Prior for dynamic regression; Model checking via test statistics

October 1, 2014

A generic time series regression prior

- Should be “unprejudiced” about the effects of X-variables on y’s. We want the data to tell us which variables matter and how big their relative effects are.

- Most economic time series are persistent. Our prior should reflect that. We don’t want it to suggest we are surprised when $\hat{y}_t = y_{t-1}$ turns out to be hard to beat as a forecast.

- We are more sure about persistence at low than at high frequencies. I.e. more confident that y_t is similar to the average of $y_{t-1}, \ldots, y_{t-4}$ than that y_t is similar to y_{t-1} when past y_t values have been oscillating.

- Big coefficients on more recent lagged values are more likely than big coefficients from more distant ones.

- Usually, we think R^2 does not grow linearly with model size. The prior should not automatically assert that models with more variables or lags are expected $a\ priori$ to have higher R^2.

Conjugate priors; dummy observations

- A conjugate prior density is one that, as a function of the parameters, behaves just like the likelihood function for an observed model.

- Such a prior can be implemented by introducing "dummy observations" at the start or end of the actual sample, though in general the resulting "likelihood" needs a scale factor to be a true joint distribution over data and parameters.

- You can think of dummy observations as "mental data".

- For a standard normal linear regression model, such a prior is normal-inverse-gamma. That is, it gives the variance parameter σ^2 an inverse-gamma distribution and the coefficient vector β a $N(\bar{\beta}, \sigma^2\Omega)$ distribution, conditional on σ^2.

The conjugate prior for regression

- It makes the precision of beliefs about β depend on the equation's residual variance.

- This may be reasonable — we expect bigger coefficients when the scale of variation in y is bigger.

- But if we have good reason to have a prior (say on substantive economic grounds) that has a spread unrelated to σ^2, we have a non-conjugate prior.

Hierarchical prior for regression

- A hierarchical prior is one with two or more layers of parameters.
- Usually at the lowest level, the parameters define a likelihood that is easy to integrate or sample from, often conjugate.
- At the higher levels, the parameters are harder to sample from. We might even just try a few values for them hoping results are insensitive to them, even though in principle we should integrate over them or sample from their distribution.
- If we condition on σ^2, a normal prior for β with fixed variance is conjugate. So we can treat σ^2 as a hyperparameter in that case.
- Often we will want to treat $\bar{\beta}$ and Ω as hyperparameters also.

The cosine transform

Define the $n \times n$ matrix F to have j'th row

$$\cos\left(\frac{\pi(k-.5)(j-1)}{n}\right) ,$$

where k indexes columns. The first row is all ones. The j'th oscillates at frequency $\pi(j-1)/n$

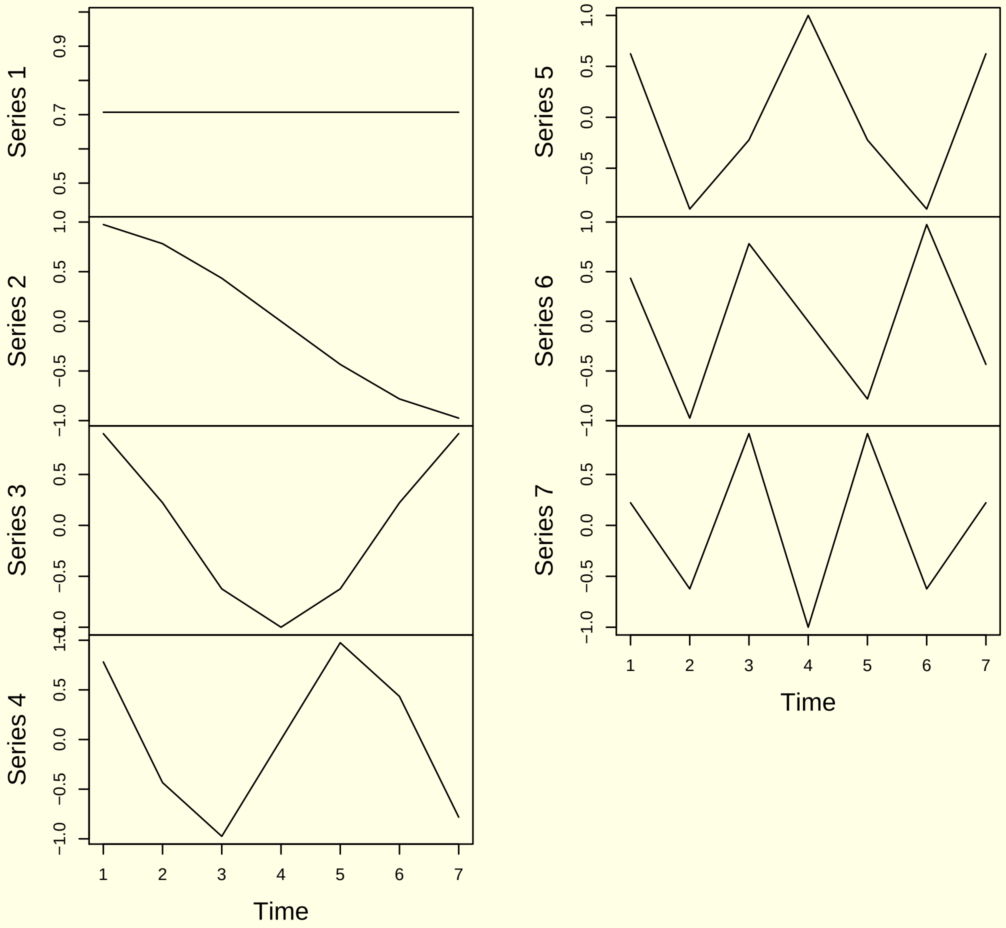
This form makes FF' diagonal, but not scalar, so $F'F$ is not diagonal. The first row is a longer vector. So we normalize by dividing the first row by $\sqrt{n}$. Then $F'F = FF' = \frac{n-1}{2}I$.

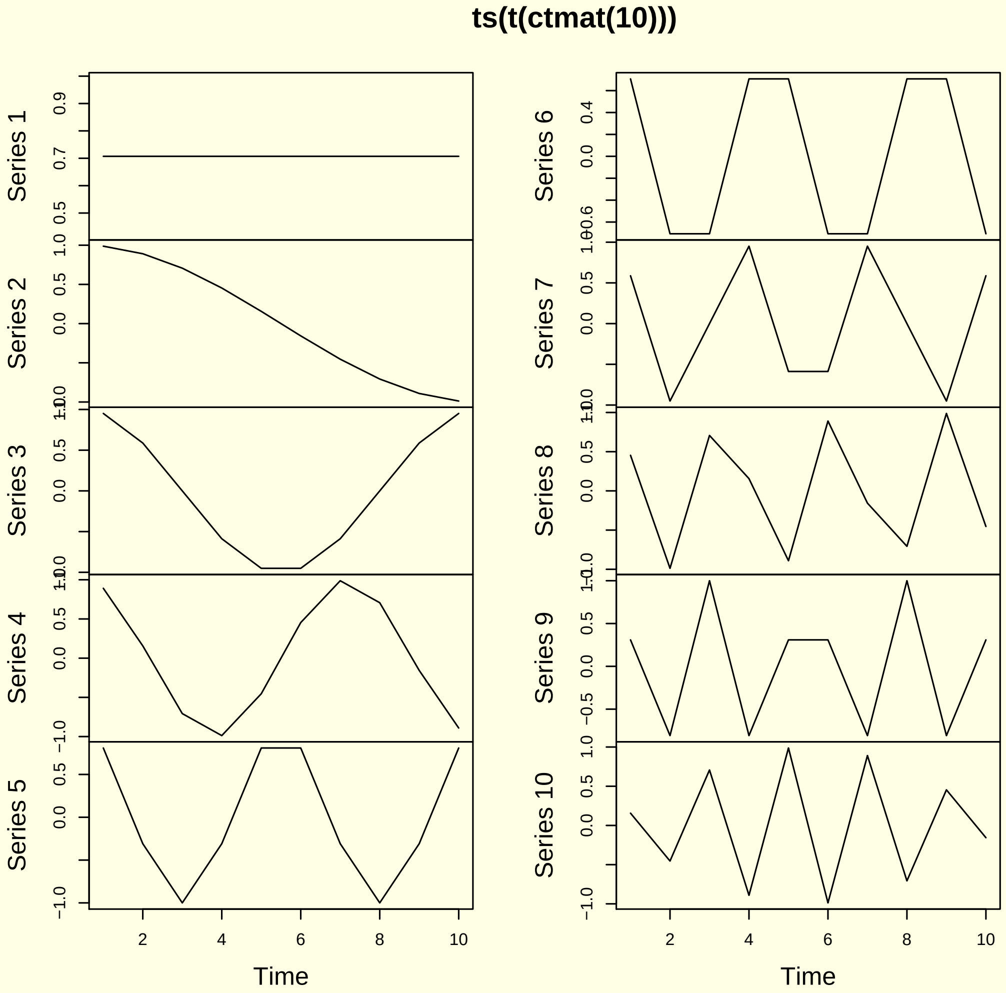
The cosine transform

```
ctmat(2), ctmat(3)
```

$$\begin{matrix} 0.7071068 & 0.7071068 \\ 0.7071068 & -0.7071068 \end{matrix} \qquad \begin{matrix} 0.7071 & .7071 & 0.7071 \\ 0.8660 & 0 & -0.8660 \\ 0.5000 & -1 & 0.5000 \end{matrix}$$

ts(t(ctmat(7)))


Series 1
0.5
0.7
0.9
Series 2
−1.0
−0.5
0.0
0.5
1.0
Series 3
−1.0
−0.5
0.0
0.5
Series 4
−1.0
−0.5
0.0
0.5
1.0
1
2
3
4
5
6
7
Time
Series 5
−0.5
0.0
0.5
1.0
Series 6
−1.0
−0.5
0.0
0.5
1.0
Series 7
−1.0
−0.5
0.0
0.5
1
2
3
4
5
6
7
Time


ts(t(ctmat(10)))
Series 1
Series 2
Series 3
Series 4
Series 5
Series 6
Series 7
Series 8
Series 9
Series 10
0.5
0.7
0.9
−1.0
−0.6
0.0
0.4
0.5
1.0
−0.5
2
4
6
8
10
Time
Time

The dynamic regression model

$$\{y \mid y_{t-1}, \ldots, y_{t-k_y}, x_{1t}, \ldots, x_{1,t-k_1}, x_{2t}, \ldots x_{2,t-k_2}, \ldots, x_{mt}, \ldots, x_{m,t-k_m}\} \\ \sim N(X_t\beta, \sigma^2)\ ,$$

where

$$X_t = (y_{t-1}, \ldots, y_{t-k_y}, x_{1t}, \ldots, x_{1,t-k_1}, x_{2t}, \ldots x_{2,t-k_2}, \ldots, x_{mt}, \ldots, x_{m,t-k_m}).$$

To deal with this as a single equation, we need to assume that if we added equations for x_{jt}'s conditional on lagged x's and y's, the likelihood components those would generate do not involve any of the same parameters that appear in this equation.

A reasonable prior

Assume we've ordered the elements of X_t as in the model description: lags of y, then lags of x_1, then lags of x_2, etc. Then the dummy observations are a square matrix, with all except the last row block diagonal. The diagonal block corresponding to lagged y's or any of the lagged x_j's is

```
diag(smooth^(0:(n-1))) %*% ctmat(lags[j]) %*% diag(damp^(0:(n-1))) * scale(j)
```

In the last row, we place a vector that is constant in positions corresponding to any given x_j and approximately a mean or normal value for that variable, together with a 1 in the lowest diagonal position, and multiply the row by `scale(nv+1)`. This asserts that the constant term should make the mean of y match the predicted value from the regression when all x's and lagged y's are at their means.

A reasonable prior

The mean of β should be 1 followed by zeros, to preserve symmetry and the idea of persistence. To make our dummy observations imply this mean, we should make our dummy observation have a current-y vector equal to the first column of the dummy X matrix.

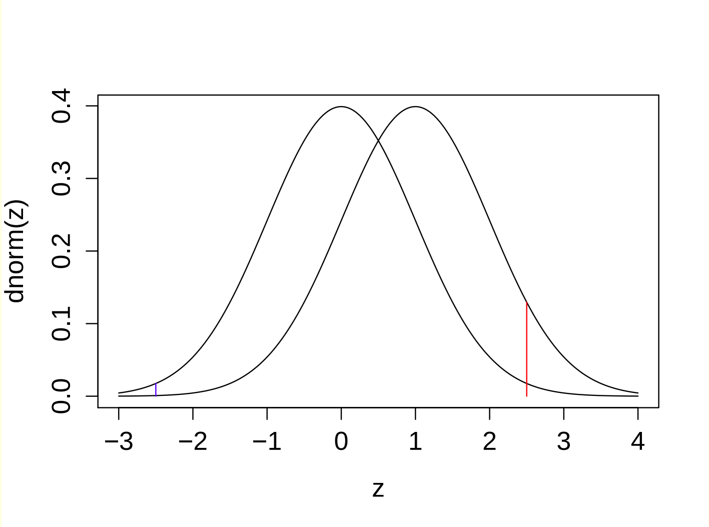
One final parameter chooses the ratio of the implied error variance in the first observation due to parameter uncertainty to the residual variance σ^2.

```
tsregPrior <- function(vlist, lags=rep(0, length(vlist)), ldv=1, scale,
   bbar=c(1,rep(0,length(unlist(lags)-1))), smooth, damp, vmeans, erratio)
```

Model “comparison” vs. model “checking” in the simplest case

Comparison Model A says $X \sim N(1,1)$, model B says $X \sim N(0,1)$.

If models have equal prior probability, posterior odds are $\phi(X-1)/\phi(X)$.

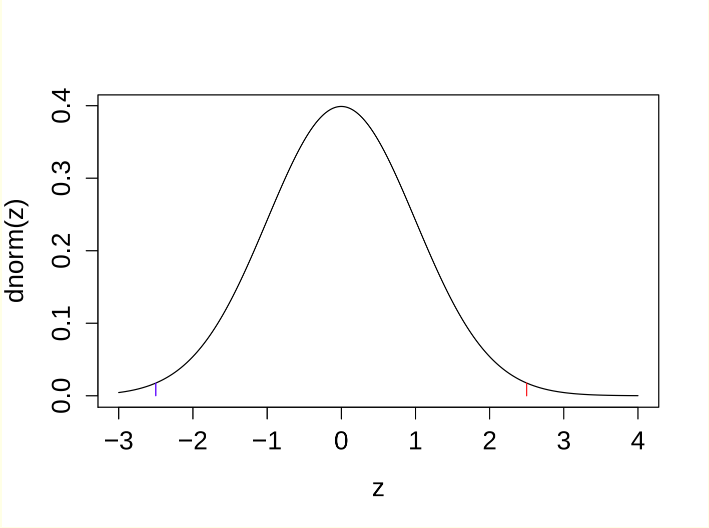


Model "comparison" vs. model "checking" in the simplest case

Checking Erase the line for the other model, just look at height of pdf. Obviously could be a mistake. Makes sense under an implicit "locally uniform" alternative.

But results can be adjusted arbitrarily by taking nonlinear, monotone transformations of the data.

Odds ratio model comparisons are invariant to montone transformations of the data.

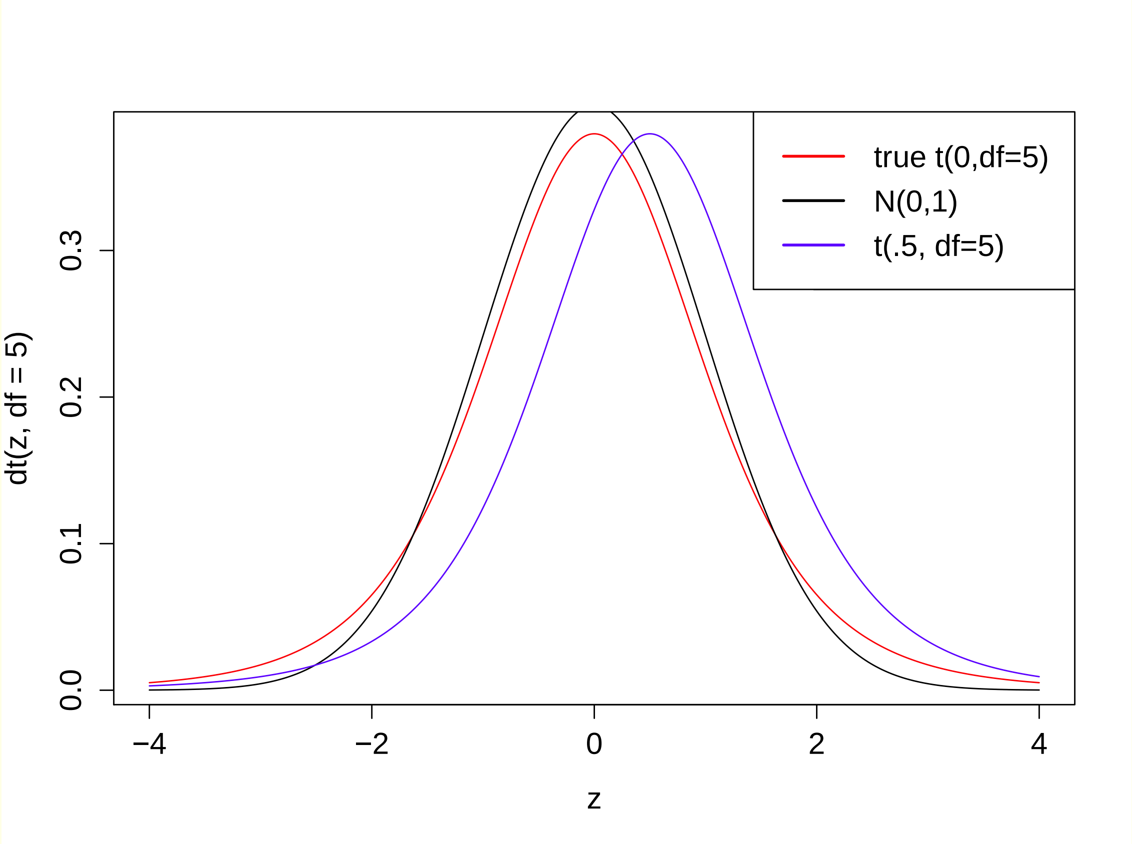

dnorm(z)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
4
z

Checking vs comparison in more complicated settings

- Likelihood-based model comparison “looks for” the aspects of the data on which the *ratios* of the model probabilities differ most sharply.
- These may be low-probability events.
- They may be aspects of the data that don’t matter much to us.
- That’s fine if one of the models is correct. If not, likelihood may pick a model that’s somewhat worse over most of the distribution of the data, but much better for some narrow range of unimportant aspects of the data.

Examples

$N(0,1)$ vs. $t(1, df = 5)$, when the truth is $t(0,5)$



Examples

$$y_t = y_{t-1} + \varepsilon_t \quad \text{vs.} \quad y_t = \rho y_{t-1} + \varepsilon_t$$

when truth is

$$y_t = .5y_{t-1} + .5y_{t-2} + \varepsilon_t \, .$$

For one step ahead the free-ρ model will be better. But for 100 steps ahead, the random walk will be better, because its forecast does not decay toward zero, and y_{t+s} does not either. The optimal n-step-ahead forecast converges quickly to $\frac{2}{3}y_t + \frac{1}{3}y_{t-1}$ as $n \to \infty$, and y_t is highly correlated with this.

Checking with test statistics

Because posterior odds may be hard to compute, or because of worries that they may pick up model failures we're not really interested in, it's appealing to look at models' implied distribution for functions of the data, or of data and parameters, that we are surely interested in. Call this thing we are interested in $T(X)$, where X is the sample data and $T(X)$ is a lower-dimensional function of X.

Limited information model comparison

- Pretend all we see is $T(x)$, so we form $\pi(\theta)p(T(x) \mid \theta_i)$ for each model i we are interested in, integrate these things over θ_i to obtain odds ratios among models.

- This would make model choice depend on how well the model explains $T(X)$, which is what interests us.

- But it clearly loses information.

- It can even lead to very wrong inference.

Example of very wrong inference

Model A: $\vec{x} \sim N(0, I_{10})$

Model B: $\vec{x} \sim N(\mathbf{1}, I_{10})$

$T(X) = \sum X_i^2$

$T(X) \sim \chi^2(10)$ under Model A, non-central $\chi^2(10, ncp = 10)$ under Model B.

Observe $X = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$ and therefore $T(X) = 10$.

Example of very wrong inference

This X is at the peak of the Model B density function. The pdf for X at this point under model B is $\phi(0)^{10} = 1.0212 \times 10^{-4}$, the product of 10 normal densities at 0. The pdf for X at this point under model A is $\phi(1)^{10} = 6.880631 \times 10^{-07}$. The Bayes factor therefore favors Model B by about 148 to 1. If we base inference just on $T(X) = 10$, though, we find that under model B the $\chi^2(10)$ with non-centrality parameter 10 has density .02784 at the observed point, while the central $\chi^2(10)$ density is .08773, implying a Bayes factor favoring Model A over Model B by 3 to 1. For $T(X)$, the most likely value under Model A is the observed value 10, while under Model B the most likely value of $T(X)$ is about 17.