A crash course of Bayesian DSGE estimation

I. Basic concepts

Takeki Sunakawa

Hitotsubashi University

May 12, 2022 @ Keio University

Introduction

- DSGE models are standard in modern macroeconomic analysis, which are used in not only academia but also policy institutions such as central banks.
- The models are useful as they fit with macroeconomic data well (Smets and Wouters, 2007).
- The models can be easily estimated by using Dynare (https://www.dynare.org/) without any programming.

Introduction, cont'd

- Even though we can estimate the models without programming, (maybe) we need to know:
  - How to solve for **rational expectation equilibrium (REE)**
  - What is **Kalman filter**, which constructs the likelihood of a given set of parameters and observables
  - What is **Bayesian inference**
  - What is **Metropolis-Hastings algorithm**, which approximates the posterior distribution of parameters
- We will go through these issues step-by-step.

A bird's-eye view: How it works?

- Taking a set of model parameters θ as given, put the REE solution and observation equation together to form a state-space representation.

$$
\begin{aligned}
y_t &= A(\theta) + B(\theta)x_t + e_t, \qquad & e_t \sim N(0, H(\theta)) \\
x_t &= P(\theta)x_{t-1} + Q(\theta)\epsilon_t, \qquad & \epsilon_t \sim N(0, S_e(\theta))
\end{aligned}
$$

- Then having the data Y, we calculate the likelihood function of the parameters, $L(Y|\theta)$, from the state-space representation using the Kalman filter.

- We conjecture a form of the prior distribution of the parameters, $p(\theta)$.

- Using the Bayes' theorem, we have the posterior distribution of the parameters.

$$p(\theta|Y) \propto p(\theta)L(Y|\theta)$$

- We use the Metropolis-Hasting algorithm, a Monte-Carlo sampling method, to approximate the shape of the posterior distribution.
- We do inferences based on the posterior distribution.

Textbooks

- Bayesian DSGE estimation
    - Herbst and Schorfheide “Bayesian Estimation of DSGE models” (compact, a bit difficult)
    - Miao “Economic Dynamics: Discrete Time (2d. ed.)” (introductory)
    - Dejong and Dave “Structural Macroeconometrics (2d. ed.)” (broad)
- Bayesian econometrics/statistics
    - Koop “Bayesian Econometrics”
    - 渡部「ベイズ統計学入門」
- Time-series models and filtering
    - 森平「経済・ファイナンスのためのカルマンフィルター入門」(intuitive)
    - Hamilton “Time Series Analysis” (very popular)
    - Durbin and Koopman “Time Series Analysis by State Space Methods (2d. ed.)” （第1版の邦訳「状態空間モデリングによる時系列入門」シーエーピー出版）(comprehensive)
    - (Maybe more)

Rational Expectation Equilibrium

Linear rational expectation models

- We want to solve the following equilibrium conditions:

$$\mathcal{A}E_t\{x_{t+1}\} + \mathcal{B}x_t + \mathcal{C}x_{t-1} + \mathcal{E}\epsilon_t = 0$$

where

- x_t is a vector of size n that collects all the endogenous model variables
- $E_t\{\cdot\}$ is the expectation operator, conditional on information available at time t
- $\mathcal{A}, \mathcal{B}, \mathcal{C}$ are $n \times n$ matrices of structural parameters
- ϵ_t is a vector of zero mean i.i.d. exogenous innovations of size m, and $\mathcal{E}$ is an $n \times m$ matrix of structural parameters

- A solution to Eq. (1) is given by

$$x_t = P x_{t-1} + Q \epsilon_t$$

where P is an $n \times n$ matrix and Q is an $n \times m$ matrix.

- Solving (1) for the REE (by assuming its uniqueness) amounts to finding the matrices P and Q.

Example

- We consider a variable q determined by the following schedule:

$$q_t = \beta(1-\rho)E_t q_{t+1} + \rho q_{t-1} - \sigma r_t + u_t$$

$$r_t = \phi q_t$$

- Substituting the latter to the former,

$$\beta(1-\rho)E_t q_{t+1} - (1+\sigma\phi)q_t + \rho q_{t-1} + u_t = 0$$

- How to solve this equation?

Undetermined coefficient method

- We assume that $q_t = aq_{t-1} + bu_t$.

- Substituting it into the equilibrium condition,

$$\beta(1-\rho)E_t(aq_t + bu_{t+1}) \\ -(1+\sigma\phi)q_t + \rho q_{t-1} + u_t = 0$$

$$\Leftrightarrow \beta(1-\rho)(aE_tq_t + bE_tu_{t+1}) \\ -(1+\sigma\phi)q_t + \rho q_{t-1} + u_t = 0$$

$$\Leftrightarrow \beta(1-\rho)aq_t - (1+\sigma\phi)q_t + \rho q_{t-1} + u_t = 0$$

$$\Leftrightarrow [\beta a(1-\rho) \ - (1+\sigma\phi)](aq_{t-1} + bu_t) + \rho q_{t-1} + u_t$$

$$\Leftrightarrow (\beta a^2(1-\rho) \ - a(1+\sigma\phi) + \rho)q_{t-1} \\ +(\beta ab(1-\rho) \ - b(1+\sigma\phi) + 1)u_t = 0$$

- This equation must hold for any q_{t-1} and u_t, which implies

$$\beta a^2(1-\rho) - a(1+\sigma\phi) + \rho = 0,$$
$$\beta a b(1-\rho) - b(1+\sigma\phi) + 1 = 0,$$

which can be solved for a and b.

- The first equation is a second-order polynomial of a, so there are two solutions.
- We pick up the solution satisfying $|a| < 1$, as it yields the stability of q_t.

Numerical example

- TBD

- Taking the equilibrium conditions as given, Dynare can solve them for the REE.
    - Usually we do log-linearization of the equilibrium conditions by hands (which can be very messy!).
    - Dynare can even do log-linearization of the equilibrium conditions (more in the next time).

A log-linearized New Keynesian model (Herbst and Schorfheide, 2015)

- The equilibrium conditions are

$$\hat{c}_t = E_t \hat{c}_{t+1} - \tau^{-1}\left(\hat{R}_t - E_t \hat{\pi}_{t+1} - \rho_z \hat{z}_t\right)$$
$$\hat{\pi}_t = E_t \hat{\pi}_{t+1} + \kappa \hat{c}_t$$
$$\hat{R}_t = \rho_R \hat{R}_{t-1} + (1 - \rho_R)(\psi_1 \hat{\pi}_t + \psi_2 \hat{c}_t) + \epsilon_{R,t}$$
$$\hat{y}_t = \hat{c}_t + \hat{g}_t$$
$$\hat{g}_t = \rho_g \hat{g}_{t-1} + \epsilon_{g,t}$$
$$\hat{z}_t = \rho_z \hat{z}_{t-1} + \epsilon_{z,t}$$

State equation

- The equilibrium conditions are summarized into

$$\mathcal{A}E_t\{x_{t+1}\} + \mathcal{B}x_t + \mathcal{C}x_{t-1} + \mathcal{E}\epsilon_t = 0$$

where

$$x_t = \left[\hat{c}_t, \hat{\pi}_t, \hat{R}_t, \hat{y}_t, \hat{g}_t, \hat{z}_t\ \right]'$$
$$\epsilon_t = \left[\epsilon_{z,t}, \epsilon_{g,t}, \epsilon_{R,t}\right]'$$

- The REE solution to the equilibrium condition

$$x_t = Px_{t-1} + Q\epsilon_t, \qquad \epsilon_t \sim N(0, \mathrm{S}_e)$$

This is **the state equation**.

Observation equation

- We have observed variables, which linked to model variables by

$$\Delta y_t^{obs} = \gamma^{(Q)} + (\hat{y}_t - \hat{y}_{t-1} + \hat{z}_t)$$
$$\pi_t^{obs} = \pi^{(A)} + 4\hat{\pi}_t$$
$$R_t^{obs} = \pi^{(A)} + r^{(A)} + 4\gamma^{(Q)} + 4\hat{R}_t$$

Explain more…

- These equations are summarized into

$$y_t = A + Bx_t + e_t, \qquad e_t \sim N(0, H)$$

This is **the observation equation**.

State-space representation

- Then we have a state-space representation:

$$\begin{aligned} y_t &= A(\theta) + B(\theta)x_t + e_t, \quad && e_t \sim N(0, H(\theta)) \\ x_t &= P(\theta)x_{t-1} + Q(\theta)\epsilon_t, \quad && \epsilon_t \sim N(0, S_e(\theta)) \end{aligned}$$

where $\theta = \left[\tau, \kappa, \psi_1, \psi_2, \rho_R, \rho_g, \rho_z, r^{(A)}, \pi^{(A)}, \gamma^{(Q)}, \sigma_R, \sigma_g, \sigma_z\right]'$.

This is a linear Gaussian state-space model, to which we can apply Kalman filter.

Kalman Filter

(based on 森平「経済・ファイナンスのためのカルマンフィルター入門」, 2019)

Kalman filter

- Using Kalman filter, we estimate a sequence (of distributions) of unobservable state variables from observable variables.
- E.g., suppose that the stock price S_t can be decomposed into its true value α_t and disturbance e_t :

$$S_t = \alpha_t + e_t$$

- How to infer the sequence of $\{\alpha_t\}$, taking that of $\{S_t\}$ as given?

State-space representation

- Suppose that α_t follows the AR(1) process:
$$\alpha_t = d + T\alpha_{t-1} + \varepsilon_t$$
- A state-space representation is given by

  Observation equation: $S_t = \alpha_t + e_t$

  State equation: $\alpha_t = d + T\alpha_{t-1} + \varepsilon_t$

  for $t = 1,2, \dots, N$,

  where $e_t \sim N(0, \sigma_e^2)$, $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$

  $E[e_t \varepsilon_s] = 0$ all s, t

  $E[e_t e_s] = 0$ and $E[\varepsilon_t \varepsilon_s] = 0$ all $s \neq t$

Initial value

- We need to know the value of α_0 initially at $t = 1$:

$$\alpha_1 = d + T\alpha_0 + \varepsilon_1$$

- Suppose

$$\alpha_0 \sim N(\hat{\alpha}_0, \hat{\Sigma}_0)$$

and $E[e_t \alpha_0] = 0$, $E[\varepsilon_t \alpha_0] = 0$ all t.

- Under this assumption, we sequentially estimate the mean and variance of α_t.
- α_t for $t = 0,1,\ldots,N$ is uncertain and its certain value is unknown *either a priori or a posteriori*.

Information set

- We compute the *conditional* mean and variance using the information available at time t.
- In the previous example, the information set at time t is given by

$$\Omega_t = \{S_1, S_2, \dots, S_{t-1}, S_t\}$$

Similarly, the information set at time $t-1$ is given by

$$\Omega_{t-1} = \{S_1, S_2, \dots, S_{t-1}\}$$

Forecasting and filtering

- We estimate the mean and variance of α_t at time t, depending on the information set:
  - One-step ahead forecasting

$$\hat{\alpha}_{t|t-1} = E[\alpha_t|\Omega_{t-1}], \qquad \hat{\Sigma}_{t|t-1} = E[\Sigma_t|\Omega_{t-1}]$$

  - Filtering

$$\hat{\alpha}_{t|t} = E[\alpha_t|\Omega_t], \qquad \hat{\Sigma}_{t|t} = E[\Sigma_t|\Omega_t]$$

  - (Smoothing)

$$\hat{\alpha}_{t|N} = E[\alpha_t|\Omega_N], \qquad \hat{\Sigma}_{t|N} = E[\Sigma_t|\Omega_N]$$

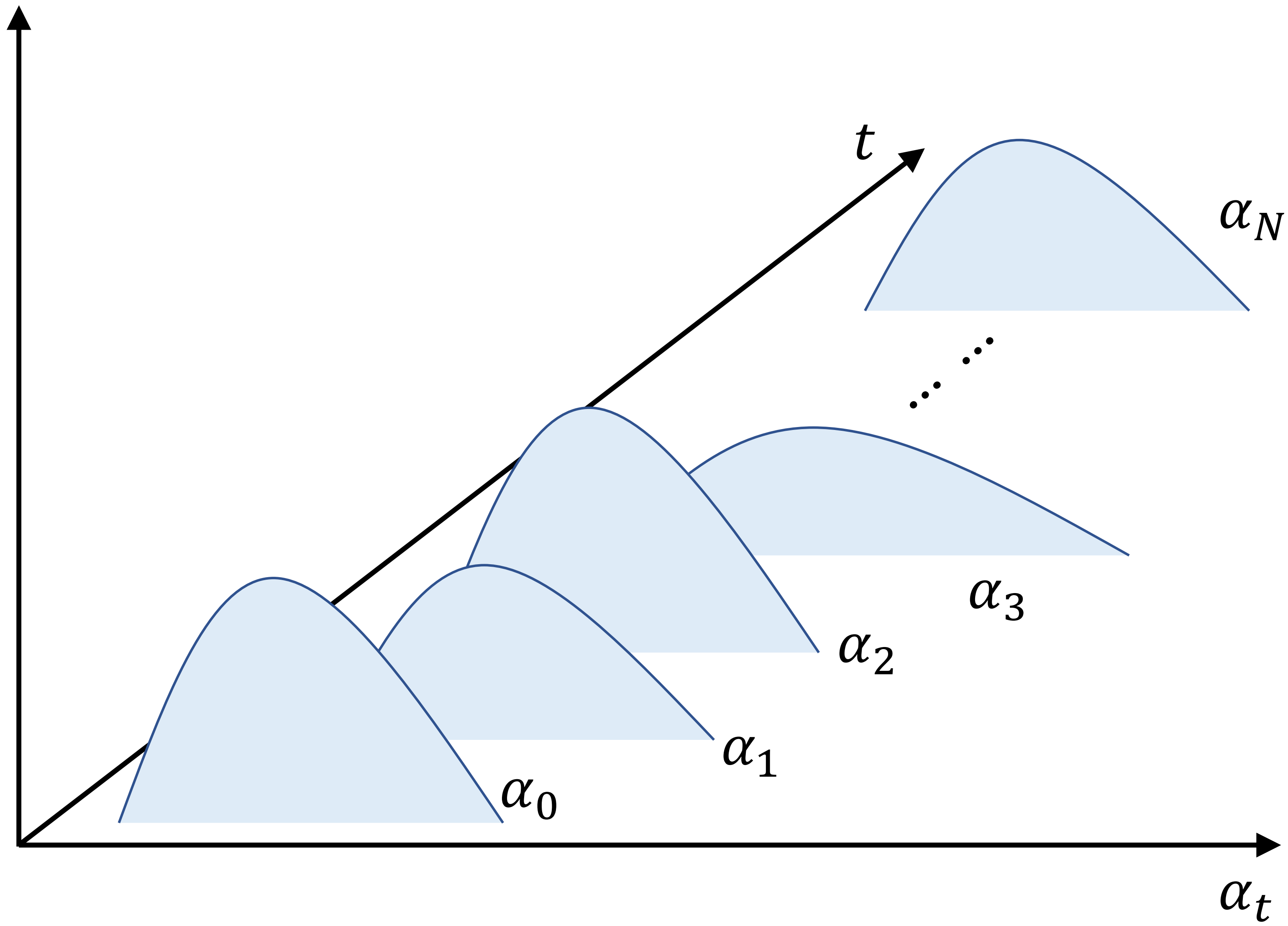
- In forecasting, we obtain the distribution at time t conditioned on the information available at time t-1

$$\alpha_t | \Omega_{t-1} \sim N(\hat{\alpha}_{t|t-1}, \hat{\Sigma}_{t|t-1})$$

- In filtering, we obtain the distribution at time t conditioned on the information available at time t

$$\alpha_t | \Omega_t \sim N(\hat{\alpha}_{t|t}, \hat{\Sigma}_{t|t})$$

The distribution of α_t evolves as time goes on.



Sequential updating

- Let's consider a simple example of sequential updating.
- Let y_t be the data at time t and μ_t be the mean of the data available until time t. That is,

$$\mu_t = \frac{1}{t}\sum_{n=1}^{t} y_n = \frac{1}{t}(y_1 + y_2 + \cdots y_t)$$

- We can also calculate the mean sequentially

$$\mu_1 = \frac{1}{1} y_1 = y_1$$

$$\mu_2 = \frac{1}{2} y_1 + \frac{1}{2} y_2 = \frac{1}{2} \mu_1 + \frac{1}{2} y_2$$

$$\mu_3 = \frac{1}{3} y_1 + \frac{1}{3} y_2 + \frac{1}{3} y_3 = \frac{2}{3} \mu_2 + \frac{1}{3} y_3$$

$$\vdots$$

$$\mu_t = \left(\frac{t-1}{t}\right) \mu_{t-1} + \frac{1}{t} y_t = (1 - K_t)\mu_{t-1} + K_t y_t$$

That is, the mean at time t is a weighted average of the previous mean at time t-1 and the new information at time t.

- Or, we can write it as

$$\mu_t = \mu_{t-1} + K_t(y_t - \mu_{t-1})$$

$y_t - \mu_{t-1}$ is a "surprise" at time t.

- We update the mean by the surprise with a weight K_t.

Sequential updating in Kalman filter

- Now, we go back to the state-space representation (a slightly more general version)

  Observation equation: $S_t = a + b\alpha_t + e_t, \quad e_t \sim N(0, \sigma_e^2)$

  State equation: $\alpha_t = d + T\alpha_{t-1} + \varepsilon_t, \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$

- We can calculate

$$\hat{\alpha}_{t|t-1} = d + T\hat{\alpha}_{t-1|t-1}$$

$$\hat{\Sigma}_{t|t-1} = T^2\hat{\Sigma}_{t-1|t-1} + \sigma_\varepsilon^2$$

$$\hat{S}_{t|t-1} = a + b\hat{\alpha}_{t|t-1}$$

$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + K_t(S_t - \hat{S}_{t|t-1})$$

$$\hat{\Sigma}_{t|t} = (1 - bK_t)\hat{\Sigma}_{t|t-1}$$

where $K_t = \frac{b\hat{\Sigma}_{t|t-1}}{b^2\hat{\Sigma}_{t|t-1}+\sigma_e^2}$ is the Kalman gain.

Forecasting

- In forecasting, we need to calculate $\hat{\alpha}_{t|t-1}$ and $\hat{\Sigma}_{t|t-1}$:

$$\alpha_t|\Omega_{t-1} \sim N(\hat{\alpha}_{t|t-1}, \hat{\Sigma}_{t|t-1})$$

- Taking $\hat{\alpha}_{t-1|t-1}$ and $\hat{\Sigma}_{t-1|t-1}$ as given, it is straightforward to derive $\hat{\alpha}_{t|t-1}$ from the state space representation.

$$\begin{aligned}\hat{\alpha}_{t|t-1} = E[\alpha_t|\Omega_{t-1}] &= E[d + T\alpha_{t-1} + \varepsilon_t|\Omega_{t-1}] \\ &= d + TE[\alpha_{t-1}|\Omega_{t-1}] + 0 \\ &= d + T\hat{\alpha}_{t-1|t-1}\end{aligned}$$

- Similarly, to derive $\hat{\Sigma}_{t|t-1}$, and $\hat{S}_{t|t-1}$,

$$
\begin{aligned}
\hat{\Sigma}_{t|t-1} &= Var[\alpha_t|\Omega_{t-1}] \\
&= Var[d + T\alpha_{t-1} + \varepsilon_t|\Omega_{t-1}] \\
&= T^2 Var[\alpha_{t-1}|\Omega_{t-1}] + \sigma_\varepsilon^2 \\
&= T^2 \hat{\Sigma}_{t-1|t-1} + \sigma_\varepsilon^2
\end{aligned}
$$

$$
\begin{aligned}
\hat{S}_{t|t-1} = E[S_t|\Omega_{t-1}] &= E[a + b\alpha_t + e_t|\Omega_{t-1}] \\
&= a + bE[\alpha_t|\Omega_{t-1}] + 0 \\
&= a + b\hat{\alpha}_{t|t-1}
\end{aligned}
$$

Filtering: Mean

- In filtering, we need to update $\hat{\alpha}_{t|t}$ and $\hat{\Sigma}_{t|t}$:

$$\alpha_t|\Omega_t \sim N(\hat{\alpha}_{t|t}, \hat{\Sigma}_{t|t})$$

- The filtering equation for the mean is

$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + K_t(S_t - \hat{S}_{t|t-1})$$

We update the mean by the surprise with a weight K_t. This looks like the equation of sequential updating for the mean:

$$\mu_t = \mu_{t-1} + K_t(y_t - \mu_{t-1})$$

- In Kalman filter, how to compute K_t?

Kalman gain

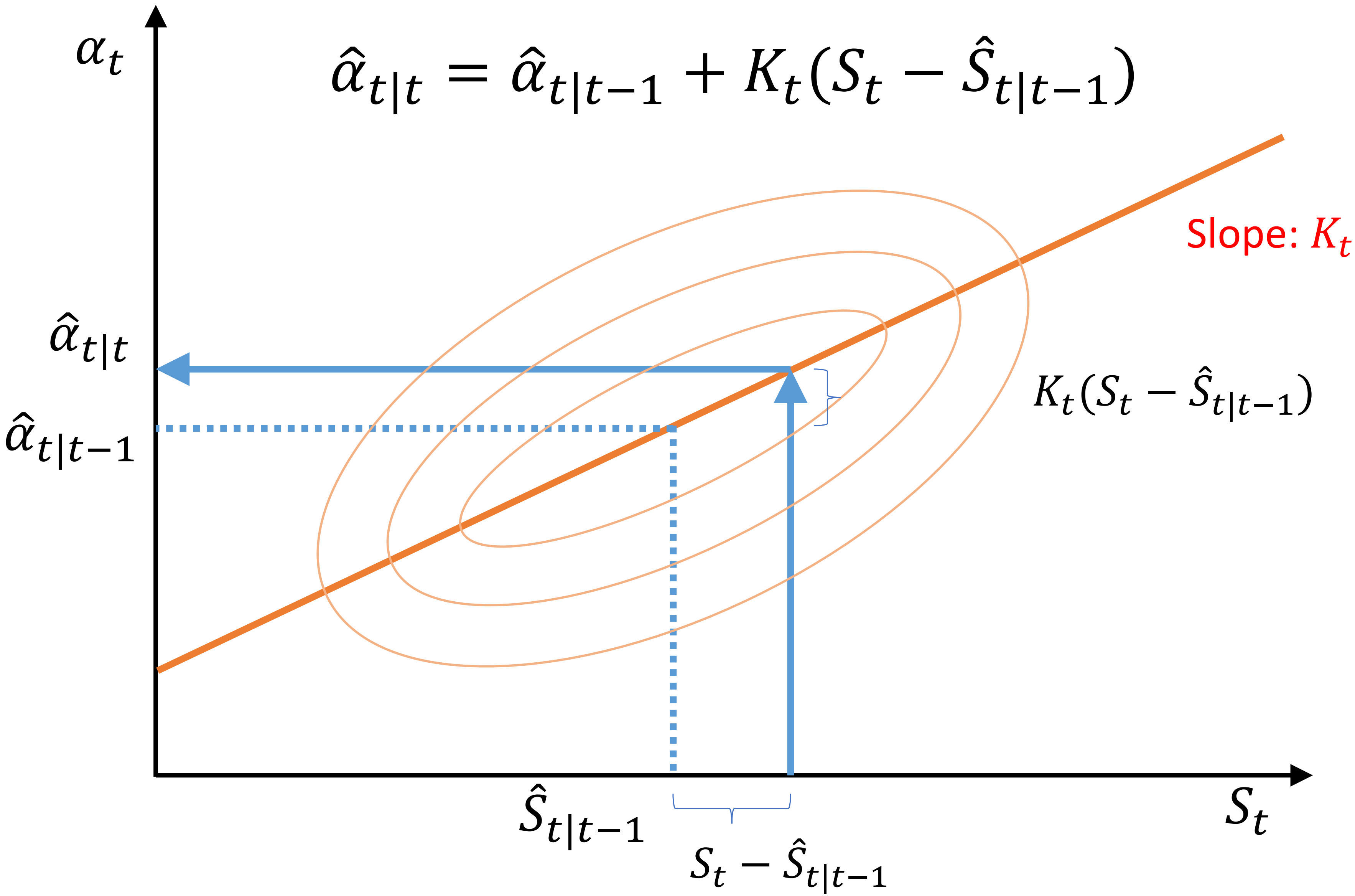
- The Kalman gain K_t can be interpreted as a regression coefficient of

$$\alpha_t = c + K_t S_t + u_t$$

where

$$K_t = \frac{Cov(\alpha_t, S_t|\Omega_{t-1})}{Var(S_t|\Omega_{t-1})} = \frac{b\hat{\Sigma}_{t|t-1}}{b^2\hat{\Sigma}_{t|t-1} + \sigma_e^2}.$$

  - $b\hat{\Sigma}_{t|t-1}$ is the covariance between α_t, S_t conditioned on the information set at time t-1
  - $b^2\hat{\Sigma}_{t|t-1} + \sigma_e^2$ is the variance of S_t conditioned on the information set at time t-1



α_t
$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + K_t(S_t - \hat{S}_{t|t-1})$
Slope: K_t
$\hat{\alpha}_{t|t}$
$\hat{\alpha}_{t|t-1}$
$K_t(S_t - \hat{S}_{t|t-1})$
$\hat{S}_{t|t-1}$
S_t
$S_t - \hat{S}_{t|t-1}$

Filtering: Variance

- The filtering equation for the variance is

$$\hat{\Sigma}_{t|t} = (1 - bK_t)\hat{\Sigma}_{t|t-1}$$

where $bK_t = \frac{b^2\hat{\Sigma}_{t|t-1}}{b^2\hat{\Sigma}_{t|t-1}+\sigma_e^2}$ takes a value between 0 and 1.

- Note that

$$\begin{aligned} Var[S_t|\Omega_{t-1}] &= b^2 Var[\alpha_t|\Omega_{t-1}] + \sigma_e^2 \\ &= b^2\hat{\Sigma}_{t|t-1} + \sigma_e^2 \end{aligned}$$

bK_t is a relative value of uncertainty in α_t.

- The larger bK_t is, the smaller is the filtered variance of α_t than the forecasted variance of α_t.

Algorithm of Kalman filter

0. Set initial values of $(\hat{\alpha}_0, \hat{\Sigma}_0)$ and parameters $(a, b, d, T, \sigma_e, \sigma_\varepsilon)$.

1. Taking $(\hat{\alpha}_{t-1|t-1}, \hat{\Sigma}_{t-1|t-1})$ as given, calculate one-step ahead forecasting at time t

$$\hat{\alpha}_{t|t-1} = d + T\hat{\alpha}_{t-1|t-1}$$
$$\hat{\Sigma}_{t|t-1} = T^2\hat{\Sigma}_{t-1|t-1} + \sigma_\varepsilon^2$$
$$\hat{S}_{t|t-1} = a + b\hat{\alpha}_{t|t-1}$$

2. Filtering at time t

$$K_t = \frac{b\hat{\Sigma}_{t|t-1}}{b^2\hat{\Sigma}_{t|t-1} + \sigma_e^2}$$
$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + K_t(S_t - \hat{S}_{t|t-1})$$
$$\hat{\Sigma}_{t|t} = (1 - bK_t)\hat{\Sigma}_{t|t-1}$$

3. Move one period ahead from t to $t + 1$ and repeat 1-2 until $t = N$.

Numerical example

- We consider the following local model:

$$S_t = \alpha_t + e_t, \qquad e_t \sim N(0, \sigma_e^2)$$
$$\alpha_t = \alpha_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

for $t = 1,2,3,4$ and data $\{S_1, S_2, S_3, S_4\} = \{4.4,4.0,3.5,3.6\}$

0. Set $\hat{\alpha}_{0|0} = 4$, $\hat{\Sigma}_{0|0} = 12$, $\sigma_e = 1$, $\sigma_\varepsilon = 2$.

1. At $t = 1$, one-step ahead forecasts are:

$$\hat{\alpha}_{1|0} = \hat{\alpha}_{0|0} = 4$$
$$\hat{\Sigma}_{1|0} = 1^2\hat{\Sigma}_{0|0} + \sigma_\varepsilon^2 = 12 + 2^2 = 16$$
$$S_{1|0} = \hat{\alpha}_{1|0} = 4$$

2. Filtering: Having $S_1 = 4.4$,

$$K_1 = \frac{\hat{\Sigma}_{1|0}}{\hat{\Sigma}_{1|0} + \sigma_e^2} = \frac{16}{16+1} = 0.941$$

$$\hat{\alpha}_{1|1} = \hat{\alpha}_{1|0} + K_1\left(S_1 - \hat{S}_{1|0}\right)$$

$$= 4 + 0.941 \times (4.4 - 4) = 4.376$$

$$\hat{\Sigma}_{1|1} = (1 - K_1)\hat{\Sigma}_{1|0}$$

$$= (1 - 0.941) \times 16 = 0.941$$

3. We repeat this procedure for $t = 2,3,4$.

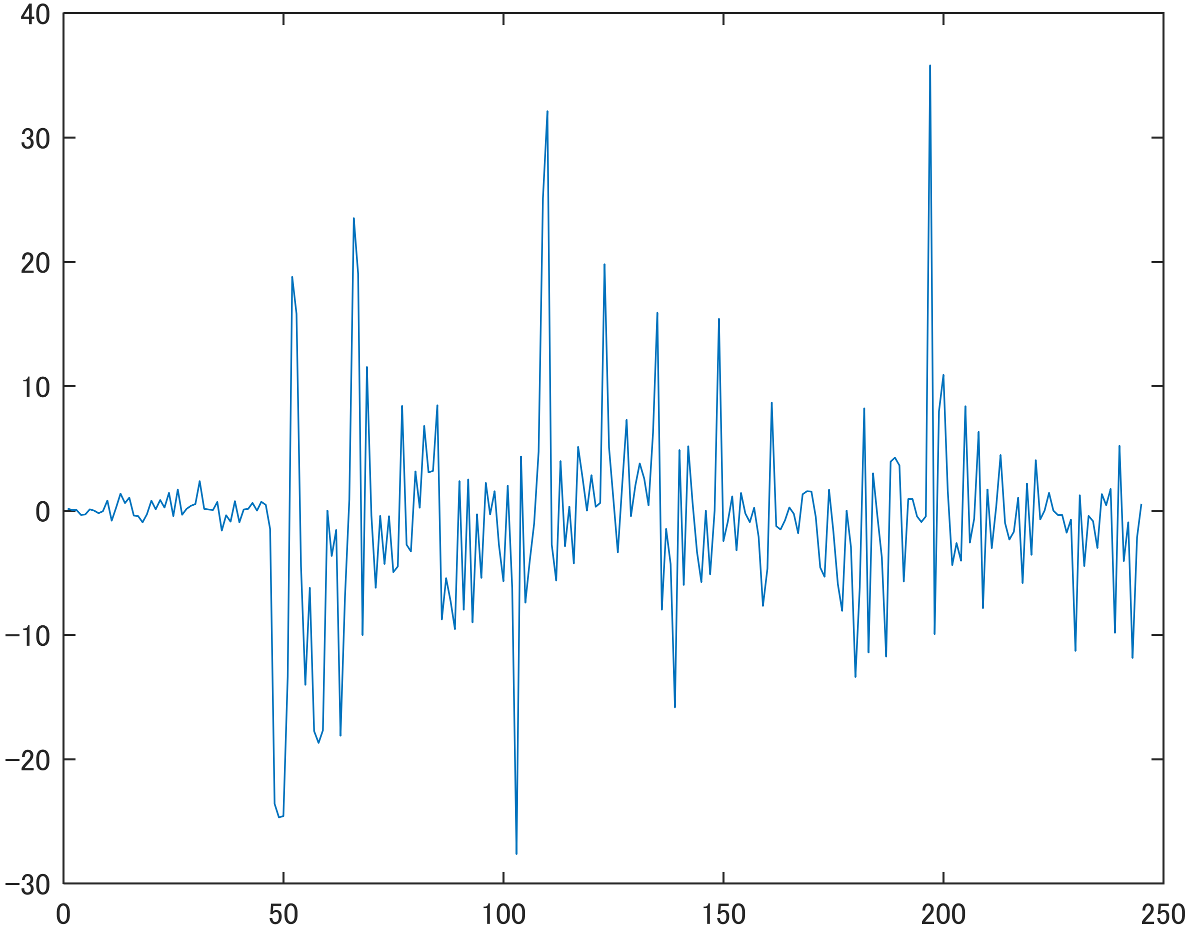
local_model.xlsx

$\hat{\alpha}_{0|0} = 4, \hat{\Sigma}_{0|0} = 12, \sigma_e = 1, \sigma_\varepsilon = 2$

time	S_t	$\hat{\alpha}_{t\|t-1}$	$\hat{\Sigma}_{t\|t-1}$	$\hat{S}_{t\|t-1}$	K_t	$\hat{\alpha}_{t\|t}$	$\hat{\Sigma}_{t\|t}$
0						4.000	12.000
1	4.400	4.000	16.000	4.000	0.941	4.376	0.941
2	4.000	4.376	4.941	4.376	0.832	4.063	0.832
3	3.500	4.063	4.832	4.063	0.829	3.597	0.829
4	4.600	3.597	4.829	3.597	0.828	4.428	0.828

Another numerical example

- The RoR on TEPCO from Jan. 4 2011 to Dec. 30 2011.



- We consider the following model:

$$S_t = \alpha_t + e_t, \qquad e_t \sim N(0, \sigma_e^2)$$

$$\alpha_t = T\alpha_{t-1} + \varepsilon_t\ , \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

and set $T = 0.3274$, $\sigma_e = 4.155$, $\sigma_\varepsilon = 5.901$

KF.m

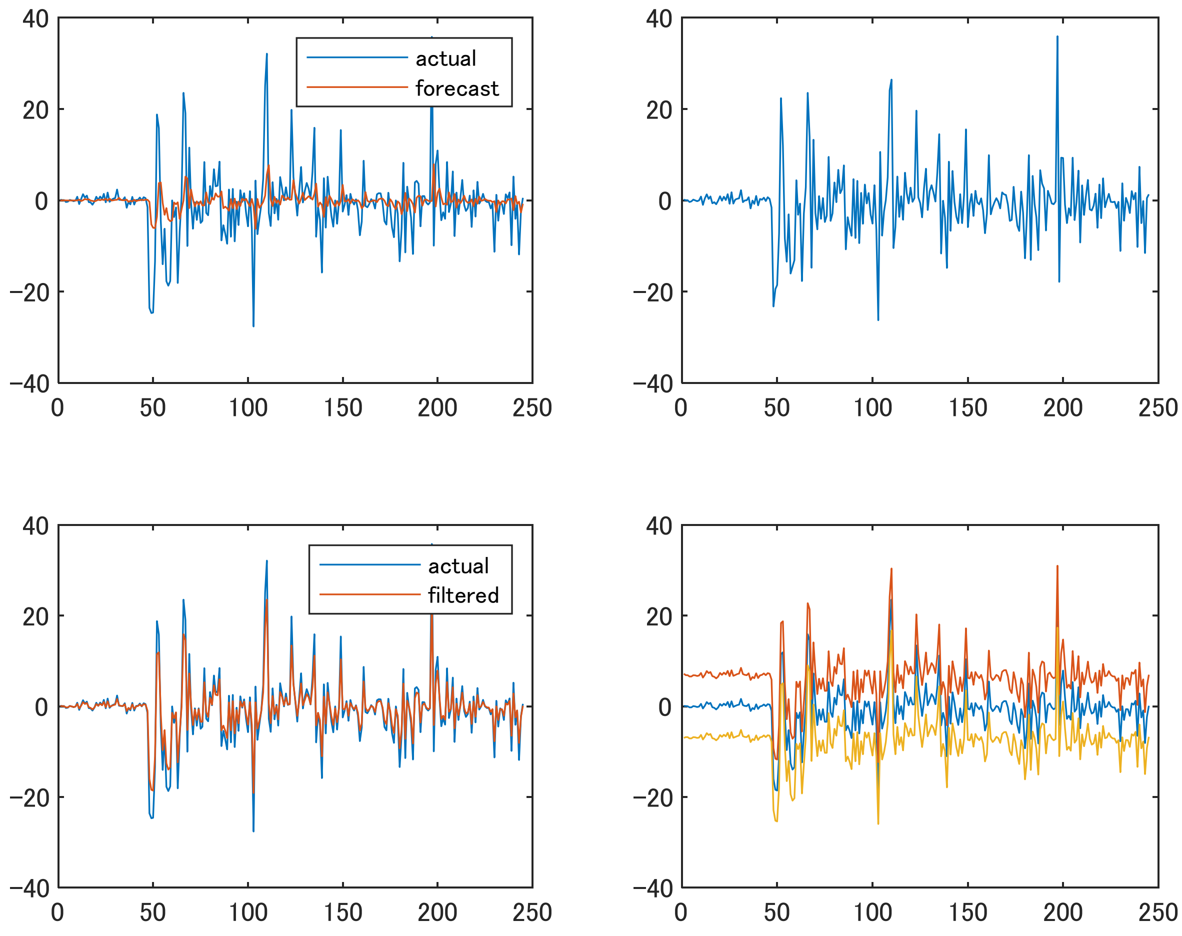
```
a_filt_prev = a0;
Sig_filt_prev = Sig0;
for t = 1:N
    a_fore(t) = T*a_filt_prev;
    Sig_fore(t) = T^2*Sig_filt_prev + sigeps^2;
    S_fore(t) = a_fore(t);

    K(t) = Sig_fore(t)/(Sig_fore(t)+sige^2);
    a_filt(t) = a_fore(t) + K(t)*(S(t)-S_fore(t));
    Sig_filt(t) = (1-K(t))*Sig_fore(t);
    a_filt_prev = a_filt(t);
    Sig_filt_prev = Sig_filt(t);
end
```

Forecasting:

$$\hat{\alpha}_{t|t-1} = T\hat{\alpha}_{t-1|t-1}$$
$$\hat{\Sigma}_{t|t-1} = T^2\hat{\Sigma}_{t-1|t-1} + \sigma_\varepsilon^2$$
$$\hat{S}_{t|t-1} = \hat{\alpha}_{t|t-1}$$

Filtering:

$$K_t = \frac{\hat{\Sigma}_{t|t-1}}{\hat{\Sigma}_{t|t-1} + \sigma_e^2}$$
$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + K_t(S_t - \hat{S}_{t|t-1})$$
$$\hat{\Sigma}_{t|t} = (1 - K_t)\hat{\Sigma}_{t|t-1}$$



actual
forecast
actual
filtered

Likelihood

- We define the one-step ahead forecasting error ν_t as
$$\nu_t = S_t - S_{t|t-1} = b\left(\alpha_t - \alpha_{t|t-1}\right) + e_t$$
- It follows a Gaussian distribution, and its mean and variance are

$$E[\nu_t|\Omega_{t-1}] = 0,$$
$$F_t = Var[\nu_t|\Omega_{t-1}] = b^2\hat{\Sigma}_{t|t-1} + \sigma_e^2$$

- Then, taking the values of (ν_t, F_t) as given, the likelihood at time t is
$$L_t = \frac{1}{2\pi F_t}\exp\left\{-\frac{{\nu_t}^2}{2F_t}\right\}$$

- Thus, we have the likelihood function with the given sequence of $\{\nu_t, F_t\}$

$$L = \prod_{t=1}^{N} \frac{1}{2\pi F_t} \exp\left\{-\frac{{\nu_t}^2}{2F_t}\right\}$$

and the log likelihood function

$$\ln L = -\frac{N}{2} \ln 2\pi - \frac{1}{2} \sum_{t=1}^{N} \ln F_t - \frac{1}{2} \sum_{t=1}^{N} \ln \frac{{\nu_t}^2}{F_t}$$

General case

- In general, we have the following state-space representation

$$y_t = A + Bx_t + e_t, \quad e_t \sim N(0, H)$$
$$x_t = Px_{t-1} + Q\epsilon_t, \quad \epsilon_t \sim N(0, S_e)$$

Now, A, B, H, P, Q, S_e are matrices and $x_t, y_t, e_t, \epsilon_t$ are vectors.

- We can calculate

$$\hat{x}_{t|t-1} = P\hat{x}_{t-1|t-1}$$
$$\hat{\Sigma}_{t|t-1} = P\hat{\Sigma}_{t-1|t-1}P' + QS_eQ'$$
$$\hat{y}_{t|t-1} = A + B\hat{x}_{t|t-1}$$
$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(y_t - \hat{y}_{t|t-1})$$
$$\hat{\Sigma}_{t|t} = \hat{\Sigma}_{t|t-1} - K_tB\hat{\Sigma}_{t|t-1}$$

where $K_t = \hat{\Sigma}_{t|t-1}B'F_t^{-1}$ and $F_t = B\hat{\Sigma}_{t|t-1}B' + H$.

- The log likelihood function is given by

$$\ln L = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{N}\ln F_t - \frac{1}{2}\sum_{t=1}^{N}\ln \frac{{\nu_t}^2}{F_t}$$

where

$$\nu_t = y_t - \hat{y}_{t|t-1}$$
$$F_t = B\hat{\Sigma}_{t|t-1}B' + H$$

Bayesian Inference

(based on chapter 3 in Herbst and Schorfheide, 2015)

What is Bayesian inference?

- The Bayesian approach regards the parameter θ as a random variable and assumes some prior knowledge on it as the form of **the prior distribution** $p(\theta)$.
- Learning about the parameter takes place by updating the prior distribution in the light of data Y. **The likelihood function** $p(Y|\theta)$ summarizes the information.

What is Bayesian inference?

- According to Bayes Theorem,

$$p(\theta|Y) = \frac{p(\theta)p(Y|\theta)}{p(Y)}$$

where $p(Y) = \int p(\theta)p(Y|\theta)d\theta$. This is called **posterior distribution**, which integrates to one.

- The formula for conditional probability is

$$p(A \cap B) = p(A)p(B|A) = p(B)p(A|B)$$

Therefore, $p(B|A) = \frac{p(B)p(A|B)}{p(A)}$.

- Bayesian inference characterizes properties of the posterior distribution.
- Unfortunately, for many interesting models, including the DSGE models, a direct analysis of the posterior is not feasible.
- All that can be done is to numerically evaluate the prior density $p(\theta)$ and the likelihood function $p(Y|\theta)$ at a given parameter θ.
- Therefore, we will use posterior sampler generating sequences of draws $\theta^i, i = 1, \ldots, N$ from $p(\theta|Y) \propto p(\theta)p(Y|\theta)$.

A simple regression model

- We begin with a simple regression model to illustrate some of the principles and mechanics.
- Consider the AR(1) model

$$y_t = \theta y_{t-1} + u_t, \qquad u_t \sim iid\mathcal{N}(0,1),$$

for $t = 1, \ldots, T$.

Likelihood

- Conditional on the initial observation y_0, the likelihood function is

$$
\begin{aligned}
p(Y_{1:t}|y_0,\theta) &= \prod_{t=1}^{T} p(y_t|Y_{0:t-1},\theta) \\
&= p(y_1|y_0,\theta) \times p(y_2|y_0,y_1,\theta) \times p(y_3|y_0,y_1,y_2,\theta) \\
&= p(u_1) \times p(u_2) \times p(u_3) \times \cdots
\end{aligned}
$$

$$
= (2\pi)^{-\frac{1}{2}}\exp\left(-\frac{(y_1-\theta y_0)^2}{2}\right) \times (2\pi)^{-\frac{1}{2}}\exp\left(-\frac{(y_2-\theta y_1)^2}{2}\right) \times \cdots
$$

$$
= (2\pi)^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}(Y-X\theta)'(Y-X\theta)\right\}
$$

where $Y_{1:t} = \{y_1, \dots, y_t\}$ and $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, X = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{T-1} \end{bmatrix}$.

Prior

- Suppose the prior distribution of the form

$$\theta \sim \mathcal{N}(0, \tau^2)$$

with a density

$$p(\theta) = (2\pi\tau^2)^{-\frac{1}{2}} \exp\left\{-\frac{\theta^2}{2\tau^2}\right\}$$

τ is a hyperparameter controlling the variance of the prior distribution.

- This is called a *conjugate prior distribution*. We expect that the posterior distribution is of the same form.

Bayes Theorem

- Recall the Bayes Theorem

$$p(\theta|Y) = \frac{p(\theta)p(Y|\theta)}{p(Y)}$$

- The posterior distribution of θ is proportional ($\propto$) to the product of prior and likelihood

$$p(\theta|Y) \propto p(\theta)p(Y|\theta)$$

Deriving the posterior

- Then the posterior distribution is proportional to (note that θ is a scalar in this case)

$$p(\theta)p(Y|\theta)$$

$$= (2\pi\tau^2)^{-\frac{1}{2}}\exp\left\{-\frac{\theta^2}{2\tau^2}\right\} \times (2\pi)^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}(Y - X\theta)'(Y - X\theta)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}(Y - X\theta)'(Y - X\theta) - \frac{\theta^2}{2\tau^2}\right\}$$

$$= \exp\left\{-\frac{1}{2}[Y'Y - \theta'X'Y - Y'X\theta + \theta'X'X\theta + \tau^{-2}\theta^2]\right\}$$

$$= \exp\left\{-\frac{1}{2}[Y'Y - X'Y\theta - Y'X\theta + X'X\theta^2 + \tau^{-2}\theta^2]\right\}$$

- Algebraic manipulation leads to

$$
\begin{aligned}
& Y'Y - X'Y\theta - Y'X\theta + X'X\theta^2 + \tau^{-2}\theta^2 \\
& = (X'X + \tau^{-2})\theta^2 - (X'Y + Y'X)\theta + Y'Y \\
& = (X'X + \tau^{-2})\left(\theta - \frac{1}{2}\frac{X'Y + Y'X}{X'X + \tau^{-2}}\right)^2 + Y'Y - \frac{1}{4}\frac{(X'Y + Y'X)^2}{X'X + \tau^{-2}} \\
& = (X'X + \tau^{-2})\left(\theta - \frac{X'Y}{X'X + \tau^{-2}}\right)^2 + Y'Y - \frac{(X'Y)^2}{X'X + \tau^{-2}}
\end{aligned}
$$

Note that $X'Y$ and $X'X$ are scalars and $X'Y = Y'X$ holds.

- Since the exponential term is a quadratic function of θ, we can *deduce* that the posterior distribution is Normal

$$\theta|Y \sim \mathcal{N}(\bar{\theta}, \bar{V}_\theta)$$

with

$$\bar{\theta} = \frac{X'Y}{X'X+\tau^{-2}}, \quad \bar{V}_\theta = (X'X + \tau^{-2})^{-1}.$$

- The pdf has the form of

$$p(\theta|Y) = \left(2\pi\bar{V}_\theta{}^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{(\theta - \bar{\theta})^2}{2\bar{V}_\theta{}^2}\right\}$$

Bayesian updating

- Define $\hat{\theta}_{mle} = (X'X)^{-1}X'Y$ and write
$$\bar{\theta} = \frac{X'X\hat{\theta}_{mle} + \tau^{-2} \cdot 0}{X'X + \tau^{-2}}$$
- Thus, the posterior mean is a weighted average of the maximum likelihood estimator and the prior mean (zero).
- The weights depend on the information content of the likelihood function, $X'X$, and the prior precision, τ^{-2}.
  - The smaller τ^2 is (i.e., the tighter the prior is), the smaller change in $\bar{\theta}$ is.

Monte-Carlo sampling methods

- In most cases, the analytical solution is not available, and we rely on sampling methods. Why?
- We abbreviate posterior distributions $p(\theta|Y)$ by $\pi(\theta)$

and posterior expectations of *objects of interest* $h(\theta)$ by

$$\mathbb{E}_\pi[h] = \mathbb{E}_\pi[h(\theta)]$$

$$= \int h(\theta)\pi(\theta)d\theta = \int h(\theta)p(\theta|Y)d\theta$$

For example, $h(\theta) = \theta$ implies $\mathbb{E}_\pi[h]$ is the mean of θ.

- We generate draws $\{\theta^i\}_{i=1}^N$ from $\pi(\theta)$ and approximate $\mathbb{E}_\pi[h]$ by

$$\bar{h}_N = \frac{1}{N}\sum_{i=1}^{N} h(\theta^i)$$

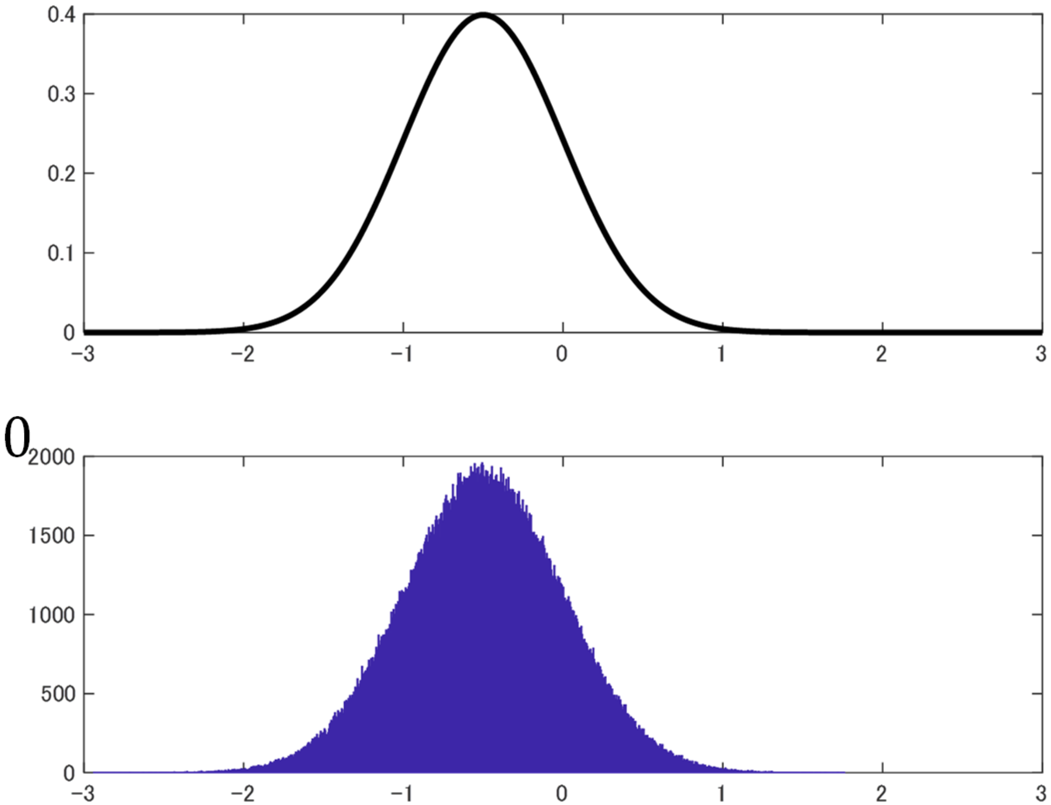
- This is numerical integration by a **Monte-Carlo sampling method** as a weighted sum of $h(\theta^i)$ approximates the integral in $\mathbb{E}_\pi[h]$.

Direct sampling

- In the simple regression model, it is possible to sample *iid* draws directly from the posterior distribution $\pi(\theta)$:

(**Direct Sampling**) For $i = 1, \ldots, N$, draw θ^i from $\mathcal{N}(\bar{\theta}, \bar{V}_\theta)$

$\bar{\theta} = -0.5, \bar{V}_\theta = 1/4, N = 500{,}000$



A posterior of a set-identified model

- Suppose that y_t follows an AR(1) with coefficient ϕ.

$$y_t = \phi y_{t-1} + u_t, \qquad u_t \sim iid\mathcal{N}(0,1),$$

for $t = 1, \dots, T$.

- The object of interest is a parameter θ, instead of ϕ, that can be bounded as

$$\phi \leq \theta \text{ and } \theta \leq \phi + 1$$

- To complete the model, we specify a prior for θ conditional on ϕ

$$\theta | \phi \sim U[\phi, \phi + 1]$$

- The joint posterior distribution of (θ, ϕ) is (from the conditional probability)

$$p(\theta, \phi|Y) = p(\phi|Y)p(\theta|\phi)$$

- $\phi|Y \sim N(\bar{\phi}, \bar{V}_\phi)$ from the previous discussion. Since $\theta|\phi \sim U[\phi, \phi + 1]$, the marginal distribution of θ is given by

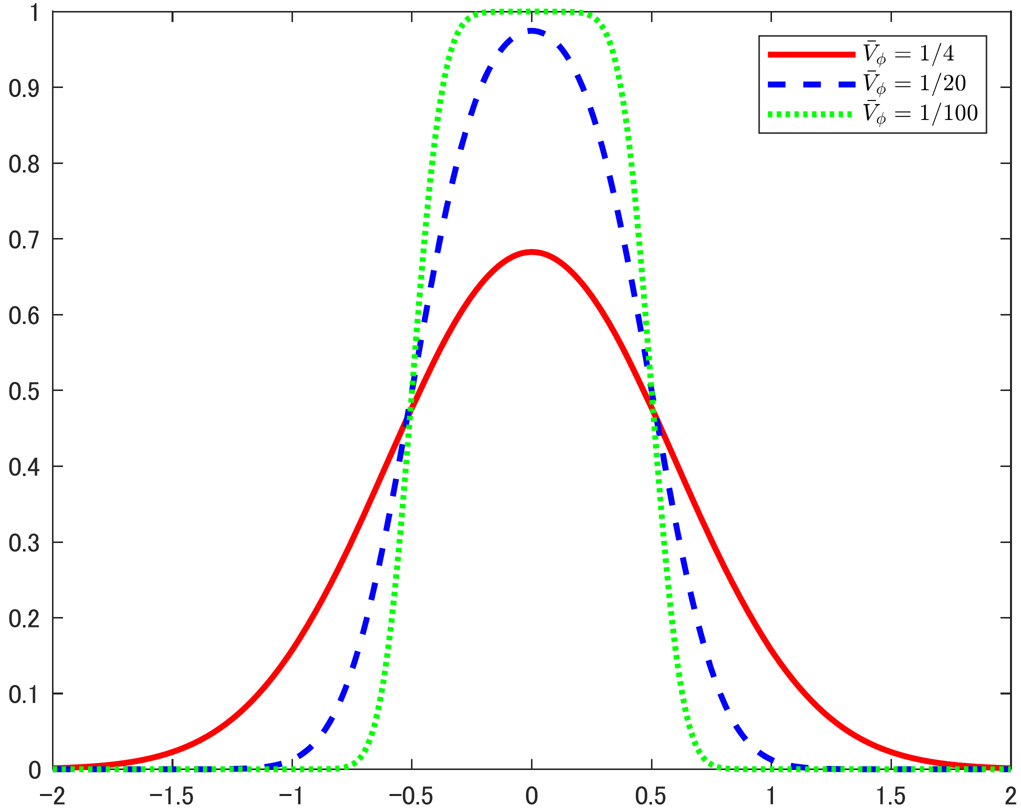
$$\pi(\theta) = \int p(\theta, \phi|Y)\, d\phi$$

$$= \int_{\theta-1}^{\theta} p(\phi|Y)p(\theta|\phi)\, d\phi$$

$$= \Phi_N\left(\bar{V}_\phi^{-\frac{1}{2}}(\theta - \bar{\phi})\right) - \Phi_N\left(\bar{V}_\phi^{-\frac{1}{2}}(\theta - (\bar{\phi} + 1))\right)$$

where $\Phi_N(x)$ is the cdf of $N(0,1)$. What are the mean of θ?

- As $\bar{V}_\phi$ decreases, the prior of θ is important and the posterior of θ looks like a step function.
- To sample iid draws from the posterior of θ, we consider importance sampling.
- (We could use the direct sampler in this case, by first sampling $\phi^i \sim N(\bar{\phi}, \bar{V}_\phi)$ and then sampling $\theta^i | \phi^i \sim U[\phi^i, \phi^i + 1]$.)



Importance sampling

(**Importance Sampling**)

1. For $i = 1, \dots, N$, draw $\theta^i \sim g(\theta)$ and compute the unnormalized importance weights

$$w^i = w(\theta^i) = \frac{f(\theta^i)}{g(\theta^i)}$$

$g(\theta)$ is called a **proposal density**. Note that the posterior density $f(\theta^i)$ and the proposal density $g(\theta^i)$ are evaluated at θ^i.

2. Compute the normalized importance weights

$$W^i = \frac{w^i}{\frac{1}{N}\sum_{i=1}^{N} w^i}$$

Then we have

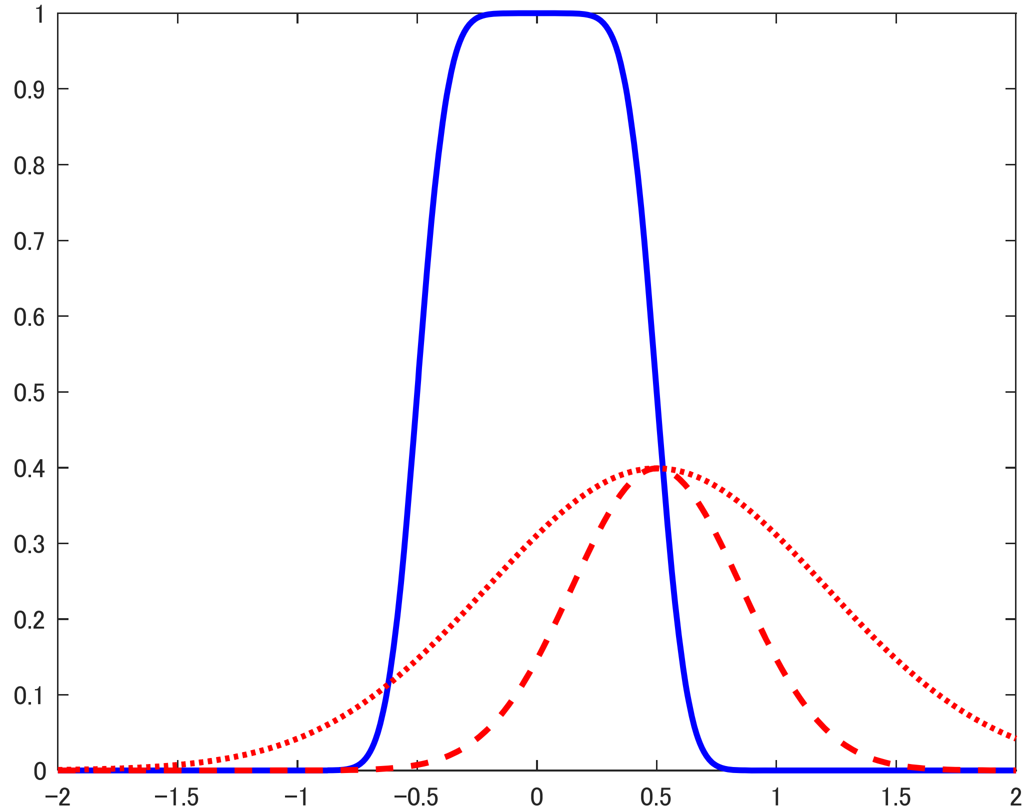
$$\mathbb{E}_\pi[h(\theta)] \approx \bar{h}_N = \frac{1}{N}\sum_{i=1}^{N} W^i h(\theta^i)$$

Two proposal densities

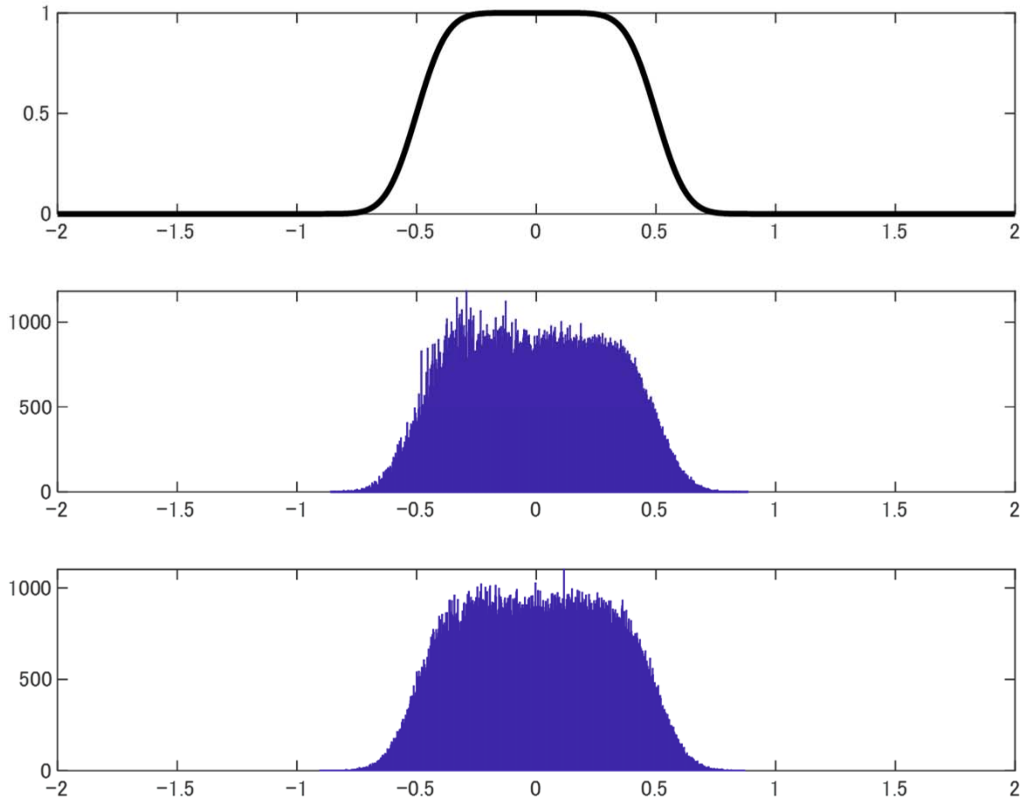
- We consider two proposal densities $g(\theta)$:

(i) “concentrated” $\theta \sim N(0.5, 0.125)$; (ii) “diffuse” $\theta \sim N(0.5, 0.5)$

- (i) assigns a very small probability to the interval $[-0.5, -0.25]$.



- We resample draws $\{\theta^i\}_{i=1}^N$ with weights $\{W^i\}_{i=1}^N$ for each of $g(\theta)$.
- The diffused (ii) looks like better in replicating the target density.



Metropolis-Hastings Algorithm

- The Metropolis-Hastings (MH) algorithm belongs to the class of Markov chain Monte Carlo (MCMC) algorithms.
- The algorithm generates a Markov chain such that the stationary distribution associated with the Markov chain is unique and equals the posterior distribution of interest.
- Example: A Markov chain

$$K = \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}$$

has the stationary distribution $\pi = [0.7143 \ 0.2857]'$.

Generic MH algorithm

(**Generic MH Algorithm**) For $i = 1, \ldots, N$:

1. Draw ϑ from a **proposal density** $q(\vartheta|\theta^{i-1})$.
2. Set $\theta^i = \vartheta$ with probability

$$\alpha(\vartheta|\theta^{i-1}) = \min\left\{1, \frac{p(\vartheta|Y)/q(\vartheta|\theta^{i-1})}{p(\theta^{i-1}|Y)/q(\theta^{i-1}|\vartheta)}\right\}$$

and $\theta^i = \theta^{i-1}$ otherwise,

where $p(\theta|Y)$ is the posterior density for a given parameter set θ.

The invariance property

- The transition kernel $K\left(\theta\middle|\tilde{\theta}\right)$ can be defined.
- The posterior distribution is an invariant distribution under the transition kernel K, that is

$$p(\theta|Y) = \int K\left(\theta\middle|\tilde{\theta}\right)p\left(\tilde{\theta}\middle|Y\right)d\tilde{\theta}$$

If $\tilde{\theta}$ is a draw from $p(\theta|Y)$, then θ is also a draw from $p(\theta|Y)$.

- The invariance property itself does not guarantee that the draws from the Markov chain $\{\theta^i\}_{i=1}^N$ converge to the posterior distribution $p(\theta|Y)$.
- In particular, one needs to ensure that
    - $K(\cdot \mid \cdot)$ has a *unique* invariant distribution.
    - The draws are not persistent so that sample averages converge to population means.
- We will examine a specific analytical example below.

An analytical example

- The parameter space is discrete and θ takes two values: τ_1 and τ_2
- The posterior distribution is two probabilities

$$\pi_l = \mathbb{P}\{\theta = \tau_l\}, \quad l = 1,2.$$

- The proposal distribution is a Markov process with

$$Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} = \begin{bmatrix} q & 1-q \\ 1-q & q \end{bmatrix}$$

where q_{lk} is the probability of drawing $\vartheta = \tau_k$ conditional on $\theta^{i-1} = \tau_l$.

Deriving K

- Remind that the acceptance probability is given by

$$\alpha\left(\vartheta \middle| \theta^{i-1}\right) = \min\left\{1, \frac{\pi(\vartheta)/q\left(\vartheta \middle| \theta^{i-1}\right)}{\pi(\theta^{i-1})/q\left(\theta^{i-1} \middle| \vartheta\right)}\right\}$$

- Suppose that $\theta^{i-1} = \tau_1$. Assuming $\pi_1 < \pi_2$,
    - With prob. $q, \vartheta = \tau_1$. The probability that this draw will be accepted is

$$\alpha(\tau_1|\tau_1) = \min\left\{1, \frac{\pi_1/q}{\pi_1/q}\right\} = 1$$

    - With prob. $1 - q, \vartheta = \tau_2$. The probability that this draw will be rejected is

$$1 - \alpha(\tau_2|\tau_1) = 1 - \min\left\{1, \frac{\pi_2/(1-q)}{\pi_1/(1-q)}\right\} = 0$$

- Thus, the prob. of a transition from $\theta^{i-1} = \tau_1$ to $\theta^{i} = \tau_1$ is equal to

$$q \times 1 + (1-q) \times 0 = q.$$

- Suppose that $\theta^{i-1} = \tau_2$. Then

$$\alpha(\tau_1|\tau_2) = \min\left\{1, \frac{\pi_1/(1-q)}{\pi_2/(1-q)}\right\} = \frac{\pi_1}{\pi_2}$$

$$1 - \alpha(\tau_2|\tau_2) = 1 - \min\left\{1, \frac{\pi_2/q}{\pi_2/q}\right\} = 0$$

and the prob. of a transition from $\theta^{i-1} = \tau_2$ to $\theta^{i} = \tau_1$

$$(1-q) \times \frac{\pi_1}{\pi_2} + q \times 0 = (1-q)\frac{\pi_1}{\pi_2}.$$

- The transition matrix is given by

$$K = \begin{bmatrix} q & 1-q \\ (1-q)\dfrac{\pi_1}{\pi_2} & 1-(1-q)\dfrac{\pi_1}{\pi_2} \end{bmatrix}$$

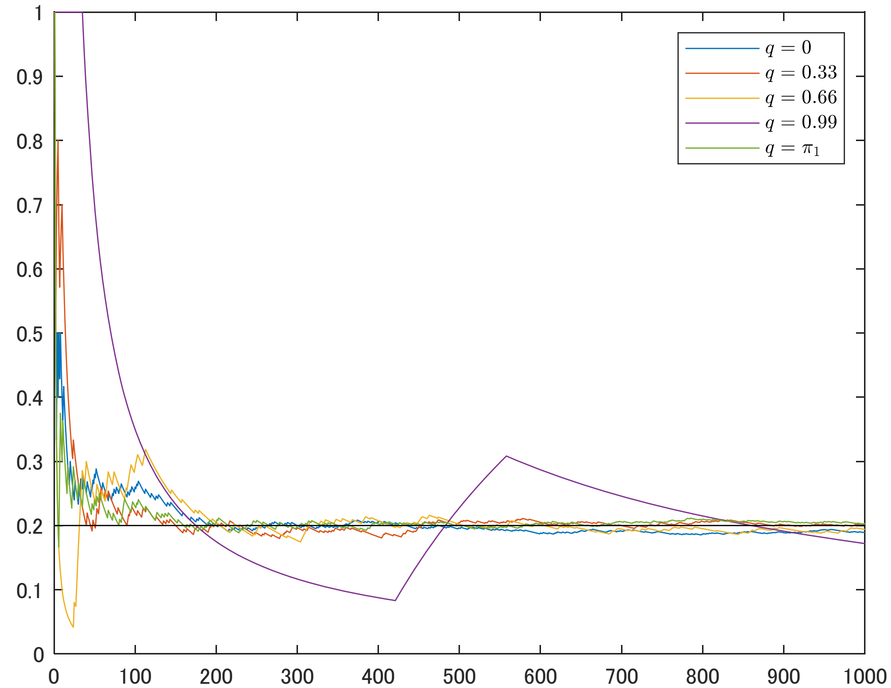
which has the stationary distribution $[\pi_1, \pi_2]$.

Persistence should be low

- The persistence of the Markov chain depends on the shape of the proposal distribution.
- **The goal of MCMC design** is to *keep the persistence as low as possible*. In this case,
  - If $q = (1-q)\frac{\pi_1}{\pi_2} \Leftrightarrow q = \pi_1$, one could obtain an iid sample.
  - If $q = 1$, $\theta^i = \theta^{i-1}$ for all i and the distribution of the chain is no longer unique.
  - If $q = 0$, the distribution of the chain remains unique, but $\theta^i = \tau_1$ is surely followed by $\theta^{i+1} = \tau_2$, and $\theta^{i+2} = \tau_2$ with probability π_2/π_1.

A Numerical Illustration

- We have a Bernoulli distribution $\tau_1 = 1$ and $\tau_2 = 0$ with $\pi_1 = 0.2$ and $\pi_2 = 1 - \pi_1$
- We vary $q = \{0, 0.33, 0.66, 0.99, \pi_1\}$ and sample draws $\{\theta^i\}_{i=1}^N$ from the Markov chain with the transition matrix K.
- The mean of $\{\theta^i\}_{i=1}^N$ quickly converges to π_1 when draws are nearly iid.



Autocorrelation of samples

- When $q = 0.99$, the chain is extremely autocorrelated. The chain is moving extremely slowly around the parameter space.
- When $q = 0.66$ or 0.33, the autocorrelation is substantially weaker.
- When $q = \pi_1$, the chain is iid.
- When $q = 0$, the chain has a negative autocorrelation.