DSGE-Models

Full information estimation
Kalman-filter, Maximum-Likelihood and Bayesian methods

Dr. Andrea Beccarini Willi Mutschler, M.Sc.

Institute of Econometrics and Economic Statistics
University of Münster
willi.mutschler@uni-muenster.de

Summer 2012

Full information estimation

1. Idea
2. Kalman-filter
    - Notation
    - Initialization
    - Recursion
    - Summary and Likelihood
3. Maximum-Likelihood
    - Idea
    - Exercise 4: An and Schorfheide (2007) via Maximum-Likelihood
4. Bayesian methods
    - Idea
    - Metropolis-Hastings-algorithm
    - Remarks
    - Exercise 5: Estimation with Bayesian methods
    - Inference, forecast, model comparison and identification
5. Discussion

Full information estimation

Idea

- Full information estimation requires a complete characterization of the data-generating-process (not only specific moments).
- Consider the linear *state-space* representation of the model:

$$
\begin{aligned}
\mathbf{x_t} &= \mathbf{F}(\boldsymbol{\mu})\mathbf{x_{t-1}} + \mathbf{G}(\boldsymbol{\mu})\boldsymbol{\upsilon}_\mathbf{t}, && \text{with } E[\boldsymbol{\upsilon}_\mathbf{t}\boldsymbol{\upsilon}_\mathbf{t}{}'] = \mathbf{V}, \quad E[\boldsymbol{\upsilon}_\mathbf{t}\boldsymbol{\upsilon}_\mathbf{s}{}'] = 0 && (1)\\
\mathbf{X_t} &= \mathbf{H}(\boldsymbol{\mu})'\mathbf{x_t} + \mathbf{e_t}, && \text{with } E[\mathbf{e_t}\mathbf{e_t}{}'] = \boldsymbol{\Sigma}_\mathbf{e}, \quad E[\mathbf{e_t}\mathbf{e_s}{}'] = 0. && (2)
\end{aligned}
$$

- Matrix **H** combines the model variables $\mathbf{x_t}$ with observable data variables $\mathbf{X_t}$
- Equation (1): *state-* or *transition-equation*
  - Corresponds to the solution of the model.
  - $\boldsymbol{\upsilon}_\mathbf{t}$ are the stochastic innovations.
- Equation (2): *observation-equation*
  - Corresponds to the measurement equations,
  - subject to possible measurement errors $\mathbf{e_t}$ in the data.

Full information estimation

Idea

- Given distributional assumptions about $\upsilon_\mathbf{t}$ and $\mathbf{e_t}$, one can derive the log-likelihood-function, $\log L(\mathbf{X}|\boldsymbol{\mu})$, analytically or numerically.
- In the linear case and considering normally distributed variables, the Kalman-filter is used to calculate the likelihood analytically.
- In the nonlinear case $\mathbf{s_t} = s(\mathbf{s_{t-1}}, \upsilon_\mathbf{t}), \mathbf{c_t} = c(\mathbf{s_t})$ and $\mathbf{X_t} = \widetilde{h}(\mathbf{s_t}, \mathbf{c_t}, \upsilon_\mathbf{t}, \mathbf{e_t}) \equiv h(\mathbf{s_t}, \mathbf{e_t})$ are considered, provided that c,s and h are functions of the vector of parameters $\boldsymbol{\mu}$. The particle-filter or the *efficient importance sampling* is then used to derive the likelihood numerically.
- There two approaches for analyzing and evaluating the log-likelihood:
  1. **the classic (frequentist) Maximum-Likelihood-method**,
  2. **the bayesian method**.

Full information estimation

1. Idea
2. Kalman-filter
   - Notation
   - Initialization
   - Recursion
   - Summary and Likelihood
3. Maximum-Likelihood
   - Idea
   - Exercise 4: An and Schorfheide (2007) via Maximum-Likelihood
4. Bayesian methods
   - Idea
   - Metropolis-Hastings-algorithm
   - Remarks
   - Exercise 5: Estimation with Bayesian methods
   - Inference, forecast, model comparison and identification
5. Discussion

Kalman-filter

Notation

We simplify and consider only the linear case and ignore possible measurement errors in the data:

- $\mathbf{X_t} = \mathbf{H}'\mathbf{x_t}$
- $\mathbf{x_{t+1}} = \mathbf{F}\mathbf{x_t} + \mathbf{G}\boldsymbol{\upsilon}_\mathbf{t+1}$
- $\boldsymbol{\upsilon}_\mathbf{i} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{V}),\ \mathbf{V} = E(\boldsymbol{\upsilon}_\mathbf{i}\boldsymbol{\upsilon}_\mathbf{i}{}'),\ E(\boldsymbol{\upsilon}_\mathbf{i}\boldsymbol{\upsilon}_\mathbf{j}{}') = 0$

Notation for the linear projection

$$\begin{aligned}
\widehat{\mathbf{x}}_{\mathbf{t|t-j}} &= E(\mathbf{x_t}|\mathbf{X_{t-j}}, \mathbf{X_{t-j-1}}, \ldots \mathbf{X_1}) \\
\mathbf{\Sigma}_{\mathbf{t|t-j}} &= E(\mathbf{x_t} - \widehat{\mathbf{x}}_{\mathbf{t|t-j}})(\mathbf{x_t} - \widehat{\mathbf{x}}_{\mathbf{t|t-j}})' \\
\widehat{\mathbf{X}}_{\mathbf{t|t-j}} &= E(\mathbf{X_t}|\mathbf{X_{t-j}}, \mathbf{X_{t-j-1}}, \ldots, \mathbf{X_1}) \\
\mathbf{u_t} &= \mathbf{X_t} - \widehat{\mathbf{X}}_{\mathbf{t|t-1}} = \mathbf{H}'(\mathbf{x_t} - \widehat{\mathbf{x}}_{\mathbf{t|t-1}}) \\
E(\mathbf{u_t}\mathbf{u_t}{}') &= \mathbf{H}'\mathbf{\Sigma}_{\mathbf{t|t-1}}\mathbf{H}
\end{aligned}$$

for $t = 1, 2, \ldots, T$ and $j = 0, 1, \ldots T$.

Kalman-filter

Initialization

- Since $\mathbf{x_t}$ is covariance-stationary, the variance is given by:

$$\underbrace{E(\mathbf{x_t}\mathbf{x_t}')}_{\equiv\boldsymbol{\Sigma}} = E\left[(\mathbf{F}\mathbf{x_{t-1}} + \mathbf{G}\boldsymbol{\upsilon_t})(\mathbf{F}\mathbf{x_{t-1}} + \mathbf{G}\boldsymbol{\upsilon_t})'\right]$$

$$= \mathbf{F}\underbrace{E(\mathbf{x_{t-1}}\mathbf{x_{t-1}}')}_{\equiv\boldsymbol{\Sigma}}\mathbf{F}' + \mathbf{G}\underbrace{E(\boldsymbol{\upsilon_t}\boldsymbol{\upsilon_t}')}_{=\mathbf{V}}\mathbf{G}'$$

$$\Leftrightarrow \boldsymbol{\Sigma} = \mathbf{F}\boldsymbol{\Sigma}\mathbf{F}' + \mathbf{G}\mathbf{V}\mathbf{G}'$$

Vectorization

The *vec*-operation stacks the rows of a $m \times n$ Matrix $\mathbf{M}$ into a $mn \times 1$ vector $vec(\mathbf{M})$. Then for arbitrary Matrices $\underset{m\times n}{A}, \underset{n\times p}{B}$ and $\underset{p\times k}{C}$:

$$vec(ABC) = (C' \otimes A)vec(B), \quad \text{with } \otimes : \text{Kronecker-product.}$$

Kalman-filter

Initialization

- Applying this formula for $\mathbf{\Sigma}$:

$$
\begin{aligned}
vec(\mathbf{\Sigma}) &= (\mathbf{F} \otimes \mathbf{F}) vec(\mathbf{\Sigma}) + vec(\mathbf{GVG}') \\
\Leftrightarrow vec(\mathbf{\Sigma}) &= (\mathbf{I} - \mathbf{F} \otimes \mathbf{F})^{-1} vec(\mathbf{GVG}')
\end{aligned}
$$

- The unconditional expectation of $\mathbf{x_1}$ is used for the initialization of the Kalman-filter, since there are is no additional information yet:

$$
\widehat{\mathbf{x}}_\mathbf{1} = \underbrace{E(\mathbf{x_1})}_{=E(\mathbf{x})} = \mathbf{F}\, \underbrace{E(\mathbf{x_0})}_{=E(\mathbf{x})} + \mathbf{G}\, \underbrace{E(\boldsymbol{\upsilon}_\mathbf{1})}_{=0} \Leftrightarrow \widehat{\mathbf{x}}_\mathbf{1} = \mathbf{0},
$$

$$
vec(\mathbf{\Sigma_{1|0}}) = E(\mathbf{x_1} - \mathbf{0})(\mathbf{x_1} - \mathbf{0})' = (\mathbf{I} - \mathbf{F} \otimes \mathbf{F})^{-1} vec(\mathbf{GVG}').
$$

Kalman-filter

Recursion

The recursion is then given by:

$$\widehat{\mathbf{x}}_{\mathbf{t+1|t}} = \mathbf{F}\widehat{\mathbf{x}}_{\mathbf{t|t}}$$

Formula for updating a linear projection (Hamilton (1994, S.99 und S.379))

$$\begin{aligned}
\widehat{\mathbf{x}}_{\mathbf{t|t}} &= \widehat{\mathbf{x}}_{\mathbf{t|t-1}} + \left[E(\mathbf{x}_{\mathbf{t}} - \widehat{\mathbf{x}}_{\mathbf{t|t-1}})(\mathbf{X}_{\mathbf{t}} - \widehat{\mathbf{X}}_{\mathbf{t|t-1}})'\right]\left[E(\mathbf{X}_{\mathbf{t}} - \widehat{\mathbf{X}}_{\mathbf{t|t-1}})(\mathbf{X}_{\mathbf{t}} - \widehat{\mathbf{X}}_{\mathbf{t|t-1}})'\right]^{-1}\mathbf{u}_{\mathbf{t}}\\
\Leftrightarrow \widehat{\mathbf{x}}_{\mathbf{t|t}} &= \widehat{\mathbf{x}}_{\mathbf{t|t-1}} + \mathbf{\Sigma}_{\mathbf{t|t-1}}\mathbf{H}\left(\mathbf{H}'\mathbf{\Sigma}_{\mathbf{t|t-1}}\mathbf{H}\right)^{-1}\mathbf{u}_{\mathbf{t}}\\
\Rightarrow \widehat{\mathbf{x}}_{\mathbf{t+1|t}} = \mathbf{F}\widehat{\mathbf{x}}_{\mathbf{t|t}} &= \mathbf{F}\widehat{\mathbf{x}}_{\mathbf{t|t-1}} + \mathbf{F}\mathbf{\Sigma}_{\mathbf{t|t-1}}\mathbf{H}\left(\mathbf{H}'\mathbf{\Sigma}_{\mathbf{t|t-1}}\mathbf{H}\right)^{-1}\mathbf{u}_{\mathbf{t}},\\
&\qquad \text{with } \mathbf{u}_{\mathbf{t}} = \mathbf{X}_{\mathbf{t}} - \widehat{\mathbf{X}}_{\mathbf{t|t-1}} = \mathbf{H}'(\mathbf{x}_{\mathbf{t}} - \widehat{\mathbf{x}}_{\mathbf{t|t-1}}).
\end{aligned}$$

Kalman-filter

Recursion

- $\mathbf{x_{t+1}} - \widehat{\mathbf{x}}_{\mathbf{t+1|t}} = \mathbf{F}\left(\mathbf{x_t} - \widehat{\mathbf{x}}_{\mathbf{t|t-1}}\right) + \mathbf{G}\upsilon_{\mathbf{t+1}} - \mathbf{F\Sigma_{t|t-1}H}\left(\mathbf{H'\Sigma_{t|t-1}H}\right)^{-1}\mathbf{u_t}$
- The *MSE:* $\mathbf{\Sigma_{t+1|t}} = E\left(\mathbf{x_{t+1}} - \widehat{\mathbf{x}}_{\mathbf{t+1|t}}\right)\left(\mathbf{x_{t+1}} - \widehat{\mathbf{x}}_{\mathbf{t+1|t}}\right)'$ is given by:

$$\mathbf{\Sigma_{t+1|t}} =$$

$$\mathbf{F\Sigma_{t|t-1}F'} + \mathbf{GVG'} - \mathbf{F\Sigma_{t|t-1}H}\underbrace{\left(\mathbf{H'\Sigma_{t|t-1}H}\right)^{-1}\underbrace{E(\mathbf{u_t u_t'})}_{=\mathbf{H'\Sigma_{t|t-1}H}}}_{=\mathbf{I}}\left(\mathbf{H'\Sigma_{t|t-1}H}\right)^{-1}\mathbf{H'\Sigma_{t|t-1}F'}$$

Mean-Sqared-Error (MSE)

$$\mathbf{\Sigma_{t+1}} \equiv \mathbf{\Sigma_{t+1|t}} = \mathbf{F\Sigma_{t|t-1}F'} + \mathbf{GVG'} - \mathbf{F\Sigma_{t|t-1}H}\left(\mathbf{H'\Sigma_{t|t-1}H}\right)^{-1}\mathbf{H'\Sigma_{t|t-1}F'}$$

Kalman-filter

Summary

Let:

- $\mathbf{K_t} = \mathbf{F\Sigma_t H}\left(\mathbf{H'\Sigma_t H}\right)^{-1}$ be the so-called gain-matrix,
- $\mathbf{\widehat{x}_t} = \mathbf{\widehat{x}_{t|t-1}}$ the linear projection,
- with $\mathbf{\Sigma_t} = \mathbf{\Sigma_{t|t-1}}$ being the *mean-squared-error*.

Then the Kalman-filter can be summarized as follows:

1. Initialization with
   - $\mathbf{\widehat{x}_1} = \mathbf{0}$,
   - $vec(\mathbf{\Sigma_{1|0}}) = (\mathbf{I} - \mathbf{F} \otimes \mathbf{F})^{-1} vec(\mathbf{GVG'})$.
2. Recursion with
   - $\mathbf{u_t} = \mathbf{X_t} - \mathbf{\widehat{X}_t} = \mathbf{X_t} - \mathbf{H'\widehat{x}_t}, \quad E(\mathbf{u_t u_t'}) = \mathbf{H'\Sigma_t H} \equiv \mathbf{\Omega_t}$,
   - $\mathbf{\widehat{x}_{t+1}} = \mathbf{F\widehat{x}_t} + \mathbf{K_t u_t}$,
   - $\mathbf{\Sigma_{t+1}} = \mathbf{F\Sigma_t F'} + \mathbf{GVG'} - \mathbf{K_t H'\Sigma_t F'}$.

Log-Likelihood

Given the gaussian assumption about the forecast error $\mathbf{u_t}$ one can derive the distribution of the data $\underset{n\times 1}{\mathbf{X_t}}$ conditional on $(\mathbf{x_t}, \mathbf{X_{t-1}}, \mathbf{X_{t-2}}, \dots)$ and set up the log-likelihood function:

Log-likelihood

$$\begin{aligned}
\log \mathcal{L}(\mathbf{X}|\boldsymbol{\mu}) &= \sum_{t=1}^{T} \log \mathcal{L}(\mathbf{X_t}|\boldsymbol{\mu}) \\
&= -\frac{nT}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log|\boldsymbol{\Omega_t}| - \frac{1}{2}\sum_{t=1}^{T}\mathbf{u_t}'\boldsymbol{\Omega_t}^{-1}\mathbf{u_t}.
\end{aligned}$$

Full information estimation

1. Idea
2. Kalman-filter
   - Notation
   - Initialization
   - Recursion
   - Summary and Likelihood
3. **Maximum-Likelihood**
   - Idea
   - Exercise 4: An and Schorfheide (2007) via Maximum-Likelihood
4. Bayesian methods
   - Idea
   - Metropolis-Hastings-algorithm
   - Remarks
   - Exercise 5: Estimation with Bayesian methods
   - Inference, forecast, model comparison and identification
5. Discussion

Maximum-Likelihood

Idea

- **Approach:** The parameters $\boldsymbol{\mu}$ are fixed and the data is a random realization of this specific parametrization.
- The *Maximum-Likelihood*-estimator $\widehat{\boldsymbol{\mu}}_{\mathbf{ML}}$ is then defined as

$$\widehat{\boldsymbol{\mu}}_{\mathbf{ML}} = \underset{\mu}{\operatorname{argmax}} \left\{ \sum_{t=1}^{T} \log \mathcal{L}(\mathbf{X_t}|\boldsymbol{\mu}) \right\}.$$

- Given some regularity conditions the *ML*-estimator is consistent, asymptotically efficient and asymptotically gaussian.
- Uncertainty and inference are based upon the assumptions that **to each realization of data there corresponds a different vector of parameters that maximizes the Likelihood**.
- Hint for the estimation of the parameters of a DSGE-model:
  - The dimension of $\mathbf{X_t}$ must be greater or equal to the dimension of the structural shocks $\boldsymbol{v}_\mathbf{t}$, or otherwise the residual term has a singular variance-covariance-matrix.
  - If not: Add measurement errors or additional shocks.

Exercise 4: An and Schorfheide (2007) via ML

Consider the following new-keynesian model:

$$
\begin{aligned}
\widehat{y}_t &= E_t[\widehat{y}_{t+1}] + \widehat{g}_t - E_t[\widehat{g}_{t+1}] - \frac{1}{\tau}(\widehat{R}_t - E_t[\widehat{\pi}_{t+1}] - E_t[\widehat{z}_{t+1}]),\\
\widehat{\pi}_t &= \beta E_t[\widehat{\pi}_{t+1}] + \kappa(\widehat{y}_t - \widehat{g}_t),\\
\widehat{c}_t &= \widehat{y}_t - \widehat{g}_t,\\
\widehat{R}_t &= \rho_R \widehat{R}_{t-1} + (1-\rho_R)\psi_1 \widehat{\pi}_t + (1-\rho_R)\psi_2 \left(\widehat{y}_t - \widehat{g}_t\right) + \epsilon_{R,t}\\
\widehat{g}_t &= \rho_g \widehat{g}_{t-1} + \epsilon_{g,t},\\
\widehat{z}_t &= \rho_z \widehat{z}_{t-1} + \epsilon_{z,t}.
\end{aligned}
$$

All variables with a $\widehat{}$ denote the logarithmic deviation from the steady-state, i.e. $\widehat{x_t} = log(x_t) - log(x)$.
The stochastic shocks are normally distributed with $E[\epsilon_{i,t}] = 0$ and $E[\epsilon_{i,t}^2] = \sigma_i^2$.

Exercise 4: An and Schorfheide (2007) via ML

- Assume you have **quarterly** data to estimate the parameters:
  - quarterly growth of GDP per capita in percent (YGR_t),
  - annualized inflation rates in percent ($INFL_t$),
  - annualized nominal interest rates in percent (INT_t).
- Model variables and observed data are linked by the following equations:

$$
\begin{aligned}
YGR_t &= \gamma^{(Q)} + 100(\widehat{y}_t - \widehat{y}_{t-1} + \widehat{z}_t),\\
INFL_t &= \pi^{(A)} + 400\widehat{\pi}_t,\\
INT_t &= \pi^{(A)} + r^{(A)} + 4\gamma^{(Q)} + 400\widehat{R}_t.
\end{aligned}
$$

- The parameter $\gamma^{(Q)}, \pi^{(A)}$ and $r^{(A)}$ are linked to the *steady-state* values of the model:

$$\gamma = \frac{A_{t+1}}{A_t} = e^{\frac{\gamma_Q}{100}} \approx 1 + \frac{\gamma^{(Q)}}{100}, \qquad \pi = e^{\frac{\pi^{(A)}}{400}} \approx 1 + \frac{\pi^{(A)}}{400}, \qquad r = e^{\frac{r^{(A)}}{400}} \approx 1 + \frac{r^{(A)}}{400},$$
$$\beta = e^{-\frac{r^{(A)}}{400}} \approx \frac{1}{1 + r^{(A)}/400}.$$

Exercise 4: An and Schorfheide (2007) via ML

Write a mod-file for the model in order to estimate it via Maximum-Likelihood.

(a) Use the simulated dataset `simdat1.mat` for the estimation. The true values are

$$\begin{array}{llll}
\tau = 2.000, & \kappa = 0.150, & \psi_1 = 1.500, & \psi_2 = 1.000, \\
\rho_R = 0.600, & \rho_z = 0.650, & \rho_g = 0.950, & \\
\sigma_R = 0.2/100, & \sigma_g = 0.8/100, & \sigma_z = 0.45/100, & \\
\pi^{(A)} = 4.000, & \gamma^{(Q)} = 0.500, & r^{(A)} = 0.400. &
\end{array}$$

1. Estimate all parameters via ML. Why does it not work?

   Hint: The nonlinear model implies that $\beta = \frac{\gamma}{r} = \frac{e^{\frac{\gamma^{(Q)}}{100}}}{e^{\frac{r^{(A)}}{400}}}$.
2. Calibrate the parameters ψ_1 and $r^{(A)}$ to their true values and estimate the other parameters. Why does it work now?

Exercise 4: An and Schorfheide (2007) via ML

(b) Use the simulated dataset `simdat2.mat` for the estimation. The true parameters are the same except now $r^{(A)} = 4$.

1. Estimate all parameters via ML. Why does it still not work?
2. Calibrate ψ_1 to its true value and estimate all other parameters. Discuss your results.

Full information estimation

1. Idea
2. Kalman-filter
    - Notation
    - Initialization
    - Recursion
    - Summary and Likelihood
3. Maximum-Likelihood
    - Idea
    - Exercise 4: An and Schorfheide (2007) via Maximum-Likelihood
4. **Bayesian methods**
    - Idea
    - Metropolis-Hastings-algorithm
    - Remarks
    - Exercise 5: Estimation with Bayesian methods
    - Inference, forecast, model comparison and identification
5. Discussion

Bayesian methods

Idea

- Experience shows that it can be pretty hard and tricky to estimate a DSGE-model via Maximum-Likelihood.
- Data is often not sufficiently informative, i.e. the likelihood is flat in some directions (identification).
- DSGE-models are always misspecified. This can lead to absurd parameter values.

Bayesian methods

Idea

- Based upon the likelihood as well: the complete characterization of the data generating process.
- **Approach:** The parameters $\boldsymbol{\mu}$ are random and the data $\mathbf{X}$ are fixed.
- The idea is to combine known information (data) with additional believes (*prior-believes*) about the parameters and to get an expression for the conditional probability of the parameters.
- Hence, one is able to put more weight on a suspected span of the parameter space.
- Bayesian methods are a bridge between calibration and the *Maximum-Likelihood*-method:

„Bayesian Inference is a Way of Thinking, Not a Basket of Methods“

Bayesian methods

Idea

- Likelihood-funktion $\mathcal{L}(\mathbf{X}|\boldsymbol{\mu})$ is a conditional density of observed data given the parameters: $\wp(\mathbf{X}|\boldsymbol{\mu}) = \mathcal{L}(\mathbf{X}|\boldsymbol{\mu})$.
- Denote $\wp(\boldsymbol{\mu})$ as the known prior density of the vector of parameters, then using Bayes-rule:

$$\wp(\boldsymbol{\mu}|\mathbf{X}) = \frac{\mathcal{L}(\mathbf{X}|\boldsymbol{\mu})\wp(\boldsymbol{\mu})}{\wp(\mathbf{X})} = \frac{\mathcal{L}(\mathbf{X}|\boldsymbol{\mu})\wp(\boldsymbol{\mu})}{\int \wp(\boldsymbol{\mu})\mathcal{L}(\mathbf{X}|\boldsymbol{\mu})\ d\boldsymbol{\mu}} \propto \mathcal{L}(\mathbf{X}|\boldsymbol{\mu})\wp(\boldsymbol{\mu}),$$

  with $\propto$ meaning „proportional to“.
- $\wp(\mathbf{X})$ is the *marginal likelihood* of the data and ultimately only a constant that normalizes the expression to unity. It is independent of the parameters.
- Removing it doesn't change the form of the posterior density $\wp(\boldsymbol{\mu}|\mathbf{X})$, it merely doesn't integrate to one.
- This non-normalized density is called *posterior-kernel* or, in logs, *log-posterior-kernel*.

Bayesian methods

Idea

- The mode is the Bayesian estimator $\widehat{\boldsymbol{\mu}}_{\mathbf{B}}$ of the true parameter vector:

$$\widehat{\boldsymbol{\mu}}_{\mathbf{B}} = \underset{\mu}{\operatorname{argmax}} \{\log \wp(\boldsymbol{\mu}|\mathbf{X})\} = \underset{\mu}{\operatorname{argmax}} \{\log \mathcal{L}(\mathbf{X}|\boldsymbol{\mu}) + \log \wp(\boldsymbol{\mu})\}$$

- Procedure: Calculate the log-likelihood with the Kalman-filter and simulate the *log-posterior-kernel* through *sampling*- or *Monte-Carlo*-methods.
- In the literature – and in Dynare – the *Metropolis-Hastings-algorithm* is commonly used.
- Inference can then be conducted via the properties of the posterior-distribution.

An and Schorfheide (2007, S. 132)

The algorithm constructs a Gaussian approximation around the posterior mode and uses a scaled version of the asymptotic covariance matrix as the covariance matrix for the proposal distribution. This allows for an efficient exploration of the posterior distribution at least in the neighborhood of the mode.

- The algorithm uses the fact that under very general regularity conditions the moments of a distribution are asymptotically normal.
- It constructs a sequence of draws (Markov-chains) from a proposal density.
- This does not need to be identical with the posterior density. It is only required that the algorithm can draw samples from the whole range of the posterior density.

- The current candidate (draw) $\boldsymbol{\mu}^*$ is dependent on the previous candidate $\boldsymbol{\mu}^{\mathbf{(s-1)}}$.
- Weights for all candidates are the same, however, they are only accepted with a certain probability α, calculated as the ratio of the *posterior-kernel* of the current to the one of the previous candidate.
- Due to this construct the algorithm tends to shift the draws from areas of low posterior probability to areas of high probability.
  - If $\boldsymbol{\mu}^{\mathbf{(s-1)}}$ is in an area of high posterior probability, it is likely that only candidates in the same area are accepted.
  - If $\boldsymbol{\mu}^{\mathbf{(s-1)}}$ is in an area of low posterior probability, it is very likely that new candidates are accepted.
- The covariance-matrix of the proposal distribution plays a major role, since it is important to set α neither too large nor to small.
- Current practice uses the covariance matrix of the mode $\widehat{\boldsymbol{\mu}}_{\mathbf{B}}$ and scales it with a factor *c* such that the average acceptance probability is between 20% and 30%.

Bayesian methods

Metropolis-Hastings-algorithm

1. Specify c_0, c and S.
2. Maximize $\log \mathcal{L}(\mathbf{X}|\boldsymbol{\mu}) + \log \wp(\boldsymbol{\mu})$ using numerical methods. $\widehat{\boldsymbol{\mu}}_{\mathbf{B}}$ denotes the mode.
3. Calculate the inverse of the Hessian evaluated at the mode, denote it with $\boldsymbol{\Sigma}_{\mathbf{B}}$.
4. Specify an initial value $\boldsymbol{\mu}^{\mathbf{(0)}}$ or draw it from $\mathcal{N}(\widehat{\boldsymbol{\mu}}_{\mathbf{B}}, c_0^2\boldsymbol{\Sigma}_{\mathbf{B}})$.

Bayesian methods

Metropolis-Hastings-algorithm

5. For $\mathbf{s} = 1, \ldots, S$:
   - Draw $\boldsymbol{\mu}^*$ from the candidate-generating distribution (proposal density) $\mathcal{N}(\boldsymbol{\mu}^{\mathbf{(s-1)}}, c^2\boldsymbol{\Sigma}_{\mathbf{B}})$.
   - Calculate the acceptance probability α:

$$\alpha \equiv \alpha\left(\boldsymbol{\mu}^{\mathbf{(s-1)}}, \boldsymbol{\mu}^*\right) = \frac{\mathcal{L}\left(\boldsymbol{\mu}^*|\mathbf{X}\right)\ \wp\left(\boldsymbol{\mu}^*\right)}{\mathcal{L}\left(\boldsymbol{\mu}^{\mathbf{(s-1)}}|\mathbf{X}\right)\ \wp\left(\boldsymbol{\mu}^{\mathbf{(s-1)}}\right)}$$

   - With probability $\min\{\alpha, 1\}$ accept the jump from $\boldsymbol{\mu}^{\mathbf{(s-1)}}$ to $\boldsymbol{\mu}^*$. In other words: If $\alpha \geq 1$, set $\boldsymbol{\mu}^{\mathbf{(s)}} = \boldsymbol{\mu}^*$.
   - With complementary probability don't accept the jump, i.e. draw a uniformly distributed random variable r between 0 and 1:
     - If $r \leq \alpha$ set $\boldsymbol{\mu}^{\mathbf{(s)}} = \boldsymbol{\mu}^*$ (jump).
     - If $r > \alpha$ set $\boldsymbol{\mu}^{\mathbf{(s)}} = \boldsymbol{\mu}^{\mathbf{(s-1)}}$ (don't jump).

6. Estimate the posterior expectation of a function $\hbar(\boldsymbol{\mu})$ with $\frac{1}{S}\sum_{s=1}^{S}\hbar\left(\boldsymbol{\mu}^{\mathbf{(s)}}\right)$.
7. If the average acceptance probability does not yield a desirable value (typically between $20\% - 30\%$) or the algorithm does not converge, change c_0, c or S.

Bayesian methods

Remarks

- Bayesian estimation of a DSGE-model requires that the number of shocks is equivalent to the numbers of observable variables.
- Common choice for priors: gaussian, (normal, shifted or inverse) Gamma, Beta or the uniform distribution.
- Choosing a proper prior one has to consider lower and upper bounds as well as the skewness and kurtosis of the distribution.
- The results can vary due to the choice of priors and their parametrization.
- Therefore one has to check the robustness of the results:
  - Try a different parametrization.
  - Try more general priors.
  - Noninformative priors.
  - Sensitivity analysis.

Exercise 5: Estimation with Bayesian methods

Consider the following simplified RBC-model (social planer problem);

$$\max_{\{c_{t+j},l_{t+j},k_{t+j}\}_{j=0}^{\infty}} W_t = \sum_{j=0}^{\infty} \beta^j u(c_{t+j}, l_{t+j})$$

$$\begin{aligned} s.t. \quad y_t &= c_t + i_t, & A_t &= Ae^{a_t}, \\ y_t &= A_t f(k_{t-1}, l_t), & a_t &= \rho a_{t-1} + \varepsilon_t, \\ k_t &= i_t + (1-\delta)k_{t-1}, & \varepsilon_t &\sim N(0, \sigma_\varepsilon^2), \end{aligned}$$

where preferences and technology follow:

$$u(c_t, l_t) = \frac{\left[c_t^\theta (1-l_t)^{1-\theta}\right]^{1-\tau}}{1-\tau}, \qquad f(k_{t-1}, l_t) = \left[\alpha k_{t-1}^\psi + (1-\alpha) l_t^\psi\right]^{1/\psi}.$$

Optimality is given by:

$$\begin{gathered} u_c(c_t, l_t) - \beta E_t \left\{ u_c(c_{t+1}, l_{t+1}) \left[A_{t+1} f_k(k_t, l_{t+1}) + 1 - \delta\right]\right\} = 0, \\ -\tfrac{u_l(c_t, l_t)}{u_c(c_t, l_t)} - A_t f_l(k_{t-1}, l_t) = 0, \\ c_t + k_t - A_t f(k_{t-1}, l_t) - (1-\delta)k_{t-1} = 0. \end{gathered}$$

Exercise 5: Estimation with Bayesian methods

(a) Write a mod-file for this model (with a sensible calibration and a steady-state block).

(b) Simulate a sample of 10000 observations for c_t, l_t and y_t using `stoch_simul` and save it in a mat-file.

(c) Define priors for α, θ and τ (or a different set of parameters).

(d) Estimate the posterior mode using the `estimation` command and a limited sample with 100 observations. How man observable variables do you need? Check the posterior mode using `mode_check`. If you get errors due to a non-positive definite Hessian, try a different optimization algorithm or change the initial values.

Exercise 5: Estimation with Bayesian methods

(e) If you are satisfied with the posterior mode, approximate the posterior distribution using the the Metropolis- Hastings-Algorithmus with 3×5000 iterations. If it does not converge to the (ergodic) posterior-distribution, repeat the algorithm without discarding the previous draws.

(f) How robust are the results regarding the specification of the priors? Repeat the estimation of the posterior-mode for different priors.

Full information estimation

1. Idea
2. Kalman-filter
   - Notation
   - Initialization
   - Recursion
   - Summary and Likelihood
3. Maximum-Likelihood
   - Idea
   - Exercise 4: An and Schorfheide (2007) via Maximum-Likelihood
4. Bayesian methods
   - Idea
   - Metropolis-Hastings-algorithm
   - Remarks
   - Exercise 5: Estimation with Bayesian methods
   - Inference, forecast, model comparison and identification
5. Discussion

Inference, forecast, model comparison and identification

Properties of the Posterior-distribution

- The posterior density combines all information about μ: information after the data is observed as well as information prior to the data.
- Bayesian estimation works for every sample size, however, it has also the following asymptotic properties:
  1. The priors become irrelevant for the determination of the posterior.
  2. The posterior converges to a degenerate distribution around the true value.
  3. The posterior is approximately gaussian.
- Using the posterior distribution one can
  - set up Bayesian confidence intervals (credibility sets),
  - calculate forecasts using the predictive-density: $\mathcal{L}(\mathbf{X_f}|\mathbf{X}) = \int \mathcal{L}(\mathbf{X_f}|(\boldsymbol{\mu}|\mathbf{X}))d\boldsymbol{\mu} = \int \mathcal{L}(\mathbf{X_f}|\boldsymbol{\mu},\mathbf{X})\wp(\boldsymbol{\mu}|\mathbf{X})d\boldsymbol{\mu}$
  - compare models.

Inference, forecast, model comparison and identification

Model comparison

- Models can differ in their prior distribution, the likelihood and the parameters.
- Bayesian approach: Calculate the probability that model i is the true model, given the data.
- Suppose there are $i = 1, 2$ models M_i with prior probability $p_i = P(M_i)$ that Modell M_i is the true model.
- Each model has a set of parameters $\boldsymbol{\mu}_{\mathbf{i}}$ with a prior distribution $\wp_i(\boldsymbol{\mu}_{\mathbf{i}})$ and a likelihood $\mathcal{L}_i(\mathbf{X}|\boldsymbol{\mu})$.
- Then the probability of model 1 being the true model given the data, is given by:

$$\begin{aligned} P(M_1|X) &= \frac{P(M_1)\mathcal{L}_1(\mathbf{X}|M_1)}{\mathcal{L}(\mathbf{X})} = \frac{p_1 \int \mathcal{L}_1(\mathbf{X}, \boldsymbol{\mu}_{\mathbf{1}}|M_1) d\boldsymbol{\mu}_{\mathbf{1}}}{\mathcal{L}(\mathbf{X})} \\ &= \frac{p_1 \int \mathcal{L}_1(\mathbf{X}|\boldsymbol{\mu}_{\mathbf{1}}, M_1)\wp_1(\boldsymbol{\mu}_{\mathbf{1}}|M_1) d\boldsymbol{\mu}_{\mathbf{1}}}{\mathcal{L}(\mathbf{X})} \end{aligned}$$

$$\text{mit } \mathcal{L}(\mathbf{X}) = p_1 \int \mathcal{L}_1(\mathbf{X}|\boldsymbol{\mu}_{\mathbf{1}}, M_1)\wp_1(\boldsymbol{\mu}_{\mathbf{1}}|M_1) d\boldsymbol{\mu}_{\mathbf{1}} + p_2 \int \mathcal{L}_2(\mathbf{X}|\boldsymbol{\mu}_{\mathbf{2}}, M_2)\wp_2(\boldsymbol{\mu}_{\mathbf{2}}|M_2) d\boldsymbol{\mu}_{\mathbf{2}}$$

Inference, forecast, model comparison and identification

Model comparison

- The expected value of the likelihood given the prior distribution is the so-called *marginal-likelihood* for model i:

$$m_i(\mathbf{X}) = \int \mathcal{L}_i(\mathbf{X}|\boldsymbol{\mu}_\mathbf{i}, M_i)\wp_i(\boldsymbol{\mu}_\mathbf{i}|M_i)d\boldsymbol{\mu}_\mathbf{i}$$

- Using this, one can calculate the *posterior-odds*:

$$PO_{12} = \frac{P(M_1|\mathbf{X})}{P(M_2|\mathbf{X})} = \underbrace{\frac{p_1}{p_2}}_{\text{Prior-Odds-Ratio}} \cdot \underbrace{\frac{m_1(\mathbf{X})}{m_2(\mathbf{X})}}_{Bayes-Faktor}$$

- Together with $P(M_1|\mathbf{X}) + P(M_2|\mathbf{X}) = 1$ one gets the *posterior-model-probabilities*:

$$P(M_1|\mathbf{X}) = \frac{PO_{12}}{1 + PO_{12}}, \qquad P(M_2|\mathbf{X}) = 1 - P(M_1|\mathbf{X}).$$

Inference, forecast, model comparison and identification

Model comparison

- The *marginal likelihood* measures the quality of a model to characterize data.
- The *Posterior-Odds* don't hint to the true model. They solely describe which model, compared to the other, has the highest conditional probability.
- A $PO_{12} >> 1$ is an indication that the data as well as the priors prefer model 1.
- Guidelines of Jeffrey (1961):
    - $1:1-3:1$ weak evidence for model 1,
    - $10:1-100:1$ strong evidence for model 1,
    - $>100:1$ decisive evidence for model 1.
- Implementation and calculation of the integrals is done by numerical MCMC- and sampling-methods, as well as Laplace- or Harmonic-Mean-approximation.

Exercise 5: Estimation with Bayesian methods

(g) Use the same dataset to estimate the parameters of a misspecified model. Use the same model, however, with a small difference:

- **a Cobb-Douglas production function** $\Leftarrow$
- or a separable utility function,
- or a model in which the household supplies inelastically one unit of labor.

Hint: *Don't forget to adjust the model equations as well as the steady-state block.*

(h) Compare the estimation of the common parameters as well as the marginal densities of the different models. Calculate the *posterior-odds* and the *posterior-model-probabilities*.

- Consider two vectors of parameters $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ for which

$$\mathcal{L}(\mathbf{X}|\boldsymbol{\mu_1}) = \mathcal{L}(\mathbf{X}|\boldsymbol{\mu_2}).$$

- If $\boldsymbol{\mu_1} = \boldsymbol{\mu_2}$, then there is identification. If, however, $\boldsymbol{\mu_1} \neq \boldsymbol{\mu_2}$, then one does not know which vector of parameters has generated the data.
- Special case: $\mathcal{L}(\mathbf{X}|\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) = \mathcal{L}(\mathbf{X}|\boldsymbol{\mu_1})$.
  - $\boldsymbol{\mu_2}$ is not identifiable through the data.
  - What are the implications for the posterior-distribution $\wp(\boldsymbol{\mu_2}|\mathbf{X})$?

$$\begin{aligned}
\wp(\boldsymbol{\mu_2}|\mathbf{X}) &= \int \wp(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}|\mathbf{X}) d\boldsymbol{\mu_1} \\
&= \left[\int \mathcal{L}(\mathbf{X}|\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) \wp(\boldsymbol{\mu_1}) \wp(\boldsymbol{\mu_2}|\boldsymbol{\mu_1}) d\boldsymbol{\mu_1}\right] \cdot [\mathcal{L}(\mathbf{X})]^{-1} \\
&= \left[\int \mathcal{L}(\mathbf{X}|\boldsymbol{\mu_1}) \wp(\boldsymbol{\mu_1}) \wp(\boldsymbol{\mu_2}|\boldsymbol{\mu_1}) d\boldsymbol{\mu_1}\right] \cdot [\mathcal{L}(\mathbf{X})]^{-1} \\
&= \int \wp(\boldsymbol{\mu_1}|X) \wp(\boldsymbol{\mu_2}|\boldsymbol{\mu_1}) d\boldsymbol{\mu_1}
\end{aligned}$$

- If the prior distribution of $\boldsymbol{\mu_2}$ is independent of $\boldsymbol{\mu_1}$, i.e. $\wp(\boldsymbol{\mu_2}|\boldsymbol{\mu_1}) = \wp(\boldsymbol{\mu_2})$, then $\wp(\boldsymbol{\mu_2}|\mathbf{X}) = \wp(\boldsymbol{\mu_2})$. Beliefs about $\boldsymbol{\mu_2}$ don't change after observing the data.
- If the independence of the priors ist not given, then the information in the data has an effect on the beliefs about $\boldsymbol{\mu_2}$ over the effect on $\boldsymbol{\mu_1}$.
- If an identification through data is not possible, then the difference between prior and posterior is due to the prior.
- In practice on has to take a stand, if parameters are identified through data or through the prior.

Full information estimation

1. Idea
2. Kalman-filter
   - Notation
   - Initialization
   - Recursion
   - Summary and Likelihood
3. Maximum-Likelihood
   - Idea
   - Exercise 4: An and Schorfheide (2007) via Maximum-Likelihood
4. Bayesian methods
   - Idea
   - Metropolis-Hastings-algorithm
   - Remarks
   - Exercise 5: Estimation with Bayesian methods
   - Inference, forecast, model comparison and identification
5. Discussion

Discussion

- More restrictive assumptions are needed compared to the limited information estimation: specification of the distribution of the schocks, i.e. the likelihood.
- Advantages of a *Maximum-Likelihood*-estimation lie in the full characterization of the data-generating-process and the exact, consistent and efficient estimation of the parameters.
- „Dilemma of absurd parameter estimates“: Problem of the ML-estimation due to wrong distributional assumptions, problems in the optimization algorithm or non-separable identifiable parameters.
- Even transformations, upper and lower bounds, etc. are only limited to help overcome this problem, when the likelihood is flat.

General problem

For a mathematical expression with many conditions and parameters, but only a limited sample, there can exist different combinations of parameters that yield the same result and a similar dataset.

Discussion

- This is where Bayesian methods come in and bridge the gap between calibration and the *ML-principle*.
- Considering priors one can incorporate additional information into a model.
- „Dilemma of absurd parameter estimates“: Even with Bayesian means it is not possible to estimate these parameters (the posterior looks almost the same as the prior), but one can assign probability such that these parameters are very unlikely.

⇒ Using priors one can exclude these absurd parameter estimates.

- Nevertheless the point of robustness and identification of the parameters remains a critical topic.

Discussion

An und Schorfheide (2006, S.124)

Once one acknowledges that the DSGE model provides merely an approximation to the law of motion of the time series (. . .), then it seems reasonable to assume that there need not exist a single parameter vector (. . .), that delivers, say, the „true“ intertemporal substitution elasticity or price adjustment costs and, simultaneously, the most precise impulse responses to a technology or monetary policy shock. Each estimation method is associated with a particular measure of discrepancy between the „true“ law of motion and the class of approximating models.