P}}_t\boldsymbol{C}_t^\top\hat{\boldsymbol{Q}}_t^{-1} \tag{42}$$

    4. Once $\boldsymbol{y}_t$ is observed, compute forecast error $\hat{\boldsymbol{q}}_t$ as

$$\hat{\boldsymbol{q}}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t \tag{43}$$

       and derive the posterior distribution

$$\boldsymbol{x}_t|Y_t \sim \mathcal{N}(\boldsymbol{x}_{t|t}, \boldsymbol{P}_{t|t}) \quad \text{where} \quad \begin{aligned} \boldsymbol{x}_{t|t} &= \hat{\boldsymbol{x}}_t + \boldsymbol{K}_t\hat{\boldsymbol{q}}_t \\ \boldsymbol{P}_{t|t} &= \hat{\boldsymbol{P}}_t - \boldsymbol{K}_t\hat{\boldsymbol{Q}}_t\boldsymbol{K}_t^\top \end{aligned} \tag{44}$$

  - Equations (40)–(44) are called the Kalman filter equations

- **Initialization for stationary state process**
  - Consider the case where $\boldsymbol{v}_t = \boldsymbol{v}$, $\boldsymbol{A}_t = \boldsymbol{A}$, and $\rho(\boldsymbol{A}) < 1$, where we define

$$\rho(\boldsymbol{A}) := \max_{\lambda \in \sigma(\boldsymbol{A})} |\lambda| \quad \text{where } \sigma(\boldsymbol{A}) \text{ is the set of all eigenvalues of } \boldsymbol{A},$$

    which makes sure that $\boldsymbol{A} \neq \boldsymbol{I}$ and $\lim_{\tau \to \infty} \boldsymbol{A}^\tau = \boldsymbol{O}$
  - Model (28) suggests that for any t and $\tau \geq 1$, we may write

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{v} + \boldsymbol{\nu}_t = \boldsymbol{A}(\boldsymbol{A}\boldsymbol{x}_{t-2} + \boldsymbol{v} + \boldsymbol{\nu}_{t-1}) + \boldsymbol{v} + \boldsymbol{\nu}_t = \boldsymbol{A}^\tau \boldsymbol{x}_{t-\tau} + \sum_{s=0}^{\tau-1} \boldsymbol{A}^s(\boldsymbol{v} + \boldsymbol{\nu}_{t-s}),$$

    which, since $\rho(\boldsymbol{A}) < 1$, may even be written as

$$\boldsymbol{x}_t = \sum_{s=0}^{\infty} \boldsymbol{A}^s(\boldsymbol{v} + \boldsymbol{\nu}_{t-s}) \quad \forall t \tag{45}$$

  - Expression (45) implies that $\boldsymbol{x}_t$ is a stationary process in the sense that

$$\mathbb{E}[\boldsymbol{x}_t] = \mathbb{E}[\boldsymbol{x}_{t-k}] \quad \text{and} \quad \mathbb{V}[\boldsymbol{x}_t] = \mathbb{V}[\boldsymbol{x}_{t-k}] \quad \forall k \tag{46}$$

  - It follows from (46) and (28) that the unconditional mean and variance must satisfy

$$\mathbb{E}[\boldsymbol{x}_t] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}_t] + \mathbb{E}[\boldsymbol{\nu}_t], \quad \mathbb{V}[\boldsymbol{x}_t] = \boldsymbol{A}\mathbb{V}[\boldsymbol{x}_t]\boldsymbol{A}^\top + \mathbb{V}[\boldsymbol{\nu}_t], \quad \forall t = 0, 1, 2, \ldots$$

    which can be solved for $\mathbb{E}[\boldsymbol{x}_t]$ and $\mathbb{V}[\boldsymbol{x}_t]$ as[2]

$$\mathbb{E}[\boldsymbol{x}_t] = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}, \quad \operatorname{vec}(\mathbb{V}[\boldsymbol{x}_t]) = (\boldsymbol{I} - \boldsymbol{A} \otimes \boldsymbol{A})^{-1} \operatorname{vec}(\boldsymbol{V}) \quad \forall t = 0, 1, 2, \ldots$$

  - Therefore, if the state process is stationary, we can use the following initial distribution:

$$\boldsymbol{x}_0 \sim \mathcal{N}(\mathbb{E}[\boldsymbol{x}_0], \mathbb{V}[\boldsymbol{x}_0]) = \mathcal{N}\left((\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}, \operatorname{vec}_{m \times m}^{-1}((\boldsymbol{I} - \boldsymbol{A} \otimes \boldsymbol{A})^{-1} \operatorname{vec}(\boldsymbol{V}))\right) \tag{47}$$

- **Initialization for non-stationary state process**
  - If the state process is non-stationary, then the strategy described above does not work, in which case we use the (approximate) uninformative prior
  - To be more precise, we put $\boldsymbol{P}_{0|0} = \kappa \boldsymbol{I}$ and use a sufficiently large $\kappa \in \mathbb{R}$

[2] Note that one can directly take the expectation of (45) and obtain

$$\mathbb{E}[\boldsymbol{x}_t] = \mathbb{E}\left[\sum_{s=0}^{\infty} \boldsymbol{A}^s(\boldsymbol{v} + \boldsymbol{\nu}_{t-s})\right] = \sum_{s=0}^{\infty} \boldsymbol{A}^s \mathbb{E}[\boldsymbol{v} + \boldsymbol{\nu}_{t-s}] = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}.$$

Similarly, directly taking the variance of (45) yields

$$\mathbb{V}[\boldsymbol{x}_t] = \mathbb{V}\left[\sum_{s=0}^{\infty} \boldsymbol{A}^s(\boldsymbol{v} + \boldsymbol{\nu}_{t-s})\right] = \sum_{s=0}^{\infty} \mathbb{V}[\boldsymbol{A}^s \boldsymbol{\nu}_{t-s}] = \sum_{s=0}^{\infty} \boldsymbol{A}^s \mathbb{V}[\boldsymbol{\nu}_{t-s}](\boldsymbol{A}^s)^\top = \sum_{s=0}^{\infty} \boldsymbol{A}^s \boldsymbol{V} (\boldsymbol{A}^s)^\top,$$

which, since $\rho(\boldsymbol{A} \otimes \boldsymbol{A}) < 1$ because of $\rho(\boldsymbol{A}) < 1$, implies

$$\operatorname{vec}(\mathbb{V}[\boldsymbol{x}_t]) = \sum_{s=0}^{\infty} \operatorname{vec}(\boldsymbol{A}^s \boldsymbol{V} (\boldsymbol{A}^s)^\top) = \sum_{s=0}^{\infty} (\boldsymbol{A}^s \otimes \boldsymbol{A}^s) \operatorname{vec}(\boldsymbol{V}) = \sum_{s=0}^{\infty} (\boldsymbol{A} \otimes \boldsymbol{A})^s \operatorname{vec}(\boldsymbol{V}) = (\boldsymbol{I} - \boldsymbol{A} \otimes \boldsymbol{A})^{-1} \operatorname{vec}(\boldsymbol{V})$$

3.3 Parameter estimation

- **Maximum likelihood estimator**
  - In case (some of) the model parameters $\boldsymbol{\theta} := (\boldsymbol{A}_t, \boldsymbol{B}_t, \boldsymbol{C}_t, \boldsymbol{w}_t, \boldsymbol{V}_t, \boldsymbol{W}_t)_{t\geq 1}$ are unknown, we estimate them as follows
  - The joint distribution of $Y_n := (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$ is

$$p(Y_n) = p(\boldsymbol{y}_n|Y_{n-1})p(Y_{n-1}) = p(\boldsymbol{y}_n|Y_{n-1})p(\boldsymbol{y}_{n-1}|Y_{n-2})p(Y_{n-2}) = \prod_{t=1}^{n} p(\boldsymbol{y}_t|Y_{t-1}), \tag{48}$$

where (41) implies

$$p(\boldsymbol{y}_t|Y_{t-1}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\hat{\boldsymbol{Q}}_t|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{y}_t-\hat{\boldsymbol{y}}_t)^\top \hat{\boldsymbol{Q}}_t^{-1}(\boldsymbol{y}_t-\hat{\boldsymbol{y}}_t)} \tag{49}$$

  - MLE of $\boldsymbol{\theta}$ is the one that maximizes the log likelihood

$$\begin{aligned}
\ln L(\boldsymbol{\theta}; Y_n) &:= \ln\left(p(Y_n)\right) = \sum_{t=1}^{n} \ln\left(p(\boldsymbol{y}_t|Y_{t-1})\right) \\
&= \sum_{t=1}^{n}\left(-\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln(|\hat{\boldsymbol{Q}}_t|) - \frac{1}{2}(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)^\top \hat{\boldsymbol{Q}}_t^{-1}(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)\right) \\
&= -\frac{np}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\left(\ln(|\hat{\boldsymbol{Q}}_t|) + (\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)^\top \hat{\boldsymbol{Q}}_t^{-1}(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)\right)
\end{aligned} \tag{50}$$

- **Example**
  - Consider the case where the evolution of state vector, $\boldsymbol{x}_t = (x_{1,t}, x_{2,t}, x_{3,t})$, is governed by the following dynamical system:

$$\begin{bmatrix} x_{1,t} \\ x_{2,t} \\ x_{3,t} \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \\ x_{3,t-1} \end{bmatrix} + \begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix} u_t + \begin{bmatrix} \nu_{1,t} \\ \nu_{2,t} \\ \nu_{3,t} \end{bmatrix}$$

where

$$\begin{bmatrix} \nu_{1,t} \\ \nu_{2,t} \\ \nu_{3,t} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}\right)$$

  - Suppose that we know that the unit step forcing is introduced after time $t = 1$:

$$u_t = \begin{cases} 1 & t \geq 1 \\ 0 & t \leq 0 \end{cases}$$

  - Assume that:
    - we can observe the value of $x_{2,t}$ for $t \geq 1$
    - the values of $x_{1,t}$ and $x_{3,t}$ are not directly observable, but we can observe the sum $\sum_{i=1}^{3} x_{i,t}$
    - there is no measurement error
  - So the measurement vector $\boldsymbol{y}_t = (y_{1,t}, y_{2,t})$ is given by

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_{1,t} \\ x_{2,t} \\ x_{3,t} \end{bmatrix} \qquad \forall t = 1, 2, \ldots, n$$

  - We want to estimate the value of $\boldsymbol{\theta} = (a_{11}, a_{21}, a_{22}, a_{23}, a_{32}, a_{33}, b, \sigma_1, \sigma_2, \sigma_3)$ based on the sample $Y_n = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ of size n

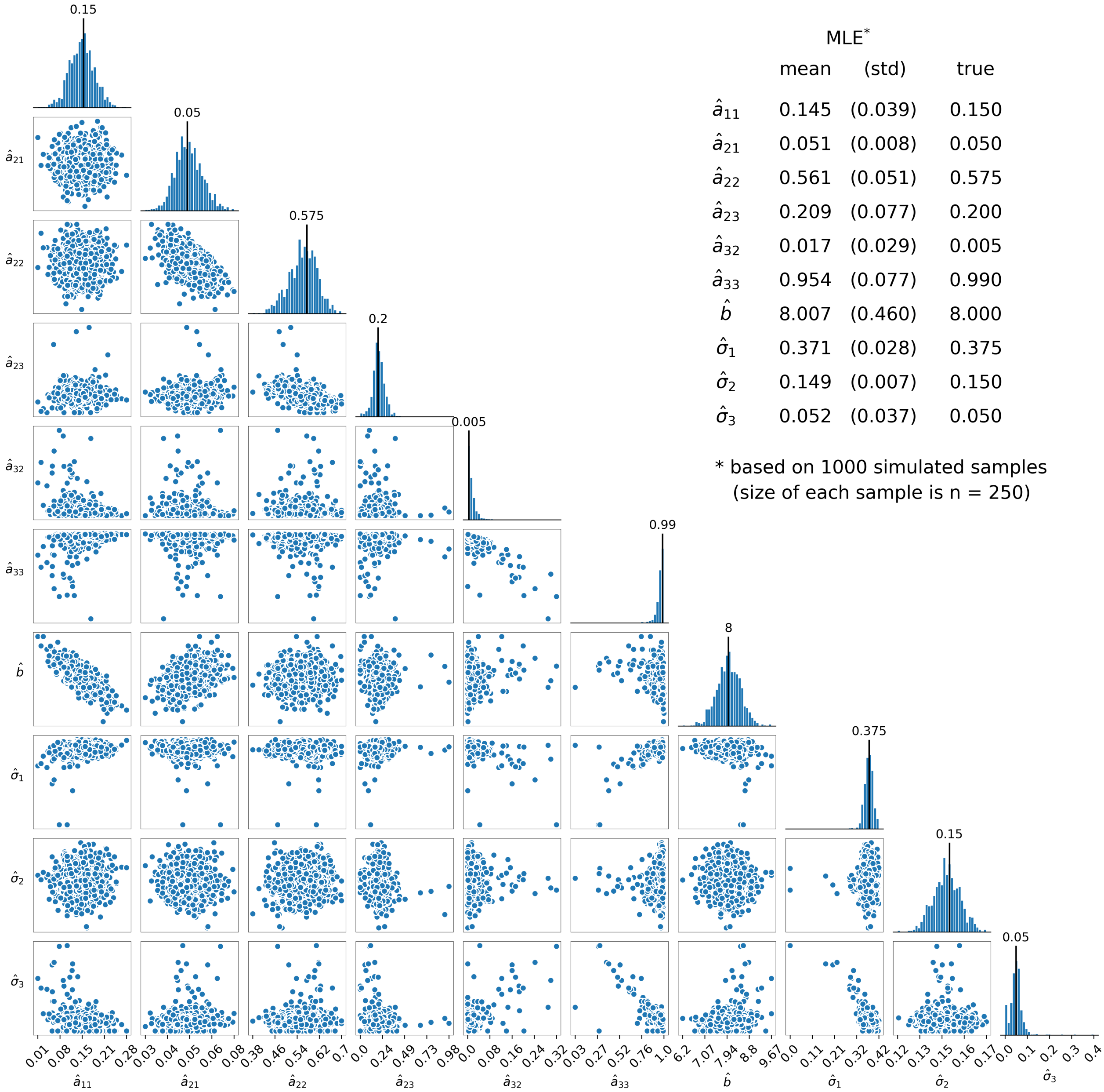

Figure 4: Pairs plot of MLE $\hat{\boldsymbol{\theta}}$ (1000 simulated samples of size $n = 250$).

- Figure 4 shows the estimated values of $\boldsymbol{\theta}$, where
  1. I first fix the true parameter values $\boldsymbol{\theta}$ as listed in the figure (where the model is stationary because $\rho(\boldsymbol{A}) < 1$)
  2. Using this true $\boldsymbol{\theta}$, I generate a simulated sample $Y_n = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ of size n:
     - randomly draw an initial state $\boldsymbol{x}_0$ based on (47) with $\boldsymbol{v} = \boldsymbol{0}$ (since $u_t = 0$ for all $t \leq 0$)
     - then randomly draw $\boldsymbol{\nu}_1$ and compute $\boldsymbol{x}_1$, which in turn determines $\boldsymbol{y}_1$
     - then randomly draw $\boldsymbol{\nu}_2$ and compute $\boldsymbol{x}_2$, which in turn determines $\boldsymbol{y}_2$
     - ...
  3. For each $\tilde{\boldsymbol{\theta}}$, I combine the sample Y_n and the Kalman filter equations (40)–(44) to compute its log likelihood $\ln(L(\tilde{\boldsymbol{\theta}}; Y_n))$ based on (50) and find the one that maximizes it:
$$\hat{\boldsymbol{\theta}} = \arg\max_{\tilde{\boldsymbol{\theta}}} \ln(L(\tilde{\boldsymbol{\theta}}; Y_n))$$
  4. I repeat Steps 2-3 for 1000 times to generate the simulated distribution of $\hat{\boldsymbol{\theta}}$