

Bayesian Unit Root Testing: A Short Simulation Study

Antoine Carnec

Introduction

Assessing whether a stochastic process is stationary is a critical part of any time series analysis, since we usually rely on theoretical guarantees that our estimators and inferences are well behaved under the assumption of stationarity. The classical test for this is the *Unit Root Test* or Dickey Fuller test, which amounts to testing the null $H_0 : \varphi = 1$ against the alternative $H_1 : \varphi < 1$ in the equation

$$y_t = \varphi y_{t-1} + \varepsilon_t. \quad (1)$$

(We assume that $\varphi \geq 0$ for simplicity, and that the variance of ε_t is known).

The test proceeds by estimating the OLS estimator of the coefficient $\hat{\varphi}$, from whence one can derive that $\hat{\varphi} - 1$ has the *Dickey Fuller* distribution, whose quantiles can be readily tabulated (Tsay, 2005). Therefore a rejection region can be defined and the test is defined.

In this paper I want to investigate empirically whether there could be a Bayesian alternative to this test. I want to investigate whether the very simple strategy of calculating the alternative hypothesis posterior $p(\varphi < 1 \mid \mathbf{y})$ by a sampling strategy can compete with the DF—a classic hypothesis testing approach. I will do this simply by simulating many time series of size T following an AR(1) process with coefficient φ and $\varepsilon_t = \mathcal{N}(0, 1)$. I will see if the Bayesian approach can be competitive with the Frequentist approach in this simple setting. So and Ll (1999) carry out the same strategy and find that their method performs about as well as the ADF test at moderate sample sizes. They use a Gibbs Sampler as their inference algorithm, therefore their approach is less general than the one presented here we will let Stan take care of inference for us.

Why be Bayesian?

Why be Bayesian in this context? The Bayesian approach is perfectly general from a modelling perspective. For example, if one wanted to design a Dickey-Fuller-type test for the model

$$y_t = f(t) + \varphi y_{t-1} + \varepsilon_t$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ is some arbitrary function, the practitioner must find a test statistics for every f proposed. The Bayesian however can easily incorporate this into their model simply by changing their likelihood function (I discuss some possible modelling extensions in the conclusion). The tradeoff is of course that Bayesian computation is difficult, and sampling strategies are only approximate methods. I don't think this is such a big problem, as the DF test also relies on an approximation—insofar as the test statistic is asymptotically DF-distributed. The Bayesian approach has no problem with low T , and the accuracy of the Monte Carlo estimator can be increased by increasing N (equivalently running the inference algorithm for longer).

What’s wrong with the unit root test?

A very small p -value only tells the statistician that there is some level of surprise. It certainly is not true that small p value in the DF test implies the rejection of the *scientific hypothesis* that $\varphi = 1$. Perhaps there is a small p value because the DF equation ? is misspecified. If we do suspect a misspecification in the model, it is not so easy to change the model that much, since the ADF test is derived for specific functional forms of ?.

A Bayesian Approach

Bayesian Statistics uses characterizes a model with *prior* probability distributions on the model parameters θ and a specifies a *likelihood* $p(\mathbf{y} \mid \theta)$, which captures how likely the data \mathbf{y} is given a certain configurations of the parameters. Having specified these, we can recover the conditional distribution of the parameters given the data we have using Bayes Theorem

$$p(\theta \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta)p(\theta)}{\int p(\mathbf{y} \mid \theta)p(\theta)d\theta} \propto p(\mathbf{y} \mid \theta)p(\theta).$$

Given that the likelihood correctly specifies the data generating distribution and that the prior accurately represents the statisticians prior beliefs, $p(\theta \mid \mathbf{y})$ captures the beliefs that the statistician should have about the parameter.

In the case of our AR(1) model, let $\theta = \varphi$. The model we use is

$$p(\varphi \mid \mathbf{y}) \propto \prod_{t=2}^T \mathcal{N}(\varphi y_{t-1}, \sigma_\varepsilon^2) \cdot p(\varphi) \quad (2)$$

where we take σ_ε^2 as observed for simplicity, $p(\varphi)$ is the prior distribution we set on φ .

Now Equation 2 gives us the posterior up to a normalising constant. In general it is very hard to calculate this normalising constant, and therefore we choose to sample from this distribution. This can easily be done with the Stan probabilistic programming language (Carpenter et al., 2017). Given a generative model such as (2), Stan allows the user to sample from the posterior using Hamiltonian Monte Carlo, which is a state of the art MCMC technique which tends to scale well with dimension¹. We include the Stan code to implement (2) in the appendix. I also set the number of samples to $N = 5000$ and use the NUTS HMC sampler.

Given a sample $\{x_n\}_{n=1}^N$ of $p(\varphi \mid \mathbf{y})$, we can estimate $\mathbb{E}_{\varphi \mid \mathbf{y}}[f]$ for any function f . This estimate is consistent and unbiased. In particular we are interested in the probability that $\varphi < 1$, which we can recover by

$$\frac{1}{N} \sum_{n=1}^N \mathbb{1}(x_n < 1) \rightarrow P(\varphi < 1 \mid \mathbf{y}), \text{ as } N \rightarrow \infty$$

Our Bayesian Hypotheses will be

$$H_0 : \varphi < 1 \quad H_1 : \varphi \geq 1.$$

We still have not yet show how know $p(H_1 \mid \mathbf{y}) := p(\varphi < 1 \mid \mathbf{y})$ leads to a decision—that is, do we reject the null? Here we reject the null if $p(H_1 \mid \mathbf{y}) > p(H_0 \mid \mathbf{y})$, this is the same decision rule used in (So and Ll, 1999).

¹many caveats could be made here, I refer the reader to Betancourt (2017) for a discussion

Prior for φ

An important question is which prior to set for φ . Generally, there are four approaches to choosing this prior. A convenient choice of prior (one which leads to closed form expressions for the posterior), a non-informative or weakly-informative prior (priors which don't enforce little to no regularization), objective priors (priors which can be deduced from theoretical arguments) and subjective priors (priors which encode the statisticians prior beliefs).

In this paper we consider two simple priors. A weakly informative prior, more specifically $\varphi \sim \mathcal{N}(\frac{1}{2}, 1)$. This prior ensures that φ has density over the space where we believe it will be between—that is between 0 and some number greater than 1—but that it does not have much mass at $\varphi = 5$ for example².

We consider a prior that is relatively strongly concentrated around 1. This reflects the fact that researchers may have a strong consensus prior belief that the time series is a random walk (using arguments such as the Efficient Market Hypothesis for example). To encode this, we set the prior to a Laplace(1, 0.1) distribution.

The priors are plotted in Figure 3.

Simulations

How should we compare the two procedures? Well, it is described in Benjamin and Berger (2019) that a given a p -value, we can calculate a tight upper bound on the posterior probability of the null hypothesis simply with

$$p^U(H_1 | \mathbf{y}, p) = \frac{BFU}{1 + BFU}, \quad BFU = \frac{1}{-ep \log(p)}$$

Hence we perform a direct comparison between posterior probabilities for the two priors with the posterior probability upper bound for $T \in \{50, 500\}$. Note that we

We plot the results in Figure 1 and 2.

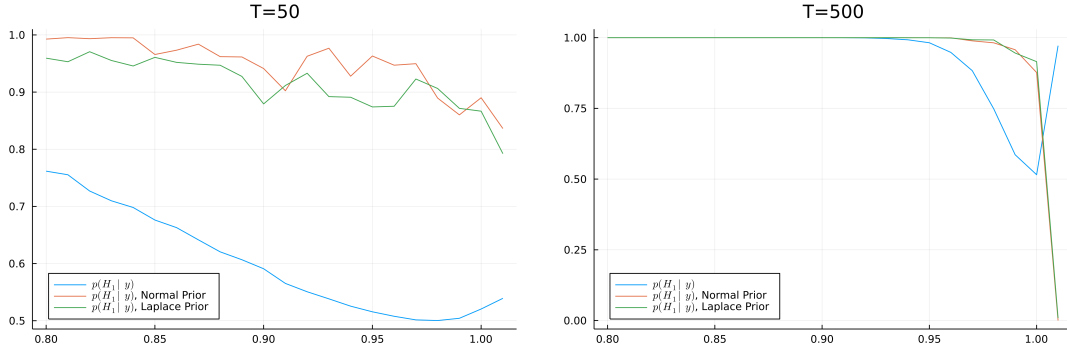


Figure 1: φ on the x axis, plotted from 0 to 1.01

Looking at $T = 50$, we can see that there are not enough data points to really bring down the posterior probability at $\varphi = 1$. This is supported by what we see when $T = 500$ which does a better job. We can see that even the upper pound of the DF test does not give the same

²If $\varphi = 5$, then simply looking at the time series would tell you that it's not stationary

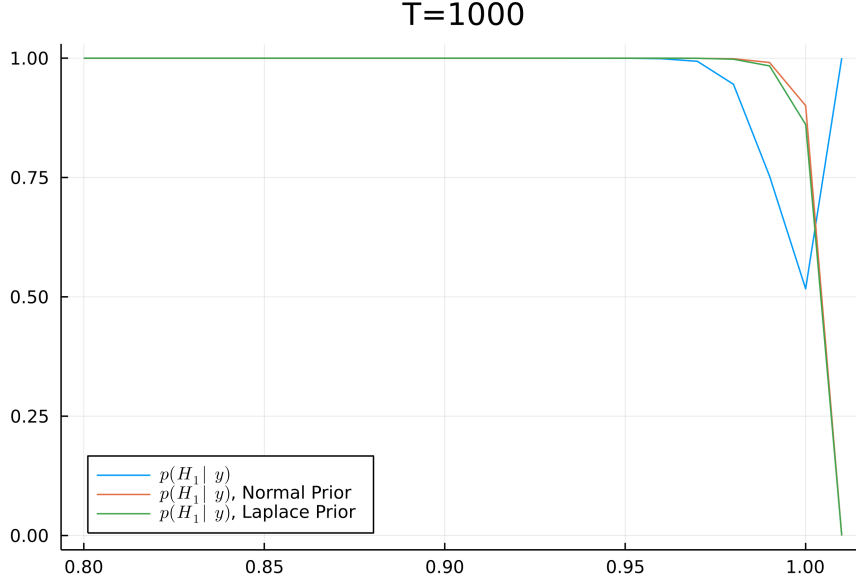


Figure 2: φ on the x axis, plotted from 0 to 1.01

confidence when $\varphi < 1$. However, there is obviously overconfidence here, since at $\varphi = 1$ the posterior distributions do not fall sufficiently fast enough—even for $T = 1000$. From these figures above we definitely cannot endorse the Bayesian approach. On the other hand, we can notice that the Bayesian methods still “get it right” at $\varphi = 0.99$ for example, improving upon the loss of power near 1 that the DF test suffers from. However, when $\varphi = 1$ the proposed method does not identify the process as stationary.

Let’s verify that last claim. Let $T = 5000$. We will simulate 50 AR(1) process with some φ and run all of the hypothesis decision rules, with an $\alpha = 0.05$. Recall that for the Bayesian methods we reject the null if $p(H_1 | \mathbf{y}) > 0.5$. We show the proportion of times the procedures made the correct choice.

	$\varphi = 0.95$	$\varphi = 0.99$	$\varphi = 1$
DF test	1	1	1
Normal Prior	1	1	0.04
Laplace Prior	1	1	0.02

Table 1: $T = 5000$, 50 AR1 processes run, $\alpha = 0.05$

As Table 1 shows, the Bayesian unit tests failed pretty miserably, while the DF test got everything right.

How could this analysis be extended?

We saw in the results that the Bayesian Hypothesis test didn’t perform very well at all. Perhaps we could employ better priors on our Hypotheses. An interesting approach might be to implement Johnson and Rossell (2010) Non Local Prior approach to Bayesian hypothesis testing. For

these special priors, the authors prove that the Bayes Factors converge at a very fast rate, even exponential under some conditions.

Another interesting idea is the following. From a decision making standpoint, we are primarily interested in $p(\varphi_1 = 1)$. However, it's difficult to estimate using simulations since we will never sample 1 exactly. A Sequential Monte Carlo sampler (Del Moral et al., 2006) could be designed, inspired by the rare event SMC techniques existing in the literature. The idea is to sample from a sequence of distributions such that the support of these distributions closes in on a small interval around 1.

Stan Code

Note, we employ a common technique when using HMC by standardizing y as a preprocessing step. This is know to (sometimes) make inference work better (Gelman et al., 1995).

Weakly Informative Prior

```
data {
  int<lower=0> T;
  real<lower=0> sigma;
  vector[T] y;
}
transformed data {
  real<lower=0> sigma_std;
  vector[T] y_std;
  y_std = (y - mean(y)) / sd(y);
  sigma_std = sigma / sd(y) ;
}
parameters {
  real phi;
}
model {
  phi ~ normal(1/2, 1) ;
  y_std[2:T] ~ normal(phi * y_std[1:(T - 1)], sigma_std);
}
```

Laplace Prior

```
data {
  int<lower=0> T;
  real<lower=0> sigma;
  vector[T] y;
}
transformed data {
  real<lower=0> sigma_std;
  vector[T] y_std;
  y_std = (y - mean(y)) / sd(y);
  sigma_std = sigma / sd(y) ;
}
parameters {
  real phi;
}
model {
  phi ~ double_exponential(1, 0.1);
  y_std[2:T] ~ normal(phi * y_std[1:(T - 1)], sigma_std);
}
```

1 Extra Figures

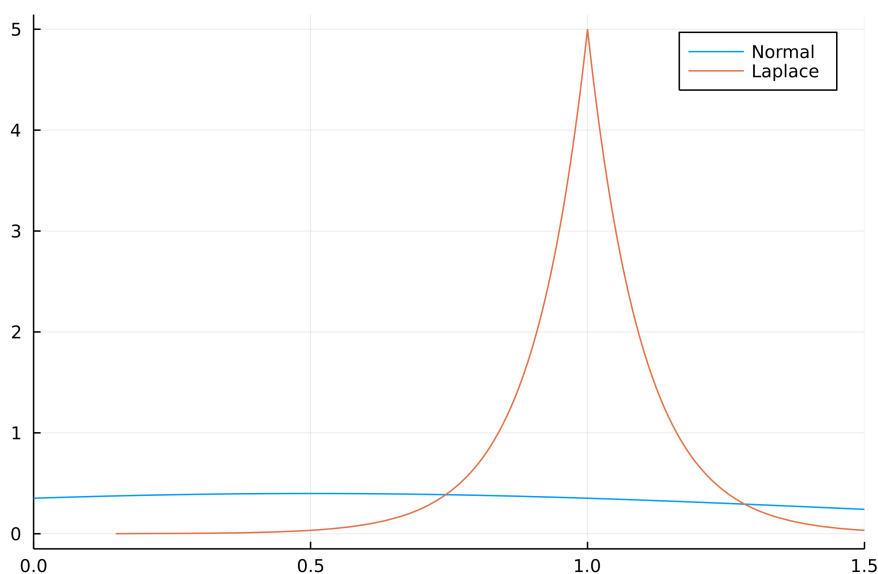


Figure 3: The two priors used

References

- Benjamin, D. J. and Berger, J. O. (2019). Three recommendations for improving the use of p-values. *The American Statistician*, 73(sup1):186–191.
- Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).
- Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.
- Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170.
- So, M. K. and Li, W. (1999). Bayesian unit-root testing in stochastic volatility models. *Journal of Business & Economic Statistics*, 17(4):491–496.
- Tsay, R. S. (2005). *Analysis of financial time series*. John wiley & sons.