Structural Macroeconometrics

Chapter IV
Bayesian Estimation

Johannes Pfeifer

Kobe University

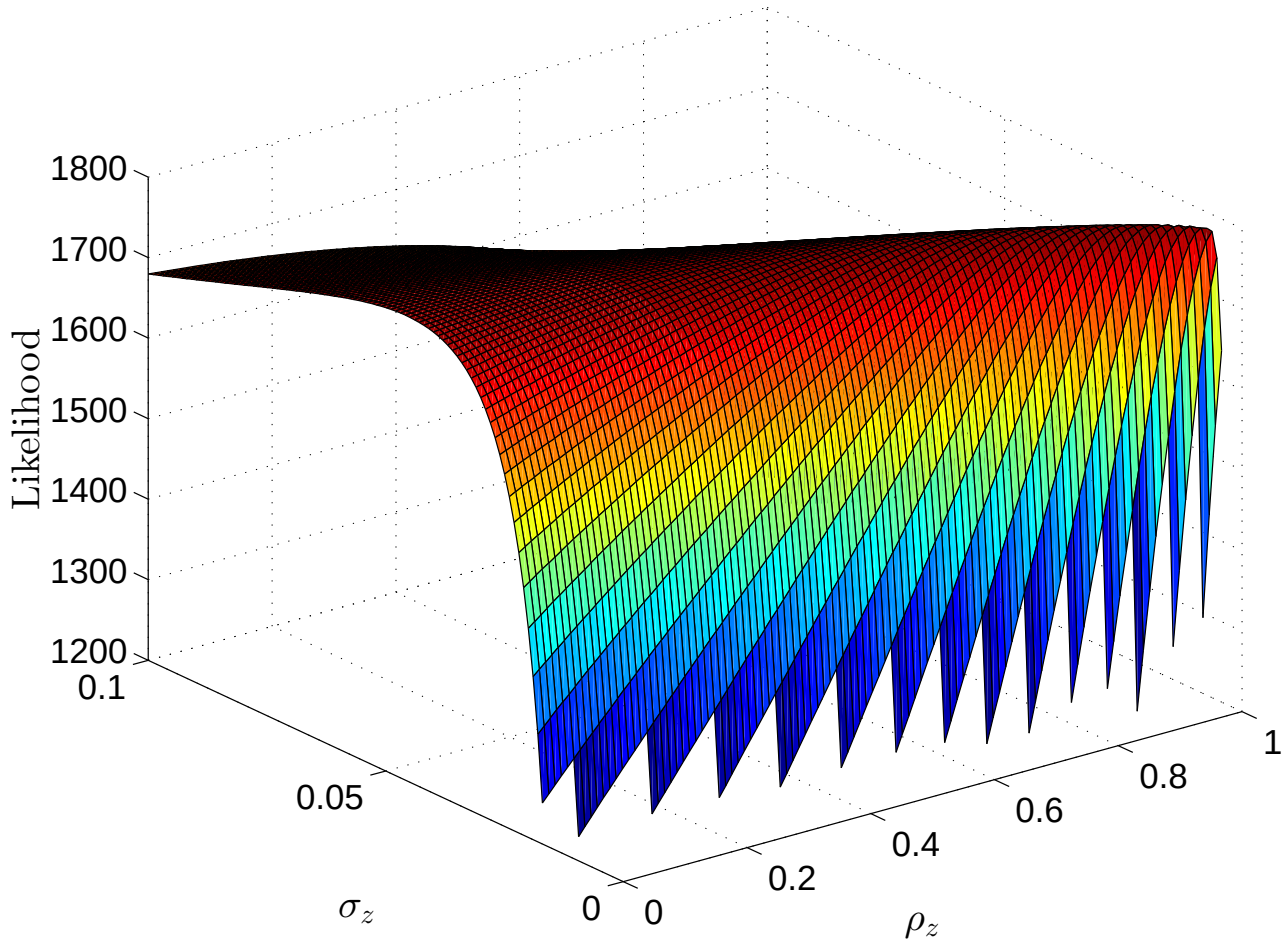
Summer 2018

Outline

Review

- Log-Linearized DSGE model solution takes state-state space form
- Typically, not all state variables are observed
- Solution: use Kalman Filter to deal with unobserved states and construct likelihood
- Result: likelihood $\mathcal{L}(y^T|\theta)$, where θ is the vector of deep parameters and y^T is the complete history of observables
- Estimating meant finding the parameters θ that maximized $\mathcal{L}(y^T|\theta)$

The Likelihood



- The likelihood is a high-dimensional object
- Even for simple models, it can be ill-behaved, showing hardly any curvature and exhibiting many local maxima

The Likelihood

- For more complicated models, you can think of it as an egg-crate
- "Dilemma of absurd parameter estimates" (An and Schorfheide 2007): ML estimates often at odds with information from outside of the model

Bayesian vs. Frequentist Philosophy

- In principle, there is a fundamental philosophical difference between Bayesian and frequentist econometrics
- Most macroeconomists are not Bayesian believers but pragmatists - up to the point that Bayesian and frequentist techniques are used in the same papers
- Adopting Bayesian techniques makes our lives easier
- Historically, Bayesian were the minority because computing power was not sufficient to apply their methods
- There are very good reasons to be Bayesian (see e.g. Berger and Wolpert 1988; Sims 2007)

Bayesian vs. Frequentist Philosophy: concept of probability

- Bayesian probability is usually associated with **degrees of belief or degrees of knowledge**
- Frequentist probability is usually based on **features of hypothetically observable systems**
  $\rightarrow$ relative frequency of an event "in the long run"
- Example: "There is a .50 probability that a fair coin will land heads."
  - Bayesian: belief evenly divided between the coin landing heads or tails
  - Frequentist: result if coin were flipped a hypothetical infinite number of times
- Frequentist analysis limited to inference about relative frequency of events in the long run
- But: long run unobservable and not in researchers' actual interest
- Bayesians can apply probability to anything that can be the subject of belief or knowledge (hypotheses, statistical parameters, entire statistical models)

Bayesian vs. Frequentist Philosophy: conditioning sets

- Frequentists estimate sampling distribution for estimated effect/parameter
  - distribution simulates observed effect if repeated infinite number of times, with only sampling error affecting results
  - data carry uncertainty via sampling distribution, while parameter has fixed "true" population value
    $\rightarrow$ **conditioning on parameter** $P(Y|\theta)$
  - Null Hypothesis Significance Testing (NHST) involves testing a hypothesis we do not believe in
  - p-value gives probability of effect, assuming Null is true
    $\rightarrow$ relies on hypothetically repeating experiment that never occurred
- Bayesians make direct probabilistic statements about effect/parameter, based on observed data
  - Parameter is treated as random/uncertain while data is taken as fixed
    $\rightarrow$ **conditioning on data** $P(\theta|Y)$

The posterior density

- Central object: **posterior probability/posterior density**

$$P(\theta|Y) \tag{1}$$

  of a parameter θ, conditional on having seen the data Y
- Posterior is fully-fledged density function!
- Allows statements like: the regression indicates that with 90% probability the fiscal multiplier is between 0.3 and 1.5
- NHST does not allow such statements!
- Rejecting the Null does not tell us anything about likely effect
  $\rightarrow$ only know the data is unlikely to come from a world where the Null is true
  $\rightarrow$ estimated value is then preferred
- Problem: how to obtain this posterior?

The Central Element: Bayes Rule

- Consider the basic laws of probability for two events A and B:

$$p(A, B) = p(A|B)\, p(B) \tag{2}$$
$$p(A, B) = p(B|A)\, p(A) \tag{3}$$

  where $p(A, B)$ is the joint probability, $p(A|B)$ is the conditional probability, and $p(B)$ the marginal probability

- Equating them results in **Bayes Rule**

$$p(B|A) = \frac{p(A|B)\, p(B)}{p(A)} \tag{4}$$

- Provides consistent rule for rational decision maker on how to update beliefs in the face of evidence!

Bayes Rule Applied

- Now apply this to a case where we want to use some data y^T to infer a parameter vector θ

$$p\left(\theta|y^T\right) = \frac{\mathcal{L}\left(y^T|\theta\right)p\left(\theta\right)}{p\left(y^T\right)} \propto \mathcal{L}\left(y^T|\theta\right)p\left(\theta\right) \tag{5}$$

- $p\left(\theta|y^T\right)$ is called the **posterior distribution**; it incorporates all we know about the parameters and incorporates data and non-data information
- $p\left(\theta\right)$ is the **prior distribution**; it is independent of the data and incorporates all non-data information on the parameters
- $p\left(y^T\right)$ is the **data density**; as it is independent of the parameters, it can be treated as a proportionality constant (except when doing model comparison)
- Equation (5) says that the **posterior is proportional to likelihood times prior**

The Problem with Bayesian Estimation

- Historically, Bayesians were the minority because computing power was not sufficient to apply their methods. Why?
- $$p\left(\theta|y^T\right) = \frac{\mathcal{L}\left(y^T|\theta\right)p\left(\theta\right)}{p\left(y^T\right)} \propto \mathcal{L}\left(y^T|\theta\right)p\left(\theta\right) \tag{5}$$
  describes a **full distribution** that often is not analytically tractable
- Say, we are interest in characterizing the posterior distribution by its mean and variance. We need to compute:
$$E\left(\theta|y^T\right) = \int \theta p\left(\theta|y^T\right)d\theta \tag{6}$$
and
$$\begin{aligned}\text{var}\left(\theta|y^T\right) &= E\left(\theta^2|y^T\right) - \left[E\left(\theta|y^T\right)\right]^2 \\ &= \int \theta^2 p\left(\theta|y^T\right)d\theta - \left[E\left(\theta|y^T\right)\right]^2 \end{aligned} \tag{7}$$
- In both cases we need to evaluate an integral that usually cannot be worked out analytically!

Getting the Mean

- Ingenious Idea: we are typically interested in integrals of the form

$$E\left(g\left(\theta\right)|y^{T}\right)=\int g\left(\theta\right)p\left(\theta|y^{T}\right)d\theta \tag{8}$$

- If we had **iid draws from the posterior**, we could simply use a law of large numbers:

$$E\left(g\left(\theta\right)|y^{T}\right)=\int g\left(\theta\right)p\left(\theta|y^{T}\right)d\theta\approx\frac{1}{S}\sum_{s=1}^{S}g\left(\theta_{s}\right)=\hat{g}_{S} \tag{9}$$

- Replace integral by sum over S draws from the posterior distribution $p(\theta|y^{T})$
- This is called **Monte Carlo Integration**

Numerical Standard Error

- How good is this estimate?
- Central limit theorem ensures that

$$\sqrt{S}\left(\hat{g}_S - E\left(g\left(\theta\right)|y^T\right)\right) \rightarrow N\left(0, \sigma_g^2\right) \tag{10}$$

- Hence,

$$E\left(g\left(\theta\right)|y^T\right) \sim N\left(\hat{g}_S, \frac{\sigma_g^2}{S}\right) \tag{11}$$

- The standard deviation

$$\sigma_g^2 = var(g(\theta)|y^T) \tag{12}$$

  can be estimated by the variance in our Monte Carlo sample
- The **Numerical Standard Error (NSE)**

$$\frac{\sigma_g}{\sqrt{S}} \tag{13}$$

  is a measure of approximation error.

Some More Jargon

- Even if you are not a Bayesian believer, you should be familiar with the different terminology
- Denote with $\omega = g(\theta)$ some vector of functions of the parameters θ, defined over some region Ω

Definition 1 (Credible Set)

The set $C \subseteq \Omega$ is a $100(1-\alpha)\%$ **credible set** with respect to $p(\omega|y)$ if

$$p(\omega \in C|y) = \int_C p(\omega|y)d\omega = 1-\alpha \tag{14}$$

- We are typically interested in the smallest one, the "Bayesian equivalent to a confidence interval"

Definition 2 (Highest Posterior Density Interval)

A $100(1-\alpha)\%$ **highest posterior density interval** for ω is a $100(1-\alpha)\%$ **credible interval** that has a smaller area than any other $100(1-\alpha)\%$ credible interval for ω

Credible Set vs. Confidence Intervals

- Bayesian credible sets are **post-experimental**: conditional upon observing the data, a $(1-\alpha)\%$ credible set contains the true parameter, which is a random variable, with $(1-\alpha)\%$ probability
- Frequentist confidence intervals are based on the long-run performance if an experiment is repeatedly performed
- They are thus inherently **pre-experimental** and treat the parameter as non-random
- CIs are constructed so that the true parameter will be contained in $(1-\alpha)\%$ of the CI constructed from the data
- Moreover, a $(1-\alpha)\%$ CI contains the true parameter value with probability $(1-\alpha)\%$ only before one has seen the data
- After the data has been seen, the probability is zero or one
- CIs do not help in putting constraints on a parameter after data is observed
- One can only say: the true parameter is either in the CI or not
- Thus, both answer fundamentally different questions

Credible Set vs. Confidence Intervals: Example

- We want to infer a parameter θ and observe two independent random variables X_1 and X_2 with

$$P(X_i = \theta - 1) = P(X_i = \theta + 1) = 0.5\,, i \in 1, 2$$

- The smallest 75% CI is

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) \text{ if } X_1 \neq X_2 \\ X_1 - 1 \text{ if } X_1 = X_2 \end{cases}$$

- When repeatedly sampling, the true θ will be in this interval in 75% of the cases, because
  - if $x_1 \neq x_2$ (half of the cases), we know for sure that $\theta = 0.5(x_1 + x_2)$ $\Rightarrow$ correct in 100% of these cases
  - if $x_1 = x_2$ (half of the cases), either $\theta = x_1 + 1$ or $\theta = x_1 - 1$ $\Rightarrow$ correct in 50% of these cases
- Bayesians, upon observing the data, are 100% sure about θ or 50%
- Does it make sense to report pre-experimental measure when it is known to be misleading after seeing the data?

Problems everywhere: Posterior Sampling

- Monte Carlo Integration sounds nice and easy, but there's a problem: how to get draws from an intractable distribution?
- This is the big topic of **posterior sampling algorithms**
- The most important ones are
  - Importance Sampling
  - Gibbs-Sampling (S. Geman and D. Geman 1984)
  - Metropolis-Hastings algorithm (Hastings 1970; Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller 1953)
- We will only consider the last one
- Gibbs-Sampler is a special case of the Metropolis-Hasting algorithm (see Gelman 1992)
- Gibbs sampling and the Metropolis-Hastings algorithm give rise to **correlated random draws from the posterior** and belong to the class of **Monte Carlo Markov Chain** algorithms

Prior distributions and subjectivity

- There is a large discussion about the "subjectivity" of priors (e.g. Berger 2006)
- We will abstract from the philosophical issues arising here
- Necessarily subjective choices like e.g. the model used tend to be more important than the prior over parameters
- But be aware: issue is contentious as "subjectivity" does not square well with the "scientific method"

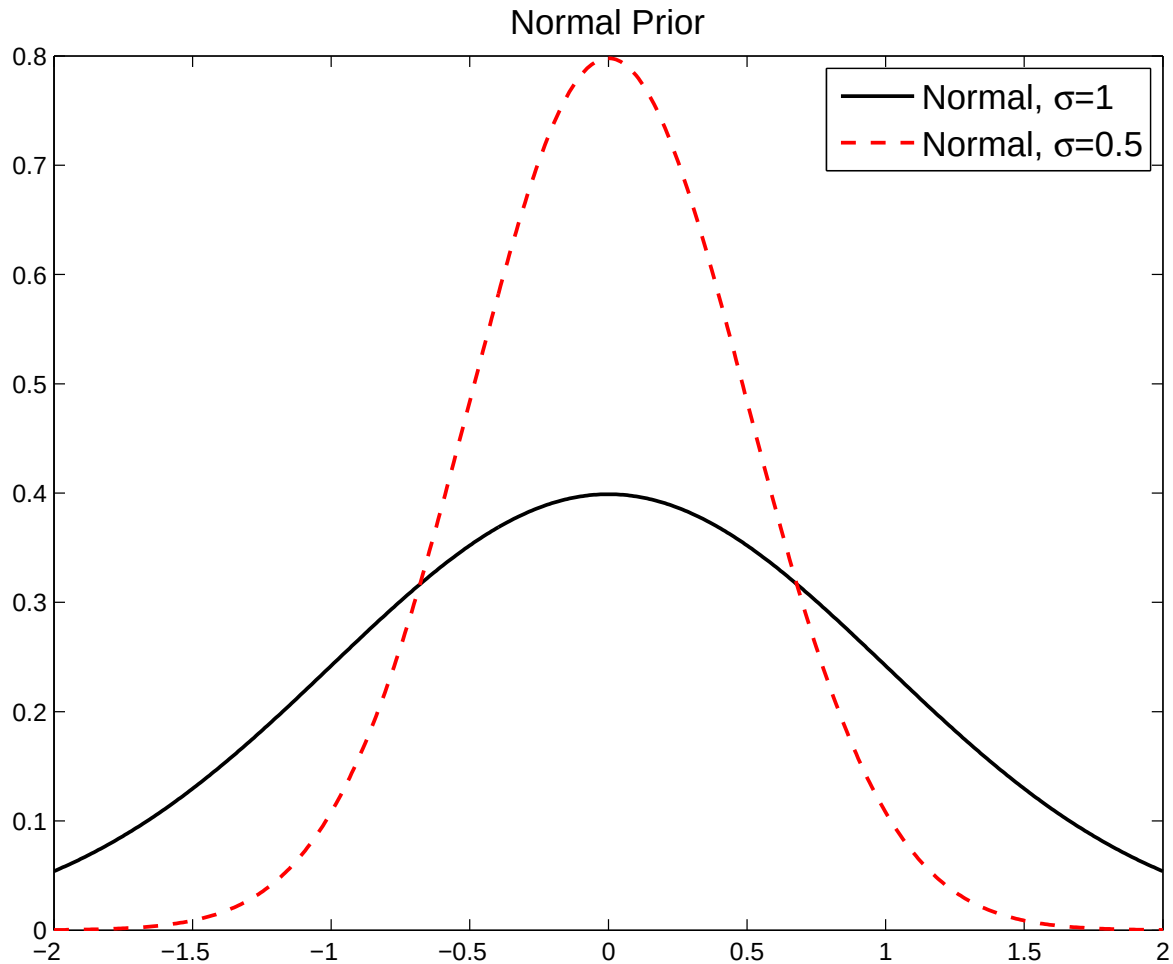
First Things First: Specifying the Prior

- We need to specify the prior distribution $p(\theta)$
- In general, this can be a full multivariate distribution
- In practice, people typically use **independent priors** (Andrle and Benes 2013, is a notable exception)
- Choosing sensible priors is hard, see Del Negro and Schorfheide (2008)
- What is the purpose of priors?
  1. Incorporating information extraneous to the sample
  2. Providing additional curvature to the likelihood function and straightening out cliffs
- Parameter range often narrows down prior choice
- After choosing the priors, you should do a **prior predictive**: check what the prior implies for the question you are asking (see Leeper, Traum, and T. B. Walker 2017)
- This way, you make sure that your estimation results are not solely driven by your prior

Endogenous Priors

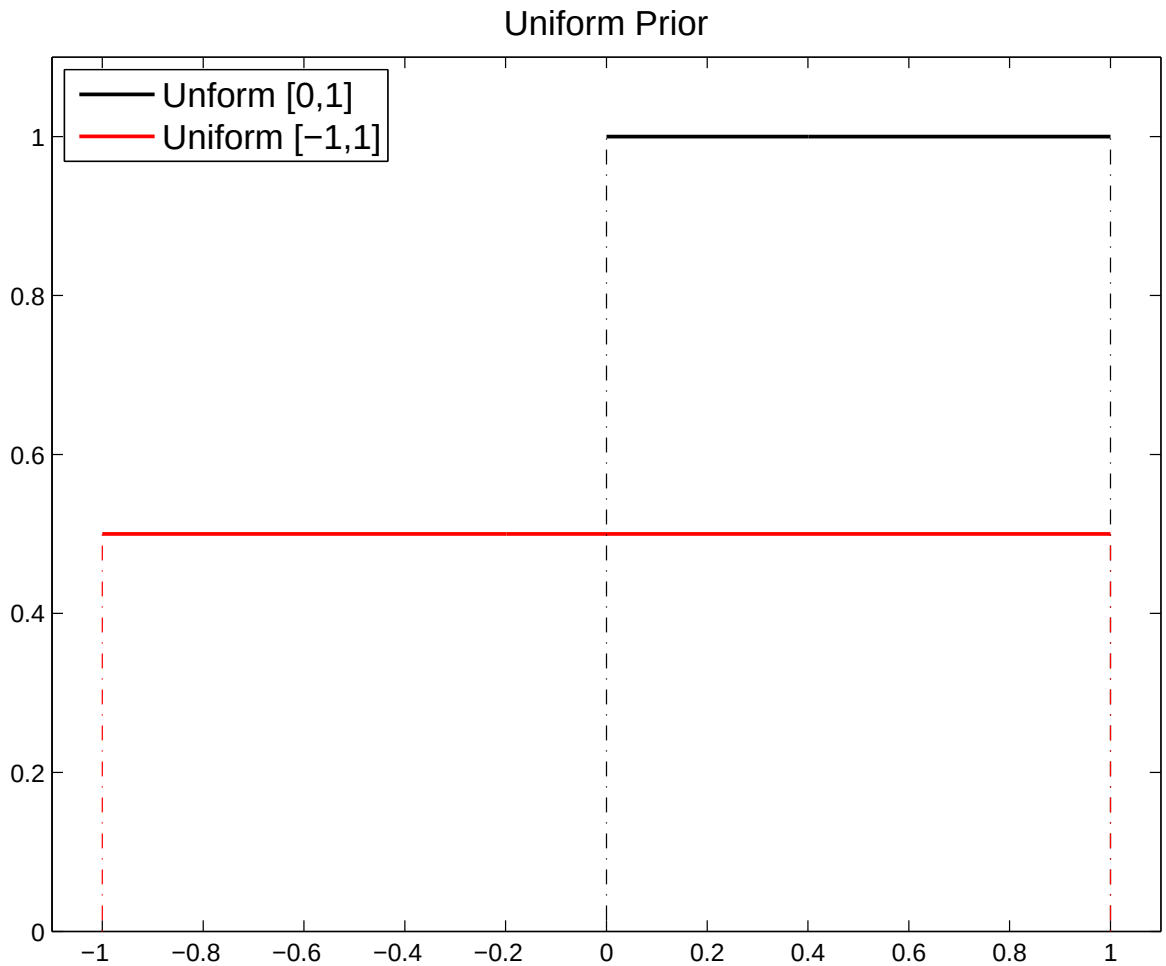
- We are doing **full information** estimation
- We are not matching moments!
- Regularly happens that estimated model implies too high variances
- Proposed solution: **endogenous priors** (Christiano, Trabandt, and Walentin 2011; Del Negro and Schorfheide 2008)
- Motivated by sequential Bayesian learning
- Starting from independent initial priors, use standard deviations observed in a "pre-sample" to update those initial priors.
- Product of the initial priors and the pre-sample likelihood of the standard deviations of the observables is used as the new prior

Normal Prior



- Unbounded support $(-\infty, \infty)$ and symmetric
- Typically used for e.g. feedback parameters where sign is unknown

Uniform Prior



- Bounded support on $[LB, UB]$

Uniform Prior

- For a variable $Y \sim U(a, b)$, the PDF is given by

$$f_U(y|a,b) = \begin{cases} \frac{1}{b-a} \text{ if } a \leqslant y \leqslant b \\ 0 \text{ otherwise} \end{cases} \tag{15}$$

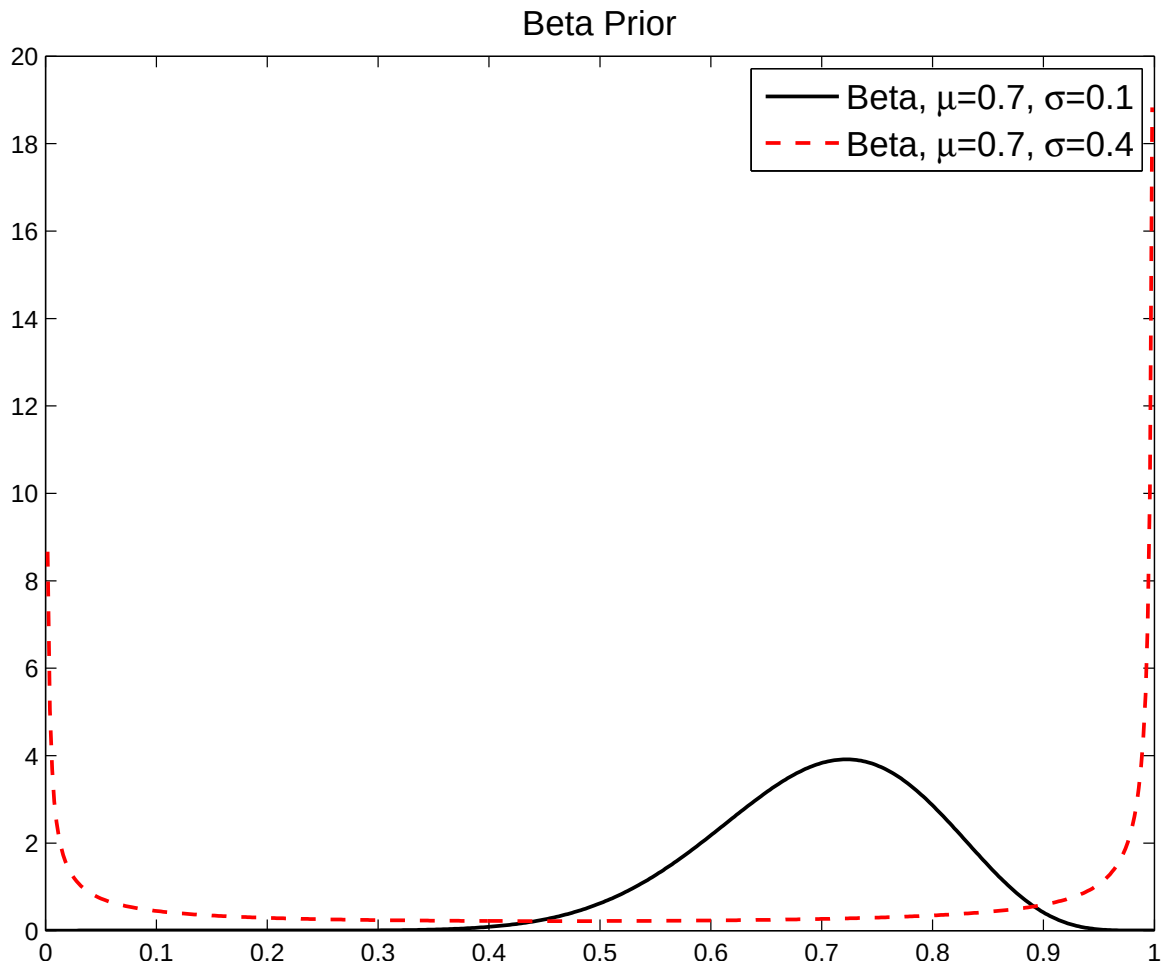
  where $-\infty < a < b < \infty$
- Moreover,

$$E(Y) = \frac{a+b}{2} \tag{16}$$

$$\text{var}(Y) = \frac{(b-a)^2}{12} \tag{17}$$

- Prior is "flat", i.e. all points are equally likely and it does not introduce curvature
- Often called **uninformative prior**
- But: it is informative in the sense that you say all parameter values in the interval are equally likely
- Beware: with bounds at infinity, prior is **improper** (cf. model comparison)

Beta Prior



- Bounded support on $[0, 1]$
- Often used for autoregressive parameters, the discount factor, Calvo parameters, etc.

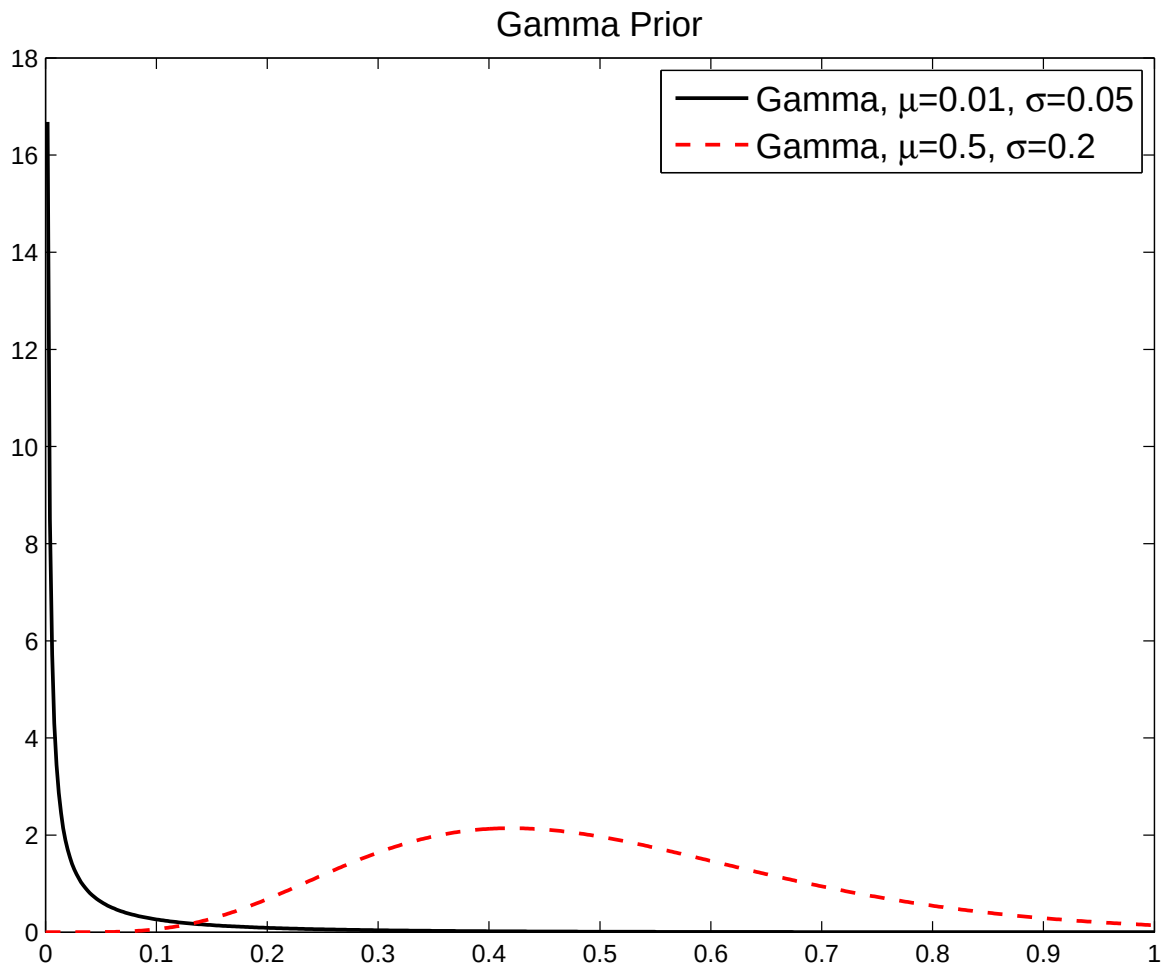
Beta Prior

- Suitable transformations allow scaling it to different support like $[-1, 1]$
- Relatively flexible in allowing for different shapes, but care is required
- Due to bounded support, a high variance can result in assigning high mass to extremes in the tails

$\Rightarrow$ mode, i.e. point of highest likelihood, far away from prior mean

- Always plot the prior distribution!

Gamma Prior



- Support $[0, \infty)$, i.e. 0 is included
- Typically used for variances and Taylor rule feedback parameters

Gamma Prior

- For a variable $Y \sim G(\mu, \nu)$, where μ is the mean and ν the degrees of freedom, the PDF is given by

$$f_G(y|\mu,\nu) = \begin{cases} \frac{1}{\left(\frac{2\mu}{\nu}\right)^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} y^{\frac{\nu-2}{2}} e^{-\frac{y\nu}{2\mu}} \text{ if } 0 \leqslant y \leqslant \infty \\ 0 \text{ otherwise} \end{cases} \tag{18}$$

  where $\Gamma()$ is the Gamma function
- Moreover,

$$E(Y) = \mu \tag{19}$$

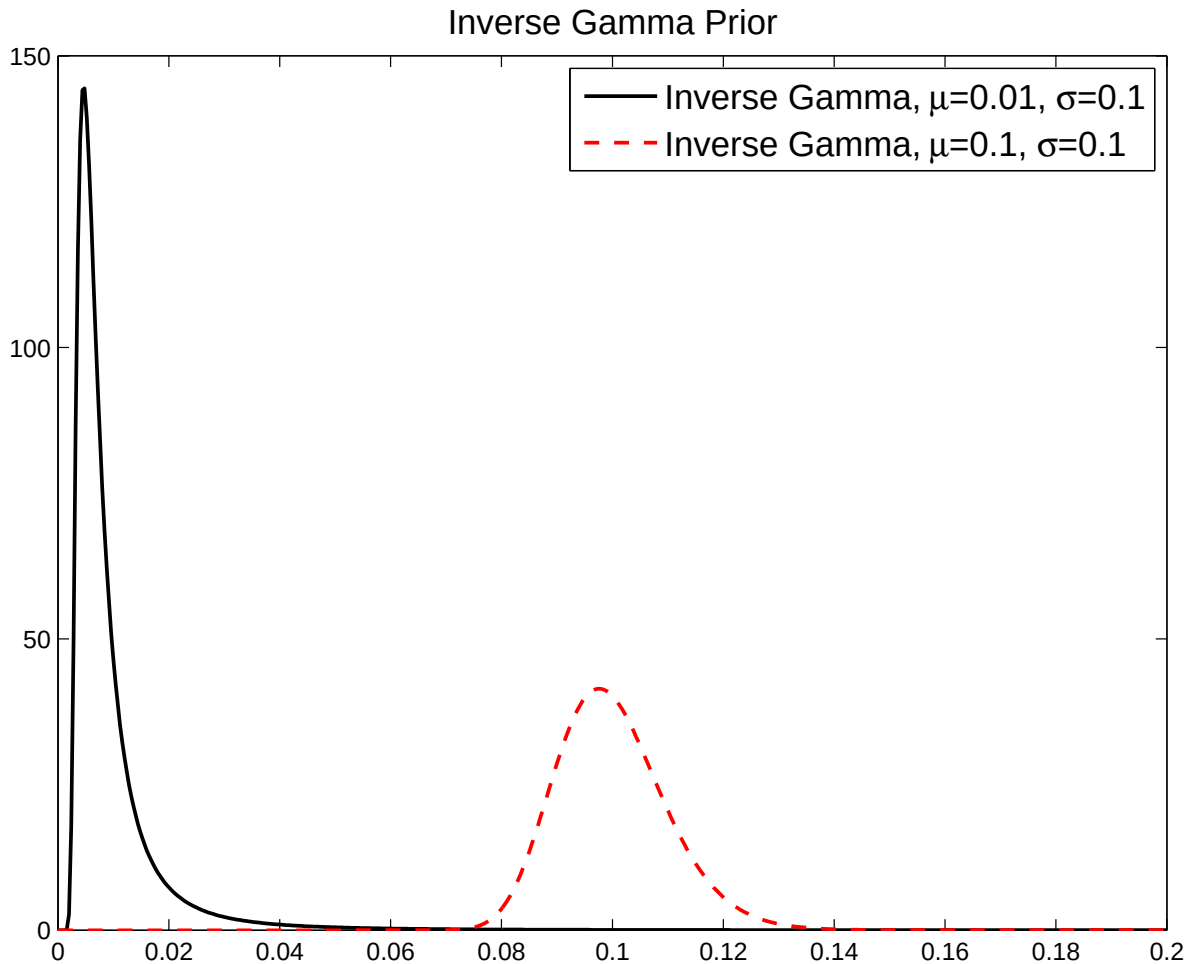
$$\text{var}(Y) = \frac{2\mu^2}{\nu} \tag{20}$$

- Beware: Matlab uses $f_G(y|a,b)$ with $a = \frac{\nu}{2}, b = \frac{2\mu}{\nu}$
- Thus, use

$$a = \frac{\mu^2}{\sigma^2} \tag{21}$$

$$b = \frac{\sigma^2}{\mu} \tag{22}$$

Inverse Gamma Prior



- Support $(0, \infty)$, i.e. 0 is not included
- Typically used for variances

Inverse Gamma Prior

- If y is Inverse Gamma, then $1/y$ is Gamma distributed
- The pdf is given by

$$f_{IG}\left(y|a,b\right) = \begin{cases} \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by} \text{ if } 0 \leqslant y \leqslant \infty \\ 0 \text{ otherwise} \end{cases} \tag{23}$$

with

$$E\left(Y\right) = \frac{a}{b} \tag{24}$$

$$\text{var}\left(Y\right) = \frac{a}{b^2} \tag{25}$$

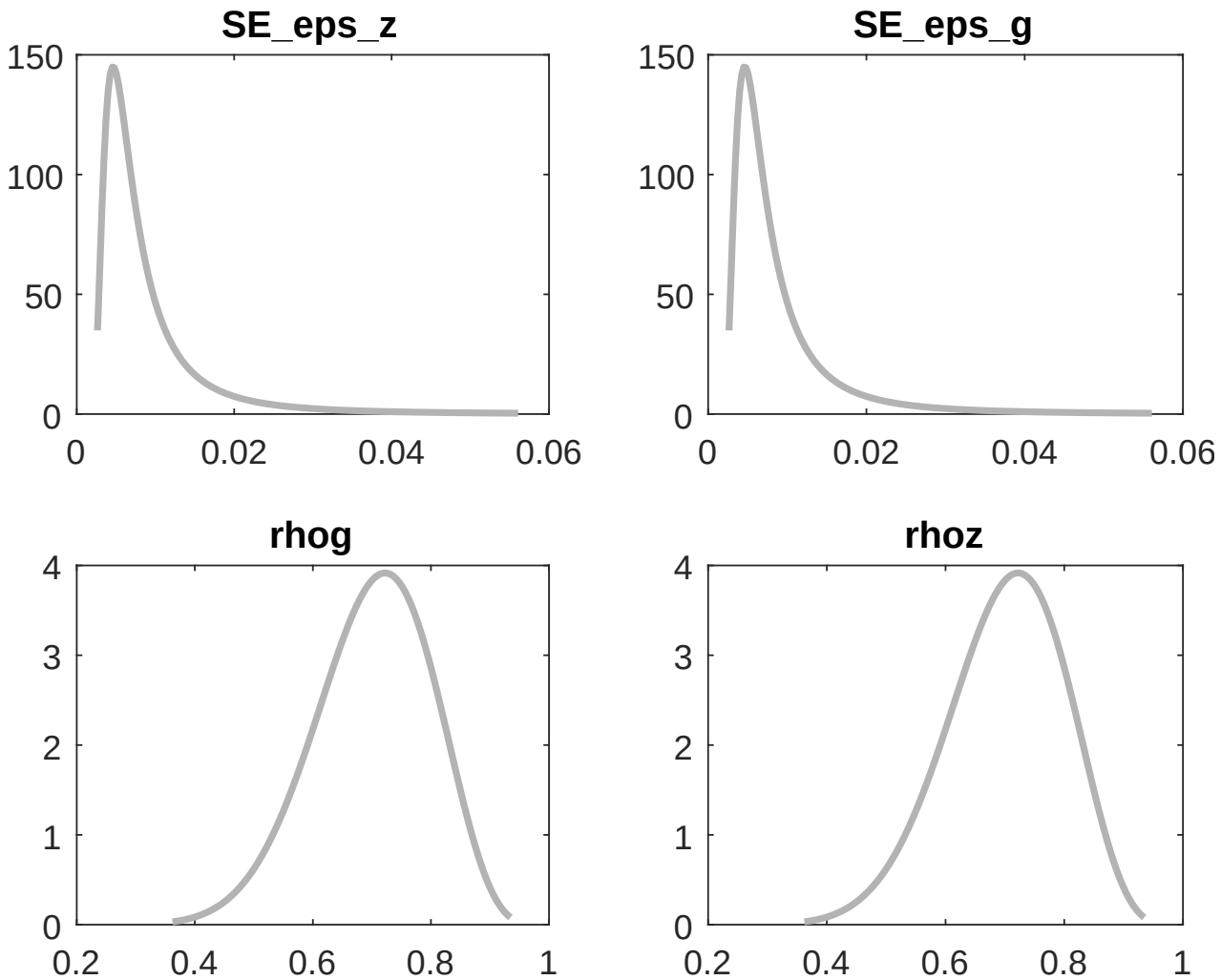
- Thus:

$$a = \frac{\mu^2}{\sigma^2} \tag{26}$$

$$b = \frac{\mu}{\sigma^2} \tag{27}$$

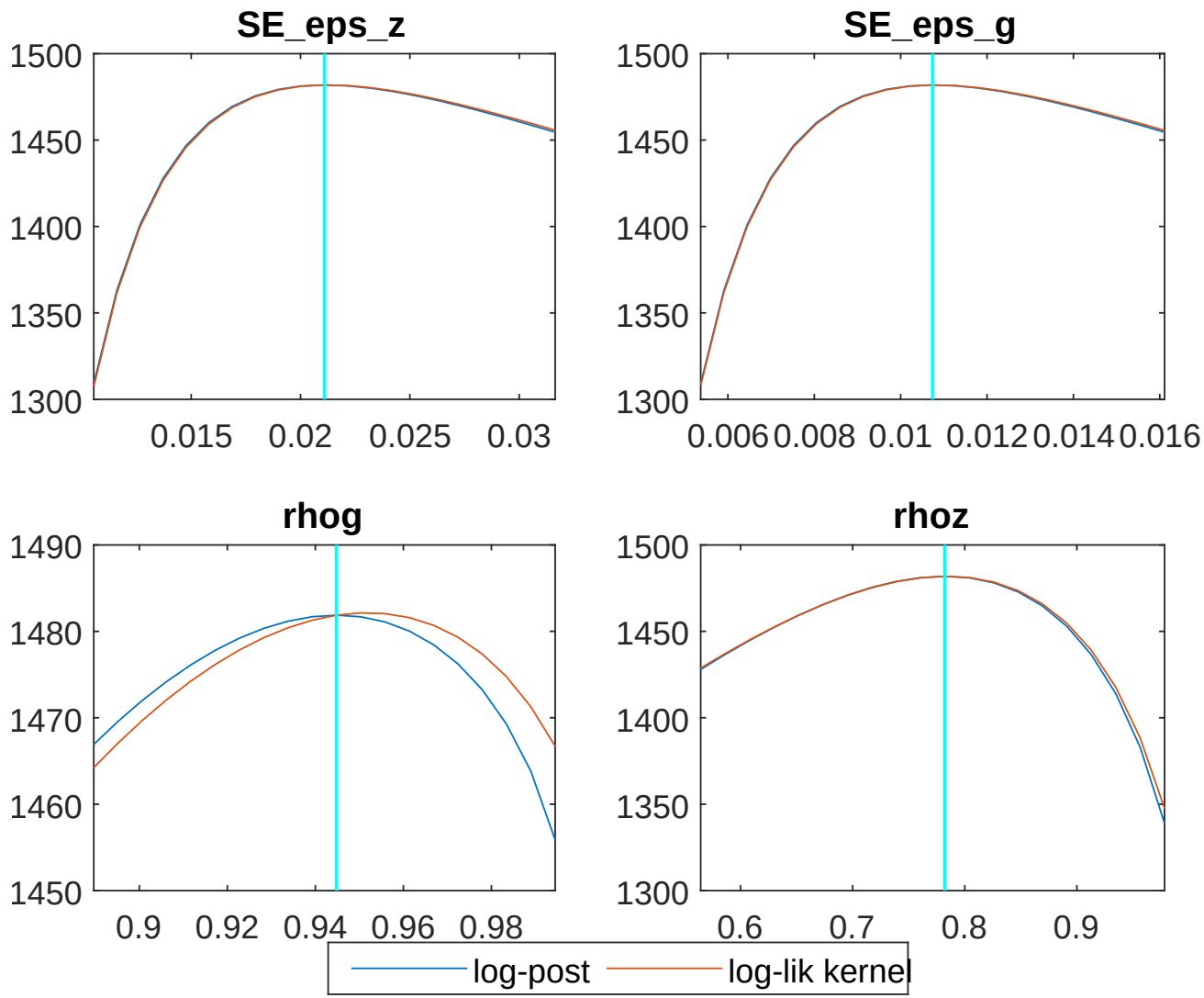
- Theoretically, a better choice for variances if you want to prevent **stochastic singularity**
- But: variance of exactly 0 is a zero probability event

Example



- Always check whether your prior looks sensible

Example



- Notice how the prior affects the posterior for `rhog`

Markov Chains: Discrete State Space

- Stochastic process x_t has **Markov property** if

$$\text{Prob}(x_{t+1}|x_t, x_{t-1}, \ldots, x_{t-k}) = \text{Prob}(x_{t+1}|x_t) \tag{28}$$

- We assume the process can be characterized by a time-invariant **Markov Chain**
- Such a Markov chain is characterized by
  1. n-dimensional state space consisting of the n basis vectors e_i of $\mathcal{R}^n$
  2. $n \times n$ transition matrix P recording probability of moving from one state to another:

$$P_{ij} = \text{Prob}(x_{t+1} = e_j|x_t = x_i) \tag{29}$$

  3. $n \times 1$ vector π_0 of initial probabilities of being in a particular state i at $t = 0$
- Markov Chain theory is interested in **existence of and convergence to invariant distribution** π, given transition matrix
- **Invariant distribution**: π is unaltered when passing through transition:

$$\pi' = \pi' P \tag{30}$$

Markov Chains: Continuous State Space

- Things are more complicated when state space is continuous
- Could try to work with densities
- But: there may be mass points
- Solution: work with **probability measures** that can handle this
- Leads to more heavy notation: e.g. transition kernel will take place of transition matrix

Markov Chains: Invariant Distribution (Continuous Case)

- Consider a Markov-Chain with continuous state-space and **transition kernel** $P(x, A)$, where
  - $\theta \in \mathcal{R}^d$
  - $A \in \mathcal{B}$, where $\mathcal{B}$ is the Borel σ-field on $\mathcal{R}^d$ (σ-algebra)
- The invariant distribution π^* is characterized by

$$\pi^*\left(d\tilde{\theta}\right) \equiv \pi\left(\tilde{\theta}\right) d\tilde{\theta} = \int_{R^d} P\left(\theta, d\tilde{\theta}\right) \pi\left(\theta\right) d\theta \,, \tag{31}$$

  i.e. $\pi(\theta)$ is unaltered when passing it through the transition kernel
- π is the density with respect to Lebesgue measure of π^*, i.e.

$$\pi^*(A) = \int_A \pi\left(\tilde{\theta}\right) d\tilde{\theta} \,, \tag{32}$$

- In our case: π^* is our known invariant distribution, but how to get $P(\theta, A)$?

Transition Kernel and the Reversibility Condition

- Suppose **transition kernel**, for some function $p(\theta, \tilde{\theta})$ is expressed as

$$P(\theta, d\tilde{\theta}) = p(\theta, \tilde{\theta}) d\tilde{\theta} + r(\theta)\delta_\theta(d\tilde{\theta}) \tag{33}$$

  - $p(\theta, \theta) = 0$
  - $\delta_\theta\left(d\tilde{\theta}\right) = \begin{cases} 1 \text{ if } \theta \in d\tilde{\theta} \\ 0 \text{ otherwise} \end{cases}$
  - $r\left(\theta\right) = 1 - \int_{R^d} p\left(\theta, \tilde{\theta}\right) d\tilde{\theta}$ is the probability of staying at θ

Theorem 3 (Sufficiency of the Reversibility Condition)

*If function $p(\theta, \tilde{\theta})$ satisfies the **reversibility condition***

$$\pi\left(\theta\right) p(\theta, \tilde{\theta}) = \pi(\tilde{\theta}) p(\tilde{\theta}, \theta)\,, \tag{34}$$

then $\pi(\cdot)$ is the invariant distribution of $P(\theta, \cdot)$

- The probability of moving from θ to $\tilde{\theta}$, with θ generated from $\pi(\cdot)$, is equal to the probability of moving from $\tilde{\theta}$ to θ, where $\tilde{\theta}$ comes from the same distribution $\pi(\cdot)$

Proof of the Sufficiency of the Reversibility Condition

Proof.

Consider the RHS of equation (31):

$$\begin{aligned}
\int P(\theta, A)\,\pi(\theta)\,d\theta &\overset{(33)}{=} \int \left[\int_A p\left(\theta, \tilde{\theta}\right) d\tilde{\theta}\right] \pi(\theta)\,d\theta + \int r(\theta)\,\delta_\theta(A)\,\pi(\theta)\,d\theta \\
&= \int_A \left[\int p\left(\theta, \tilde{\theta}\right) \pi(\theta)\,d\theta\right] d\tilde{\theta} + \int_A r(\theta)\,\pi(\theta)\,d\theta \\
&\overset{(34)}{=} \int_A \left[\int p\left(\tilde{\theta}, \theta\right) \pi\left(\tilde{\theta}\right) d\theta\right] d\tilde{\theta} + \int_A r(\theta)\,\pi(\theta)\,d\theta \\
&= \int_A \left(1 - r\left(\tilde{\theta}\right)\right) \pi\left(\tilde{\theta}\right) d\tilde{\theta} + \int_A r(\theta)\,\pi(\theta)\,d\theta \\
&= \int_A \pi\left(\tilde{\theta}\right) d\tilde{\theta}
\end{aligned}$$

□

Constructing such a Transition Kernel

- How to find such a function $p(\theta, \tilde{\theta})$ satisfying the reversibility condition?
- Consider a **candidate-generating density/proposal density** $q(\theta, \tilde{\theta})$ with $\int q(\theta, \tilde{\theta}) d\tilde{\theta} = 1$
- If the process starts at θ, the density generates a value $\tilde{\theta}$ from $q(\theta, \tilde{\theta})$
- If $q(\theta, \tilde{\theta})$ satisfies the reversibility condition, we are done, but it usually does not
- Say without loss of generality that for some $\theta, \tilde{\theta}$

$$\pi(\theta) q\left(\theta, \tilde{\theta}\right) > \pi\left(\tilde{\theta}\right) q\left(\tilde{\theta}, \theta\right) \tag{35}$$

- We move from θ to $\tilde{\theta}$ too often and from $\tilde{\theta}$ to θ too rarely
- Idea: introduce **probability of move** $\alpha(\theta, \tilde{\theta})$ that such a move is made
- If no move is made, we stay at θ

Reweighting

- Transition from θ to $\tilde{\theta}$ happens according to

$$p_{MH}\left(\theta,\tilde{\theta}\right)=q\left(\theta,\tilde{\theta}\right)\alpha\left(\theta,\tilde{\theta}\right),\theta\neq\tilde{\theta} \tag{36}$$

- How to construct α?
- From (35) we know that we move from $\tilde{\theta}$ to θ too rarely

$\Rightarrow$ set $\alpha(\tilde{\theta},\theta)=1$

- From reversibility condition (34) follows

$$\pi\left(\theta\right)q\left(\theta,\tilde{\theta}\right)\alpha\left(\theta,\tilde{\theta}\right)=\pi\left(\tilde{\theta}\right)q\left(\tilde{\theta},\theta\right)\alpha\left(\tilde{\theta},\theta\right)=\pi\left(\tilde{\theta}\right)q\left(\tilde{\theta},\theta\right) \tag{37}$$

- Thus, α

$$\alpha\left(\theta,\tilde{\theta}\right)=\begin{cases}\min\left[\frac{\pi\left(\tilde{\theta}\right)q\left(\tilde{\theta},\theta\right)}{\pi\left(\theta\right)q\left(\theta,\tilde{\theta}\right)},1\right] & \text{if } \pi\left(\theta\right)q\left(\theta,\tilde{\theta}\right)>0\\ 1 \text{ otherwise}\end{cases} \tag{38}$$

The Desired Transition Kernel

- What happens if there is non-zero possibility of staying at θ:

$$r(\theta) = 1 - \int_{R^d} q\left(\theta, \tilde{\theta}\right) \alpha\left(\theta, \tilde{\theta}\right) d\tilde{\theta} \tag{39}$$

- Thus, the required transition kernel is given by

$$P_{MH}(\theta, d\tilde{\theta}) = q(\theta, \tilde{\theta})\alpha(\theta, \tilde{\theta})d\tilde{\theta} + \left[1 - \int_{R^d} q(\theta, \tilde{\theta})\alpha(\theta, \tilde{\theta})d\tilde{\theta}\right] \delta_\theta\left(d\tilde{\theta}\right) \tag{40}$$

- Due to satisfied reversibility condition, $\pi(\theta)$ is its invariant density
- Remarks
  - the M-H algorithm is specified by its proposal density $q(\theta, \tilde{\theta})$
  - If candidate $\tilde{\theta}$ is rejected, the next draw is θ
  - Calculating $\alpha(\theta, \tilde{\theta})$ does not involve the normalizing constant of $\pi(\cdot)$ as it appears in numerator and denominator
  - For symmetric proposal densities with $q(\theta, \tilde{\theta}) = q(\tilde{\theta}, \theta)$, probability of move is $\pi(\tilde{\theta})/\pi(\theta)$; jump always "uphill" and "downhill" with some probability (cf. Simulated Annealing)

Applying the Theory to our DSGE Model

- How does this help us? Remember

$$p\left(\theta|y^T\right) = \frac{p\left(y^T|\theta\right)p\left(\theta\right)}{p\left(y^T\right)} \propto \mathcal{L}\left(y^T|\theta\right)p\left(\theta\right) \tag{5}$$

- Thus, if $\pi(\theta) = p(\theta|y^T)$, we can use the following algorithm to generate draws from the posterior and we don't even need $p\left(y^T\right)$
- Mild regularity conditions assure convergence to posterior when starting from arbitrary point
  - **irreducibility+positive recurrence**: if θ and $\tilde{\theta}$ are in domain of posterior, it must be possible to move from θ to $d\tilde{\theta}$ in a finite number of iterations with positive probability
  - **aperiodicity**: the number of moves required to get from θ to $d\tilde{\theta}$ is not required to be the multiple of some integer
- Typically satisfied when $q(\theta, \tilde{\theta})$ has positive density on same support as posterior

Summary: Metropolis Hastings-Algorithm

- Start with a vector θ_0
- Repeat for $j = 1, \ldots, N$
  - Generate $\tilde{\theta}$ from $q(\theta_{j-1}, \cdot)$ and u from $\mathcal{U}(0,1)$
  - If $\tilde{\theta}$ is valid parameter draw (steady state exists, Blanchard-Kahn conditions satisfied etc.) and $u < \alpha(\theta^{j-1}, \theta^{j})$ set $\theta_j = \tilde{\theta}$
  - Otherwise, set $\theta_j = \theta_{j-1}$ (implies setting $\pi(\tilde{\theta}) = 0$ if draw invalid)
- Return the values $\{\theta_0, \ldots, \theta_N\}$
- After the chain has passed the **transient stage** and the effect of the starting values has subsided, the subsequent draws can be considered draws from the posterior

$\Rightarrow$ **burnin** required that assures remaining chain has **converged**

The Random-Walk Metropolis Hastings Algorithm

- As long as the regularity conditions are satisfied, any proposal density will ultimately lead to convergence to the invariant distribution
- However: speed of convergence may differ significantly
- In practice, people often use the **Random-Walk Metropolis Hastings** algorithm where

$$q\left(\theta, \tilde{\theta}\right) = q_{RW}\left(\tilde{\theta} - \theta\right) \tag{41}$$

and q_{RW} is a multivariate density

- The candidate $\tilde{\theta}$ is thus given by the old value θ plus a random variable increment

$$\tilde{\theta} = \theta + z, z \sim q_{RW} \tag{42}$$

Choosing a Proposal Density

- Often, one uses

$$q_{RW} = \mathcal{N}(0, c^2\Sigma) \text{ or } q_{RW} = t_\nu(0, c^2\Sigma) \tag{43}$$

- Thus, one needs a **scaling matrix** Σ and a **scaling factor** c
- Note: with symmetric density, the scaling probability simplifies to

$$\alpha\left(\theta_{j-1}, \tilde{\theta}\right) = \min\left[\frac{L\left(Y^T \middle| \tilde{\theta}\right) p\left(\tilde{\theta}\right)}{L\left(Y^T \middle| \theta_{j-1}\right) p\left(\theta_{j-1}\right)}, 1\right] \tag{44}$$

- Idea: construct Gaussian approximation to posterior and use asymptotic covariance matrix as scaling matrix
- Allows efficient evaluation around mode
- Note: asymptotically, the prior plays no role and the posterior will only depend on the likelihood
- Koop (2003) only uses likelihood function, while An and Schorfheide (2007) use posterior; we side with the latter

Asymptotic Normality

- Using normal approximation is justified, because with regularity conditions, posterior of θ will be **asymptotically normal** (see e.g. Crowder 1988; Kim 1998; A. M. Walker 1969)
- But convergence may be slow
- Normal approximation can often be improved by **natural re-parametrization** (Adolfson, Lindé, and Villani 2007): bring bounded parameters to unbounded support:
  - log transformation of positive parameters
  - logit-transformation for parameters on unit interval

Technical Considerations

- Usual procedure: use numerical optimizer to find **posterior mode** of log-posterior $p(\theta|Y^T)$
- In practical applications, non-derivative based optimizers seem to perform better
- Finding the mode is hard and time-intensive; try (sequence of) different optimizers
- In some sense, finding the mode is not important as long as the regularity conditions are met (positive definite scaling matrix)
- Wherever you start, asymptotically the MCMC sampler will spend most time at the mode and get there
- MCMC is a quite inefficient optimizer (cf. Simulated Annealing)
- Thus: if you only have finite time, try to get as close as possible
- Problems of not having found the mode can often be seen as a slow drift in the parameters and posterior density

Technical Considerations: the Scaling matrix

- Set the scaling matrix Σ to the **inverse Hessian** at the posterior mode:

$$\Sigma = \text{var}\left(\hat{\theta}\right) = I\left(\theta\right)^{-1} = \left(-E\left[\frac{\partial^2 \log p\left(\theta|Y\right)}{\partial\theta\partial\theta'}\right]\right)^{-1} \tag{45}$$

- `csminwel` will provide an estimate of the Hessian as one of its outputs
- In practice, having fatter tails often works better: might want to use t-distribution instead of normal
- Sounds easy, but creates a lot of problems!
- Theoretical Inverse Hessian at the true mode is positive definite, but the numerical one at the conjectured mode often is not
- Solution: various dirty tricks like
  - use Jordan decomposition to decompose matrix, set the eigenvalues smaller than or equal 0 to some small number, and then recompose the matrix
  - use **generalized Cholesky** (see e.g. Gill and King 2004)
  - Try different step sizes for numerical evaluation of derivatives (An and Schorfheide 2007)

Choosing the Scaling Factor

- The scaling factor affects the behavior of the M-H Chain through:
  - **Acceptance Rate**: percentage of times a move is made
  - Region of the sample space covered by the sampler
- Consider a case where the sampler has converged and the area around the **mode** is sampled
- If the scaling is too wide, many implausible parameter vectors far away from the mode will be proposed and rejected (accept. prob. low)
- If the scaling is too small, many likely parameter vectors close to the mode will be proposed and accepted (accept. prob. high)
- Low probability regions will be undersampled $\Rightarrow$ will take the sampler a long time to traverse the support of the density
- In both cases, there tends to be a **high autocorrelation in the draws**
- Roberts, Gelman, and Gilks (1997): if target and proposal are normal densities, the optimal acceptance rate is 45% in the univariate case and 23% for infinitely many parameters (and already 25% for 6 parameters)

Acceptance Rate Too Low

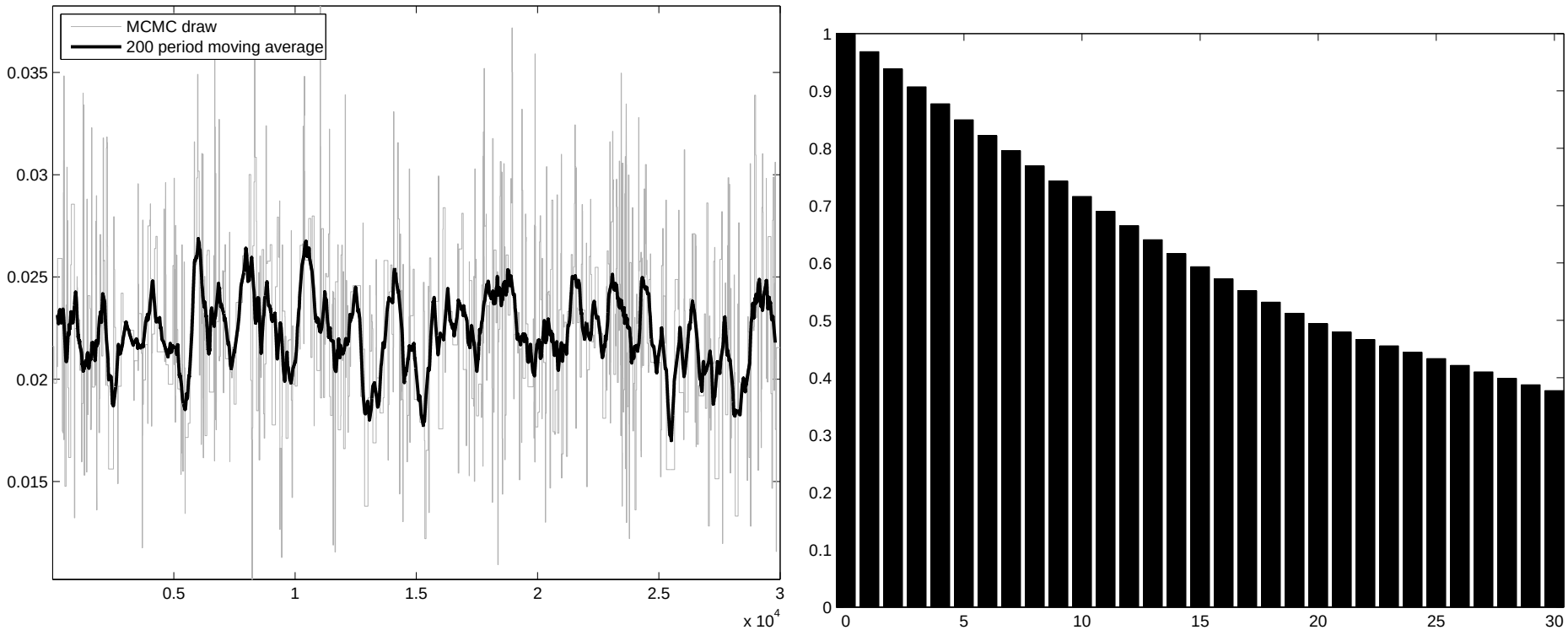


Figure 1: Trace and Autocorrelation Plot: `eps_z`

- Acceptance Rate of 2.5%
- Bad mixing and autocorrelation function only decays slowly

Acceptance Rate Too High

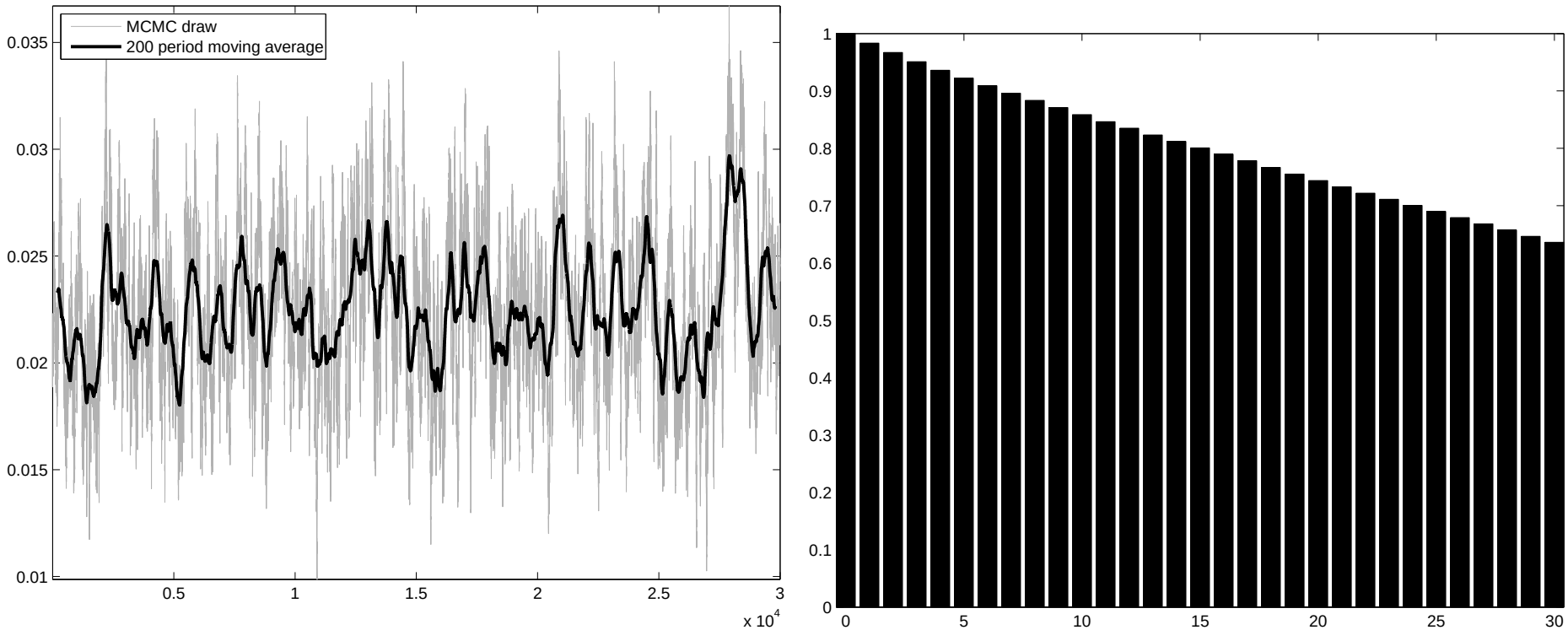


Figure 2: Trace and Autocorrelation Plot: `eps_z`

- Acceptance Rate of 85%
- Bad mixing and autocorrelation function only decays slowly

Acceptance Rate on Target

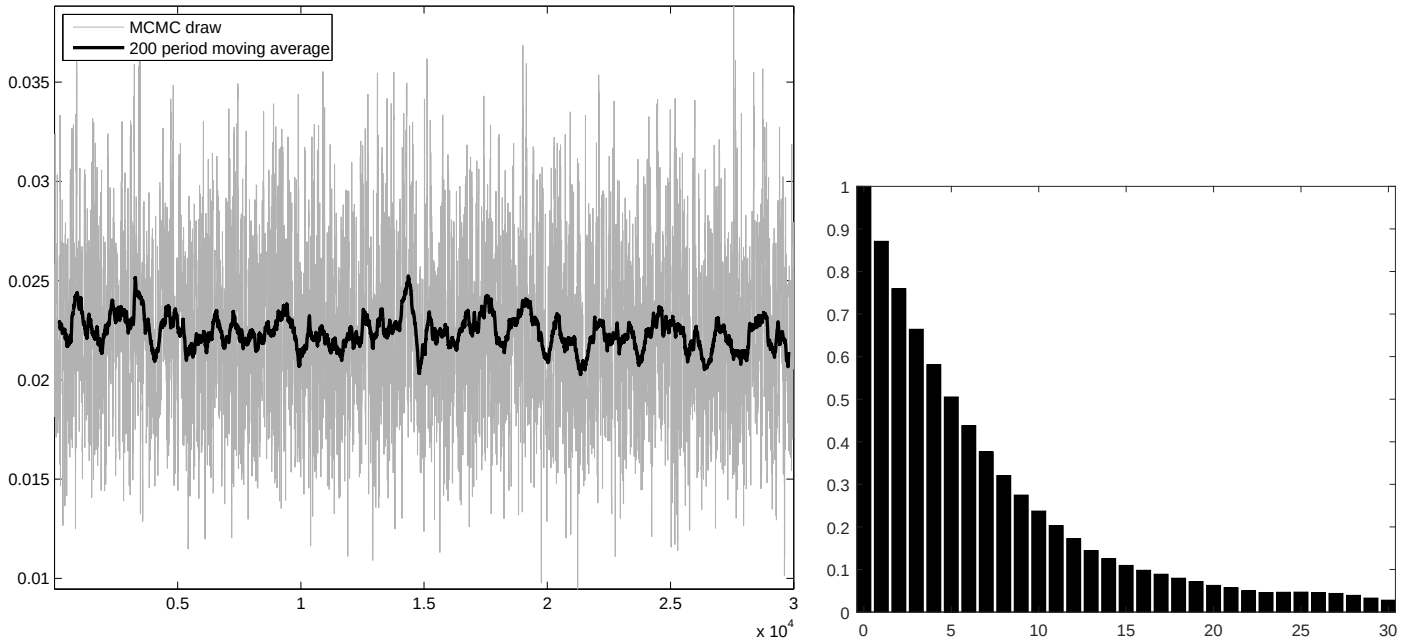


Figure 3: Trace and Autocorrelation Plot: `eps_z`

- Acceptance Rate of 21%
- Good mixing and autocorrelation function decays relatively fast

Efficiency

- Ideally, we want iid draws from the posterior
- MCMC only delivers correlated draws (hence: Markov Chain)
- While high autocorrelation in the draws may signify problems with the scaling, even with the "optimal" acceptance rate, there might be high autocorrelation
- In this case, it might be necessary to try different proposal density

$\Rightarrow$ e.g. Tailored Random Block Metropolis Hastings (Chib and Ramamurthy 2010)

Monitoring Convergence

- We only want to consider draws after the transition kernel has converged to the invariant distribution
- How to monitor that?
  - Geweke (1992) convergence diagnostics: requires single MCMC
  - Brooks and Gelman (1998) convergence diagnostics: requires at least two MCMC

Geweke (1992) Convergence Diagnostics

- Idea: if we have sufficient number of draws from posterior, the first S_A draws after a burnin of S_0 should be similar to the last S_C draws
- By leaving out S_B draws in the middle, draws in S_A and S_C should be independent
- In practice, often $S_A = 0.1S_1$, $S_C = 0.4S_1$, where S_1 is the number of draws after the burnin
- To test similarity, test the means (or any other statistics) $g_{S_i} = E(g(\theta)|Y^T), i \in \{A, C\}$:

$$CD_{GWK} = \frac{\hat{g}_{S_A} - \hat{g}_{S_C}}{\frac{\hat{\sigma}_A}{\sqrt{S_A}} + \frac{\hat{\sigma}_C}{\sqrt{S_C}}} \tag{46}$$

- This is a two-sample t-test and thus asymptotically

$$CD_{GWK} \to N(0, 1) \tag{47}$$

- Problem: estimate of numerical standard error $\hat{\sigma}_i$ needs to take correlation in draws into account
- Use Newey and West (1987)-type estimator that tapers spectral density

Example

Geweke (1992) Convergence Tests, based on means of draws 100000 to 120000 vs 150000 to 200000.
p-values are for Chi2-test for equality of means.

Parameter	Post. Mean	Post. Std	p-val No Taper	p-val 4% Taper	p-val 8% Taper	p-val 15% Taper
SE_eps_z	0.023	0.004	0.000	0.299	0.292	0.262
SE_eps_g	0.011	0.001	0.681	0.915	0.915	0.921
rhog	0.944	0.009	0.024	0.517	0.520	0.507
rhoz	0.756	0.069	0.024	0.568	0.560	0.540

- Higher tapers correcting for serial correlation suggest convergence

Brooks and Gelman (1998) Convergence Diagnostics

- Idea: wherever the MC starts, it should converge to the same invariant distribution
- Start multiple chains from **overdispersed draws** and see whether they yield similar posterior draws after burnin
- The univariate convergence diagnostics are based on comparing **pooled** and **within** MCMC moments (ANOVA)
- Consider a statistic g with variance σ and having J chains with N draws each after discarding a burnin
- Denote means with bars
- The **variance within a chain** is given by

$$\sigma_j^2 = \frac{1}{N-1} \sum_{n=1}^{N} (g_{nj} - \bar{g}_j)^2 \tag{48}$$

Getting the Variances

- The **between sequence variance** B/N is given by

$$\frac{B}{N} = \frac{1}{J-1}\sum_{j=1}^{J}(\bar{g}_j - \bar{g})^2 \tag{49}$$

- Here, B/N is the square of the standard error of the mean and B the actual variance estimate of g
- The (average) **within sequence variance** is given by

$$W = \frac{1}{J(N-1)}\sum_{j=1}^{J}\sum_{n=1}^{N}(g_{jn} - \bar{g}_j)^2 = \frac{1}{J}\sum_{j=1}^{J}\sigma_j^2 \tag{50}$$

- The variance σ^2 can now be estimated by a weighted average of the two:

$$\hat{\sigma}^2 = \left(1 - \frac{1}{N}\right)W + \frac{1+J}{NJ}B \tag{51}$$

- Because we use overdispersed starting points, this is an overestimate of σ^2, but it is consistent

Potential Scale Reduction Factor

- The statistic of interest is the **potential scale reduction factor**, i.e. the ratio between the pooled variance estimate and the within-chain variance estimate:

$$\hat{R} = \sqrt{\frac{\hat{\sigma}^2}{W}} \tag{52}$$

- If $\hat{R}$ is large, either the pooled variance estimate $\hat{\sigma}^2$ can be decreased by further simulations or the within chain variance will increase due to it not yet having made a full tour through the target distribution
- If $\hat{R}$ is close to 1, each of the J chains of N draws is close to the target distribution
- At convergence, three properties should hold
  - $\hat{R}$ should be close to 1
  - The pooled variance $\hat{\sigma}^2$ should stabilize with convergence as the chains were started from an overdispersed distribution
  - The same should hold true for the within chain variance, which should be smaller than $\hat{\sigma}^2$
- All three conditions can be monitored graphically

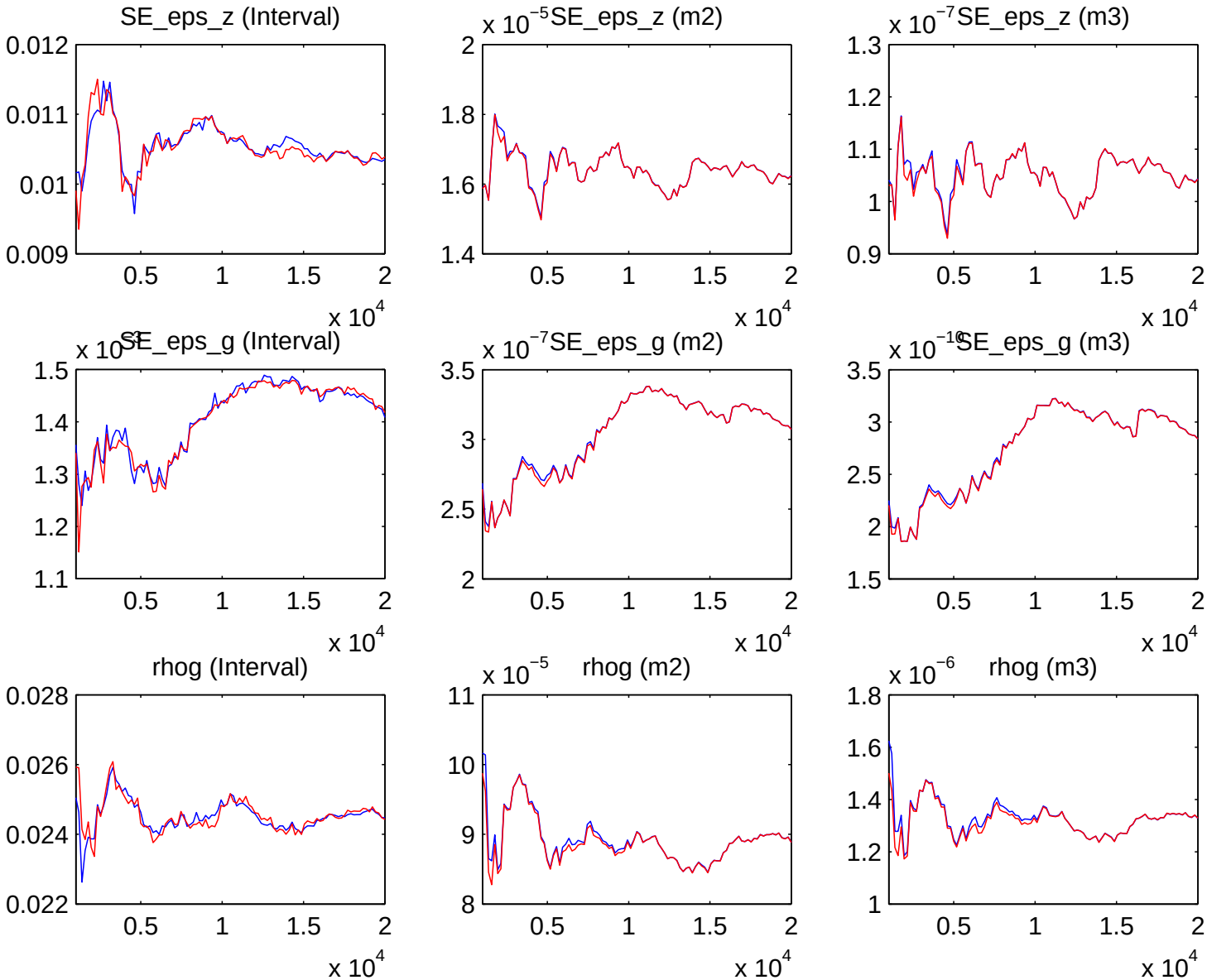
A Non-Parametric Version

- But: previous approach assumes normality by looking at means and variances
- Alternative (also used in Dynare): take length of the $1-\alpha$ percentile for each of the chains and for the pooled draws
- Compare the interval for the pooled draws with the average from the individual chains

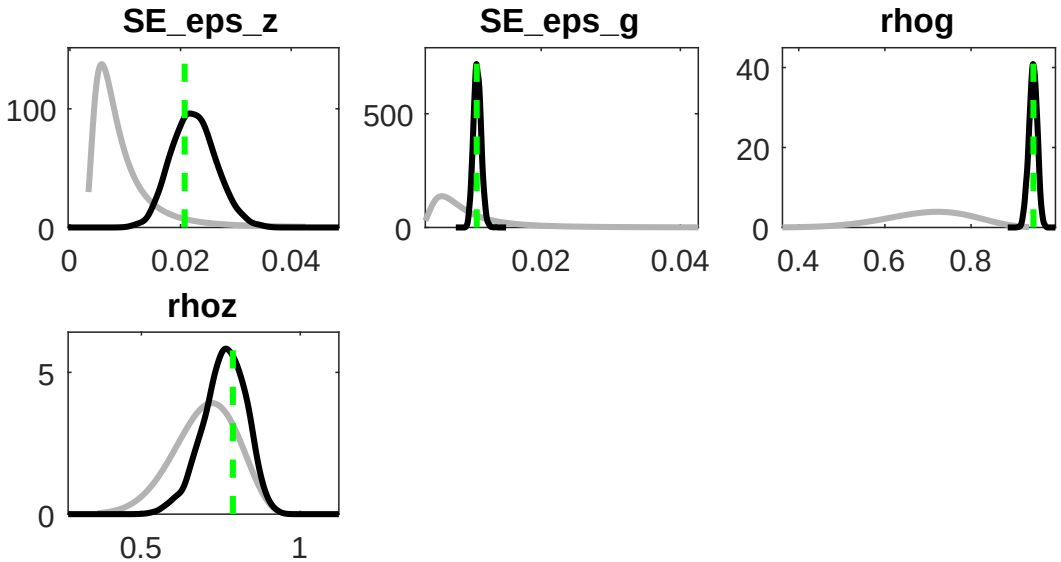
$$\hat{R}_{interval} = \frac{\text{length of total sequence interval}}{\text{mean length of within sequence interval}} \tag{53}$$

- $R_{interval}$ is a **potential scale reduction factor** based on percentiles
- The same convergence criteria apply

Example


SE_eps_z (Interval)
x 10^−5 SE_eps_z (m2)
x 10^−7 SE_eps_z (m3)
x 10^−5 SE_eps_g (Interval)
x 10^−7 SE_eps_g (m2)
x 10^−10 SE_eps_g (m3)
rhog (Interval)
x 10^−5 rhog (m2)
x 10^−6 rhog (m3)
x 10^4

Prior vs. Posterior



- Priors (grey) and posteriors (black) differ: data seems to be informative for updating
- Exception: ρ_z where we saw the flat likelihood
- When the prior is (almost) equal to the posterior, there are two cases
  - Data is uninformative
  - Data is informative, but coincides with chosen prior

Bibliography I

Adolfson, Malin, Jesper Lindé, and Mattias Villani (2007). "Bayesian analysis of DSGE models - some comments". *Econometric Reviews 26* (2-4), 173–185.

An, Sungbae and Frank Schorfheide (2007). "Bayesian analysis of DSGE models". *Econometric Reviews 26* (2-4), 113–172.

Andrle, Michal and Jaromir Benes (2013). "System priors: formulating priors about DSGE model's properties". IMF WOrking Paper WP/13/257.

Berger, James Orvis (2006). "The case for objective Bayesian analysis". *Bayesian Analysis 1* (3), 385–402.

Berger, James Orvis and Robert L. Wolpert (1988). *The likelihood principle*. Ed. by Shanti S. Gupta. Institute of Mathematical Statistics.

Brooks, Stephen P. and Andrew Gelman (1998). "General methods for monitoring convergence of iterative simulations". *Journal of computational and graphical statistics 7* (4), 434–455.

Bibliography II

Chib, Siddhartha and Edward Greenberg (1995). "Understanding the Metropolis-Hastings algorithm". *The American Statistician 49* (4), 327–335.

Chib, Siddhartha and Srikanth Ramamurthy (2010). "Tailored randomized block MCMC methods with application to DSGE models". *Journal of Econometrics 155* (1), 19–38.

Christiano, Lawrence J., Mathias Trabandt, and Karl Walentin (2011). "Introducing financial frictions and unemployment into a small open economy model". *Journal of Economic Dynamics and Control 35* (12), 1999–2041.

Crowder, Martin (1988). "Asymptotic expansions of posterior expectations, distributions and densities for stochastic processes". *Annals of the Institute of Statistical Mathematics 40* (2), 297–309.

Del Negro, Marco and Frank Schorfheide (2008). "Forming priors for DSGE models (and how it affects the assessment of nominal rigidities)". *Journal of Monetary Economics 55* (7), 1191–1208.

Bibliography III

- Gelman, Andrew (1992). "Iterative and non-iterative simulation algorithms". *Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface*, 433–438.
- Geman, Stuart and Donald Geman (1984). "Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images". *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6* (6), 721–741.
- Geweke, John (1992). "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments". *Bayesian Statistics*. Ed. by José M. Bernardo, James O. Berger, A. Philip Dawid, and Adrian F. M. Smith. Vol. 4. Oxford: Clarendon Press, 641–649.
- Gill, Jeff and Gary King (2004). "What to do when your Hessian is not invertible alternatives to model respecification in nonlinear estimation". *Sociological methods & research 33* (1), 54–87.
- Hastings, W. Keith (1970). "Monte Carlo sampling methods using Markov chains and their applications". *Biometrika 57* (1), 97–109.

Bibliography IV

Kim, Jae-Young (1998). "Large sample properties of posterior densities, bayesian information criterion and the likelihood principle in nonstationary time series models". *Econometrica 66* (2), 359–380.

Koop, Gary (2003). *Bayesian econometrics*. Chichester: John Wiley & Sons.

Leeper, Eric M., Nora Traum, and Todd B. Walker (2017). "Clearing up the fiscal multiplier morass". *American Economic Review 107* (8), 2409–54.

Ljungqvist, Lars and Thomas J. Sargent (2012). *Recursive macroeconomic theory*. 3rd ed. Cambridge, MA: MIT Press.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (1953). "Equation of state calculations by fast computing machines". *The Journal of Chemical Physics 21*, 1087–1092.

Bibliography V

- Newey, Whitney K. and Kenneth D. West (1987). "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix". *Econometrica 55* (3), 703–708.
- Roberts, Gareth O., Andrew Gelman, and Walter R. Gilks (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms". *The annals of applied probability 7* (1), 110–120.
- Sims, Christopher A. (2007). "Bayesian methods in applied econometrics, or, why econometrics should always and everywhere be Bayesian". Mimeo. Princeton University.
- Walker, A. M. (1969). "On the asymptotic behaviour of posterior distributions". *Journal of the Royal Statistical Society. Series B (Methodological) 31* (1), 80–88.
- Zyphur, Michael J. and Frederick L. Oswald (2013). "Bayesian estimation and inference". *Journal of Management 41* (2), 390–420.