A Test for the Distributional Comparison of Simulated and Historical Data

Valentina Corradi[1] and Norman R. Swanson[2]

[1]Queen Mary, University of London and [2]Rutgers University

November 2003

Abstract

In this note, we apply the testing methodology of Corradi and Swanson (2003a), which is based on a distributional generalization of White's (2000) reality check in order to examine the "distributional fit" of alternative real business cycle models. Parameter calibration and estimation error is found to matter much less than distributional assumptions, when some simple illustrative macroeconomic models are considered.[1]

1 Introduction and Testing Framework

The comparison of distributional characteristics of simulated versus actual data is important in a number of areas, including the analysis of alternative real business cycle models and the specification of models for economic scenario generation in the insurance and banking industries. In recent years, Kolmogorov type distance measures for testing distributional equality have been extended to the case of dependent observations and/or parameter estimation error (see e.g. Andrews (1997), Inoue (2001), Bai (2003), Corradi and Swanson (2003b)), and the papers cited therein. However, the papers cited above primarily address the case in which the objective is to test for the correct specification of some aspects of a given candidate model. In many cases, it is crucial to allow all models to be approximations, and thus misspecified. This suggests evaluating the relative degree of misspecification of a given group of competing models based on the comparison of empirical distribution functions of historical and simulated data.

Without loss of generality, consider the joint empirical distribution of (actual and model-based) current and previous period values of a variable of interest, X_t, for m models, and set model 1 as the benchmark. Let $Y_t = (X_t, X_{t-1})$, $Y_{j,n}(\widehat{\theta}_{j,T}) = (X_{j,n}(\widehat{\theta}_{j,T}), X_{j,n-1}(\widehat{\theta}_{j,T}))$ be the historical and the simulated data under model j, respectively (note that data are simulated from models containing

estimated parameters). Also let $F_0(u;\theta_0)$ be the distribution of Y_t evaluated at u and $F_j(u;\theta_j^\dagger)$ be the distribution of $Y_{j,n}(\theta_j^\dagger)$, where $\theta_j^\dagger$ is the probability limit of $\widehat{\theta}_{j,T}$, taken as $T \to \infty$, and where $u \in U \subset \Re^2$, possibly unbounded. Accuracy is measured in terms of the squared (approximation) error, so that we choose Model 1 over Model 2 if

$$\int_U \left(\left(F_1(u;\theta_1^\dagger) - F_0(u;\theta_0) \right)^2 \right) \phi(u)du < \int_U \left(\left(F_2(u;\theta_2^\dagger) - F_0(u;\theta_0) \right)^2 \right) \phi(u)du,$$

where $\int_U \phi(u)du = 1$ and $\phi(u) \geq 0$. For any evaluation point, this measure defines a norm and is a typical goodness of fit measure.[2] The null hypothesis of interest is:

$$H_0 : \max_{j=2,\dots,m} \int_U \left(\left(F_0(u;\theta_0) - F_1(u;\theta_1^\dagger) \right)^2 - \left(F_0(u) - F_j(u;\theta_j^\dagger) \right)^2 \right) \phi(u)du \leq 0$$

and the alternative replaces '$\leq$' with '$\rangle$'. Thus, under H_0, no model can provide a better approximation (in a squared error sense) to the distribution of Y_t than the approximation provided by model 1. Test statistic is $\sqrt{T}Z_{T,S}$, where $Z_{T,S} = \max_{j=2,\dots,m} \int_U Z_{j,T,S}(u)\phi(u)du$, and

$$Z_{j,T,S}(u) = \frac{1}{T}\sum_{t=1}^{T}\left(1\{Y_t \leq u\} - \frac{1}{S}\sum_{n=1}^{S} 1\{Y_{1,n}(\widehat{\theta}_{1,T}) \leq u\}\right)^2 - \frac{1}{T}\sum_{t=1}^{T}\left(1\{Y_t \leq u\} - \frac{1}{S}\sum_{n=1}^{S} 1\{Y_{j,n}(\widehat{\theta}_{j,T}) \leq u\}\right)^2,$$

with $\widehat{\theta}_{j,T}$ an estimator of $\theta_j^\dagger$ that satisfies standard assumptions, and where S denotes the length of the simulated samples.

Then, under strict stationarity and standard regularity conditions concerning moments and smoothness of distributions, assuming that $T, S \to \infty : T/S \to \delta$, $0 < \delta < \infty$, and where for simplicity $F_j(u) = F_j(u;\theta_j^\dagger)$, it turns out that

$$\max_{j=2,\dots,m} \sqrt{T}\int_U \left(Z_{j,T,S}(u) - \left((F_0(u) - F_1(u))^2 - (F_0(u) - F_j(u))^2 \right) \right) \phi(u)du \xrightarrow{d} \max_{j=2,\dots,m} \int_U Z_j(u)\phi(u)du,$$

Further, when $T, S \to \infty : S/T^2 \to 0$ and $T/S \to 0$, then:

$$\max_{j=2,\dots,m} \sqrt{T}\int_U \left(Z_{j,T,S}(u) - \left((F_0(u) - F_1(u))^2 - (F_0(u) - F_j(u))^2 \right) \right) \phi(u)du \xrightarrow{d} \max_{j=2,\dots,m} \int_U \widetilde{Z}_j(u)\phi(u)du,$$

where $Z_j(u)$ and $\widetilde{Z}_j(u)$ are a zero mean Gaussian processes with well defined covariance kernels. Notice that when $T/S \to 0$, simulation error vanishes, so that the covariance kernel of $\widetilde{Z}_j(u)$ does not reflect the contribution of simulation error. Thus, when all competing models provide an

---

[2]Another measure of distributional accuracy available in the literature (see e.g. White (1982) and Vuong (1989)), is the KLIC. Our approach is an alternative to the KLIC that should be viewed as complementary in some cases, and preferred in others. (e.g. measuring accuracy over a specific region cannot be done in an obvious manner using the KLIC, but is easily done using our measure.

approximation to the true joint distribution that is as accurate as that provided by the benchmark, then the limiting distribution is a zero mean Gaussian process with a covariance kernel that reflects the contribution of parameter estimation error, the time series structure of the data and, for $\delta > 0$, the contribution of simulation error. On the other hand, when all competing models are less accurate than the benchmark model, the statistic diverges to minus infinity. Finally, under the alternative, $\int_U \left((F_0(u) - F_1(u))^2 - (F_0(u) - F_j(u))^2 \right) \phi(u) du > 0$ for some j, so that $\sqrt{T} Z_{T,S}$ diverges to infinity. Therefore, the test has correct asymptotic size if all models are equally good, is conservative when some model is strictly dominated by the benchmark, and has unit power under the alternative. This result is analogous to that found by White (2000) and can be seen as a distributional generalization of it. Our approach can also in principle be modified to allow for the evaluation of predictive densities, when the simulated data in our context are viewed as predictions, and when the starting values for the simulations are carefully chosen, based on the historical record. However, we leave discussion of such an extension to future research, and the reader is referred to Diebold, Gunther and Tay (1998), Diebold, Hahn and Tay (1999), Clements and Smith (2002), Giacomini (2002), Corradi and Swanson (2003c), and the references cited therein for further discussion of predictive density evaluation.

Valid bootstrap critical values are obtained in the following manner. Begin by resampling b blocks of length l, $bl = T - 1$, from the actual sample, and using the resultant series, say Y_t^* to compute $\widehat{\theta}_{j,T}^*$, for $j = 1, ..., m$. Then, $\widehat{\theta}_{j,T}^*$ plugged in, and a sample is simulated under model j, $j = 1, ..., m$. Let $Y_{j,n}(\widehat{\theta}_{j,T}^*)$, $n = 2, ..., S$ denote the series simulated in this manner, for $j = 1, ..., m$. Now, note that when $\delta = 0$ (vanishing simulation error) there is no need to resample the simulated series. Otherwise, we resample the simulated series, forming $Y_{j,n}^*(\widehat{\theta}_{j,T}^*)$, $j = 1, ..., m$, $n = 2, ..., S$ denote the resampled series under model j. In this more general case, we form bootstrap statistics $Z_{T,S}^{**} = \max_{j=2,...,m} \int_U Z_{j,T,S}^{**}(u) \phi(u) du$, where

$$
\begin{aligned}
Z_{j,T,S}^{**}(u) \quad =& \frac{1}{T} \sum_{t=1}^{T} \left(\left(1\{Y_t^* \le u\} - \frac{1}{S} \sum_{n=1}^{S} 1\{Y_{1,n}^*(\widehat{\theta}_{1,T}^*) \le u\} \right)^2 - \left(1\{Y_t \le u\} - \frac{1}{S} \sum_{n=1}^{S} 1\{Y_{1,n}(\widehat{\theta}_{1,T}) \le u\} \right)^2 \right) \\
&- \frac{1}{T} \sum_{t=1}^{T} \left(\left(1\{Y_t^* \le u\} - \frac{1}{S} \sum_{n=1}^{S} 1\{Y_{j,n}^*(\widehat{\theta}_{j,T}^*) \le u\} \right)^2 - \left(1\{Y_t \le u\} - \frac{1}{S} \sum_{n=1}^{S} 1\{Y_{j,n}(\widehat{\theta}_{j,T}) \le u\} \right)^2 \right).
\end{aligned}
$$

The analogous bootstrap statistic for the case where $Y_{1,n}(\widehat{\theta}_{1,T}^*)$ is used in place of $Y_{j,n}^*(\widehat{\theta}_{j,T}^*)$ above (i.e. the case of $\delta = 0$) is denoted by $Z_{T,S}^*$. Finally, compute the bootstrap statistic, $\sqrt{T} Z_{T,S}^{**}$

$\left(\sqrt{T}Z^*_{T,S}\right)$. Perform B bootstrap replications (B large) and compute the quantiles of the empirical distribution of the B bootstrap statistics. Reject H_0 if $\sqrt{T}Z_{T,S}$ is greater than the $(1-\alpha)th$-quantile. Otherwise, do not reject. Now, for all samples except a set with probability measure approaching zero, $\sqrt{T}Z_{T,S}$ has the same limiting distribution as the corresponding bootstrapped statistic, when $\int_U \left((F_0(u) - F_1(u))^2 - (F_0(u) - F_j(u))^2\right)\phi(u)du = 0$ for all $j = 2,...,m$. In this case, the above approach ensures that the test has asymptotic size equal to α. On the other hand, when one (or more) competing models is (are) strictly dominated by the benchmark, the approach ensures that the test has an asymptotic size between 0 and α. Finally, under the alternative, $Z_{T,S}$ diverges to (plus) infinity, while the corresponding bootstrap statistic has a well defined limiting distribution. This ensures unit asymptotic power. Note that the suggested bootstrap procedure mimics the limiting distribution of the statistics in the least favorable case for the null, thus leading to conservative inference (see Corradi and Swanson (2003a).

2 A Stylized Stochastic Growth Model with One Shock

As an illustration, consider a stylized closed economy stochastic growth model with no government intervention, fixed labor supply, zero capital depreciation, and one shock (see e.g. Christiano (1990)). Assume that a representative agent maximizes $E_0\left(\sum_{t=0}^{\infty}(1-\tau)^{-1}C_t^{(1-\tau)}\beta^t\right)$, subject to $C_t + K_t - K_{t-1} = \pi_t K_{t-1}^{\alpha}$, where the technology shock π_t is generated as $\log \pi_t = \varkappa + \log \pi_{t-1} + \epsilon_t$, $\epsilon_t \sim iid(0,\sigma_\epsilon^2)$. Here, C_t and K_t are consumption and capital stock, β is the discount rate factor, τ measures risk aversion, and α is the capital share of income. Here, $X_t = C_t + (K_t - K_{t-1}) = \theta_t K_{t-1}^{\alpha}$, where X_t is real output. Further, we can estimate the variance of the technology shocks as discussed in Christiano and Eichenbaum (1992). In particular, we set

$$\widehat{\sigma}_\epsilon^2 = \frac{1}{T}\sum_{t=1}^{T}(\log \pi_t - \log \pi_{t-1})^2 = \frac{1}{T}\sum_{t=1}^{T}(\Delta \log X_t - \alpha \Delta \log K_{t-1})^2,$$

In the sequel, we use annual U.S. per capita data for the period 1933-2002 (from the Bureau of Economic Analysis). Additionally, we consider linear-quadratic and loglinear-quadratic approximations, from Christiano (1990). Let λ be the unique root, satisfying $|\lambda| \leq 1$, such that: $\lambda^2 - \Psi\lambda + \beta^{-1} = 0$, where $\Psi = 1 + \beta^{-1} + \left(\frac{1-\alpha}{\tau}\right)(1-\beta)\frac{C^*}{K^*}, \frac{C^*}{K^*} = \frac{(\beta^{-1}-1)}{\alpha}$. Then, for $q = \frac{\beta^2}{\tau}(\beta^{-1}-1)C^*$, $C^* = K^*(\beta^{-1}-1)/\alpha$ and $K^* = \left(\frac{\alpha\beta}{1-\beta}\right)^{\frac{1}{1-\alpha}}$, a linear-quadratic approximation ("Approximation

Scheme 1") yields:

$$K_t - K_{t-1} = (1 - \lambda)K^* + (\lambda - 1)K_{t-1} + \frac{q\lambda}{1 - \beta\lambda}\log(\theta_t)$$

Alternatively, a loglinear-quadratic approximation ("Approximation Scheme 2") yields:

$$K_t = K^{*(1-\lambda)}\exp\left(\frac{q}{K^*}\frac{\lambda}{1 - \beta\lambda}\log\theta_t\right)K_{t-1}^\lambda.$$

Given this setup, we evaluate the joint distribution of $Y_t = (\Delta \log X_t, \Delta \log X_{t-1})$. Simulated data are constructed by first estimating σ_ϵ^2 from the data, and then, given an assumption concerning the distribution of ϵ_t, generating $\log(\pi_t)$, K_t, and Y_t using various calibrated values for α, β, and τ; as well as approximation schemes 1 and 2.

Some illustrative results are given in Tables 1 and 2. The first column of numerical entries report $\sqrt{T}Z_{T,S}$ for $T = 70$ (corresponding to the 70 annual U.S. per capita output growth observations used for the period 1933-2002), and for $S = \{T, 2T, 5T, 10T\}$.[3] All statistics are based on use of 20 uniformly distributed values for u across the range of the actual data. Since Y_t is two dimensional, this requires evaluating the statistics over a grid of 20x20 values. The next four columns of numerical entries contain 5% and 10% bootstrap critical values constructed using both $\sqrt{T}Z_{T,S}^{**}$ and $\sqrt{T}Z_{T,S}^{*}$ (i.e. $T/S- > \delta > 0$ and $T/S- > 0$, respectively). For the case where $S = T$, we set the bootstrap block length equal to $l1 = 2$, $l2 = 5$, and $l3 = 7$. For all other cases, where $S = aT$, say, we set li equal to "a" times the corresponding value of li when $S = T$, for $i = 1, 2, 3$. Thus, "a" takes the values 2,5, and 10, depending on the magnitude of S. Bootstrap empirical distributions are constructed using 100 bootstrap replications (i.e. $B = 100$). In addition to test statistics and critical values, various summary statistics (including *% ¡0*, *min*, *max*, *mean*, and *standard error*) constructed using the simulated data are given in the remaining columns of the tables. Corresponding values based on the actual data are: *%¡0 = 0.213; min = -0.128; max = 0.158; mean = 0.027*; and *standard error = 0.044*. In all, four different varieties of models are examined, including:

Part I - Comparing Approximation Schemes: Two models are compared. Set $\{\alpha = 0.4, \beta = 0.95, \tau = 0.5\}$, $\widehat{\sigma}_\epsilon^2 = 0.0025$, *the benchmark model uses approximation scheme 1 and the alternative uses approximation scheme 2, and technology shocks are normally distributed with variance=$\widehat{\sigma}_\epsilon^2$.*

Part II - Comparing Approximation Schemes: As Part I except the benchmark model assumes that technology shocks are drawn from a $1/3t_3$ *distribution and the alternative assumes normality with*

---

[3]Note that in all case $S/T > 0$, however, when for example $T = 70$ and $S = 700$, then simulation error is likely to be negligible with respect to parameter estimation error.

variance=$\widehat{\sigma}_{\epsilon}^2$. *Part III - Comparing Different Parameterizations: Three models are compared. Set* $\{\alpha = 0.4, \ \beta = \{0.95, 0.90, 0.99\}, \ \tau = 0.5\}, \ \widehat{\sigma}_{\epsilon}^2 = 0.0025.$ *All models uses approximation scheme 1 and assume that technology shocks are normally distributed with variance=$\widehat{\sigma}_{\epsilon}^2$. The model with* $\beta = 0.95$ *is the benchmark, while the models with* $\beta = 0.90$ and $\beta = 0.99$ *are the alternatives. Part IV - Comparing Different Parameterizations: Three models are compared. Set* $\{\alpha = 0.4, \ \beta = 0.95,$ $\tau = \{0.5, 0.3, 0.7\}.$ *As Part III except that the benchmark has* $\tau = 0.5,$ *while the models with* $\tau = 0.3$ and $\tau = 0.7$ *are the alternatives.* In Part II, the unreasonable distributional assumptions in the benchmark model are imposed in order to illustrate the power of the test to reject when the alternative is clearly superior to the benchmark model.

Turning to the tabulated results, notice in Part I that the null hypothesis is only rejected for two of twelve (l, S) combinations at a 5% test level, suggesting that approximation scheme 2 is not preferred to approximation scheme 1. This points to a conclusion that there is not too much to choose between a linear-quadratic and loglinear-quadratic approximation. This result is not surprising, given the statistics reported in the last 10 columns of the table, where it is seen that summary measures of output growth rates simulated using either approximation closely mimic analogous summary measures based on the actual historical record. In contrast to the results of Part I, notice that the statistics reported in Part II indicate strong rejection of the benchmark model, regardless of whether bootstrap critical values are constructed using $\sqrt{T}Z_{T,S}^*$ or $\sqrt{T}Z_{T,S}^{**}$. This, again, is not surprising, given the benchmark model in this example. In accord with the fact that this distributional assumption is nonsensical, notice that the summary statistics reported for "Sim1" (i.e. the benchmark) in the last 10 columns of the table are not at all near the analogous values based on the actual data. With regard to Parts III and IV, it is interesting to note that for the parameter values considered, there appears to be little to choose between the models characterized by different values for the calibrated parameters. This result is supported by the fact that summary statistics based on simulated data are very close to analogous statistics based on the actual data, as in Part I. Thus, the test seems to indicate a certain robustness of our simple illustrative RBC model to modest changes in the parameters used for calibration. Of course, such conclusions may not necessarily hold up when our test is applied to more complicated and realistic RBC models.

In summary, we feel that the testing approach discussed above may be of use in a variety of empirical settings where macroeconomic and/or financial data are being examined, and "optimal" simulation models are being selected.

3 References

Andrews, D.W.K., (1997), A Conditional Kolmogorov Test, *Econometrica*, 65, 1097-1128.

Bai, J., (2003), Testing Parametric Conditional Distributions of Dynamic Models, *Review of Economics and Statistics,* 85, 531-549.

Christiano, L.J., (1990), Solving the Stochastic Growth Model by Linear-Quadratic Approximation and by Value Function Iteration, *Journal of Business and Economic Statistics, 8, 23-26.*

Christiano, L.J., and M. Eichenbaum, (1992), Current Real Business Cycles Theories and Aggregate Labor Market Fluctuations, *American Economic Review,* 82-430-450.

Clements, M.P. and J. Smith, (2002), Evaluating Multivariate Forecast Densities: A Comparison of Two Approaches, *International Journal of Forecasting,* 18, 397-407.

Corradi, V., N.R. Swanson, (2003a), Evaluation of Dynamic Stochastic General Equilibrium Models Based on Distributional Comparison of Simulated and Historical Data, *Journal of Econometrics*, forthcoming.

Corradi, V. and N.R. Swanson, (2003b), Bootstrap Conditional Distribution Tests in the Presence of Dynamic Misspecification, *Journal of Econometrics*, forthcoming.
Corradi, V., N.R. Swanson, (2003c), The Block Bootstrap for Parameter Estimation Error In Recursive Estimation Schemes, With Applications to Predictive Evaluation, Queen Mary, University of London, University of Exeter and Rutgers University.

Diebold, F.X., T. Gunther and A.S. Tay, (1998), Evaluating Density Forecasts with Applications to Finance and Management, *International Economic Review,* 39, 863-883.

Diebold, F.X., J. Hahn and A.S. Tay, (1999), Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High Frequency Returns on Foreign Exchange, *Review of Economics and Statistics*, 81, 661-673.

Giacomini, R. (2002), Comparing Density Forecasts via Weighted Likelihood Ratio Tests: Asymptotic and Bootstrap Methods, Working Paper, University of California, San Diego.

Inoue, A., (2001), Testing for Distributional Change in Time Series, *Econometric Theory*, 17, 156-187.

Vuong, Q. (1989), Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses, *Econometrica,* 57, 307-333.

White, H., (1982), Maximum Likelihood Estimation of Misspecified Models, *Econometrica,* 50, 1-25.

White, H., (2000), A Reality Check for Data Snooping, *Econometrica*, 68, 1097-1126.