Estimation of DSGE models

Tamás K. Papp
tkpapp@gmail.com

Institute for Advanced Studies, Vienna

April 28, 2018

Overview

- goal of the course: estimating DSGE models
- toolkits exist, Dynare is very commonly used, we will use it too
- the process is simple, but far from automatic
- things can break very easily
- fixing problems and understanding the results requires some theoretical/practical background
- this course is 90% about giving the general background from a theoretical and practical perspective, 10% about the mechanics of Dynare
- the latter you can easily pick up from the manual once you understand the background (Griffoli 2013)
- this course is an introduction, serves as a starting point
- after this course, you should read the literature and replicate papers

If you only read a single article, make it the handbook chapter of Fernández-Villaverde, Rubio-Ramírez, and Schorfheide (2016).

Course plan

elements Bayesian statistics for this part, the recommended textbook is Gelman et al. (2013)

numerical methods Markov Chain Monte Carlo (MCMC)

the Kalman filter † forming a likelihood for DSGE models

using Dynare † some practical advice

detailed discussion of papers † Christiano, Eichenbaum, and Evans (2005) and Smets and Wouters (2007)

†: these parts have no or incomplete slides as I was using the blackboard, papers, or running examples. For the Kalman filter, see Särkka (2013). The data appendix of Smets and Wouters (2007) has data and Dynare code for replication.

The physicist's twins

1. sonogram: a physicist is pregnant with twins, both are boys
2. what is the probability that they are identical or fraternal?
3. doctor: $1/3$ of twin births are identical, $2/3$ are fraternal
4. we know that identical twins are always same sex
5. fraternal twins same sex with $1/2$ probability

$$p(\text{iden} \mid \text{same}) = \frac{p(\text{iden}, \text{same})}{p(\text{same})} = \frac{p(\text{same} \mid \text{iden})p(\text{iden})}{p(\text{same})} =$$

$$\frac{p(\text{same} \mid \text{iden})p(\text{iden})}{p(\text{same} \mid \text{iden})p(\text{iden}) + p(\text{same} \mid \text{frat})p(\text{frat})} = \frac{1 \cdot 1/3}{1 \cdot 1/3 + 1/2 \cdot 2/3} = \frac{1}{2}$$

We can also do this with a table (for discrete examples).

Concepts in this example

prior The probabilities from the doctor, presumably from a large number of observations. It is a distribution, which is a function of the states which integrates to 1: here

$$p(\text{iden}) = \frac{1}{3} \qquad p(\text{frat}) = \frac{2}{3}$$

data twin pregnancy, both boys

likelihood Probability of same sex conditional on identical/fraternal. Also a distribution, in full

$$p(\text{same} \mid \text{iden}) = 1 \qquad p(\text{diff} \mid \text{iden}) = 0$$
$$p(\text{same} \mid \text{frat}) = \frac{1}{2} \qquad p(\text{diff} \mid \text{frat}) = \frac{1}{2}$$

posterior the distribution we obtain from the exercise, here

$$p(\text{iden} \mid \text{same}) = \frac{1}{2} \qquad p(\text{frat} \mid \text{same}) = \frac{1}{2}$$

Concepts in general: prior, likelihood, posterior

The fundamental question: after I have seen the data, what do I think about the parameters that generated it?
Data y, parameters θ. They have a joint distribution

$$p(\theta, y) = p(\theta)p(y \mid \theta)$$

You can also do

$$p(\theta, y) = p(y)p(\theta \mid y)$$

Then

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

aka "Bayes's rule". Also note that

$$p(y) = \int_\theta p(y, \theta)d\theta = \int_\theta p(\theta)p(y \mid \theta)d\theta$$

so you only need to know $p(y)$, $p(y \mid \theta)$.

Bayes's rule

$$p(\theta \mid y) \propto p(\theta)p(y \mid \theta)$$

where

1. $p(\theta)$ is the prior: what you know (assume) about the parameters before you have seen the data.
2. $p(y \mid \theta)$ is the likelihood: summarizes the data generating process (ie your model).
3. $p(\theta|y)$ is the posterior: what you think about the parameters after you have seen the data.
4. the notation $\propto$ reads as "proportional to", ie up to a constant (usually irrelevant for numerical methods). We can always find the constant by integrating.

Bayesian inference in practice

Commonly used models $p(y \mid \theta)$ which are either convenient to use, or have desirable (numerical) properties. Latter used to be a key question before numerical methods, now less relevant. We will see some examples from statistics, but DSGE models are in general intractable analytically.

What to do with the posterior? We usually sample a collection of parameter values θ numerically from the distribution. Methodology: Markov Chain Monte Carlo (MCMC).

The art of prior distributions. The less data we have, the more priors influence the result. Ideally, we would prefer to avoid this. But macro data is usually shorter and less informative than we would like; also, structural parameters may be weakly identified.

Example: univariate normal, known variance

We have n observations y_i, $i = 1, \ldots, n$, with distribution

$$y_i \sim N(\mu, 1)$$

where N is the normal distribution, μ is its mean, and we assume that the variance is known.
Recall that the density of $N(\mu, \sigma^2)$ is

$$p(y_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)$$

Here $\sigma^2 = 1$, and we ignore the constant,

$$p(y_i \mid \mu) \propto \exp\left(-\frac{1}{2}(y_i - \mu)^2\right)$$

Example: univariate normal, known variance (cont)

We use $y = \{y_i\}_{i=1}^n$ for the whole sample. Since it is IID,

$$p(y \mid \mu) = \prod_{i=1}^{n} p(y_i \mid \mu)$$

is the likelihood for the whole data.
Logarithms give us analytical convenience (also better numerically):

$$\log p(y \mid \mu) = \sum_{i=1}^{n} \log p(y_i \mid \mu) = -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2 + C$$

For now, use a flat (improper) prior

$$p(\mu) \propto 1$$

This is "improper" because it is not a distribution (integral is ∞).
Then the (log) posterior is

$$\log p(\mu \mid y) = -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2 + C$$

Example: univariate normal, known variance (cont)

Introduce the sample mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Then

$$\log p(\mu \mid y) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2 = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \mu)^2 =$$

$$-\frac{1}{2}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 + 2(\bar{y} - \mu)\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})}_{\equiv 0} + n(\bar{y} - \mu)^2\right) = -\frac{1}{2}n(\bar{y}-\mu)^2 + C$$

We recognize this as a univariate normal distribution, ie

$$\mu \mid y \sim N(\bar{y}, 1/n)$$

Example: univariate normal, known variance (cont)

What to take away from this example?

1. if you have n observations with mean $\bar{y}$, after seeing them your posterior distribution is

$$\mu \mid y \sim N(\bar{y}, 1/n)$$

   This is a distribution, not a point estimate.
2. It is centered on the sample mean.
3. The standard deviation is $n^{-1/2}$, decreasing with the square root of the sample size. Asymptotically, as $n \to \infty$, it would go to 0.
4. Here, we recognized the posterior distribution as a member of a known family. Generally, this need not be the case.

Improper prior distributions

- Bayes's rule works for distributions, ie densities (or more generally, measures) that integrate to 1 over the domain.
- When this does not hold, we may still obtain a proper posterior distribution as in the example above. But this does not work generally.
- Checking that a distribution is proper is equivalent to integrating it. However, our usual methods (MCMC) break down for improper distributions, so we can't use them.
- It is best to avoid improper prior distributions in practice.
- Generally, use weakly informative priors. For example, suppose that

$$\mu = \text{unemployment rate} \in [0, 1]$$

  A prior $\mu \sim N(0, 10^2)$ would not constrain the parameter.
- With weakly informative priors, it helps to standardize data.
- In economics, with little data, sometimes we need to use more informative priors. More on this later.

Conjugate priors: example

We continue with the univariate normal example. Instead of using the improper prior, let's assume

$$\mu \sim N(\mu_0, \tau^2)$$

so that

$$\log p(\mu) = -\frac{(\mu - \mu_0)^2}{2\tau^2} + C$$

The (log) posterior is

$$\log p(\mu \mid y) = -\frac{1}{2}(n + 1/\tau^2)\left(\mu - \frac{n\bar{y} + \mu_0/\tau^2}{n + 1/\tau^2}\right)^2 + C$$

or equivalently,

$$\mu \mid y \sim N\left(\frac{n\bar{y} + \mu_0/\tau^2}{n + 1/\tau^2}, \frac{1}{n + 1/\tau^2}\right)$$

Conjugate priors: example (cont)

To summarize, from prior

$$\mu \sim N(\mu_0, \tau^2)$$

we obtain the posterior

$$\mu \mid y \sim N\left(\frac{n\bar{y} + \mu_0/\tau^2}{n + 1/\tau^2}, \frac{1}{n + 1/\tau^2}\right)$$

Both the prior and the posterior come from the same distribution family. We say that this is a conjugate prior for this likelihood function.

1. Conjugate priors used to be important because of analytical convenience before the computational revolution of Bayesian statistics.
2. Not all models have a conjugate prior, and there is no strong reason to restrict ourselves when using MCMC.

Asymptotics

Still with the same prior and likelihood, consider the posterior

$$\mu \mid y \sim N(\mu_1, \sigma_1^2) \qquad \text{with} \qquad \mu_1 = \frac{n\bar{y} + \mu_0/\tau^2}{n + 1/\tau^2}, \quad \sigma_1^2 = \frac{1}{n + 1/\tau^2}$$

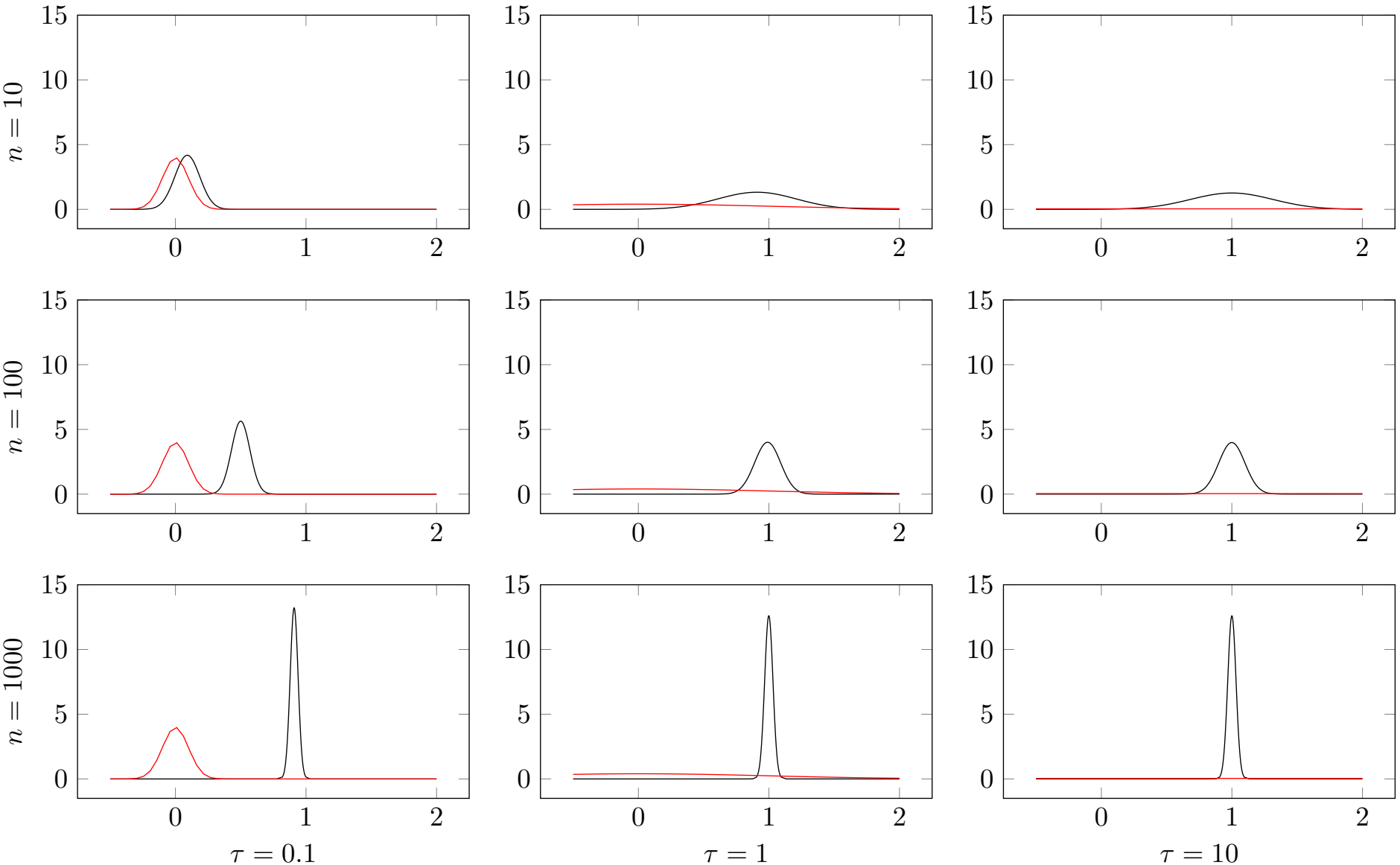
naming the posterior mean and variance μ_1 and σ_1^2, respectively.

1. the posterior mean is a weighted average of the prior mean μ_0 and the sample mean $\bar{y}$, with weights n and $1/\tau^2$,
2. as $n \to \infty$, $\mu_1 \to \bar{y}$ and $\sigma_1^2/n \to 1$
3. in the limit $\tau^2 \to \infty$, we obtain the flat prior $p(\mu) \propto 1$

There are results in Bayesian statistics about asymptotic normality and consistency, similar to frequentist statistics. They are nice to know, but usually of limited value in macroeconomics (where we have little data).

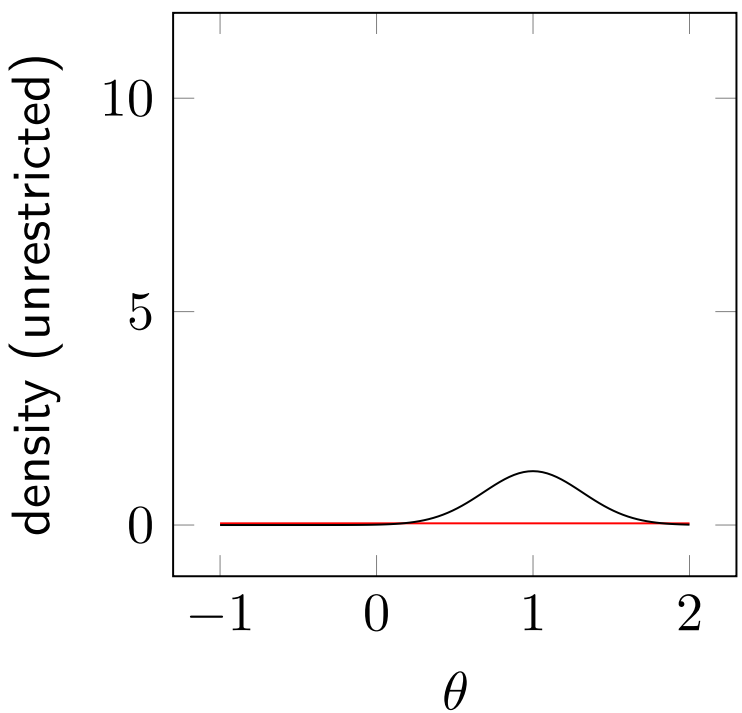
Some prior-posterior combinations (still the same model)

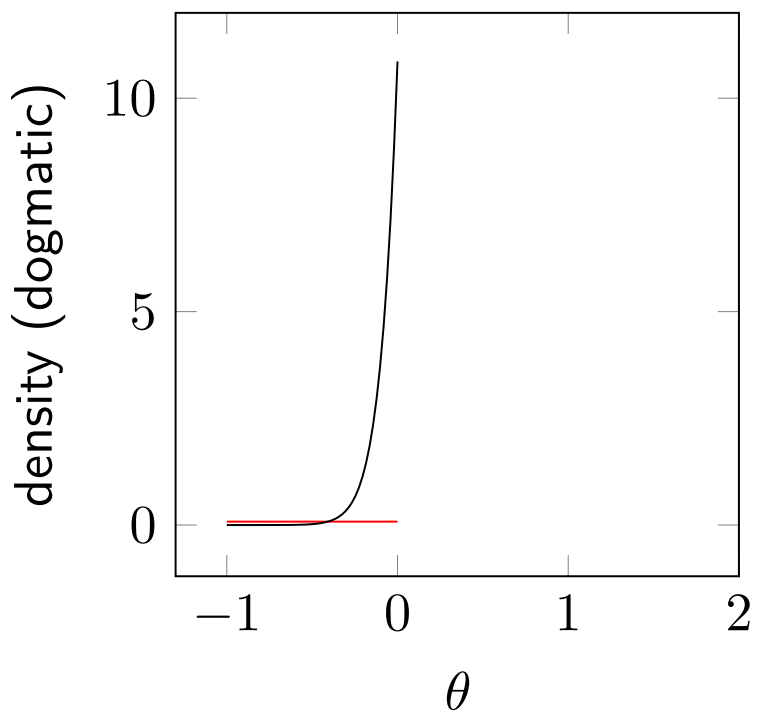
Let $\bar{y} = 1$, $\mu_0 = 0$, vary n and τ. Legend: posterior ——, prior ——.



Dogmatic priors

- a dogmatic prior assigns zero density (probability) to otherwise feasible regions of the parameter space
- from Bayes's rule, $p(\theta) = 0 \quad \Rightarrow \quad p(\theta \mid y) = 0$
- even asymptotic convergence does not work for dogmatic priors
- suppose $\bar{y} = 1$, $n = 10$, prior is $\theta \sim N(0, 10^2)$ truncated to $(\infty, 0]$





- prefer informative priors, check your model

Monte Carlo methods

1. Assume we have obtained an unnormalized posterior

$$f(\theta) \propto p(\theta \mid y)$$

   for a specific dataset y. f is generally intractable.
2. We would like to draw $\theta_s \sim f$, $s = 1, \ldots, S$ from this distribution.
3. Then for a function $h(\theta)$,

$$\mathrm{E}[h(\theta) \mid y] = \int h(\theta) p(\theta \mid y) d\theta \approx \frac{1}{S} \sum_{s=1}^{\infty} h(\theta_s)$$

4. these are simulation ("Monte Carlo") methods, usually with a computer

Sampling from distributions

- for known distributions, this is relatively easy
- for univariate distributions,

$$F(\theta) = \int_{-\infty}^{\theta} f(\tilde{\theta}) d\tilde{\theta} \quad \Rightarrow \quad F^{-1}(U) \sim f \quad \text{if } U \text{ is uniform on } [0, 1]$$

- for general distributions, most effective methods construct a Markov chain that has a stationary distribution f
- other methods exist, they are less relevant for DSGE models
- efficient methods are an active research area in Bayesian statistics (we talk about efficiency below)

Markov chains

1. a stochastic sequence x_t
2. Markov property:

$$p(x_{t+1} \mid x_t, x_{t-1}, \dots) = p(x_{t+1} \mid x_t)$$

3. there are Markov chains in discrete and continuous time, with discrete and continuous state spaces for x
4. in Bayesian statistics, we use discrete time, continuous state sequences (mostly); for the examples, I use discrete states
5. most of the time we care about the stationary distribution $\pi(x)$

$$\pi(x') = \int \pi(x) p(x' \mid x) dx$$

   and hope that actual samples from the chain are representative of π
6. in MCMC, we construct p for a particular π

Discrete example

Two states, 1 and 2. $p(2 \mid 1) = \alpha$, $p(1 \mid 2) = \alpha\beta$. Transition matrix:

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \alpha\beta & 1-\alpha\beta \end{bmatrix}$$

Looking for π such that

$$\pi = \pi P, \qquad \sum \pi_i = 1$$

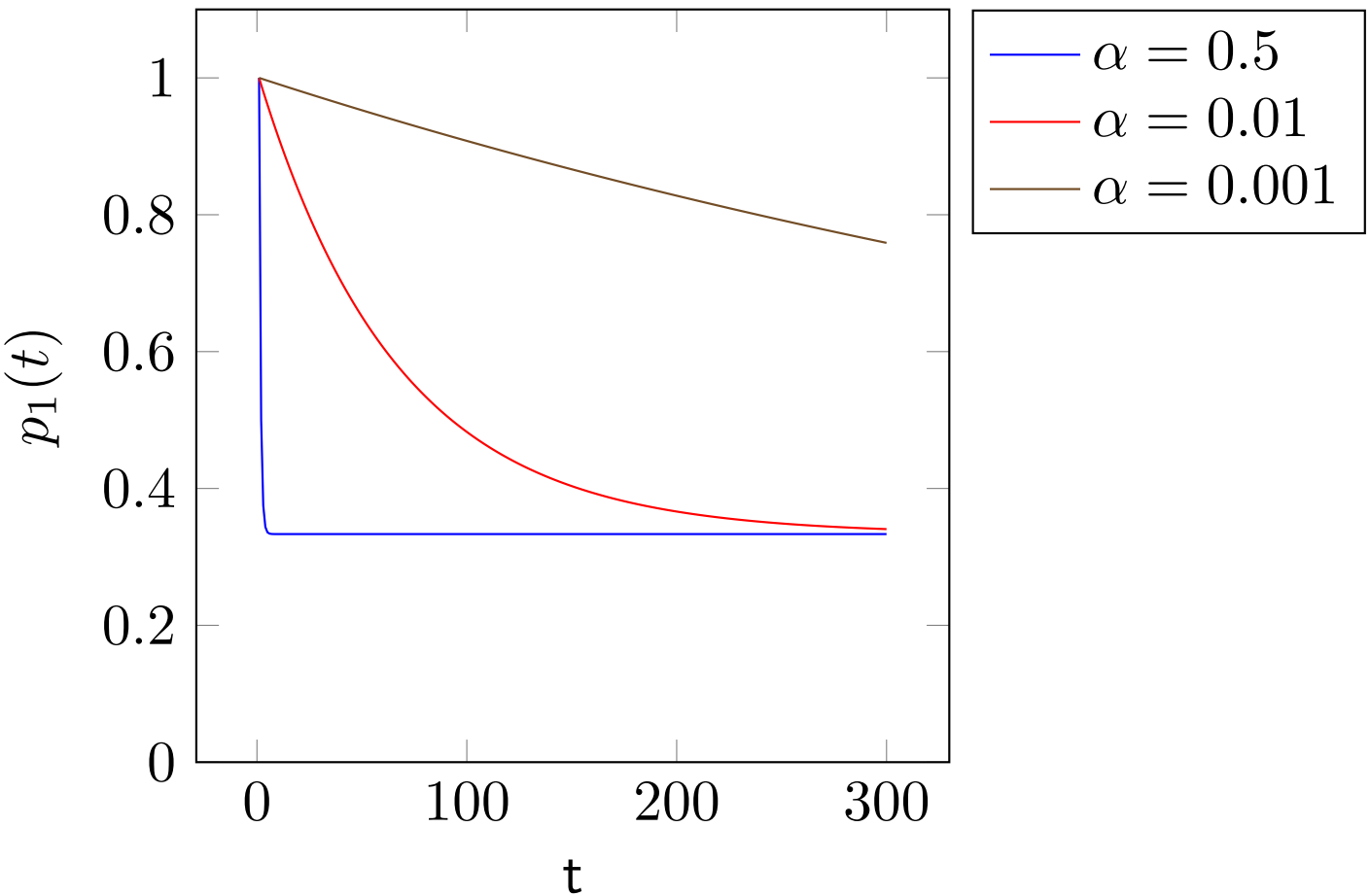
Doing the algebra,

$$\pi_1 = \pi_1(1-\alpha)+(1-\pi_1)\alpha\beta \quad \Rightarrow \quad \pi_1(\alpha+\alpha\beta) = \alpha\beta \quad \Rightarrow \quad \pi_1 = \frac{\beta}{1+\beta}$$

independent of α (this was intended, and is specific to this example).

Convergence and mixing

- Under some technical conditions, the distribution of x_{t+N} converges to π as $N \to \infty$.
- Here, we don't go into technical details, just illustrate how this works in practice. Continue with our example with $\beta = 1/2$, starting $x_0 = 1$,



Pathologies: reducibility

- The chain is irreducible if it is possible to get from any state to any state.
- It is easy to show that all distributions are stationary distributions for

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

- Practically, once the above chain is in a state, it is stuck there. When $\alpha \approx 0$ in the previous example, the chain is irreducible in theory, but may not mix well in practice.
- Let P_1 and P_2 be transition matrixes, and

$$P = \begin{bmatrix} (1-\epsilon_1)P_1 & \epsilon_1 I \\ \epsilon_2 I & (1-\epsilon_2)P_2 \end{bmatrix}$$

- If $\epsilon_1 = \epsilon_2 = 0$, chain has mixing within, but not between the two sets of states. For $\epsilon_1, \epsilon_2 \approx 0$, mixing is again problematic in practice. Something similar can easily happen when estimating DSGE models.

Pathology: periodicity

If $\exists k$ and a state such that the chain returns to the same state in k number of steps. Example (all states have period 3):

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Pathology: transience

If there are states for which there is a non-zero probability what we never return. Example:

$$\begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \end{bmatrix}$$

Neither of the above is a concern in practice for MCMC with common setups.

Detailed balance

Detailed balance holds for a distribution π and a transition matrix P if

$$\pi_i p_{ij} = \pi_j p_{ji} \qquad \text{for all states } i, j$$

If we have an irreducible, aperiodic, non-transient Markov chain, and detailed balance holds, then π is the steady state distribution of P.

Proof (sketch)

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$$

so $\pi P = \pi$.

We use detailed balance as a proof technique in MCMC.

Metropolis algorithm

Uses a proposal distribution $J(\theta^* \mid \theta)$ that is symmetric

$$J(\theta_a \mid \theta_b) = J(\theta_b \mid \theta_a)$$

to sample from a density f (for concreteness, think of $\theta_a \sim N(\theta_b, \Sigma)$).

Algorithm

1. use a starting point θ_0, for which $f(\theta_0) > 0$
2. at step t, sample a proposal θ^* from $J(\theta^* \mid \theta^t)$, and let

$$r = \frac{f(\theta^*)}{f(\theta^t)}$$

3. let

$$\theta^{t+1} = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^t & \text{otherwise} \end{cases}$$

Why the Metropolis algorithm works

We show detailed balance. WLOG assume $f(\theta_b) \geq f(\theta_a)$, and let $P(\theta_a, \theta_b)$ denote the transition kernel from θ_a to θ_b. Then

$$f(\theta_a)P(\theta_a, \theta_b) = f(\theta_a)J(\theta_b \mid \theta_a)$$

and

$$f(\theta_b)P(\theta_b, \theta_a) = f(\theta_b)J(\theta_a|\theta_b)r = f(\theta_b)J(\theta_a|\theta_b)\frac{f(\theta_a)}{f(\theta_b)} = f(\theta_a)J(\theta_a \mid \theta_b)$$

Then symmetry of J implies detailed balance.

Metropolis-Hastings

This can be extended to non-symmetric J with a correction factor. This is known as the Metropolis-Hastings algorithm.

Random walk Metropolis-Hastings (RWMH)

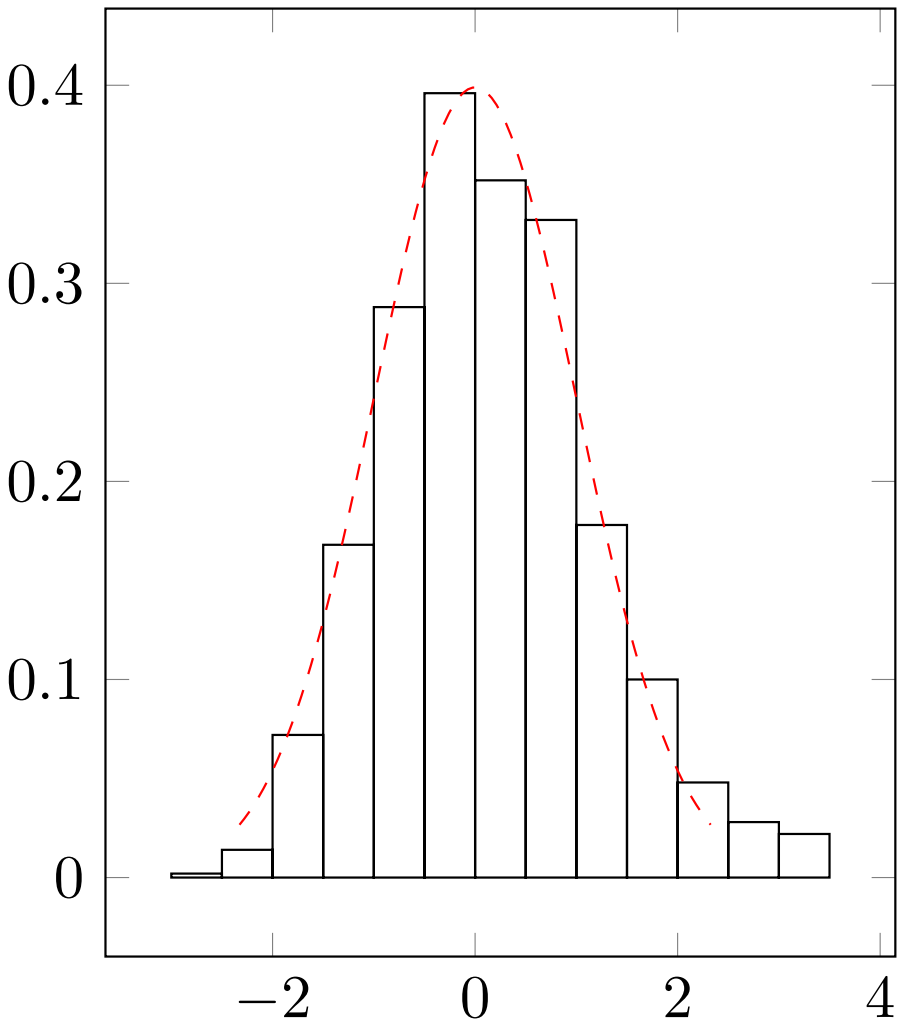
- use a multivariate normal proposal

$$\theta^* \sim N(\theta, \Sigma)$$

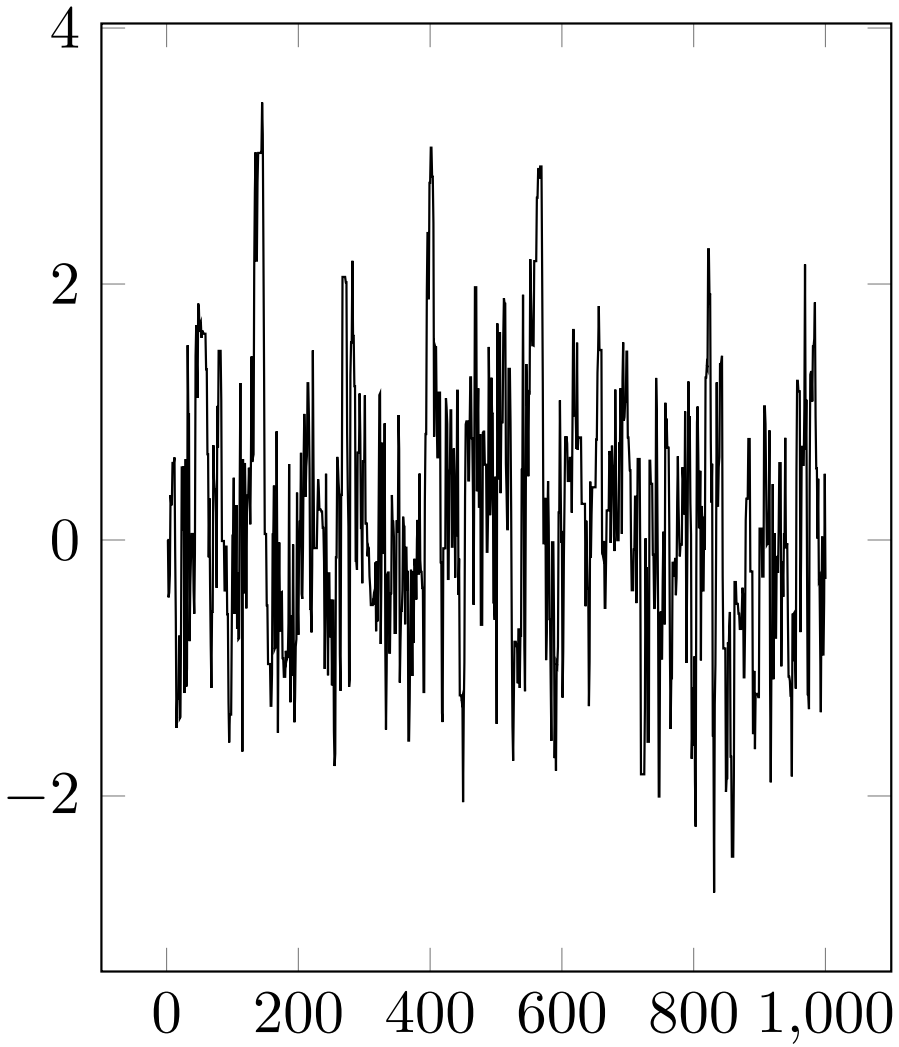
  which is symmetric by construction.
- what should Σ be?
- this depends on the distribution, and involves various trade-offs

$f = N(0,1),\ J = N(0,1)$

densities: $J = N(0,1)$

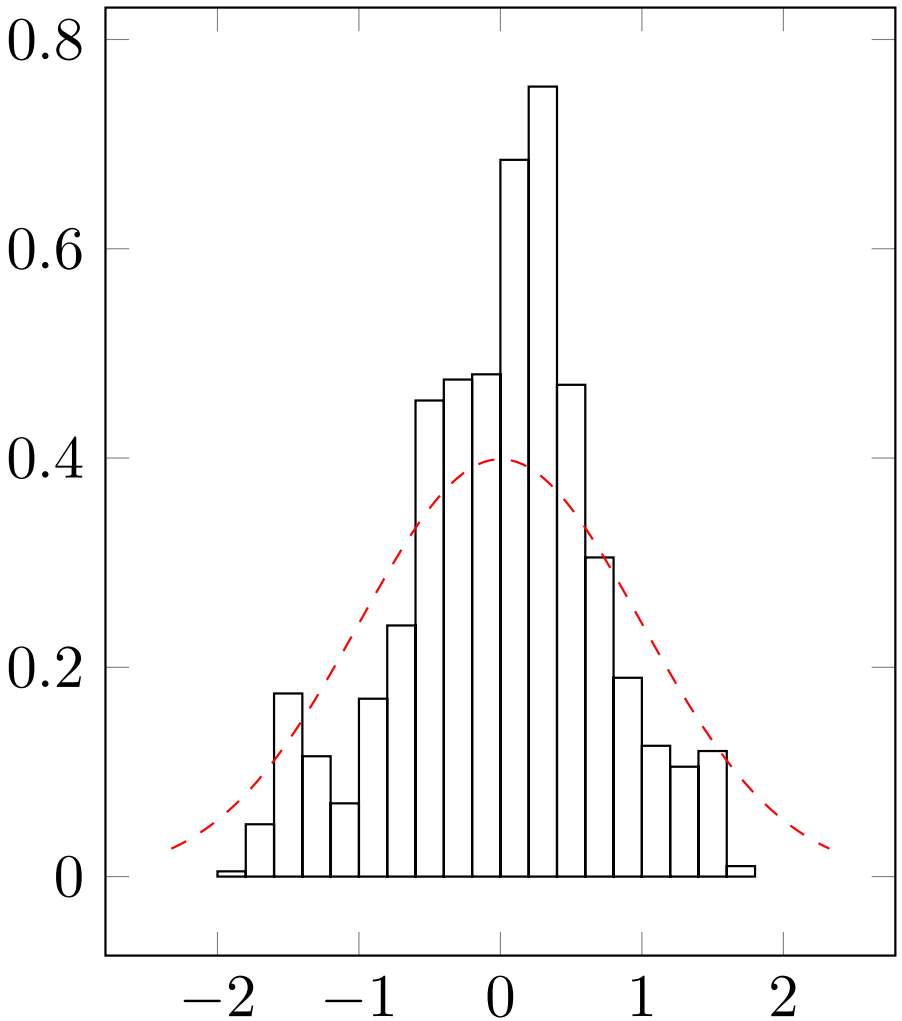


chain, A = 0.706

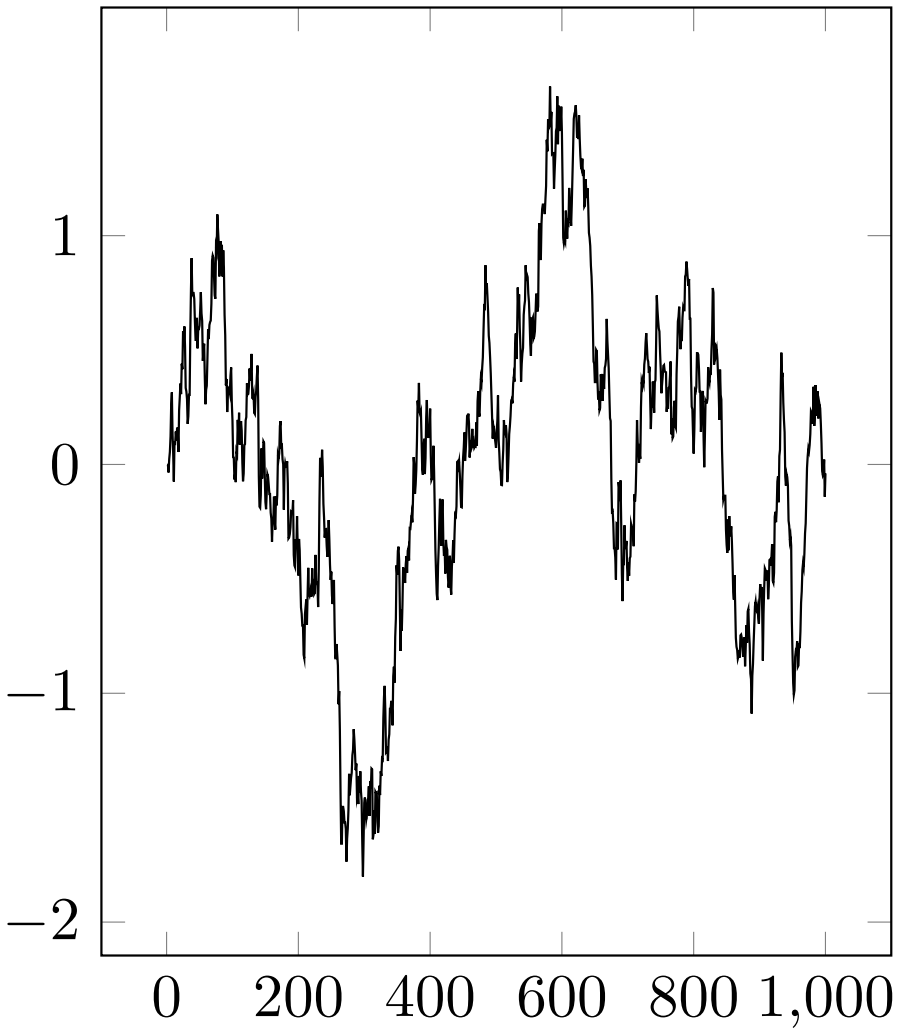


$f = N(0,1),\ J = N(0,0.1)$

densities: $J = N(0,0.1)$

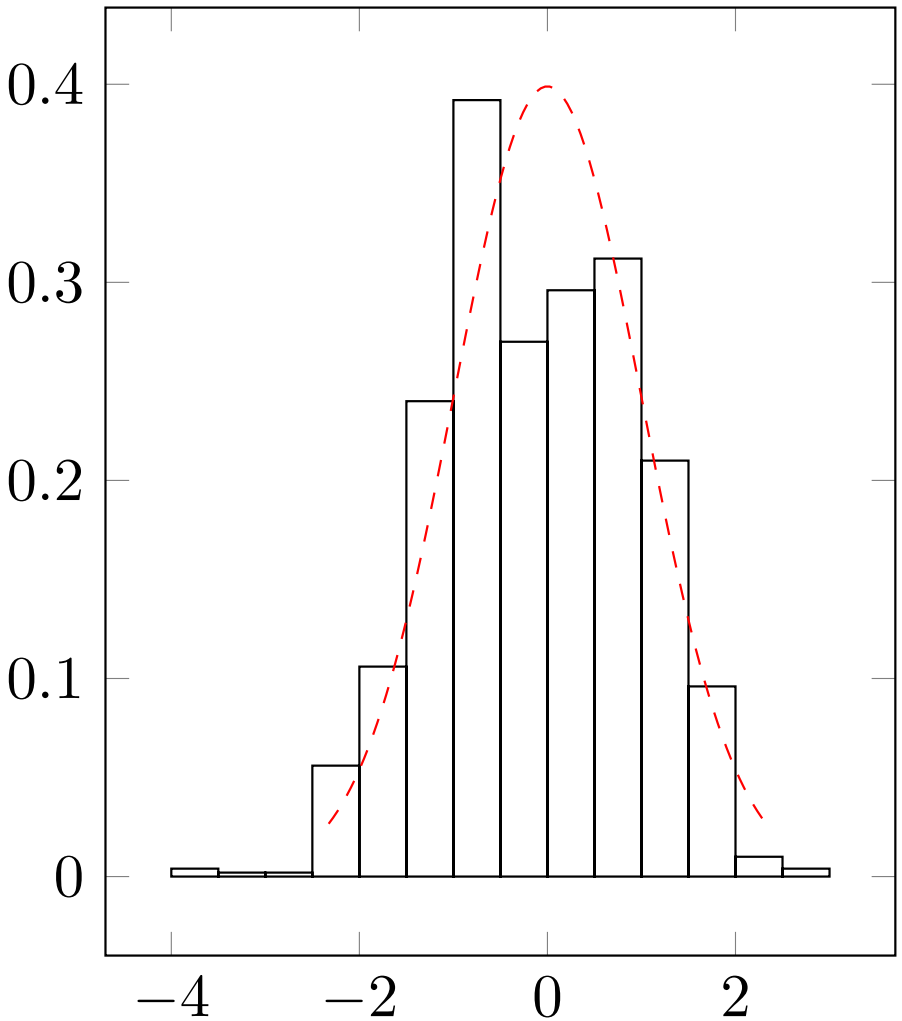


chain, A = 0.974

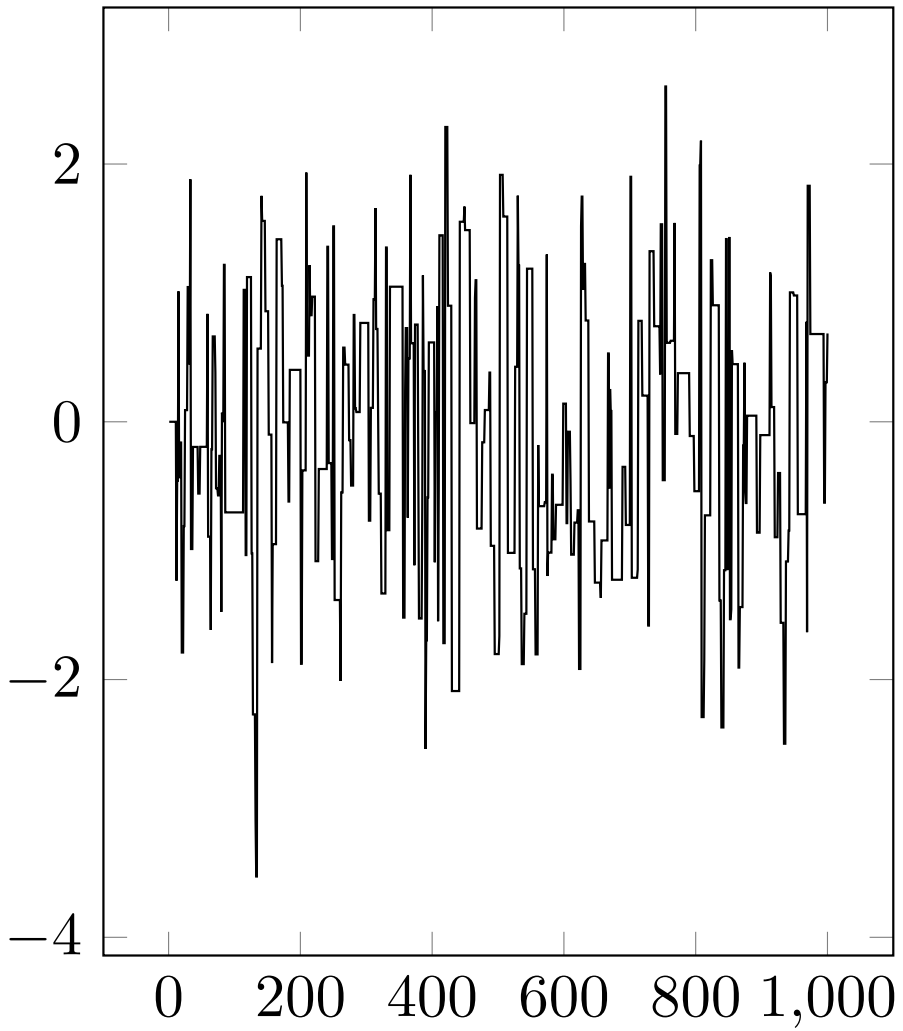


$f = N(0,1)$, $J = N(0,5)$

densities: $J = N(0,5)$

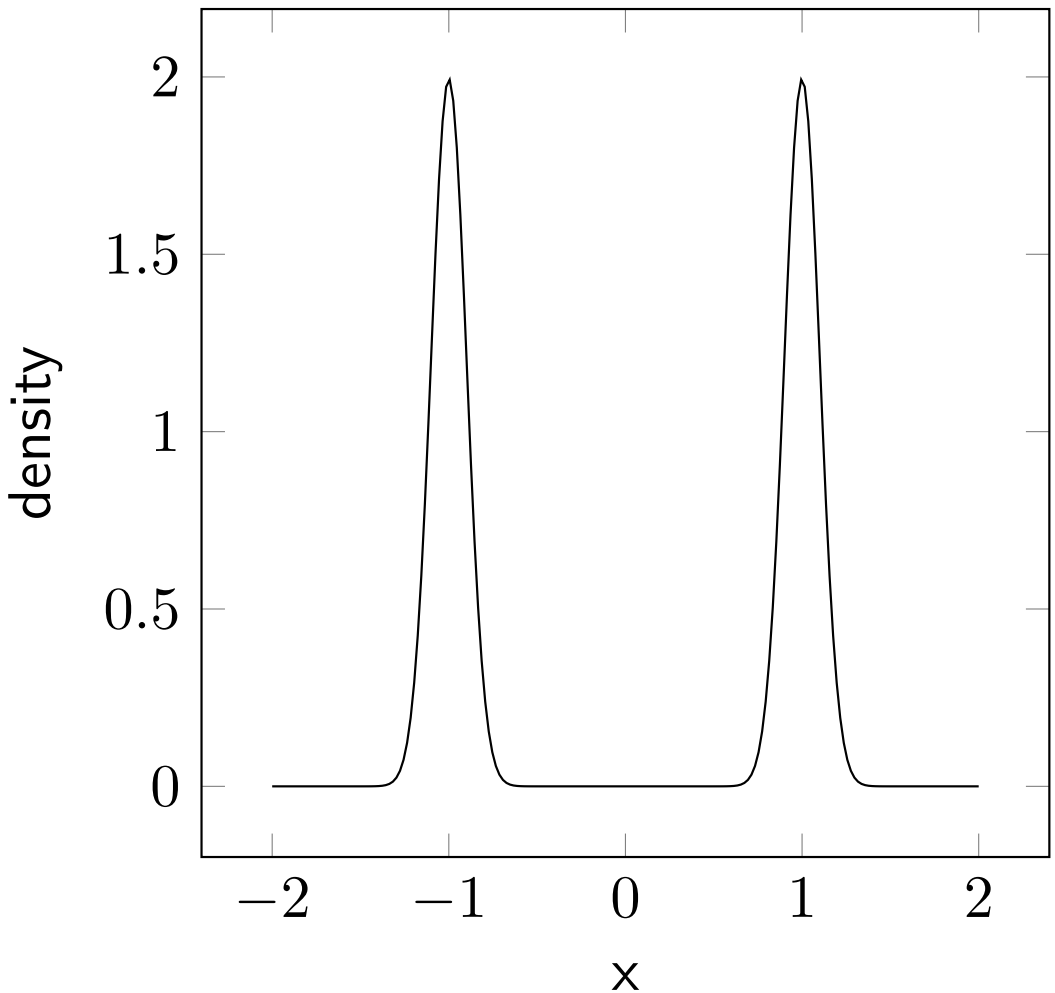


chain, A = 0.235



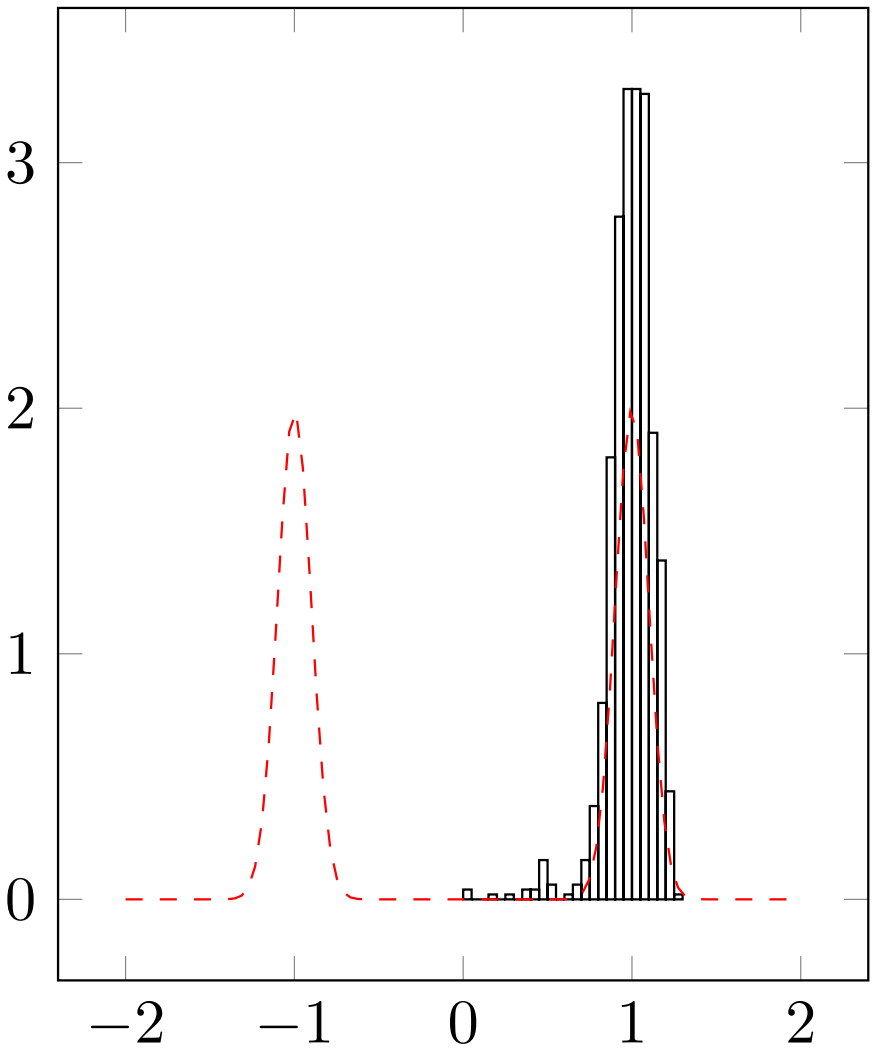
Bimodal example with a "valley"

$$x = \begin{cases} \sim N(-1, 0.1) & \text{with probability } 1/2, \\ \sim N(1, 0.1) & \text{with probability } 1/2 \end{cases}$$

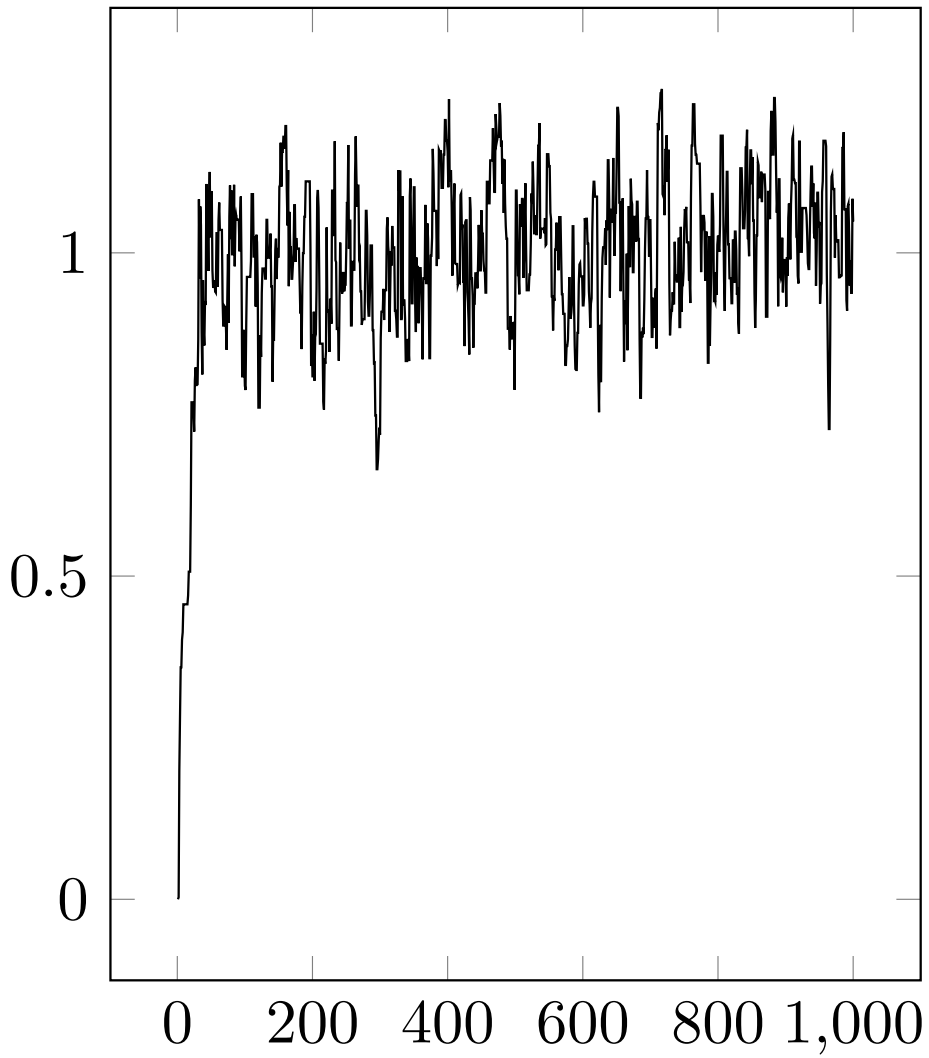


Bimodal example, $J = N(0, 0.1)$

densities: $J = N(0, 0.1)$

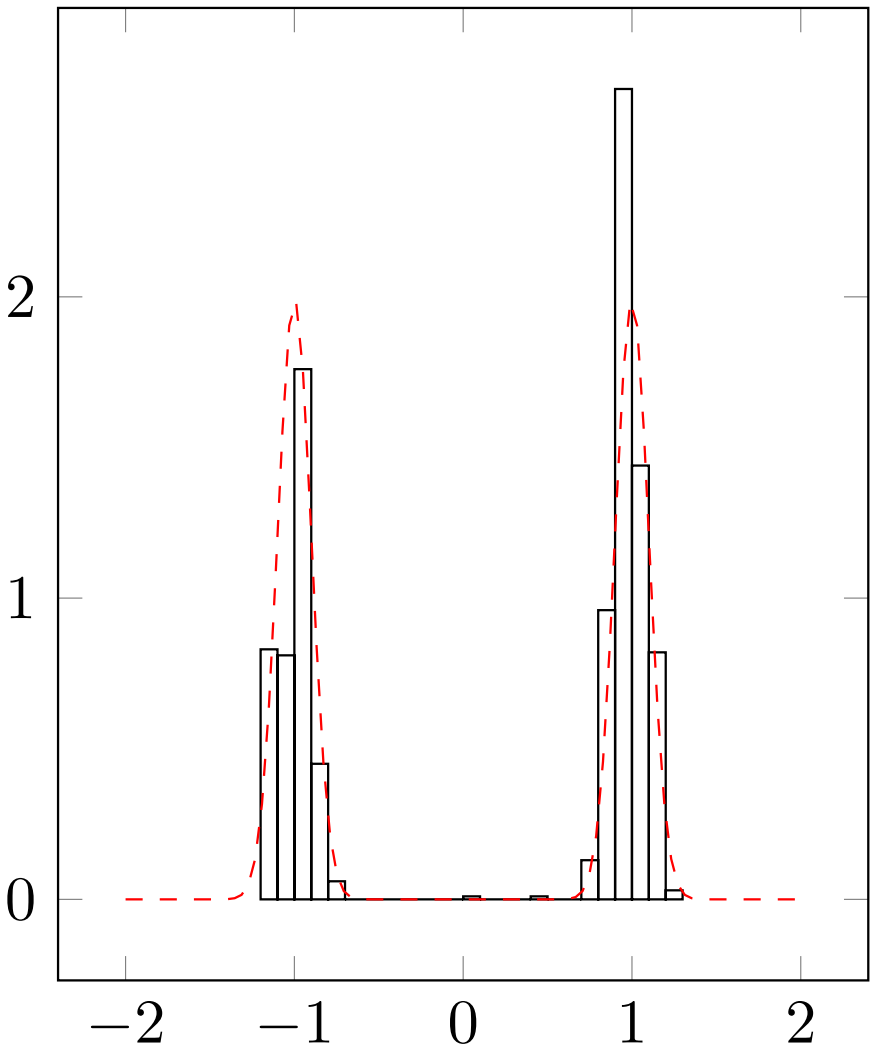


chain, A = 0.712

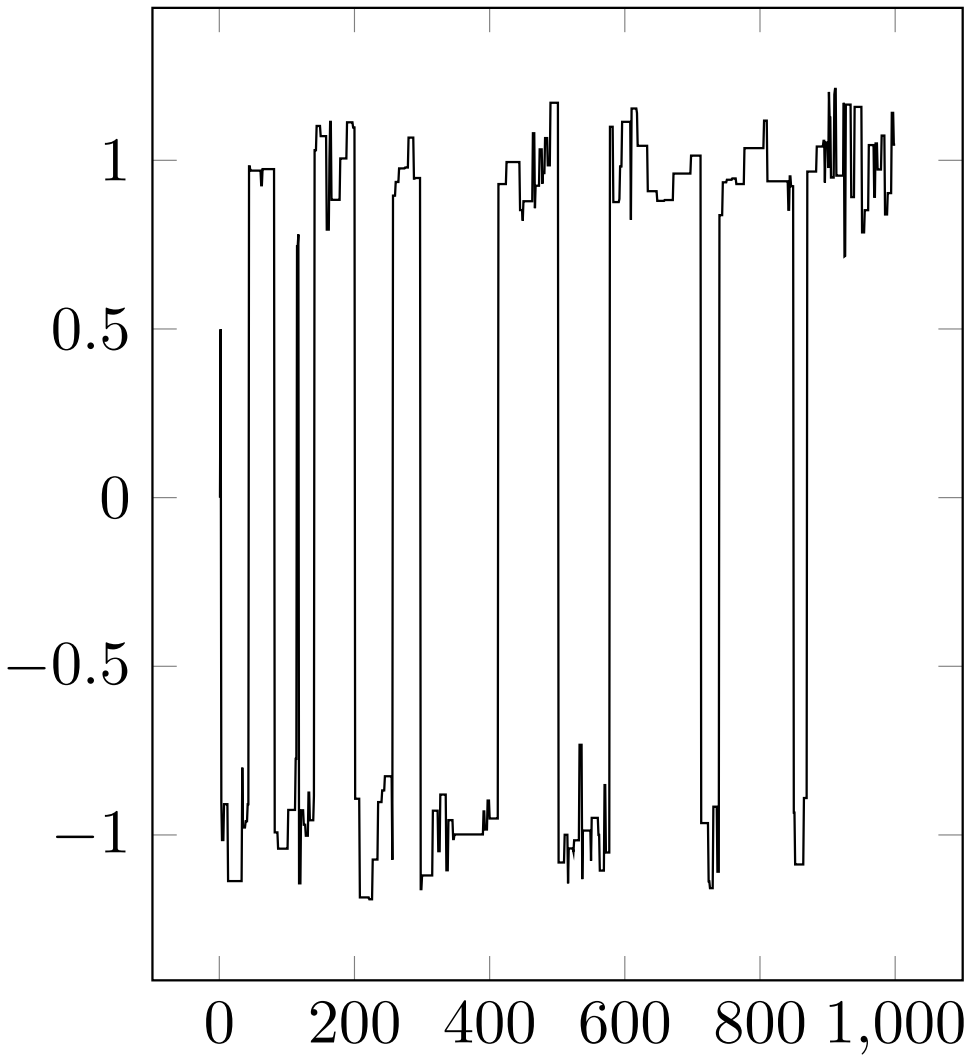


Bimodal example, $J = N(0, 1)$

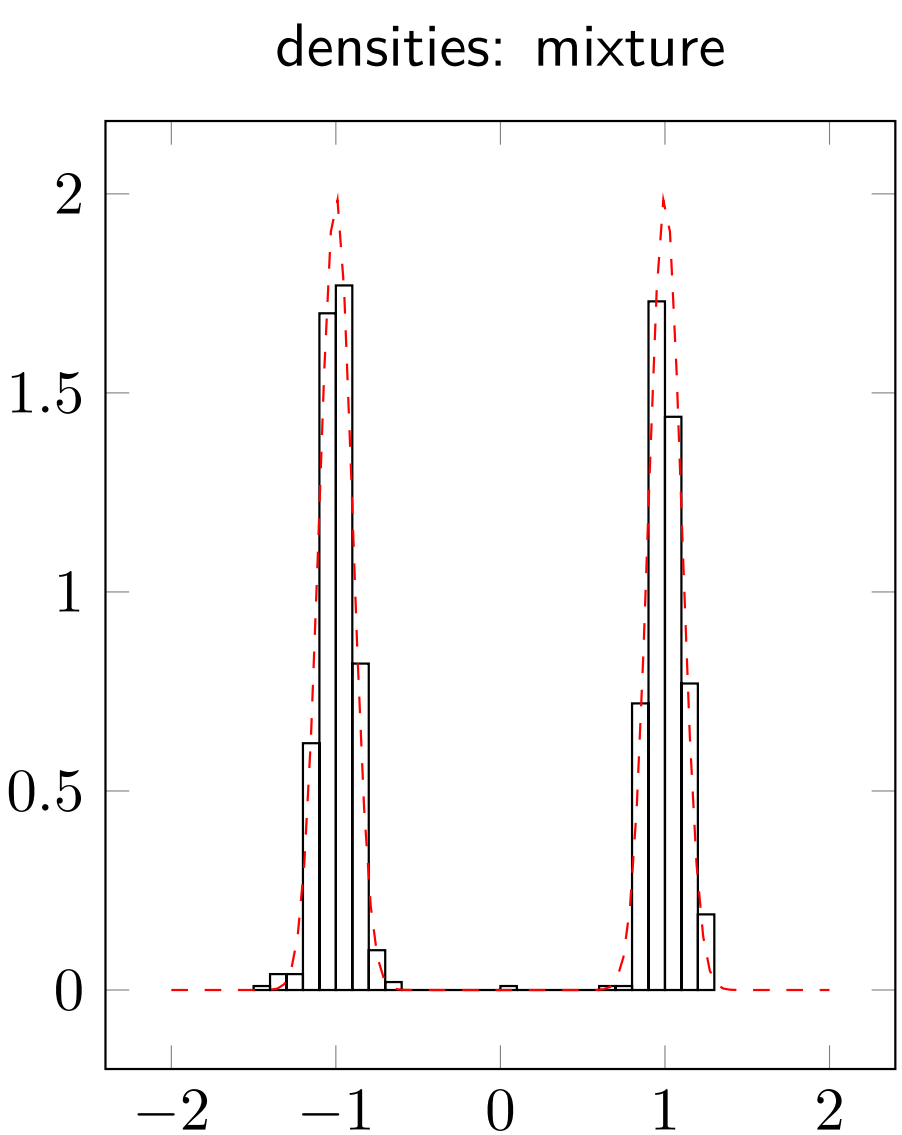
densities: $J = N(0, 1)$

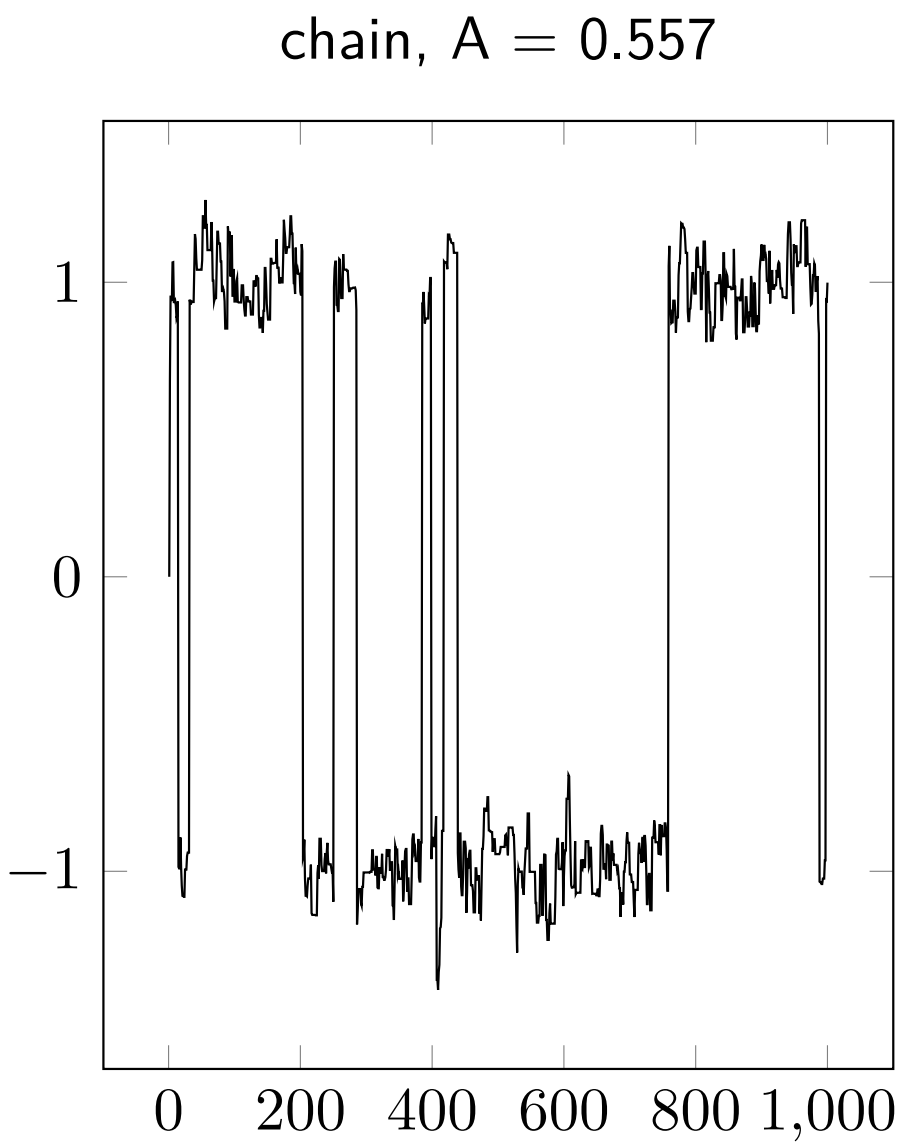


chain, A = 0.155



Bimodal example, $J = 0.3 \cdot N(0, 1.5) + 0.7 \cdot N(0, 0.1)$





Commonly used solutions for selecting Σ

Approximation at the mode

1. find a (local) maximum (mode)
2. obtain the Hessian of the log posterior
3. use this for calculating Σ (essentially pretending that the distribution is locally normal)
4. scale this for better acceptance

This is quick, but can easily fail (cf bimodal example).

Adaptive algorithms

1. choose an initial Σ, either I or with local approximation
2. sample using this, monitoring acceptance rate and variance of sample
3. adjust Σ accordingly

Adaptation phase samples cannot be used as it violates detailed balance.

Effective sample size

Suppose you have n independent random draws x_i from some distribution, and want to approximate the mean as

$$E[x] = \bar{x} = \frac{1}{n} \sum_{i=1}^{N} x_i$$

Then

$$\mathrm{Var}(\bar{x}) = \frac{\mathrm{Var}(x)}{n}$$

However, if x_i are not independent, but autocorrelated, then n is replaced by

$$n_{\mathsf{eff}} = \frac{n}{1 + 2\sum_{t=1}^{\infty} \rho_t}$$

where ρ_t are autocorrelations. These can be estimated from the sample.

n_{eff}/n is an excellent diagnostic tool

Assessing mixing: $\hat{R}$

Suppose you started m MCMC chains, each for length n, for some scalars x_{ij} for $i = 1, \ldots, n$, $j = 1, \ldots, m$. Let

$$\bar{x}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad \bar{x}_{\cdot\cdot} = \frac{1}{m}\sum_{j=1}^{m} x_{\cdot j} \qquad s_j^2 = \frac{1}{n-1}\sum (x_{ij} - \bar{x}_{\cdot j})^2$$

$$B = \frac{n}{m-1}\sum_{j=1}^{m} (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2 \qquad W = \frac{1}{m} s_j^2$$

for variance between and within chains. Estimate the total variance as

$$V = \frac{n-1}{n} W + \frac{1}{n} B$$

and define

$$\hat{R} = \sqrt{\frac{V}{W}}$$

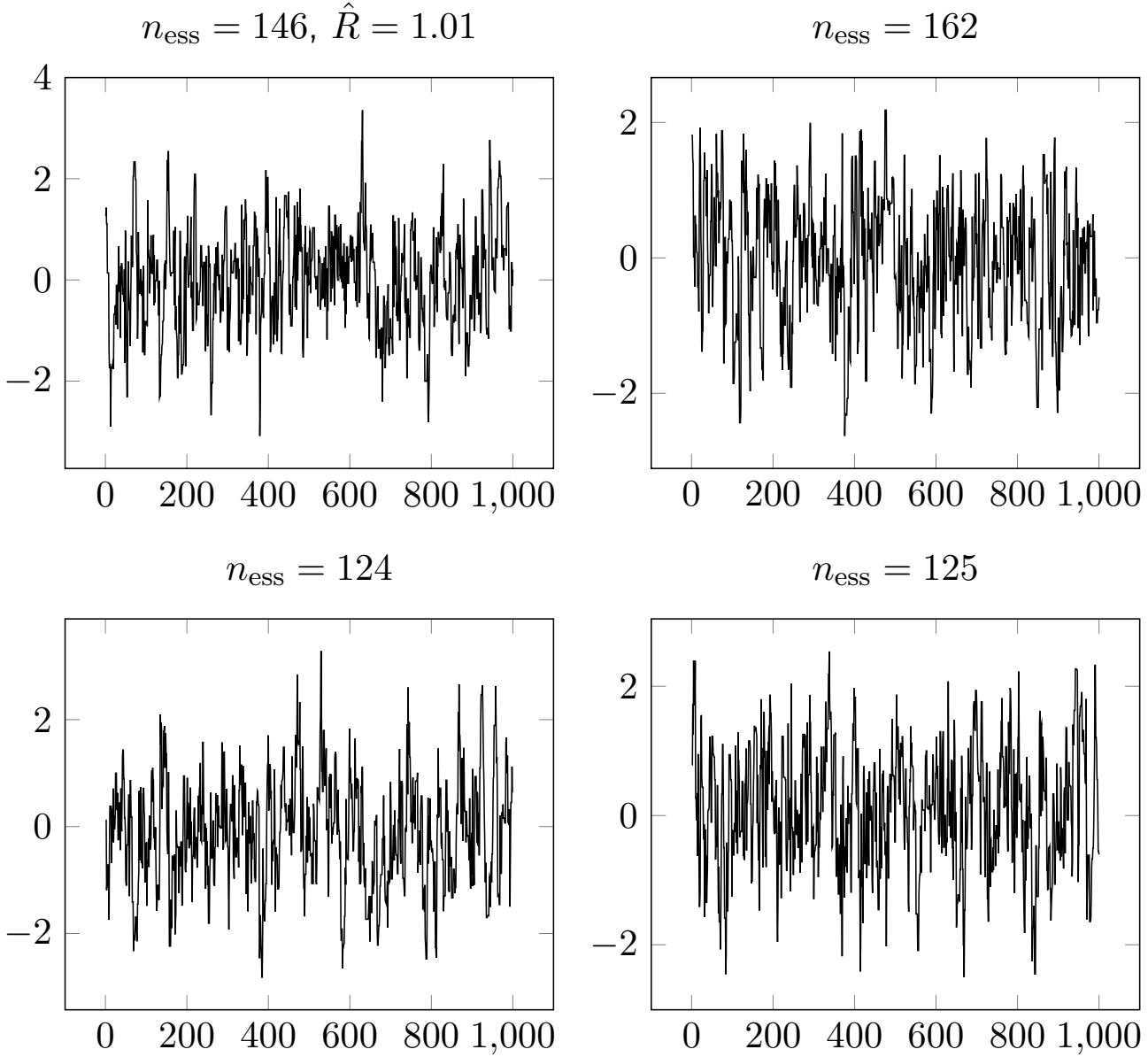
Note $\hat{R} \geq 1$, and as $n \to \infty$ we have $\hat{R} \to 1$.

Strategy with multiple chains

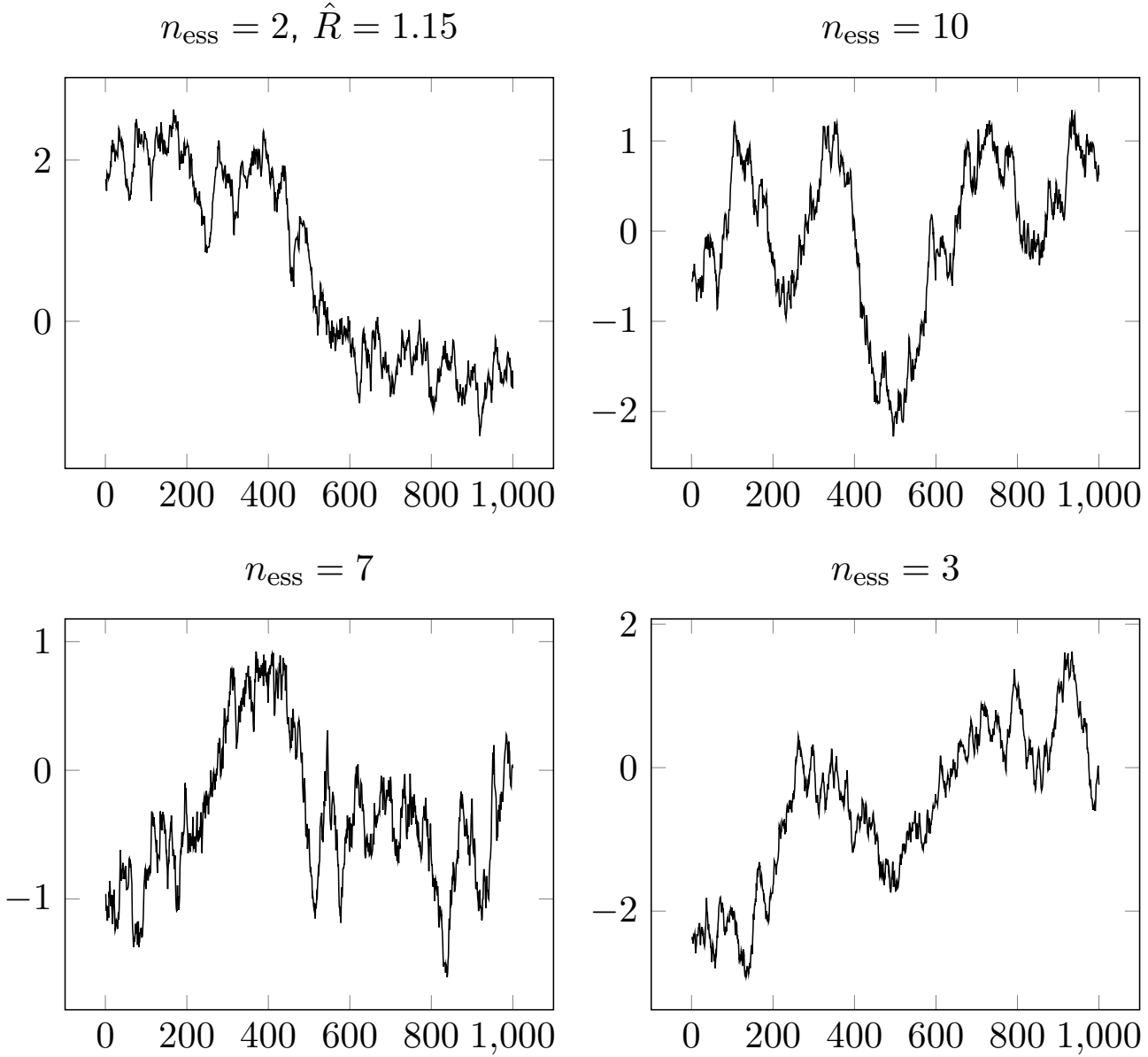
1. start 3–5 overdispersed chains
2. calculate n_{eff}/n
3. calculate $\hat{R}$, suspect problems if larger than 1.05

We continue our example with the $N(0, 1)$ distribution, sampling a normal J with standard deviation σ.

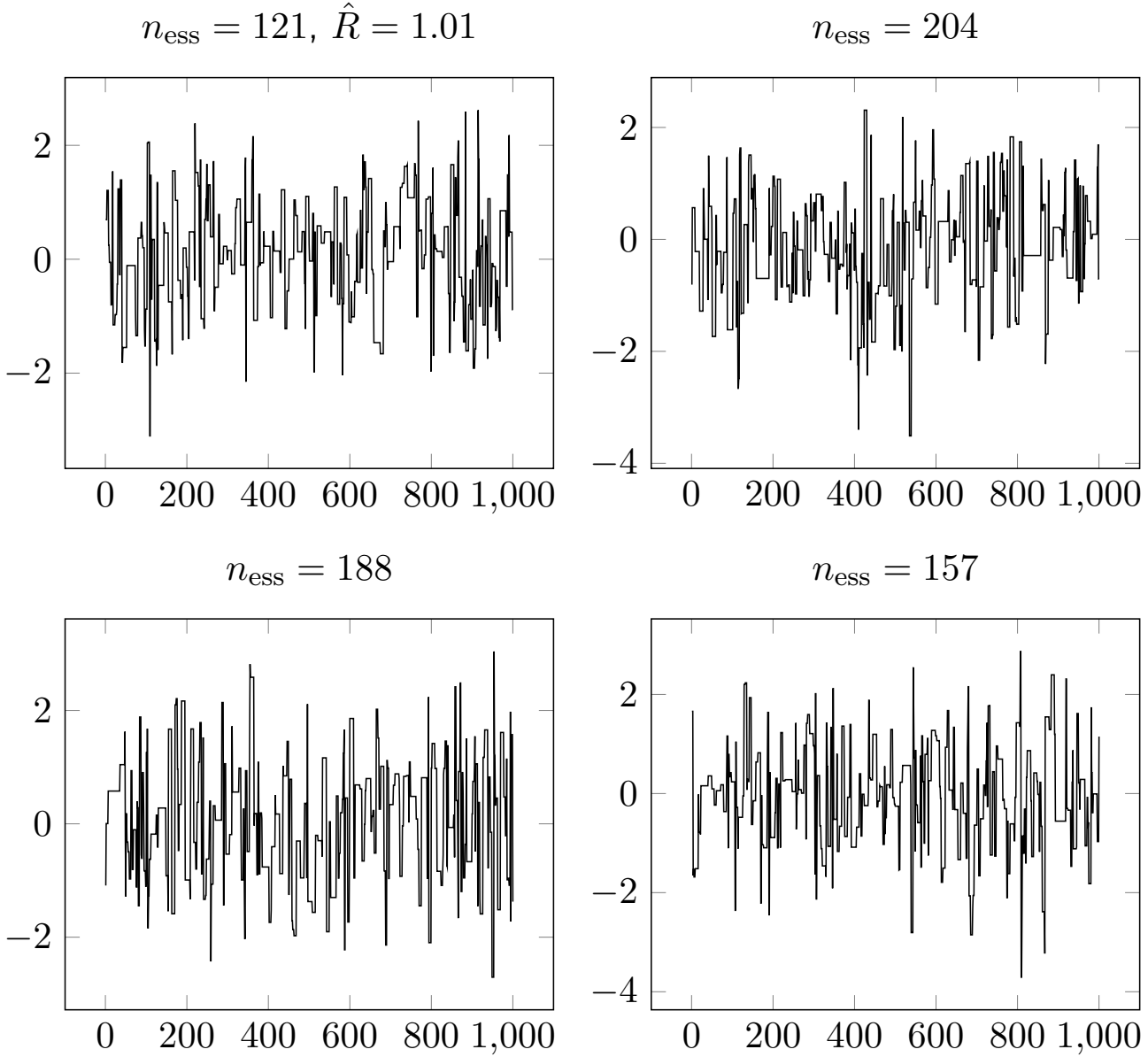
$\hat{R}$ and effective sample size with $\sigma = 1$



$\hat{R}$ and effective sample size with $\sigma = 0.1$



$\hat{R}$ and effective sample size with $\sigma = 5$



RWMH in practice

- even after careful tuning, it may be difficult to get good mixing and convergence from RWMH
- both can easily get much worse if the dimension of the parameter space increases
- many sharp local modes can be very problematic
- "folk theorem of statistical computing" by Andrew Gelman:

  > *When you have computational problems, often there's a problem with your model.*

- better algorithms exist, using local information (derivatives)
- various reparametrization tricks may speed up convergence
- but these are not yet easy to use for DSGE estimation

Debugging MCMC

Use fake/simulated data. This should be the first step in any estimation problem by default, even when not debugging. Systematic approach:

1. pick a θ,
2. simulate $y \mid \theta$,
3. sample the posterior $\theta \mid y$,
4. compare quantiles (systematic) or means (exploratory)

- fix/unfix parameters one by one
- poor convergence and mixing is usually a warning sign that something is wrong
- proceed from simpler models to more complex models
- Dynare specific: pay attention to warnings

The "dangers" of Bayesian inference

- provided you have a proper posterior and a good MCMC algorithm (or are willing to wait long enough), you can estimate any model on any remotely compatible data
- even if the model does not remotely resemble the data
- usually, models fit some features of the data well, others not so well
- need to decide if those are important, to improve or replace the model
- in social science, no models are "true", all are approximations
- convenience motivates modeling choices, this may or may not be problematic
- trade-offs between model convenience and accuracy
- never stop at estimating the model, spend at least as much time on model checking (especially posterior predictive checks)
- understanding results in the context of the literature

Posterior odds

- used in Bayesian econometrics (esp. DSGE), with some problems
- two competing models M_1 and M_2
- relative posterior probabilities

$$\frac{p(M_2 \mid y)}{p(M_1 \mid y)} = \frac{p(M_1)}{p(M_2)} \cdot \text{Bayes factor}(M_2, M_1)$$

- captures the relative effect on the posterior independently of the prior
- only applicable for two models

$$\text{Bayes factor}(M_2, M_1) = \frac{p(y \mid M_2)}{p(y \mid M_1)} = \frac{\int p(\theta_2 \mid M_2) p(y \mid \theta_2, M_2) d\theta_2}{\int p(\theta_1 \mid M_1) p(y \mid \theta_1, M_1) d\theta_1}$$

- can work well if we really have two discrete models
- doesn't work well when the models are part of an inherently continuous family; can't use with improper priors, sensitive to limit priors
- not very robust to priors and likelihoods in a continuous setting
- recommendation: embed models in a continuous setting instead

Posterior predictive checks

- idea: compare features of simulated data to the actual data
- very versatile and powerful tool
- easy to apply: just simulation
- key idea: test statistic $T(y, \theta)$, a function of the data (which can be simulated), and the estimated parameters
- recall classical p-value

$$\Pr(T(y^{\text{rep}}) \geq T(y) \mid \theta)$$

- we generalize this to a Bayesian p-value

Posterior predictive checks (cont)

- define joint distribution for replicated data and parameter

$$p(y^{\text{rep}}, \theta \mid y) = \overbrace{p(y^{\text{rep}} \mid \theta)}^{\text{likelihood/model}} \underbrace{p(\theta \mid y)}_{\text{posterior}}$$

- this is usually done with simulation: for each posterior draw θ_i, simulate a y_i^{rep}
- Bayesian p-value is defined as

$$p_B = \Pr(T(y^{\text{rep}}, \theta) \geq T(y, \theta) \mid y)$$

over the joint distribution $p(\theta, y^{\text{rep}} \mid y)$, ie

$$p_B = \int \int \left[T(y^{\text{rep}}, \theta) \geq T(y, \theta)\right] p(y^{\text{rep}} \mid \theta) p(\theta \mid y) dy^{\text{rep}} d\theta$$

- convention: when $p_B > 0.5$, use $1 - p_B$

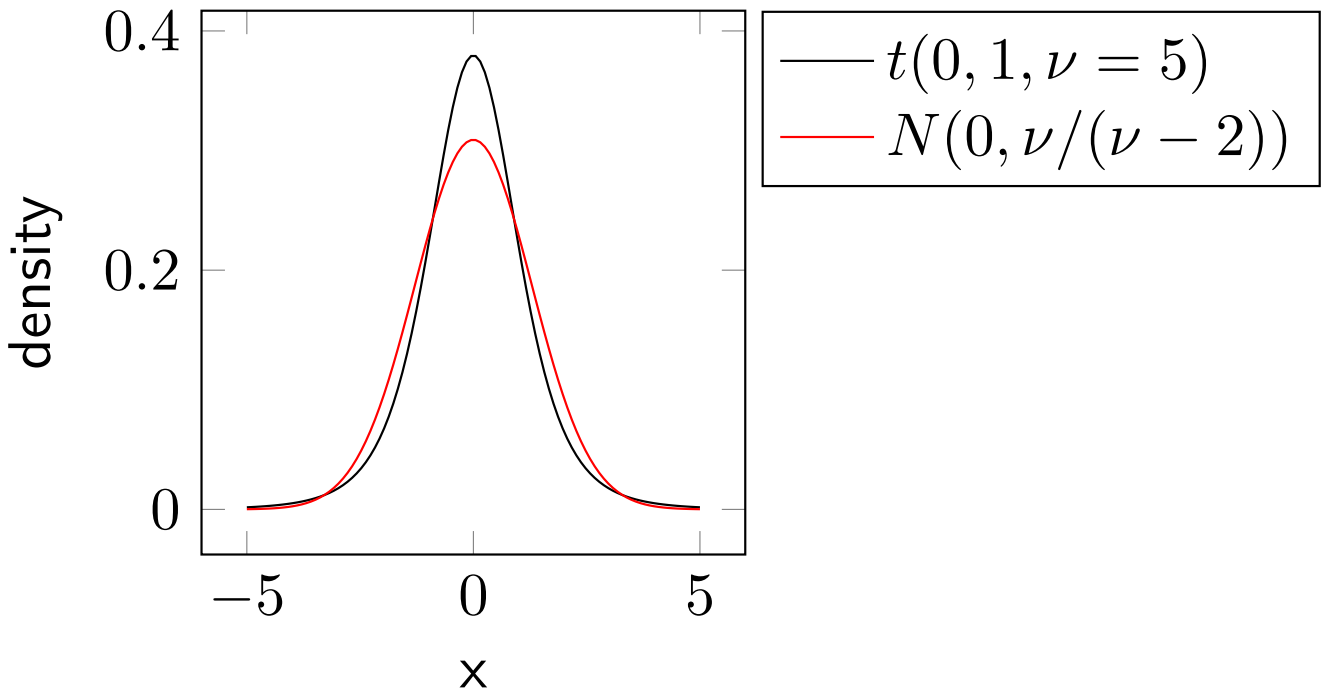
Posterior predictive checks: example

- generate IID data $i = 1, \ldots, 100$ from

$$y_i \sim t(0, 1, 5)$$

- estimate the misspecified model

$$y_i \sim N(\mu, \sigma^2)$$

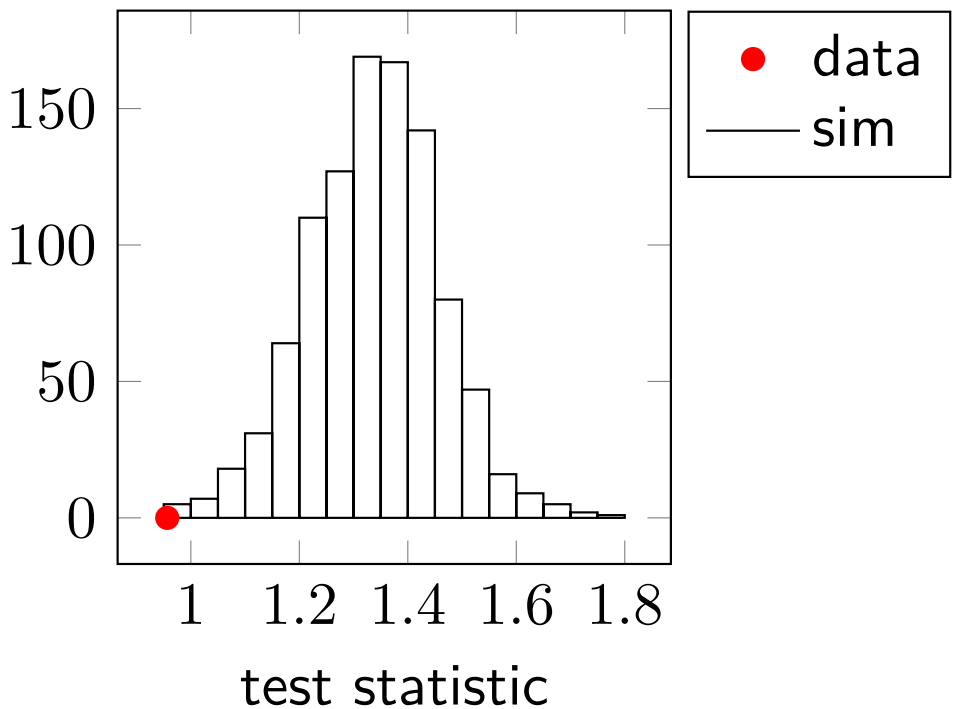


Picking a test statistic

For data y, define

$$T(y) = \frac{q_{75\%} - q_{25\%}}{\mathrm{std}(y)}$$

This is normalized by the scale, yet the quantiles should be informative. It is also independent of θ.



In this case, $p_B \approx 0$.

Practical advice on posterior predictive checks

- consider multiple test statistics
- 200 replications are OK, 1000 are plenty: if p_B is small (≤ 0.05), the exact value is not that relevant
- save computation: for each replication, calculate the multiple Ts
- plot T as a histogram (when independent of θ) or a scatter plot

Be creative with salient features of the data

- the probability of "large" movements in time series
- T can come from point estimates of simple models on the data
- when does a VAR impulse response peak?
- approximate half lives

Dynare

- a software suite that can solve, simulate and estimate DSGE and similar models
- http://www.dynare.org/, free and open source software
- can automate mechanical steps ...
- ...but requires an understanding of the methodology
- many prominent researchers contribute cutting-edge methods
- has an extensive manual (200 pages) and user guide, these should be studied in detail
- additional model examples with data at http://macromodelbase.com/
- friendly forums at https://forum.dynare.org/
- some papers that use Dynare have code in an appendix, replication is a good way to learn

Running Dynare

- primarily with Matlab (a bit faster) or Octave (free software)
- model and operations are described by a model file
- then call it as

  `dynare modelfile.mod [options]`
- this generates `model.m` and similar interim files
- warning and error messages are usually very informative
- results end up in global variables like `oo_`, see the documentation on its structure

Frequently used distributions built into Dynare

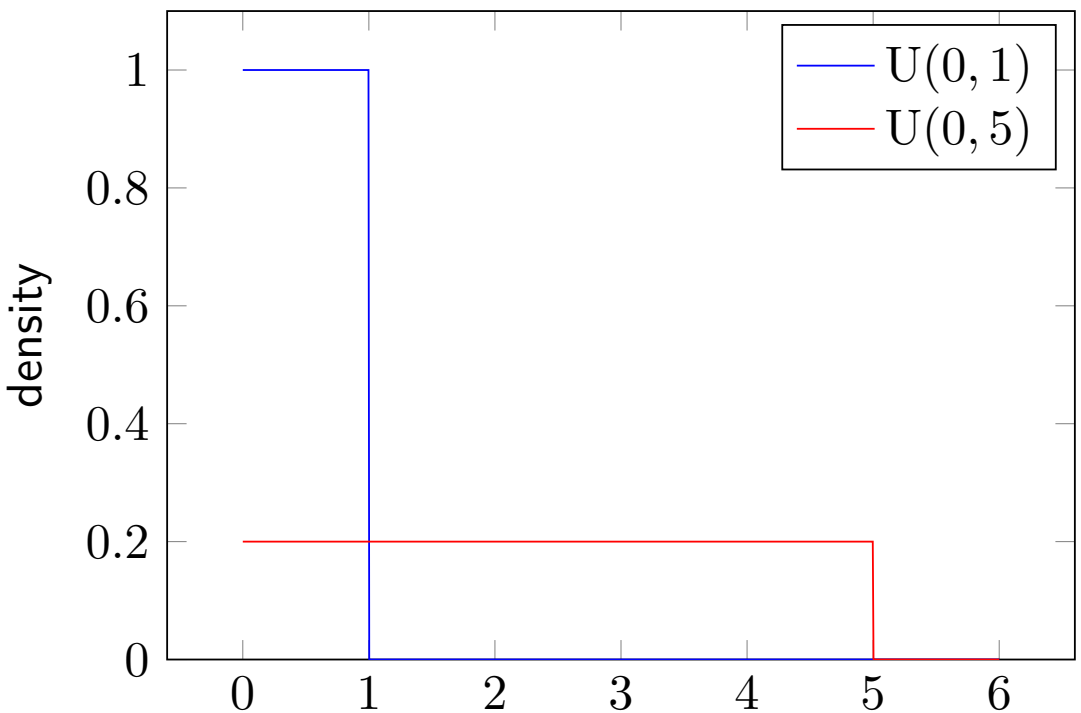
uniform	$U(a,b)$
normal	$N(\mu,\sigma^2)$
beta	Beta(α,β)
gamma	Gamma(α,θ)
inverse gamma	InvGamma(α,β)

Note: alternative parametrizations exist, eg $N(\mu,\sigma)$. Always check documentation of the relevant software, sometimes you will need to translate.

Uniform distribution $U(a, b)$

Useful for vague priors. Reasonable option for initial exploration.

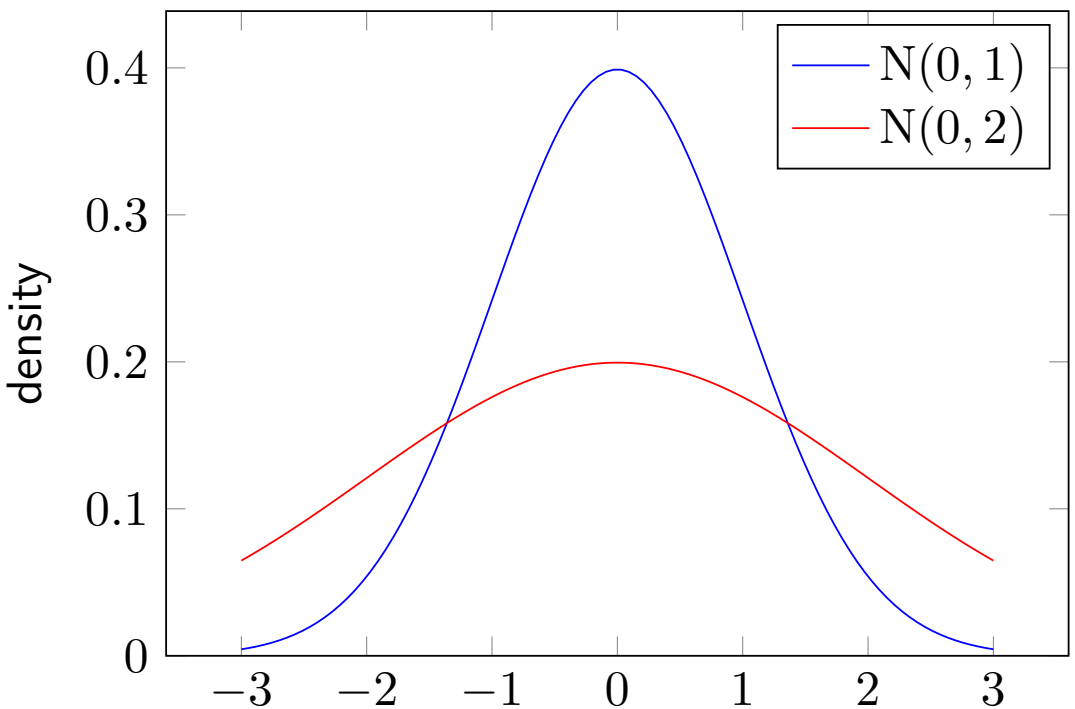
$$f(x; a, b) = \frac{[a \leq 1 \leq b]}{b - a} \qquad \mathrm{E}[x] = \frac{a + b}{2} \qquad \mathrm{std}[x] = \frac{b - a}{\sqrt{12}}$$



Normal distribution $N(\mu, \sigma)$

Location μ, scale $\sigma > 0$. Useful for vague priors on $\mathbb{R}$. See also: multivariate normal.

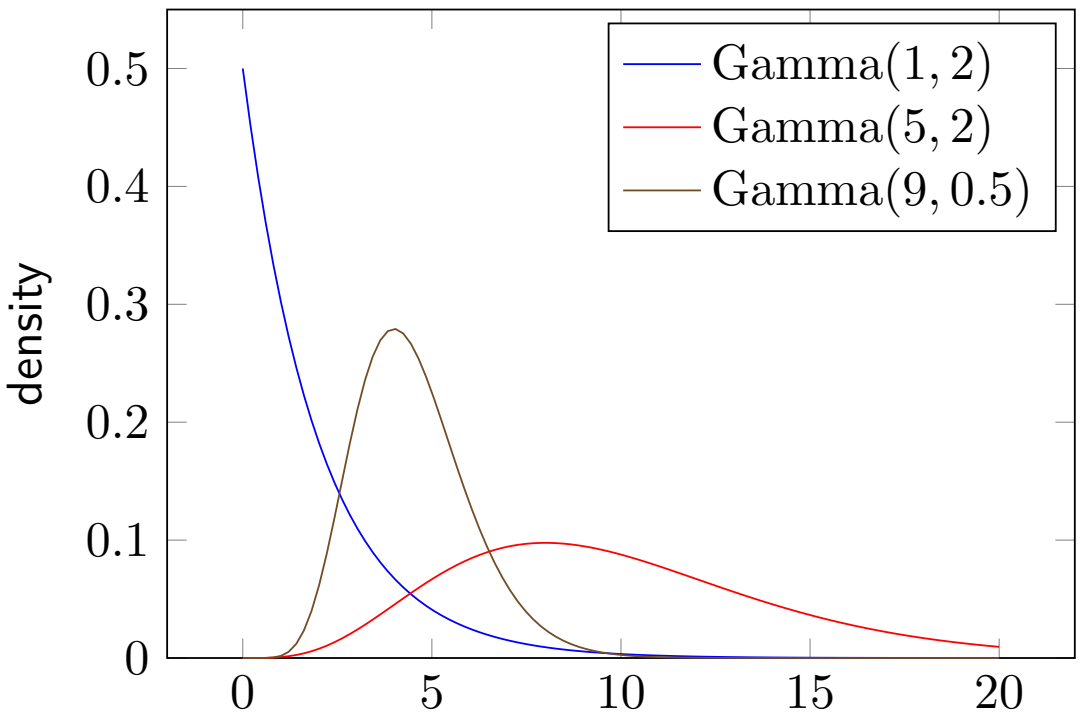
$$f(x; \mu, \sigma) \propto \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \qquad \mathrm{E}[x] = \mu \qquad \mathrm{std}[x] = \sigma$$



Gamma distribution Gamma(α, θ)

Shape α, scale θ. For priors on $\mathbb{R}_+$.

$$f(x; \alpha, \theta) \propto x^{\alpha} \exp(-x/\theta)[x > 0] \qquad \mathrm{E}[x] = \alpha\theta \qquad \mathrm{std}[x] = \sqrt{\alpha}\theta$$

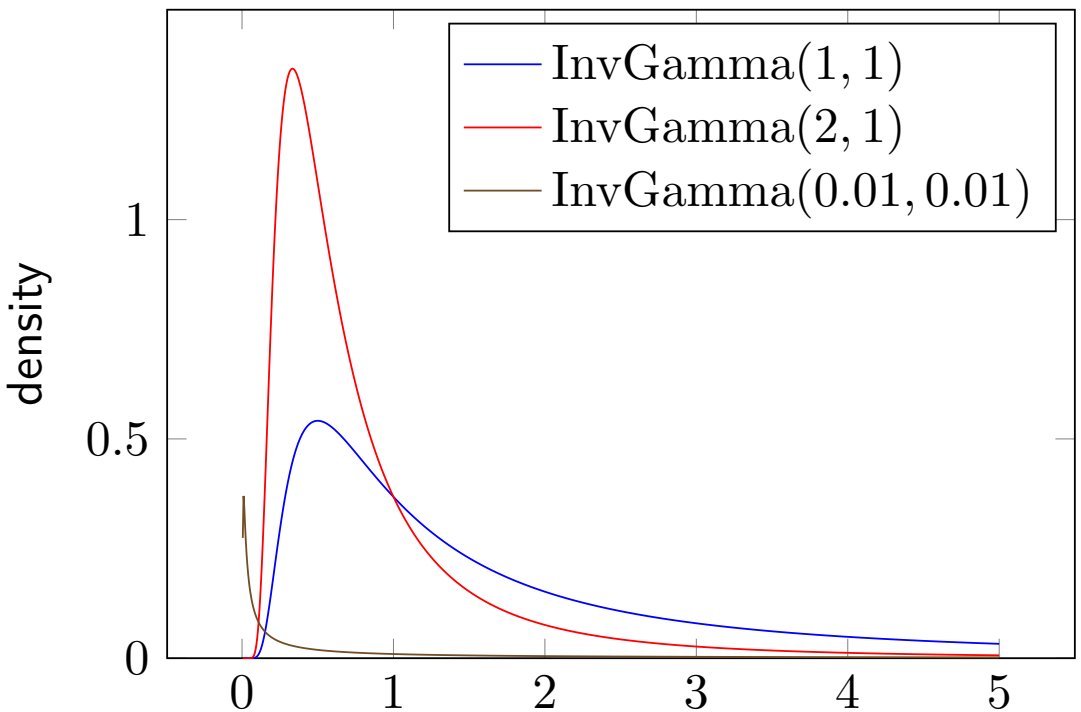


Inverse gamma distribution InvGamma(α, β)

Shape $\alpha > 0$, scale $\beta > 0$. Some examples use InvGamma$(0.01, 0.01)$ for vague variance priors, avoid this, inference can be very sensitive to it (Gelman 2006). When $x \sim$ Gamma(α, β), $1/x \sim$ InvGamma(α, β).

$$f(x; \alpha, \beta) \propto x^{-\alpha-1} \exp(-\beta/x)[x > 0]$$

$$\mathrm{E}[x] = \frac{\beta}{\alpha - 1} \quad (\alpha > 1) \qquad \mathrm{std}[x] = \frac{\beta}{(\alpha - 1)\sqrt{\alpha - 2}} \quad (\alpha > 2)$$

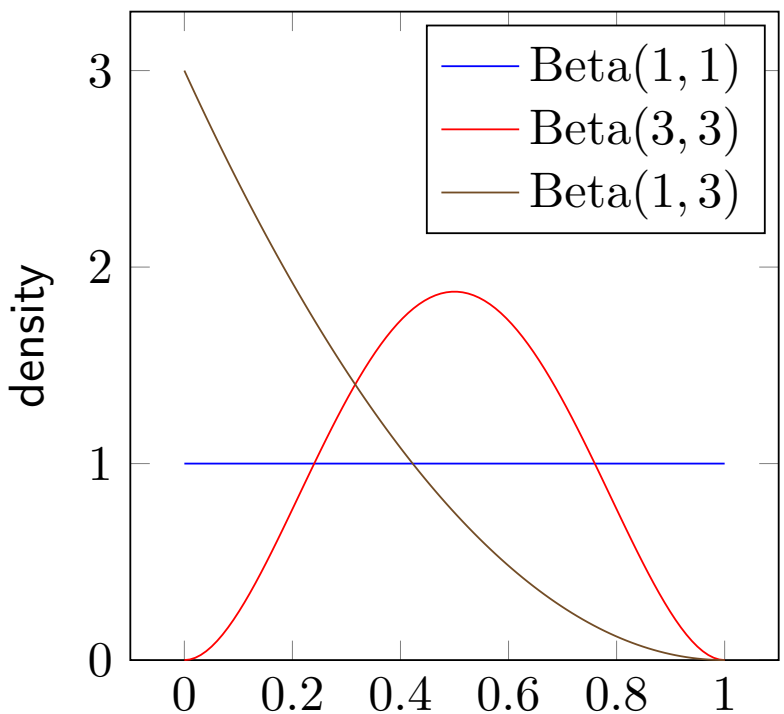


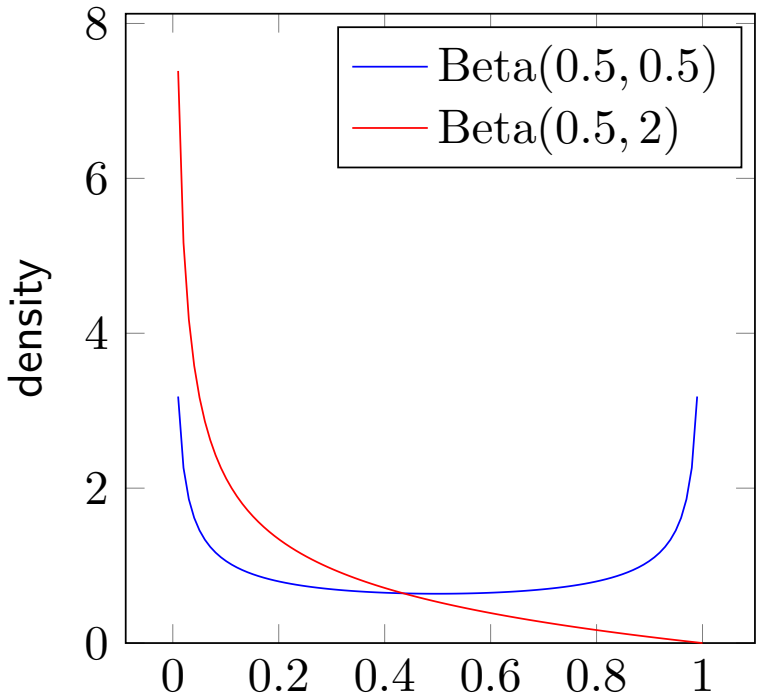
Beta distribution Beta(α, β)

Shape $\alpha, \beta > 0$. Useful for parameters on intervals (eg $[0, 1]$, or transform). Multivariate generalization: Dirichlet.

$$f(x; \alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}[0 \leq x \leq 1]$$

$$\mathrm{E}[x] = \frac{\alpha}{\alpha+\beta} \qquad \mathrm{std}[x] = \frac{\sqrt{\alpha\beta}}{(\alpha+\beta)\sqrt{\alpha+\beta+1}}$$





Further reading

Särkka (2013) on the Kalman filter and smoother

Sims and Zha (1998) on Bayesian VAR

Herbst and Schorfheide (2015) on Bayesian DSGE estimation in general

Sims, Waggoner, and Zha (2008) for marginal likelihood calculations

Berger et al. (1988) for a detailed discussion of statistical principles

Credits for materials used from other sources

- The physicist's twins example is from Efron (2005).
- Discussion from $\hat{R}$ and n_{eff} follows Gelman et al. (2013).

References

- Berger, James O et al. (1988). "The likelihood principle". In: *Lecture notes-Monograph series* 6, pp. iii–199.
- Christiano, Lawrence J, Martin Eichenbaum, and Charles L Evans (2005). "Nominal rigidities and the dynamic effects of a shock to monetary policy". In: *Journal of Political Economy* 113.1, pp. 1–45.
- Efron, Bradley (2005). *Modern science and the Bayesian-frequentist controversy*. Division of Biostatistics, Stanford University.
- Fernández-Villaverde, Jesús, Juan Francisco Rubio-Ramírez, and Frank Schorfheide (2016). "Solution and estimation methods for DSGE models". In: *Handbook of Macroeconomics* 2, pp. 527–724.
- Gelman, Andrew (2006). "Prior distributions for variance parameters in hierarchical models". In: *Bayesian analysis* 1.3, pp. 515–534.
- Gelman, Andrew et al. (2013). *Bayesian data analysis*. CRC Press.
- Griffoli, Tommaso Mancini (2013). *DYNARE User Guide*. Tech. rep. Dynare.
- Herbst, Edward P and Frank Schorfheide (2015). *Bayesian estimation of DSGE models*. Princeton University Press.
- Särkka, Simo (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.
- Sims, Christopher A, Daniel F Waggoner, and Tao Zha (2008). "Methods for inference in large multiple-equation Markov-switching models". In: *Journal of Econometrics* 146.2, pp. 255–274.
- Sims, Christopher A and Tao Zha (1998). "Bayesian methods for dynamic multivariate models". In: *International Economic Review*, pp. 949–968.
- Smets, Frank and Rafael Wouters (2007). "Shocks and frictions in US business cycles: A Bayesian DSGE approach". In: *American Economic Review* 97.3, pp. 586–606.